SHA571: Understanding and Visualizing Data

What you'll do

- **Identify and collect data that can be used to inform a decision**
- **Recognize and mitigate potential bias when generating a data sample**
- **Develop statistical summaries and data visualizations to understand how variables impact outcomes**
- **Evaluate decisions by looking at key performance measures and determining their implications for stakeholders**

Course Description

Important business decisions require justification, and while we often have data that can help us make those decisions, the skill with which we analyze the data can make the difference between a good and bad outcome. This course, developed by Professor Chris Anderson, is designed to move learners beyond making decisions focused solely on averages. In this course, you will develop a working familiarity with the grounding principles of data analysis. You will learn to derive the greatest benefit possible from the data available to you while ensuring that the conclusions you draw remain valid. You will apply a decision-making framework within which you'll interact with the data to achieve the best outcome.

This course includes valuable tools and help sheets for data handlers along with the insight and perspective you need as a data consumer. While this course is not a replacement for a full-length statistics course, you will have a basic grounding in many statistics concepts by the time the course is over. You should be able to complete this course without

any prior knowledge of statistics.

Chris Anderson
Associate Professor, School of Hotel Administration, Cornell University

Chris Anderson is an associate professor at the Cornell School of Hotel Administration. Prior to his appointment in 2006, he was on the faculty at the Ivey Business School in London, Ontario Canada. His main research focus is on revenue management and service pricing. He actively works with industry, across numerous industry types, in the application and development of revenue management, having worked with a variety of hotels, airlines, rental car and tour companies as well as numerous consumer packaged goods and financial services firms. Anderson's research has been funded by numerous governmental agencies and industry partners, and he serves on the editorial board of the *Journal of Revenue and Pricing Management* and is the regional editor for the *International Journal of Revenue Management*. At the School of Hotel Administration, he teaches courses in revenue management and service operations management.

Author Welcome

So you're in the situation and you have some thoughts that perhaps data can help you make more informed decisions. That's going to be the focus of the scores. We're going to look at the nature, use, and limitations of our data Understand where it comes from. Think about how we might collect future data. And then given where that data came from or how we might collect future data, what are the limitations on the inferences we might make, any causal statements we might draw from that data? We're going to spend some time summarizing that data, visualizing that data And then using those summaries, and visualizations to link that back to our decision, to make a more data informed decision. We're going to go through a series of examples through out the course. Keep a focus on the high level of those situations, understand the nature of the situation, and how we're trying to tackle that with our use of data. Right, so enjoy the course and I hope you have lost of improved data informed decisions

going forward.

Table of Contents

Meet Your Class

Module 1: Gather and Qualify Data

Module 2: Visualization and Analysis

10. Identify Visualizations for Statistics
11. Watch: Visualizing Two Categorical Variables
12. Watch: Visualizing Two Quantitative Variables
13. Watch: Visualizing a Categorical Variable Together with a Quantitative Variable
14. Watch: Interpreting Initial Visualizations and Identifying Issues
15. Activity: Interpret Statistics and Visualizations
16. Analyze Based on Data Summaries and Visualizations
17. Course Project, Part Two: Identify Data Summaries and Visualizations
18. Module Two Wrap-up: Visualization and Analysis

Module 3: Bring the Data into the Decision

1. Module Three Introduction: Bring the Data into the Decision
2. Watch: Using a Decision-Making Framework
3. Watch: Interacting with the Data-Model Abstraction
4. Read: Case Study: The Billboard Effect—Using a Model to Determine Value
5. Tool: Data Qualification Checklist
6. Determining Data Viability
7. Watch: Working with Dashboards
8. A Decision With or Without Data
9. Course Project, Part Three: Data and Your Decision
10. Tool: "Understanding and Visualizing Data" Action Plan
11. Module Three Wrap-up: Bring the Data into the Decision
12. Read: Thank You and Farewell

Meet Your Class

Back to Table of Contents

Meet Your Class

Using the discussion below, please tell us about yourself; we're eager to learn more about you as well as your classmates. What do you hope to learn from the course? What is your profession? Where are you located? Please respond in a text format or as a video using the film strip icon that is available once you click Reply.

(If posting a video response, we recommend that you do not use your cell phone, as most do not use Flash software required to convert the recording.)

Back to Table of Contents

Module 1: Gather and Qualify Data

Back to Table of Contents

Module One Introduction: Gather and Qualify Data

Whether you are awash in data from your ongoing business processes or considering how you might capture some of this crucial information, you need to begin by deciding what you hope the data will do for you. Clarity about your goals should drive any efforts to collect and use data.

In this module, you will consider how outcomes and stakeholders drive data needs. You will also learn about bias in data collection and practice recognizing it.

Back to Table of Contents

Watch: Evaluating the Problem

In order to make a decision using data, you need to be sure you properly frame your problem. Who are your stakeholders? What are the desired outcomes? How will we know if a decision makes sense?

In this video, Professor Anderson outlines the process that begins with defining the problem and ends with making a data-based inference that informs a decision. He introduces a fictional scenario that will be threaded throughout this course: that of the Snow Hawk Tire company and its upcoming marketing campaign.

Transcript

All right, so initially, we're going to focus on the statistical process towards making a decision. Right? So trying to use data to help us in this complicated setting where we're looking for some insight about the underlying problem. And so when we're trying to use this statistical process to tackle this decision, step one is, you know, who's involved? Who are the key stakeholders? Right? And then once we've determined who those are, then what are their metrics of interest? So what are their KPI's, or key performance metrics? Indicators across those stakeholders. So what are they concerned about, when do they know whether or not you've made a good decision or not. Once we've outlined who's involved and what their interests are then we can go through this process of collecting and analyzing data, right? So part of this data collection process is to understand what you want to draw this data from. We're going to talk about the population, ultimately we're going, you know, to narrow down that population, and focus on a subset, what we're going to refer to as a sample. And how we might collect that sample. Either historic data. We might execute some experiments, or launch some surveys. We collect our data. And now we're going to understand that data through a process of descriptives and visualization. So, how to describe it, what might happen, what's the range of things that could happen. And then if I could look at that visually versus through a series of metrics does that tell me anything about what's going on in the situation. And then once I understand what's going on in this situation, then we're

going to make inference. So we have this sample, we're going to measure it, we're going to visualize it.

And ultimately, we're going to infer from that what we think is going to happen, you know, when we instigate this decision or execute this setting. So it's about what happens tomorrow. Versus what has happened historically with this sample. And then we can, now we're in this position to make a decision, right? We have a sense who's involved, what their interests are. Collect some data appropriate to that setting. Analyze it and then make some inference about whether or not that analysis is representative of what we're looking at and then we can make that decision. Now, that decision may be that I don't have enough information. I may have to apply, rinse, and repeat and go back again. Collect more data to further analyze this problem. The example we're going to work through is Snow Hawk tires. And so, Snow Hawk tires is a tire manufacturer, and they're interested in developing a marketing campaign to increase the sale of snow tires. Has a catchy title, right, pay no dough if it doesn't snow. Right? Where rebates to consumers are linked to the amount of snowfall. The idea here is that consumers may choose not to buy snow tires because they think they don't need them, because it's not going to snow that much, right? So if we can remove that variability around snowfall into the purchase decision then maybe consumer would purchase more tires, right? And so our goal is to see whether not this make sense. From a data standpoint, who do we thinks involved? The firm, the consumers? Right? What kind of data do we need to tackle this instance? So what are going to be these KPIs, how are we going to measure those with the data, and then ultimately decide whether or not we should go ahead with this campaign or maybe tweak its parameters? Or maybe only use this campaign in certain areas, and what are the characteristics of those areas? And so we're going to go through this in general from start to finish, as we try and look at the different elements of this statistical decision-making process.

Back to Table of Contents

Watch: Analyzing Cases across Variables

The type of data we collect can be informed by how we intend to use it. We have some control over how we design our data collection scheme, and the decisions made up front will have an impact on the conclusions we can draw based on that data. Professor Anderson draws a distinction between quantitative and categorical data and discusses some of the properties and implications of each data type.

Transcript

So data is at the core of our statistical decision-making process. And it becomes critically important to understand how that data is organized, the sources of data, how it's sort of created and how it's sort of coded, right? And we're going to walk through that in this session, right? And so typically data is organized into a data set, right? That data set has a series of cases, or units, right? And for each of those cases we've measured a series of variables, right? So that variable might be sales. And I might have that sales information for customer A, customer B, customer C, right? So a series of different customers are my cases and I will have sales or returns or date from last purchase for each of those customers, right? And so, if we think about this in our two-dimensional spreadsheet world, right? It kind of helps to visualize it. You can think about cases being rows within that spreadsheet, right? And then variables being columns, right? So each row is a different customer, or a different month, or a different week, or a different business unit. And then each column is a different variable that we've measured for each one of those cases.

So let's pop over to a spreadsheet and look at an example, right? This example is from airline purchases, right? And so each row or each case is purchased by a consumer, right? And then we have a series of variables that describe that purchase, right? And so that purchase may be of economy versus a business class ticket. That purchase may have been made by a male or a female. And at the same time, that purchase was made so many days before departure of that given flight. And sometimes we present the same piece of information in different ways,

right? So, we have these two columns here where we've presented the same information. Here we have advanced purchase as 106 days versus here we've put that into a bucket of 91 days or longer. And so that becomes important when we think about how we organize data, and how we present that data. Typically, we think of two types of variables when we think about data. Categorical and quantitative, right? So a quantitative variable is easiest to think about as a number, right? It has numerical properties. Two is twice as big as one, right? Whereas a categorical variable, you know, puts things into buckets, right? And the buckets have no size relevance, right? They're basically just different categories which we've compartmentalized our data. So if we go back to our airline example. So here we can look at the data organized across a series of different variables. Some of those variables are quantitative and some of those variables are categorical, right? So obviously gender would be categorical and these days before purchase or advanced purchase is quantitative.

But as I talked about earlier, we've also presented that information categorically, where the seven is now transformed into this bucket of four to seven days. And so just because information is quantitative, we may present it categorically, right? Here the argument would be if the consumer who purchases seven days is not all that different from the consumer who purchased five days. So I'm going to put them into one bucket, and I'm going to compare that bucket versus people who purchased one to three days, because I think they're different, right? So we have some choice in how we design these variables, and how we present that information, right? If we look at a second example, this example is of book purchases. So here we have each row is from an individual consumer, right? And then we've captured some information about that consumer, right? And so that consumer that we've captured some information on how much they've spent with us, how many times they've purchased with us. But interesting here when we look at gender, we've coded gender here as a one or a zero, right? So yes, gender is a category. We've assigned those categories numerical values. It's critically important though that when we analyze that, we treat it as a categorical variable not a quantitative variable, because those ones and zeroes have no numerical purpose. They could have been A or B, they're just two different categories that we're using to separate, in this case, gender.

So, in summary, when we think about data, data is everywhere, lots of different topics, different settings. We're going to focus on an individual data set. That data set, you know, like our spreadsheet, is going to have a series of rows being cases, right? Going to have a series of columns of variables. Those variables are of different types, categorical or quantitative. And then we're going to present and summarize those variables in order to help us make that decision, right? It's critical for us to understand whether or not we have categorical or quantitative variables, because how we present it and how we make inference is going to differ across those different variable types.

Back to Table of Contents

Identify Useful Data Types

In this quiz, you will answer questions based on data from a fictional scenario.

Scenario: White Manufacturing has collected a sample of recent times to completion, in hours, for jobs that involve distinct manufacturing steps to produce a required number of completed units. For some jobs, they may have expedited the job by giving it higher priority at one or more stages of manufacturing in order to meet a customer deadline.

The data collected are shown in the table below.

You must achieve a score of 100% on this quiz to complete the course. You may take it as many times as needed to achieve that score.

Back to Table of Contents

Watch: Sampling for Inference

When we want to make an informed prediction about what will happen in the future, our prediction or inferred outcome is often based on data from past outcomes. Professor Anderson describes the relationship between a data sample and the population from which it is drawn, and he discusses some implications and limitations of working with a data sample.

Transcript

So, while data is at the core of our statistical decision-making process, right, we have to have some way to collect that data and then imply some results from the analysis that we perform on that data. And sampling is really at the core of that process. Right? And so you can think of, I mean, the issue of importance. You know, we're going to call this our population, right? And so this is the thing we're really trying to make our decision on. We've got this general, all-encompassing population, right, and so you can think of that as snowfall, if we're trying to think about a campaign linked to snowfall, then our population would be all past, present and future snowfall for a given area. But obviously we can't collect all that information and so we collect a subset of that. If we're lucky we have a large subset of this historical snowfall. Right? That subset is going to be our sample. Right? And we're going to perform some analysis on the variables within that sample. And then, once we've done that, we're going to make some inference about the population, right? So, all past and present and future snow fall based upon some measurements of the sample of historic snowfall, right? And so that kind of, you know, paints this population sampling picture, right? Where we've got this general population, there's a subset of information that I can measure. I'm going to sample from the subset. I'm going to perform some analysis of the variables within that sample, and then make some inference back to the population.

So a classic, well-documented example is of the 1948 US Presidential election. So the Chicago Tribune went to press on the evening of the election announcing that Dewey was going to be the new president of the United States, when in fact he was not, right? Harry Truman was elected

president in 1948. And what happened was there was a series of polling going on during the election, right? And as a result of that sampling they were trying to determine who was going to get the popular vote. And we could think of lots of reasons time zones across the US, who they were sampling. The net of this was that who they sampled and how they sampled and how many they sampled was not representative of the total US population and so their inference about who was president was incorrect. So our goal is to make sure that when we go about this process of sampling the population, that we do it appropriately so the inference we make is going to be consistent with behaviors of that population.

Back to Table of Contents

Watch: Recognizing Sampling Bias

There is no guarantee that a decision based on data will be a good decision. If the data sample you collect does not accurately represent the population, your inferences may not be valid and your conclusions can't be trusted. Professor Anderson examines several sources of bias in data collection and offers practical advice for preventing or minimizing the effects of bias in your sample.

Note: In this video Professor Anderson exposes some of the mathematical calculations as he works through an example. You need not be too concerned with the mathematical details, especially if you are not working in a technical capacity. It is enough to follow in a general way the concept being discussed.

Transcript

So the inferences we make from a sample is only as informative as that sample is representative. If that sample is not representative, we think of it as being biased. And so sampling bias exists in many different ways, but it basically results in the generalizations from the sample not necessarily being representative or informative about the population. So in an ideal world, you could think of a sample. It looks just like the population, only smaller, right? It's smaller largely owing to time and monetary considerations. We can't look at everything so we look at a subset. In an ideal world we want every observation in the population to have the same chance as being in the sample, right? We think of that as what we call random sampling. Where each member of the population has the same chance of being in the sample itself, right? And so you could think of that as drawing names from a hat, right? We typically use technology to make that random picking from the population into the sample.

So a classic example comes from the 2008 US Presidential election where a Gallup Poll was constructed of a random sample of 2,847 Americans. Of those Americans, 52% said that they supported Obama. The election happens and 53% of the US electorate voted for Obama.

Right? So here we see the value of sampling is that the small subset of 2,800 persons was representative and informative about who is going to win the election, right? And so you mean random samples have the same characteristics, you know, say averages, of our variables as our population, right? A non-random sample may suffer from some sampling bias, and those averages may be different, right? So only with that random sample can we truly make inferences or generalizations about the population, right? But, the realities of sampling may result in, maybe the random samples not ideal or maybe not even feasible, right? So think about snowfall. We're going to launch a campaign as a function of snowfall.

Now obviously that's going to be in the future, so our population really is all past, present and future snowfalls. Well, I don't have a crystal ball so I have no idea how to collect information on future snowfall so that's automatically out. And then I may not want to just randomly select from history, I may want to use it all. But maybe I don't want to use it all. I don't want to use snowfall from 100 years ago because it's not representative owing to global warming. So we need to think about some of the potential sources of bias and try to avoid them, right? And so, you could think about any kind of firm trying to look at quality of service, right? And so we're going to send out an email to a subset of our consumers to ask something about their most recent purchase or stay with us,. right? And so, we may not simply just send out this simple email to our consumers. Obviously that subset of consumers may not be representative of all the population of potential consumers. But it's at least reachable, because those are the ones you have information on, right? Now, we may want to not just stop after we send out this initial email. We may want to send out subsequent emails to those who haven't responded yet, to make sure we avoid any kind of non-responder bias, right? To make sure that even though, you know, we had a smaller sample we drew from, let's make sure that that is as representative as possible of those potential future customers as well. There's all, so even if we have a random sample there still could be other sources of bias, right? How we word our questions, right? We're asking customers to rate our service on a one to five scale, right? Well is one better, or five better? Right? So how do we, you know, make any inference from that when the respondent doesn't necessarily know what we're asking, right? What's the context which in we're collecting that data or surveying customers.

So there's an infinite number of possibilities that could be impacting how you created this data set. And so the key here is for you to think critically about how that data was collected. You need to sort of ask these questions about how it was created, right? And could any of those ways at which it created lead to this sample being non-representative of the population? And, you know, making your inferences less valid. This is one of the easiest ways for you to instantaneously become more statistically minded, right? Because you're critically assessing where this data came from, and do I think it's going to be representative of who I'm trying to draw that inference to.

Back to Table of Contents

Watch: Relationships between Variables

Professor Anderson stresses the importance of randomization when sampling the population in both experimental and observational data-gathering scenarios. He illustrates the difference between causation, in which one variable changes in response to another, and correlation, in which two variables move together without evidence of one variable influencing the other. He examines several aspects of data that show the importance of randomization when collecting data.

Transcript

All right, so more often than not in our data-driven setting we're going to have information on multiple variables. And we might be interested in whether or not there's a relationship between some of those variables. So we typically say that variables are associated or related if one variable tends to move with the other one. We can take that association one step further and talk about causation or causally associated if one variable's movement causes the other one to move. There's not just that they're moving together, it's the fact that this one's movement caused the other to move. And so one of the difficulties that we face if A and B are potentially associated is that there maybe some other plausible reason why B is moving, right? So B may also be related to other unobserved variables. And we typically talk about those as confounds. So we have a confounding variable as an unobserved variable that may be impacting our observed variables. And so the presence of a confound limits what we can say about how A and B are causally associated. We can still talk about them being associated, but it limits how we can talk about them being causally associated. If we want to talk about A and B being causally associated we need to think about how we collected our sample. And we need to differentiate between simply a observational study, or a collection of historic data, versus an experiment, right? So an experiment is a setting where we manage the treatments, right? We assign the treatments, and we observe the outcomes. Whereas observational might just be a set of data of historic consumer behavior. And so part of this experimental process is going to hinge upon randomization, right? We're going to need randomization for causation, just like we need

randomization to talk about inference, right?

So ideally I want to have a sample which is where each member of the sample is equally likely to be selected from the population. Then I can talk about inference from the sample towards the population. In our experimental setting we're going to want to make sure that these treatments I assign to variable A are randomly assigned to members of that sample such that I can make some inference on variable B, right? Without randomization, then the presence of confounds is going to limit what we can say about how A is causing B, right? So think of it this way, I've drawn a sample from my population and then I'm going to randomly assign members of that sample to see two different advertisements, ad A and ad B. And then I'm going to measure their perceived brand loyalty, because I've randomly assigned who's going to see ad A versus ad B. Then I can talk about the impact of that ad upon loyalty even if there was other things that I haven't measured which may be impacting loyalty. Because that other things I haven't measured is, you know, equally likely to be distributed across those two different treatments. Whereas if I had a database and in that database I had a listing of whether or not this consumer had seen this ad or that ad then I can't really talk about the impact of that ad upon loyalty because I didn't know how those ads were assigned to each of those consumers, right? So again, randomization is key both in the selection of our sample from our population and in the assignment of treatments to members of that sample, right? Just like I can't make inference to the population, I can't make inference to causation without random assignment of treatments, right? So that's, again, the critical part being randomization.

Back to Table of Contents

Tool: Data Quality and Bias Mitigation

- Data Quality and Bias Mitigation Tool

Use this tool to assess the fitness of your data for making decisions. It will help you recognize sources of bias in your data and make appropriate adjustments.

Back to Table of Contents

Watch: Obtaining a Random Sample

As we've seen, the randomization of data is critical in both the selection of the sample from the population and in the assignment of treatments to members of the sample. So how do you go about obtaining a random sample? Professor Anderson explains how we can use different methods of sampling to make our samples more representative and mitigate sampling bias.

If you need a random sample of existing data, you may wish to download these step-by-step instructions for obtaining a random sample from data in Microsoft Excel.

Transcript

All right. So, now I want to focus on the steps we need to take to generate this representative random sample. At the other end of the spectrum, we have this convenience sample, right? So, it's a sample that is constructed out of convenience. Limiting time, effort. Perhaps I use my judgment in what data I put in my sample or who I ask to participate or maybe consumers have voluntarily filled out some email survey, right? There's some steps I've taken to limit the access of the population to being in the sample. If I think of a random sample, I have this list of the population and I don't limit access, every member of the population has the same chance of being selected into my sample.

We can think of stratified random sample, as I have a series of groups or strata that perhaps I want to make some generalizations across and so I want to make sure that my sample is representative across those strata. So I take my population and I subdivide it into these groups and then I select from those groups such that each member of that subgroup has the same chance of being selected into that sub-sample. And then I combine those sub-samples together to create this sample. And then we have a series of other forms of sampling which are sort of random, they have random elements. And they help us reduce bias by making our sample more representative. One of those is systematic sampling where I have this list of the population and then I randomly pick where I start in

that list. And then after I do that, I sort of select every tenth person on the list to be in my sample. That is different than basically asking every tenth customer that walks by to complete my survey. Right? Because I didn't start with this list of the population, I just had this convenience sample of people walking by. We can think of cluster sampling as similar to our stratified random sampling, except we have these set of groups, so the population is divided into these groups and then we might randomly select a series of those groups to participate in our study or be part of our sample in that we take the entire group into our sample.

So we have a series of groups, we randomly select a subset of those groups. And then those entire groups or clusters become part of our sample. We can think of multistage sampling as combining different methods together. Right? So I might have a series of clusters, right? And then I pick set of those clusters and then I randomly sample from those clusters that I've selected. So I might take the country, divide it into a series of regions. I might pick three of those ten regions and then randomly sample from those three regions that I've selected. So our multistage sampling, our cluster sampling, and our systematic sampling all have some element of randomness to it that help us reduce bias and make our sample more representative. Any time we start to add a convenience nature back to any of those methods, then that's going to potentially induce bias and make our sample and the inference we make from that less representative of that underlying population.

Back to Table of Contents

Recognize Sampling Bias

In 1936, the popular weekly magazine *Literary Digest* published the results of a poll predicting the outcome of the American presidential election. For the first time in five consecutive elections, the *Literary Digest* prediction was wrong. In fact, its prediction was off by an embarrassing 19 percent. This was all the more embarrassing because *Literary Digest* had boasted that its results were based on a better than one-in-five response rate from a pool of 10 million people surveyed. Its method involved sending mailers to households whose addresses were obtained through telephone books, automobile registrations, and magazine subscriber lists. In the years since this historic polling failure, many researchers have studied what went wrong with the *Literary Digest* poll.

In this quiz, you will use what you know about data sampling to answer questions about ways *Literary Digest* could have achieved more accurate results.

You must achieve a score of 100% on this quiz to complete the course. You may take it as many times as needed to achieve that score.

Back to Table of Contents

Case Study: The Poll That Made Gallup

Literary Digest and Gallup

In 1936, the American presidential election was a turning point in the fortunes of two polling enterprises. At this time the well-respected *Literary Digest*, despite its poor performance in predicting the results of the 1936 presidential election, was still considered an authoritative source of information about the will of the people.

The *Literary Digest* method for collecting data was simple and straightforward. It printed surveys and mailed them to millions of households. But it was critical to their success that the households receiving the surveys were representative of the voting public. It was also critical that those responding to the survey would have a range and distribution of opinion that was similar to the opinions of those who didn't respond.

George Gallup, founder of modern polling and the Gallup Poll, predicted that Roosevelt would win and that the *Literary Digest* would get it wrong—and he was correct on both accounts. Gallup regularly conducted polls of a sample of around 2,000 people. Each person was chosen in a random manner to represent a larger group, including all classes, races, and regions. This approach has come to be known as quota sampling. Instead of relying on mail-in ballots, Gallup would send pollsters to talk with potential voters in person. The pollsters would interview the public at work, at home, or on the street. Gallup predicted the correct presidential outcome with a sample of 50,000 respondents, which was a tiny fraction of the *Literary Digest's* response rate of more than 2 million. Gallup's prediction of the outcome error of the *Literary Digest* was also determined by using a small sample. He compared results using one representative from among *Literary Digest* readers for each representative of the voting public.

Back to Table of Contents

Watch: Moving from an Idea to a Data-Gathering Plan

Professor Anderson returns to the marketing campaign of the fictional Snow Hawk Tire company to illustrate the process of using data to tackle a problem. As you will see, this involves asking questions about the data gathered and seeing how the limitations of that data could influence decisions.

Transcript

All right, so let's talk about how to go from our idea to our data plan. You know, one of the key things to realize is that we're always going to have imperfect information, right? If we waited until we had exactly what we needed in order to make this decision, we'd probably never get started. Then we'd never make a decision. So part of this process is just getting some inertia behind getting started. Go down this process of data-driven decision making given the limitations of the data you might have to start with. Get some insight, make that decision, maybe circle back, you know, revisit some of those limitations, collect other data, whatever it might be. But the critical thing here is we need to get started. And so let's put that in context.

Let's look at Snow Hawk tires. The idea here being we're trying to look at evaluating a marketing campaign to increase the sale of snow tires. Our campaign has this catchy title, right? Pay no dough if it doesn't snow. And the idea being here is that there is quite often this barrier to the purchase of snow tires where the consumer has questions about the actual need for snow tires. So is it going to snow this year? Do I need new snow tires? If it's not going to snow, perhaps I can get by without new tires. Right, so we're trying to come up with a campaign that removes that uncertainty about future snowfall from the purchase decision. Right, so now if I no longer have that uncertainty, then I'm more apt to purchase tires and Snow Hawk could grow its sale of snow tires. And so let's look at developing this campaign for next year. So we're looking at our up and coming winter driving season. And our campaign is going to have a series of rebates. And those rebates are going to be

linked to the amount of snow next year. So if it snows less than 20%, say, of the historic average then we'll give consumers back 100% of their purchase price. Now if it snows, say, between 20% and 30% We'll give them a rebate of 75%. Say if it snows between 30% and 40% we'll give them a rebate of 50%. Now if it snows more than 40%, our assumption here is that they needed the tires, and they're going to pay retail.

And so the question is, what kind of data do we need to evaluate this campaign? And so, it's relatively easy for us to grab some historic data. So, we're going to go back and grab the last 75 years of snowfall data. We're going to do that for two cities. One city that typically receives a lot of snowfall, that would be Rochester, New York. And another city Vancouver, British Columbia, which may or may not receive a lot of snowfall given it's on the Northern West Coast. Now the data that's easiest for us to collect is calendar year data. Basically snow fall from January through December. Keep in mind that a winter driving season probably starts in the fall, you know, October, November, December and goes into the spring January, February, March. So its spans across two calendar years, right? But our data is for just one calendar year. So we've already seen some potential limitations in our data. But we're going to use that data to sort of tackle this problem, see whether or not we've got the right sample size, if we think there's going to be any issues across calendar year versus snowfall year. But ultimately play around with some of those limitations of our data, and see whether or not that changes our decision. Right, so we're going to focus on looking at that data in our subsequent sessions.

Back to Table of Contents

Course Project, Part One: Draft a Data-Collection Plan

Once you have data, you can analyze and think more deeply about a decision. But what data do you need? Where will the data come from? And how can you be confident that a decision based on your data will be better than a decision without it? In this part of the project, you will choose a situation or decision and then define key aspects of a data-gathering plan by answering questions about the data you plan to use in your decision.

Project instructions:

Download the Course Project document and save it to a convenient location on your computer. Open it and complete **Part One—Draft a Data-Collection Plan**.

Do not hesitate to contact your instructor if you have any questions about the project. You will add to this document as the course proceeds and will submit it to the course instructor at the end of the course.

This is a required part of the final course project; completion and submission of all parts of the course project will be required at the end of the course in order to achieve credit.

Before you begin:

Before starting your work, please review the **rubric** (list of evaluative criteria) for this assignment and eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Module One Wrap-up: Gather and Qualify Data

The way you frame your problem or decision will have a significant impact on how effectively you can use data to inform it. In this module, you took the important first step in planning a sound data-driven decision. You articulated an important decision, identified key stakeholders, and determined which key performance indicators were important to these stakeholders. You considered what data you could use along with factors that could ensure or limit the effectiveness of the data. As you continue, you will look more closely at how you will analyze and package the data you've identified in order to get the greatest insight from it.

Back to Table of Contents

Module 2: Visualization and Analysis

Back to Table of Contents

Module Two Introduction: Visualization and Analysis

We're all familiar with the concept of an average, and to many the idea of a standard deviation will be familiar. But when is it appropriate to use these in making judgments, and when isn't it? To make good data-based decisions, you need to be able to see the meanings of relationships between variables or trends in variables. For instance, when do you compare the absolute magnitude of two values, and when is it more useful to look at the proportion of them? It's critical to understand how to structure data summaries and how to compose visualizations in a way that emphasizes features of the data that will be useful.

In this module, you will explore summary statistics and commonly used visualizations. You will also learn how to generate them, fine-tune them, and draw conclusions from them.

Back to Table of Contents

Watch: Summarizing Data

Recall that we have a sample taken from the population, consisting of quantitative and qualitative data. Now Professor Anderson will present ways to unpack and describe the meaning of those two types of data. He will introduce two critical concepts that we will be using in this course: central tendency and spread.

Transcript

All right, so we have this sample from our population. That sample consists of quantitative and categorical data. We're going to look at different ways to describe those two types of data as part of our inference-making puzzle. So, for our quantitative data, right? We're really going to focus on, sort of, two aspects. One is called central tendency, right? So that is what things tend to happen, right? So what is the usual outcome? And then we also were going to talk about spread or risk or uncertainty. So what are the ranges of possible outcomes? Right? For this spread part, we're going to, in addition to summarizing it, we're also going to visualize it.

Right, so we're going to use some graphs and figures to help us understand that distribution of possible outcomes. For categorical data, life is a bit simpler, right? Because if we're trying to numerically sort of describe a categorical situation there's obviously less we can say numerically. And so what we're going to do is focus on basically proportions or fractions. So how does our data fall out across these different categories. And we'll look at that both, in say a tabular form. But also in figures, to get a sense for that visual presentation of those proportions.

Back to Table of Contents

Watch: Using Summary Statistics—Central Tendency

Previously, Professor Anderson introduced the concept of central tendency, which refers to likely outcomes, or what we might expect to happen. Now he will further explore central tendency: how we can think about it in more than one way and how it relates to our efforts to use summary statistics to inform decisions.

Transcript

So when looking at a quantitative variable, one of the key things we're quite often interested in is what's a typical observation or what might we except to happen, right? Or an average what happens? This idea of central tendency. And so the average, or the mean, is one of the classical sort of summary measures of central tendency where we simply have the sum of all observations divided by the number of observations. Another measure of central tendency is the median. Where the median is the data point in the middle of our ordered sample. So if we take our sample and sort it from smallest to largest, what is the value right in the middle? So the reason we like to have more than one measure of central tendency is this point of what we call resistance.

So because all our observations go into the calculation of the mean, if we had some outliers those outliers might impact the sort of value of that mean. Whereas if I have all my observations and one of my extreme values, it might be an outlier, if I exclude that from my sample, if I have appreciably large sample, then the data point in the middle is still going to be relatively close to where it was with the inclusion of that extreme event. And so we think of the median as being resistant to those extreme events and potentially the mean or the average being a function of those extreme events.

So let's take a set of recent box office receipts as an example. So we're looking at a series of recent movies and what was their annual revenue? So Star Wars was a gangbuster, right, generated over $1 billion in revenue. If we include Star Wars as part of our sample, then we have an

average revenue per movie of 55 million. Now, if we drop Star Wars from our sample because it was so large, then our sample average drops 10% down to 50 million, right? So, that sample average is susceptible to that extreme event. Whereas if we focus on the median, if we look at all movies, right? All those recent movies, and order them from smallest to largest and look at the one in the middle. And if we drop Star Wars off and then look at the one in the middle of the remaining movies. And that remaining movie is still right around 20 million, so the median is right around 20 million with or without the inclusion of Star Wars.

So we think of the median as a measure of central tendency, which is resistant to extreme values whereas the mean, a more typical measure of central tendency, but influenced by those extreme values. So if you thought you might, your sampling process, or your data itself might have some of these extreme events, then the median would adjust for some of those extreme events in its description of that central tendency.

Back to Table of Contents

Watch: Using Summary Statistics—Measures of Spread

The term spread in the context of statistics refers to the range of possible outcomes and the risk associated with that range, as Professor Anderson explains here. When you are working with summary statistics and trying to interpret meaning, it will be helpful to understand some of the ways to think about spread.

If you need to work with summary statistics, you may wish to download these step-by-step instructions for generating a summary statistics table in Microsoft Excel. The instructions guide you through creating a table that includes minimum, maximum, mean, and standard deviation.

Transcript

When looking at a quantitative variable, spread is one of the critical measures of describing that variable. So typically, what's the set of possible outcomes, or understanding the risk associated with that particular quantitative variable. Standard deviation is probably the most well-used measure of spread for a quantitative variable. So just like the mean is the average outcome in our sample, right, the standard deviation is really like the average about the average, right? So on average, how far are individual observations away from the average? So it's a nice description of the spread amongst our sample. And so the larger the standard deviation, the bigger the average about the average, the more spread there is in our data. And we typically compare the standard deviation to the average, right? So if the standard deviation is much bigger than the average, then there's a lot of variation in our data.

If the standard deviation relative to the average is smaller, then there's not a lot of variance in our data, or not a lot of spread in our data. We have other measures of spread. A classic one is range, right? The difference between the largest and smallest value. Now when we talk about the range for the standard deviation we're using our entire sample, right? So because we're using our entire sample, those summary measures are a function of extreme events or outliers. So if we have had

some outliers, then the range, the difference between the biggest and the smallest, or the average spread above the average, both of those metrics are going to be susceptible to those extreme events. And so we are possibly looking for measures of spread, which are resistant to those outliers.

And so one of those is what we call our interquartile range, right? So just like we have our max and our min, right? Taking the ends of our data, and the median being the middle, we can further subdivide our data into quarters, or quartiles, and think of the third quartile as the point that's halfway between the mean, and the max, and the first quartile is the data point that's halfway between the min, and the mean. And then we call this thing interquartile range, which is the difference between the first quartile and the third quartile. And so because we have excluded the ends, right? That interquartile range is not a function of those extreme events and so as a measure of spread, but is resistant to those extreme events, right? And so it's nice to sort of have both measures of spread, which are inclusive of all the data. But also measures of spread which are exclusive of some of those extreme events.

Back to Table of Contents

Watch: Making Decisions with Summary Statistics

Professor Anderson will now discuss using summary statistics to make a data-driven decision regarding one quantitative variable. He begins by imagining a fictional scenario—we are working for an airline and trying to use data to manage insurance costs—to show how we can use statistics, compiled over time and related to several variables, to help inform the best choice.

Note: In this video Professor Anderson exposes some of the mathematical calculations as he works through an example. You need not be too concerned with the mathematical details, especially if you are not working in a technical capacity. It is enough to follow in a general way the concept being discussed.

Transcript

All right, so let's focus on making a data-aided decision for a single quantitative variable. So our setting is insuring our fleet of aircraft. So we're a regional airline looking to manage our costs. On a regular basis, some of our planes receive incidental damage, perhaps hitting a bird while flying. And occasionally, or very occasionally they incur major damage. Planes are very expensive, so therefore we're looking at some insurance policies to help us mitigate our costs as a function of that damage. Because these different policies that we are going to look at they may have different structures, different premiums, different, coverage, different deductibles.

So again, our population would be all past and future, sort of damage costs, right? Costs of damage to our fleets. We're going to have a sample, which we're going to use to make inference in our decision. Our sample's going to be the last 20 years. And what would have been our costs, had we deployed each of these insurance policies across those 20 years. As a baseline, let's compare those policy costs to actually having no insurance, or self insurance, right? So that'll be like kind of our baseline. So here we have our four policy costs and our costs associated with self insurance. We have four summary statistics, the minimum cost,

the maximum cost, the mean cost, and the spread of those costs or the standard deviation of those costs. Let's look at a couple of policies.

All right, let's look at policy A. Specifically it has an average cost of 89.5 million and a standard deviation of 4.6. If we look at policy C it has an average cost of 51.3 and a standard deviation of 3.1. So here we could say that policy C dominates policy A because the 51 is smaller than 89 And the 3 is smaller than the 4.5. Right? So on average it costs less and it also has less risk. So we would say it dominates A so let's just toss A away. It's not longer in our set because C is much better. And then let's look at policy B. Policy B has an average cost of 38.5 And a standard deviation of about 15. If we compare that to self insurance, self insurance costs less than half, right? So 18.3 million. And it's only a little more risky, so 18.8 versus 15. Basically, policy B is not much of an insurance policy. So it has similar risk to no insurance but has a much higher cost, right? So, we would probably also throw out policy B because it's really not much of a policy. So now we have policy C, D, and self insurance.

Now, again self insurance has this max cost of almost $138 million. Odds are we probably can't afford that on a regular basis, we may go bankrupt, we're probably not going to look at self insurance, because it's just too risky. So let's shelf self insurance for now. And so let's now focus down on policy C, and D. So policy C and D have similar minimum costs. You know 49.4 versus 42. Similar maximum cost 88, I mean 87.9 versus 84.6. And their average cost are also similar 51.3 versus 50. All right? But their standard deviation or spread is quite a bit different, right? Policy C at 3.1 versus Policy D at 6.6. Right? So although Policy D is just a little bit cheaper on average, it's considerably more risky. So policy C is probably our best, viable policy. So it does a really good job at reducing risk, it is the least risky policy, right? 3.1 is the smallest of all policies. And its costs are reasonable at 51.3. So, just focusing on these four summary statistics provides a fair bit of insight into which policy is most efficient for our carrier.

Back to Table of Contents

Using Summary Statistics

In this quiz, you will answer questions related to the use and interpretation of summary statistics.

You must achieve a score of 100% on this quiz to complete the course. You may take it as many times as needed to achieve that score.

Back to Table of Contents

Tool: Visualizations Guide

- Visualizations Guide

Use this guide to choose the visualizations most appropriate for the data you are working with.

Back to Table of Contents

Watch: Interpreting Visualizations

Understanding how to use visual tools to illustrate and illuminate quantitative variables offers a number of benefits: it can help us more quickly decipher meaning and can help us decide what the connections are between different variables. Professor Anderson examines this more closely through a narrated example here.

If you need to create visualizations, you may wish to download one of the following step-by-step instruction sheets for generating visualizations in Microsoft Excel.

Create a histogram.
Create a dot plot.
Create a box-and-whisker plot.

Transcript

Okay, so let's focus on visualizing our quantitative variables, right? And so you remember that classic saying, a picture is worth a thousand words? So we're going to use pictures or visualizations to help us augment our understanding from summary statistics. So our visualizations are going to help us understand the range and likely value of outcomes. So let's use airline ticket purchases as an example, and here we're going to focus on how many days prior to departure did the consumer purchase their ticket. The first visualization we're going to look at is a histogram. And so, for a histogram what we have is a series of bars, right? Across our horizontal axis we've taken our variable of interest and put it into a series of bins or buckets. Right? So basically the seven indicates all consumers who purchase a ticket seven days or less prior to departure. The 14 is all consumers who purchased a ticket from seven to 14 days. The height of those bars are the number of those said purchases.

And so the histogram helps us visualize over this potential series of days before departure when, or most often when, are tickets purchased. So,

we see a lot of activity towards the end, and they sort of decay down with this sort of little bump around two months prior to departure. A dot plot is like a histogram but just with the refined X axis, right? So here we focus on those last 30 days, right? And our X axis is subdivided into those 30 days. The height of our dots are simply the number of people who purchased at zero days, one day, two days, et cetera. So here we're sort of having a refined look at that distribution. And now our airline could use this refined look to sort of get a sense of, okay, what's driving consumers to purchase right around three weeks before departure and then this big flurry, 10 or 11 days? And then why are so many purchasing the day of? Right? So on zero, even prices are probably high, but there's a lot of demand then, so maybe prices could be higher. So lots of inference you might pull out from that. A boxing whisker plot is our third type of visualization, and really all it is a visual presentation of some summary statistics. Right?

So we have our median in the middle, and then our interquartile range, the first quartile to the third quartile, around that is a box. And then, we have these whiskers reaching out to the min and the max, right? So, it's just a visual presentation of some of those summary statistics we calculated earlier. So if we look at that visual representation for our airline example, we see here how tight the first quartile and the median are, right? So there's not a lot of difference between the middle and that upper 25%, right? There's much different, a much larger difference for that next 25 as we go from the median to the third quartile. And then there's this huge jump for that last 25% as we go from Q3 to the max. Right? So while it's just a visual of those summaries, it is informative and helps us look at the relative size of those numbers. So to summarize, I mean these visualizations of quantitative data, whether it's a histogram, a dot plot, or a box and whisker plot, are all about gaining some visual sense of the range and likelihood of potential outcomes. You know, sometimes we refer to this as skewness. So how do things sort of happen over that range? When do they concentrate it? Are they all concentrated, or does it tail off? Or are they concentrated early on and then dropping off? So getting a sense of that distribution of possible values.

Back to Table of Contents

Watch: Summarizing and Visualizing Categorical Data

Bar charts offer an easily digestible method of summarizing and visualizing a categorical variable. They can be especially helpful when you seek to understand proportions, frequencies, or how the data is partitioned across categories. In this video, Professor Anderson uses sample survey data related to consumers' cellphone usage to explore the ways that using bar charts can lead to new insights.

Transcript

All right so let's now focus on summarizing and visualizing a categorical variable. So I guess kind of by definition there is less to say quantitatively about a categorical variable. Our main focus really is on proportions or frequencies. So how is our data partitioned across the different categories that we're measuring? So, let's focus on a simple survey. So in 2011 Neilson randomly sampled a series of US adults and asked them for their phone usage, right? Specifically, their cellphone usage, were they using a smartphone or feature phone? If they were using a smartphone, what type were they using, right? And so we can take that data and we can describe the proportions across those different categories of phone use.

And so we could first look at sort of feature phones versus smartphones, right. So what fraction, or what proportion of cell phone users are using a smartphone versus a feature phone. And then we could dig deeper. Of that 40% that are using a smartphone, what fraction are using Android versus BlackBerries versus iOS. So, we could just subdivide that category into a series of sub categories and present both of those visually, right? We do that visually with this, what we call a pie chart or a pie graph, right? Because it's very visual. Very easy for us to see. The relative sizes of the pies communicate a lot, Not just the data points. We could also look at another way you can visualize these proportions. That quite often is done with bar charts. Bar charts are nice if we have, say, this category moving over time, right?

So here we have, again, a different survey. This survey was conducted in three years, 2011, 2012 and 2013. And we have a set of bars for each of those years, right? And we can see how the type of phone use is evolving over time. So now we're comparing not just the bars across the three categories, but the bars across time. So bar charts are nice to sort of see how categories are evolving, in this case with time, but they could be evolving across another category as well. One thing to keep in mind is even though a bar chart looks a lot like a histogram they are in fact, different, right? A bar chart is for categorical data, where the X axis is the actual categories, as defined by the categories, and has no numeric value. Whereas a histogram, the X axis is defined by your data, right? So you subdivide that data into whatever sort of bins or buckets, and that is what's translated into that X axis, right?

So categorical bar charts, quantitative histogram, X axis has meaning in a histogram, a quantitative meaning, sorry. And in a categorical sense, the bar chart also has meaning, but it's predefined by the actual categories. So in summary, you know, less to say about categoricals, it's all about proportions or frequencies or percentages, right? We could show those as a table, but more often than not, we use some sort of visualization because it's easier for consumers of our data to understand those pie charts or those bar charts.

Back to Table of Contents

Identify Visualizations for Statistics

In this quiz you will answer several questions related to choosing visualizations for data.

You must achieve a score of 100% on this quiz to complete the course. You may take it as many times as needed to achieve that score.

Back to Table of Contents

Watch: Visualizing Two Categorical Variables

Two common tools for visualizing categorical variables are bar charts and frequency charts. In this video, Professor Anderson discusses sample data presented in both a bar chart and in a frequency chart. This illustrated example helps conceptualize how to look at one category in relation to another to see what insights you can draw from them and, importantly, what some of the other questions are that you might need to ask.

If you need to create a chart for two categorical variables, you may wish to download these step-by-step instructions for generating grouped and stacked bar charts in Microsoft Excel.

Note: In this video Professor Anderson exposes some of the mathematical calculations as he works through an example. You need not be too concerned with the mathematical details, especially if you are not working in a technical capacity. It is enough to follow in a general way the concept being discussed.

Transcript

All right, so now let's talk about how to look at two categorical variables together, right? So let's focus initially on bar charts. So just like I could use a bar chart to look at one categorical variable I can look at groupings of a bar chart, or a grouped bar chart, to look at how two categorical variables are interacting.

So, let's take our cell phone example. So we've surveyed consumers over three subsequent years. Each one of those surveys has asked them for what is the operating system on their cell phone. And so in our grouped bar chart, our first grouping looks at one of those operating systems and how it has moved over time, right? So what proportion of consumers are using that OS and how that proportion has changed over time. And then our next grouping looks at a different OS and then how that OS has moved over time, right? So, grouped bar charts are a great way to look at maybe these two categories; focusing on one category,

and how it changes with the other.

We can look at a stacked column chart as another way to look at two categorical variables. And what we have here is, again, a column chart. But the column is split into proportions across one of our categories, so you could think of that column as being broken down across the different cell phone operating systems, right, so in the proportions. And then we have an adjacent column and that adjacent column is those same categories. But those proportions for the next year, right? So now we can visualize how the categories together are moving from one time period to the next, right, or from one category to the next, right? So a grouped bar chart is great for looking at how one category is changing with the other category, and a stacked column chart is a great way to look at how all three of these categories or four categories, however many in that first category, are moving with the other categorical variables. Seeing them moving together versus just this guy and then the next one.

Another way to look at two categorical variables is a frequency chart. So instead of a graphical visual, we have a tabular visual. And in our frequency chart, we have the frequencies of the category pairs. All right, so let's say for example we surveyed 120 of our employees. Right? Those 120 employees are broken down into three categories, right? Frontline, sales, and administrative, so across their job type. And then we've asked them whether or not they're in favor of a wage freeze, or layoffs, right?

So we have two categorical variables, what is our job type? And then what is our preferred method of cost reduction? So pay freeze, or layoffs. And so the frequency chart becomes a nice visual for us to sort of ask some questions. We see here that, 43 out of those 120 employees prefer a wage freeze versus layoffs. But what's interesting is that if we focus on frontline staff, right, 8 out of 40, right, or 25% of those employees in favor that's pay freeze, right, versus 32 out of 40 or 75% favoring layoffs. But then that's flipped if we look at our administrative staff, right, where 5 out of 20 of those are in favor of layoffs, right? The bulk of them, the 15 or the 20, would much prefer a wage freeze, right? So we can start to ask some questions about why would our administrative staff look at a wage freeze different than our frontline staff, right? So a great way to calculate those proportions and start to ask those sort of inquisitive questions.

Watch: Visualizing Two Quantitative Variables

Scatter plots allow you to visualize the association between two quantitative variables. In this video, Professor Anderson examines sample scatter plots and shows the benefits they offer. Note how easily you can see how the variables are related to one another, how they move together, and what these associations mean.

Transcript

All right, so here we're going to talk about looking at two quantitative variables, right? So just like with the single quantitative variable, I can talk about some summary statistic, in this case we're going to talk about correlation, or I can talk about a visual or graphic, here we're going to focus on a scatter plot. When we're talking about two quantitative variables, we're really focusing on the association. How are they related, how are they moving together. So a positive association means that as one goes up, the other tends to go up, or if one goes down, the other tends to go down. A negative association means that when one goes up the other goes down. And two variables are not associated, if basically one goes up, and the other could up or down, right? So there's no relationship between how these two quantitative variables are moving, right?

So let's look at a simple example, right? So we're looking at an example from the shop floor. We have a series of different jobs that have gone through our process, right? So, each case or each row in our data set is a different job, that job has a series of attributes or a series of variables, right? So, how many particular pieces that we're working on, or how many parts are part of that job. How many operations that job went through in the floor, whether or not we expedited it through the shop floor and then lastly, how long it took to complete that job. And so, we could look at correlations specifically between time as one of our quantitative variables and then either parts or observations. And we see here that the correlation between in time and parts is 0.74 and the correlation between time and operations is 0.23, right? So a much, so a stronger correlation with parts but still a positive correlation with operations. Typically, those

correlations range from minus 1 to 1. Right, so the closer we are to 1 the stronger we talk about. The closer to minus 1, strong but a negative association. Now, we can look at those correlations visually. Right, so here is what we call a scatter plot. Where we still have our variable of interest lets call it time on the vertical axis and on our horizontal axis we have one of our other variables of interest. We could have parts. So we have a visual of how time varies with parts. So we get a sense for if we have more parts in our job it takes longer to complete. The nice part about scatter plots is we can also look at associations between two quantitative variables. And how they differ for a categorical variable, right?

So remember one of our variables was expedited or not, did we rush this job through the shop floor? And so here we have that same scatter plot but we have two different series. One series for the jobs that were rushed in a different series than the jobs that were not rushed, and so now we can see the relationship between these two quantitative variables and how they differ across this category. So we're simply having different scatter plots or different lines for the different categories on our scatter plot, right? So a nice way to visualize some of those associations between those two quantitative variables.

Back to Table of Contents

Watch: Visualizing a Categorical Variable Together with a Quantitative Variable

Different visualizations show us different things, and they enlighten us in different ways. Now Professor Anderson offers some examples of the ways in which you might examine different variables together, across categories, and some of the advantages of doing so.

Transcript

So sometimes we're looking at both quantitative and categorical variables together. And so we can look at that quantitative variable from a summary measure standpoint and look at those summary measures across the different categories of that categorical variable. So an instance, look at means across the different categories. Or we could look at some visualizations and compare those visualizations across the different categories. So look at box plots or histograms of those quantitative variables across the different categories. So let's look at a couple of examples to illustrate. So here we're trying to evaluate four insurance policies. Each of those four policies has a different structure. That structure results in different costs to the firm. And so if we were trying to decide if policy A was better than policy D or vice versa we could simply look at some summary statistics for those two policies. Now we see that policy D is less expensive than policy A, which is great, but it also is a little bit more risky because it has a higher standard deviation than policy A. So just looking at those two summary statistics, it's not necessarily clear whether policy D is better than policy A.

So we move on and look at some visualizations of the cost across those two different policies, right? So here we have categories of policy. We're going to compare those visualizations across policy A and policy D where we look at here a box-and-whisker plot of those expenses. And so we see here now it's very clear what's going on with those costs. If we look at the middle part of the box-and-whisker plot which is between the first and third quartile, for policy D that is considerably lower than that corresponding region for policy A. But more importantly if we look at the max whisker for policy D it is actually lower than the min whisker for

policy A, right? So the risk we thought we saw in the standard deviation is really not there at all, because in fact that policy D is much less riskier, because its max is much smaller than even the min cost of policy A. If we look at another example, here we're looking at mobile ad performance. So, we have an app on a mobile phone, and when customers are in that app we show them different ads. And so we have a potential new ad server we're trying to try out. And so we do a series of experiments. So we randomly send a series of consumers over to see ad B. Other consumers are seeing ad A. And we look at, on a daily basis, what fraction of consumers click on that ad for both ad B and ad A. And so we do that over a series of days and we can look at some summary statistics for those ad performances, right?

So our two categories here are A and B, for treatment A versus treatment B. And what we see here is that treatment A has a higher click-through rate. So on average a higher fraction of consumers who see ad A click that ad then those that see ad B. So at some level it looks like this new proposed ad server is less effective than the old one. But when we dig a little bit deeper and look at some histograms we see some interesting confounds, right? Some issues here. Where all of a sudden, the range of outcomes for ad server B is much tighter. The histogram of potential click-through rates is much closer to the mean for ad B than it is for A, right? So, while on average more people click-through when they see ad A, consumer behavior is much more consistent with platform or ad b. So this sort of enlightens us. Perhaps we want to dig a little bit deeper into why ad B is underperforming but consistently doing so. And so, depending on what you're trying to get at different visualizations will show different things, right? So, our histogram is a little more refined look than our box-and-whisker plots and we could go one step further and even compare dot plots across these different categories.

Back to Table of Contents

Watch: Interpreting Initial Visualizations and Identifying Issues

Professor Anderson uses a fictional example to demonstrate how you could work with summary statistics and initial visualizations to identify where to focus your attention. He discusses the case of a marketing campaign for the fictional Snow Hawk tire company, with sample statistics to show the kinds of issues that can surface at this point.

Note: In this video Professor Anderson exposes some of the mathematical calculations as he works through an example. You need not be too concerned with the mathematical details, especially if you are not working in a technical capacity. It is enough to follow in a general way the concept being discussed.

Transcript

Let's focus on interpreting our initial visualizations and summary statistics, and use those to help us identify issues that we want to dig into deeper. Right, so Snow Hawk is looking at a marketing campaign for the upcoming winter driving season. And they're trying to incentivize consumers to buy more snow tires by removing the uncertainty of snowfall from the purchase decision. So if it snows less than say 20% of the historic average snowfall, we'll fully refund your tire purchase. So we have a sample, that sample is the last 75 years of annual snowfall. We have that data for two cities, Rochester and Vancouver, all right? We can describe that quantitative variable of snowfall using some summary statistics across our two categories, Rochester and Vancouver, where our min, our max, our average snowfall and then we have our standard deviation as some summary of the uncertainty in that snowfall. To really understand that uncertainty in snowfall we plot histograms. So we have a histogram for Rochester and another histogram for Vancouver. What we see here when we look at these histograms is some very different behavior. So Rochester we tend to have a lot of massive snow fall in and around that core average. So a lot of. I mean a lot of the snowfall tends to be very centralized, right? And as a function of that we have a smaller standard deviation relative to the average. If we look at Vancouver there

is a lot more spread in that snowfall. And we see that spread because the standard deviation of snowfall in Vancouver is basically very similar to the mean of the snowfall in Vancouver, right? So we see the similarities of the mean and the standard deviation of Vancouver and we see that the standard deviation in Rochester is much smaller than the mean, right? So a lot more variance in snow fall in Vancouver than Rochester. So we would expect probably more rebates in Vancouver because there's lot more uncertainty in snowfall.

So now, using our sample, we can come up with our critical points to evaluate our campaign. So what's the average snowfall over the last 75 years for Rochester? What's that same average for Vancouver? And then what's our critical rebate levels, the 20, the 30, and the 40. So what's 20% of that average, 30% of that average, and 40% of that average for each of our two cities. Then we can go back to our data and count how many times it snowed less than 20% in Rochester. How many times it snowed between 20 and 30% in Rochester. How many times between 30 and 40. And the same thing for Vancouver. Right. These become the frequencies of our different rebate levels. If we take these frequencies and divide by the total size of the sample that we're using to calculate those frequencies, we have the relative frequencies with the percent of time these different rebates were provided. If we sum up these three percentages for each city, that would be the percent of consumers receiving a rebate across both cities. If we actually take the percent of consumers receiving a rebate, and multiply that by the rebate level, whether that's 100, 75, or 50, we do that across all three levels, we'll get the actual average rebate provided to consumers. So now we have these two sort of summaries of our program, what fraction of consumers receive a rebate? And on average, what is that rebate? And we'll notice here that the rebates are much smaller in Rochester versus Vancouver. So both the number of consumers receiving a rebate and the average rebates themselves. That's a function of the increased variance in the snowfall in Vancouver.

Remember the standard deviation relative to the average was much bigger in Vancouver or Rochester so we have a wider range of snowfalls and a potential bigger impact of our program. Now the question is, is that what we're looking for, right? Is the program a good idea, right? So did we use the right sample to evaluate our program? Did we use the right

sample to set the critical program characteristics? That being these average snowfalls that we link the 20% and 30% and 40% to. So what are consumers going to think of that sample we used to set that critical number? So at some level we need to sort of go back and forth with what our sample is, what years we used to put into that sample and how that impacts our program.

Back to Table of Contents

Activity: Interpret Statistics and Visualizations

In this activity, you will use what you know about summary statistics to make arguments for and against a hypothesis. You will present one of these arguments in a discussion later in this course.

The data shown were gathered as part of a gender discrimination lawsuit against Harris Bank. The claim was made that the bank compensated female employees at a significantly lower rate than male employees with comparable education and experience.

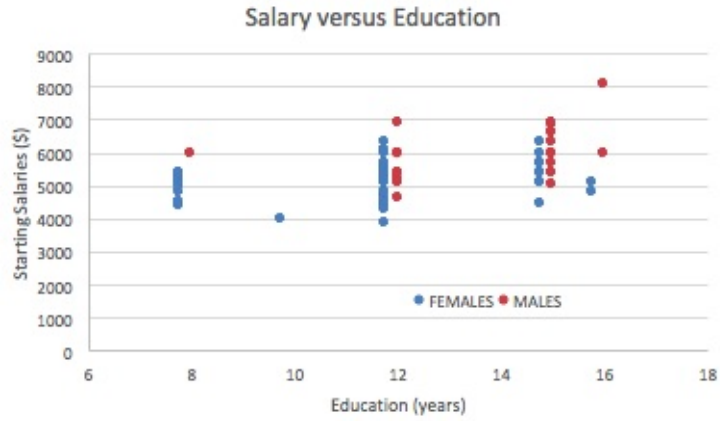
Using the summary statistics and/or the visualizations, make an argument in favor of the claim. Then, make an argument against the claim.

Record your two opposing arguments. You will share one of them in a class discussion.

Summary Statistics
Graph: Salary versus Education
Graph: Salary versus Experience

Summary statistics

Gender	Average Starting Salary per Month	Average Education (years)	Average Experience (months)
female	$5138.90	12.0	100
male	$5956.90	13.5	103



Back to Table of Contents

Analyze Based on Data Summaries and Visualizations

Instructions:

You are required to participate in both of the discussions in this course. You will not be able to see other students' posts to this discussion until you have made a post yourself.

Discussion topic:

Earlier in this module, you made arguments for and against discrimination based on data used in a lawsuit against Harris Bank. In this discussion, you will present one of your arguments and support it by referring to data summaries and/or visualizations.

Create a post in the discussion board that:

States whether you support or refute the claim that Harris Bank paid female employees significantly less than male employees with comparable experience
Makes a compelling argument in favor of your position
Refers to specific features of the data summaries or visualizations as part of the argument

To participate in this discussion:

Click **Reply** to post a comment or reply to another comment. Please consider that this is a professional forum; courtesy and professional language and tone are expected. Before posting, please review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Course Project, Part Two: Identify Data Summaries and Visualizations

Project instructions:

Open the Course Project document you started in the first part of the course and complete the assignment for **Part Two—Identify Data Summaries and Visualizations**.

Do not hesitate to contact your instructor if you have any questions about the project. You will add to this document as the course proceeds and will submit it to the course instructor at the end of the course.

This is a required part of the final course project; completion and submission of all parts of the course project will be required at the end of the course in order to achieve credit.

Before you begin:

Before starting your work, please review the **rubric** (list of evaluative criteria) for this assignment and eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Module Two Wrap-up: Visualization and Analysis

You have an abundance of choices when it comes to analyzing and visualizing data. Some ways of expressing the characteristics of your data set will serve you better than others. The more experience you gain working with data, the better versed you will become in making the best choices when preparing your data for analysis.

This module introduced the two basic data characteristics of central tendency and spread. You considered these characteristics as you examined different options for summary statistics and visualizations. You practiced selecting and interpreting both summary statistics and data visualizations. You constructed arguments based on data visualizations, and you considered which summary statistics and visualization could help you better understand a decision-making situation in which you are involved.

Back to Table of Contents

Module 3: Bring the Data into the Decision

Back to Table of Contents

Module Three Introduction: Bring the Data into the Decision

If your data has been gathered without bias and is presented in a way that exposes the features that are most important to your decision, you are ready to bring the data into your decision process.

In this module, you will engage in a process that tests your data against the real world problem you are trying to solve. In part, this involves reconsidering whether your data can help you gain the needed insights. At the same time, you'll explore how an interactive dashboard can be used to manipulate parameters of the data to understand the situation more clearly in terms of the key performance indicators.

Back to Table of Contents

Watch: Using a Decision-Making Framework

Whenever you need to make a decision, you typically use an analytical process that includes some form of information gathering and a method of coming to an opinion or conclusion. In the video below, Professor Anderson discusses a data-driven decision-making process in which a model is used to analyze the data and insights are drawn by manipulating that model.

Transcript

All right, so now I want to talk a little bit about our data approach to decision making. So what is the sort of data-driven decision making process? If you think about making decisions in general, we think about, you know, generating your options, whatever that might be, some sort of brainstorming, or path finding, whatever that might be. And ultimately you're going to conduct some analysis, and then make some recommendations. So think of that as just a general approach to making a decision.

What we do now in the data-driven decision making process, is we focus on that analysis step. And in that analysis step, we really have these kind of three parts. I'll call it data, models and insight. So taking that data that we've extracted from our population, right, then building some models that we used to analyze that data, right, visualize it per say, and then ultimately using our knowledge of the context and our insight about the setting to feed back and make that recommendation. The models we've talked about so far are largely, you know, descriptive and visual in nature, right? But obviously, we could have lots more refined models. Those models could focus on probabilities that come out of our histograms, right? We could focus on really refining those associations and making some predictions about future outcomes as a function of those causal relationships that we've sort of measured.

And, ultimately, where we would go even further down this process, and come up with some optimal decisions through some sort of prescriptive approach, where I've used some mathematical tools, to help me decide,

all right, given the relationships, what should I do? But all that modeling is within the context of your insight about the setting, right? Your intuition about how things are linked, your questioning of the data, questioning of its validity. Linking that back to the context, right? And how we relate to that initial question you were asking, right? So data-driven decision making is no different than decision making in general. It's just what we do in that analysis part of the process.

Back to Table of Contents

Watch: Interacting with the Data-Model Abstraction

Now Professor Anderson will discuss the data-model abstraction and how you can draw connections between the abstraction and your own specialized knowledge and insights. By interacting with the abstraction, you're really asking probing questions: What are the insights to be drawn? Do the outcomes make logical sense? And how can you apply these insights back to making a decision?

Transcript

All right. So now, I want to focus on this abstraction, right, of our data model insight framework from our setting, right? One of the key things to realize is that the data is not going to make the decision for you, right? That ultimately we're putting the data and the models in the hands of the individual with the insight in order to interpret, ask the right questions and ultimately make that decision, right? So we have, I mean, we're going to formulate these questions we're going to want to ask, right? We're going to have some insight into how we might measure, right? What we might measure and how we might measure. Those KPIs as they relate to that question. Then we go through this process of collecting data, right? As we talked about earlier, we're going to sample that population, right? And, depending upon whether or not we're focused on association versus causation, we may perform some experimentation on that sample.

Right, the tools we focused on so far are, you know, relatively descriptive in our analysis, right? So we're describing that data, we're visualizing it, trying to get a sense for some of the uncertainty and those, you know, central tendencies. But ultimately, we're going to focus on the insight that we can pull from that visual understanding of the data and how it relates back to those KPIs, right? Such that we can ask the right questions, right? We can think about, so we have this abstraction, right? We have this data models. We have some output from our models, right? We're going to visualize that data, ultimately perform we typically call sensitivity analysis, so how sensitive are some of those outputs from the sample itself and what subset of the sample I use. Are the outcomes from my valid sensical? Do they make sense? Some kind of what we refer to as

validation. Getting a sense for. I mean logically, given my insight from the setting, does this makes sense? Once we've confirmed that, we can think about pulling this back towards the real world in making this decision, right. So this data model part, this abstraction, right, is really hinging upon your knowledge and insights and how well this abstraction really links back to the real world. It's that how well it links is going to limit the degree to which that decision you make is relevant to your underlying setting or the real world.

Back to Table of Contents

Read: Case Study: The Billboard Effect—Using a Model to Determine Value

Online Travel Agent Impact on Non-OTA Reservation Volume

A carefully crafted data model generates insight into a question. In this article, Professor Anderson uses statistical summaries to examine the effect of an online travel agent (OTA) on average daily reservations at hotels. This article shares insights gained during a three-month experiment in which hotel properties were added to and removed from the Expedia.com travel website. The underlying question was, Does listing a hotel with an OTA provide a benefit to the hotelier beyond the reservations made through the OTA, and if so, how great is the benefit?

To read about the experiment, download The Billboard Effect.

Back to Table of Contents

Tool: Data Qualification Checklist

Use this checklist along with the data-models-insights framework to help you determine whether the data you have is suitable to answer the questions you are interested in.

Back to Table of Contents

Determining Data Viability

In this quiz, you will consider a fictional scenario. You'll be provided some historical data and several questions related to that data. You will select one of the questions to focus on and discuss how the available data does or does not support answering the question.

This is a practice quiz. Your response will not be reviewed by instructors, and the results will not affect your grade in this course.

Back to Table of Contents

Watch: Working with Dashboards

When you need to make a data-driven decision, you want to visualize in an interactive way the data influencing that decision. Typically you will create a dashboard that consists of relevant summary statistics and data visualizations, a summary of key performance indicators, and levers you can manipulate to make changes that reflect the tunable parameters in your decision. If constructed usefully, this dashboard allows you to experiment with your data until you find an acceptable outcome.

In this video, Professor Anderson uses a dashboard to consider how best to fulfill expectations of stakeholders in the fictitious Snow Hawk Tire campaign.

If you need to create a dashboard for your data, you may wish to download one of the following step-by-step instruction sheets. Each sheet guides you through a procedure that contributed to building the dashboard you see in the video.

Create a pivot table.
Create a slicer.

Transcript

All right, so now let's focus on this interplay between our model and our decision. Right, keeping things in our data model insights framework, we're going to want to focus on this sort of interaction between our model and that decision. All right, so let's put ourselves in that setting. So, let's pop over to our spreadsheet, and kind of walk through this process. So, in our spreadsheet, we're tackling our marketing campaign at Snow Hawk, right? And so, our spreadsheet is sort of structured so we're going to have a series of inputs, and from those inputs, get at some outputs, right? So if we focus on here one of the key inputs to our campaign is average annual snowfall. That average annual snowfall is going to be a function of the sample we choose to describe snowfall across our two cities. Even with that sample we're still going to have some parameters

for our campaign. Those parameters being what are the rebates, 100% or 75% whatever they might be, and when did they kick in as certain percentages of those historic average annual snowfalls.

At the bottom of our spreadsheet, we have some visuals to help us understand the variance in snowfall, right? We can control how those visuals are presented by changing the bins for our axis. And these visuals are nice to help us get a sense of when we look at these average snow falls, where does that average fall for Vancouver? Where does it fall for Rochester? And then, you know, how much spread is there in and around that average, right. So a nice visual of that snowfall, that communicate why there may or may not be a payout, or a rebate in each of these different cities. At the end of the day, the key outcome of our model here, our Key Performance Indicators, are going to be, what fraction of consumers receive a rebate, right? So, that's what consumers are focused on. And then from Snow Hawk's standpoint, what is that average rebate we have to provide to those consumers, right? So, we have these two indicators. One for the firm, one for the consumers, right?

And so what we might do now is take our base case, right, and maybe copy that down into our spreadsheet and look at how, as we change some of these inputs, those KPIs vary as a function of some of those campaign parameters. So once we have our base case copied, it's going to facilitate us comparing and contrasting sort of different outcomes of the structure of our campaign. Right, where we can compare the percentage of consumers who received rebates versus our base case as a function of maybe adjusting some of the parameters of our campaign. Right? So, for argument's sake, we might think that there was a subset of years where it snowed a lot and they were outliers of typical snowfall. And so we might exclude those from our sample, right? So we might simply exclude some of these years from our sample. As we exclude those years, our average snowfalls will decrease. As those average snowfalls decreases, then those critical levels decrease, and we're now providing consumers with fewer rebates. Right? So, our fraction of consumers who receive rebates decreases a bit and, potentially, those average payouts. So maybe that's going in the wrong direction. So, just because those years might be outliers doesn't mean we want to exclude them from our sample if their inclusion aligns our campaign with our key stakeholders. We could look at different ways to adjust those payouts.

For example, we could change our snowfall levels. So instead of the smallest payout happening when snowfall was only 40% of average, what if it was 50% of average? So of course now we've raised the bar for the lowest rebate and in essence we've kind of decreased the bar, right? And so now more consumers are going to be receiving rebates. So as more consumers receive rebates our percent with rebates has increased. Right? But maybe in line with that that our average rebates are not what we're looking for, right? And so we may, in concert with changing the snowfall levels when the rebates kick in, is actually changing the rebates themselves and drop those rebates a little bit. All right, so to make the outcomes more in line with one of our other key stakeholders being the firm, right? So what's critical here is that the actual percentage of consumers receiving rebates is still the same. But the average rebates received are now perhaps more in line with our base case.

And so this highlights, you know, part of this sort of process we're going to go through with our model as it links back to our decision. We're going to have this back and forth between our statistics and our analysis of those statistics, and the structure of our decision given our insight of that setting, right? And sort of tweaking some of the parameters, tweaking that sample, getting a sense for how everything works together, right? So we're really trying to focus on how everything works together. We're going to sort of look at different attributes of our decision such that we're prepared for that meeting where somebody might ask some of those tough questions. So what about this? What about if we change those parameters? We can easily be ready for those questions or if need be, be ready to answer them in a relatively easy, sort of, dashboard framework.

Back to Table of Contents

A Decision With or Without Data

Instructions:

You are required to participate in both of the discussions in this course.

Discussion topic:

Recall a decision you have made in which data was involved, or for which you wish you had been able to obtain useful data.

Create a post in the discussion board that describes:

A decision you made long enough ago that there has been a clear outcome
How data informed the decision, or how you feel it could have informed the decision
A summary of the good and/or bad outcomes that resulted
How you might approach the decision differently considering what you have learned in this course

Your example should come from your own personal or work experience.

To participate in this discussion:

Click **Reply** to post a comment or reply to another comment. Please consider that this is a professional forum; courtesy and professional language and tone are expected. Before posting, please review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Course Project, Part Three: Data and Your Decision

In previous parts of the project, you identified a decision-making question, stakeholders who care about that question, and data you hope will answer that question. You made a preliminary assessment of how best to visualize that data. In this module, you saw how all the planning and preliminary work comes together through the data-models-insights framework. Now it's time to reassess all your thinking that has led up to this point and to create a working model.

Project instructions:

Open the Course Project document you started in the first part of the course, and complete the assignment for **Part Three—Data and Your Decision**.

When you have completed Part Three, take a moment to review our earlier work. When you are satisfied with your work, submit your completed project document for review and grading by your instructor.

Do not hesitate to contact your instructor if you have any questions about the project.

This is a required part of the final course project; completion and submission of all parts of the course project is required in order to achieve credit.

Before you begin:

Before starting your work, please review the **rubric** (list of evaluative criteria) for this assignment and eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Tool: "Understanding and Visualizing Data" Action Plan

This Action Plan can guide your efforts on the job

Now that you have had a chance to review the content in this course related to collecting data, visualizing it in a number of different ways, and using it to inform your decisions, you will make a plan for implementing these ideas within your workplace.

You can use the Action Plan here to outline a plan for yourself that will guide your efforts within your own organization. The action plan on this page follows traditional "SMART" methodology to help you identify steps to take on the job that are specific, measurable, action-oriented, realistic, and time-based. You may choose to use it now as a tool for yourself, as a means of demonstrating to your manager or to peers how this course will influence your efforts on the job, or you may choose to save it and use it to guide your future work.

Back to Table of Contents

Module Three Wrap-up: Bring the Data into the Decision

The quality of your data-based decision relies on more than just how carefully the data were collected. The data you have must be relevant to the question. The data must provide enough detail and the right kind of detail to give insight. It can take some time and experimentation to tease out the useful insights, and in some cases this process can cause you to reframe your statement of the problem.

In this module, you practiced using the data-models-insights framework and applied it to a decision-making question of your own. You took a critical look at whether the data available were sufficient to answer the question posed. You saw how dashboards can be constructed and used, and hopefully you were able to plan or implement a dashboard of your own. Now you have all the tools needed to make the best decision your data will allow, and you have a working understanding of the criteria you need to consider when deciding how much confidence to place in your data-based decision.

Back to Table of Contents

Read: Thank You and Farewell

Chris Anderson
Associate Professor
School of Hotel Administration
Cornell University

Congratulations on completing *Understanding and Visualizing Data*. I hope this course has deepened your understanding of data, has increased your comfort working with data in the decision-making process, and has given you tools and a framework that you can use to meet the decision-making needs of your organization.

From all of us at Cornell University and eCornell, thank you for participating in this course.

Sincerely,

Chris Anderson

Back to Table of Contents

Glossary

Box-and-whisker plot

A graph used to depict a range of data that shows a line running from the minimum to the first quartile, a box from the first quartile to the median, another box from the median to the third quartile, and a line running from the third quartile to the maximum.

Case

An object in a collection of data (e.g. if the data are a collection of shells, a case is an individual shell).

Categorical variable

A variable that can take on a limited, or fixed, number of possible values (e.g. a ball's color would be a categorical variable if the ball can only be red, blue, green, and purple).

Central tendency

The measurable clustering of values in a statistical distribution.

Cluster sampling

A sampling technique in which the population is divided into smaller groups, or clusters, and a simple random sample is taken of the clusters. This is different from a stratified random sample because the whole cluster is sampled, rather than a sample from within the cluster.

Confound

An extraneous variable that could also cause a correlation between the dependent variable and the independent variable (e.g. shark attacks are more common when beach ice cream sales increase. Is this because ice cream causes shark attacks, or because hot weather leads to both increased ice cream

consumption and increased swimming? In this case, hot weather is the confound).

Convenience sample

A sample made up of the portion of the population that is easiest to reach (e.g. a poll on the O'Reilly Factor website, where Bill O'Reilly asks all of his listeners to go to his website and answer the poll).

Data validation

Verifying that each datum in a set of data is an accepted value (e.g. checking to make sure that all answers to "how many children do you have" are non-negative integers because you cannot have -5 children, or half a child, etc.)

Dependent variable

The variable in an experiment that is being measured (e.g. in an experimental drug trial, the dependent variable is the effect of the drug because experimenters are measuring the effects of the drug).

First quartile

The middle value in a set of data between the minimum and the median (e.g. if the data is {1,2,3,4,5}, then 2 is the first quartile because it is between 1 and 3).

Grouped bar graph

Charts designed to show different sub-groups within a category (e.g. in a bar chart of the different types of recycling used in locations around the city, the categories would be the locations in the city, and the sub-groups would be the types of recycling used). It is a bar chart in which each bar is a histogram of the different sub-groups of that bar.

Histogram

A diagram that uses rectangles to show the frequency of data values within successive numerical intervals.

Independent variable

The variable in an experiment being adjusted by the experimenter to measure its effects on the dependent variable (e.g. in an experimental drug trial, the drug is the independent variable because experimenters are controlling how much is taken).

Inference

A claim about the properties of a population based on statistical analysis.

Interquartile range (IQR)

The range of values between the first and third quartiles.

Mean

The central value of data, calculated by adding up all the values and dividing the sum by the number of different values (e.g. {1, 4, 5}, the mean is 3.33).

Median

The central value of data, calculated by taking the middle value of the data (e.g. {1,4,5} the median is 4).

Multistage sampling

A sampling method in which sampling is carried out in stages (e.g. taking a cluster sample of the elements chosen by a preliminary cluster sample).

Pie graph

A circular chart that is divided into slices or radiuses, in which the angle created by each radius is proportional to the percentage of the population that it represents.

Pivot table

A data summarization tool that automatically sorts and gives summary statistics for data within a table.

Population

The pool from which a statistical sample is drawn and about which inferences are made.

Quantitative variable

A variable that can be measured in terms of numbers (e.g. the age of children in a household).

Random sample

A sample from a population in which members are chosen randomly.

Sample

A set of data selected from a population, from which inferences are being drawn.

Sensitivity testing

Sensitivity is the proportion of correctly identified positive results in a test (for example, if 100 people are known to have AIDS and 96 test positive, then the test has 96% sensitivity). Sensitivity testing is the measuring of the sensitivity of a test.

Slicer

A quick method of filtering data on a pivot table by the type of data needed (e.g. on a pivot table showing the different types of vegetable oils, a slicer would narrow it down to olive oil).

Spread

A measure of how much variation exists within a sample.

Stacked column graph

A bar graph in which each bar is a tower of smaller bars, in which the height of each smaller bar represents the proportion of the total bar that the smaller bar represents (a combination between a bar graph and a pie chart).

Stratified random sample

A sample in which the population is divided into homogeneous clusters, or stratums, and a random sample is taken from within each stratum. This is different from a cluster sample because a random sample is taken from within each stratum, rather than of a whole stratum.

Summary statistics

Numbers derived from a set of data that give a quick description, or summary, of the data.

Systematic sampling

A sampling method in which members of the sample are selected from the population at a specific interval (e.g. going through a phone book and selecting the first, fourth, seventh, tenth, and so forth, names for the sample).

Third quartile

The middle value between the median and the maximum (e.g. in a set of {1,2,3,4,5}, the third quartile is 4 because it is between 3 and 5).

Two-way table

A table showing the distribution of one variable in rows and another in columns, used to visualize the association between the two variables.

Type

A specific instance within a categorical variable (e.g. for the categorical variable automobiles, one type could be minivan).

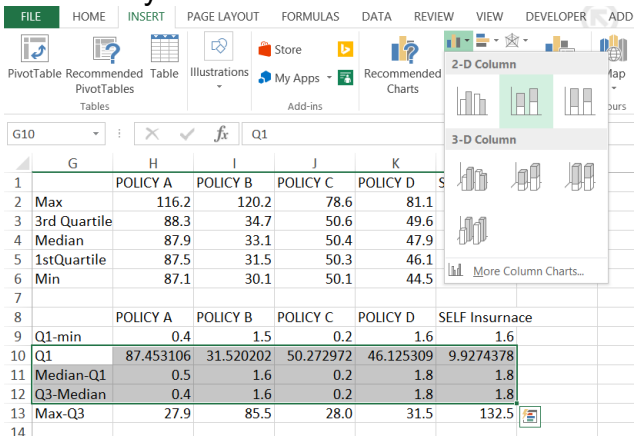
Create a Box and Whisker Plot

Excel Step-by-Step How-to for Windows

Excel for Mac Instructions on page 8

Instructions: Use this guide to create a box and whisker plot.

Data requirement: five number summary, quantitative data
Sample Data: multiple insurance policies

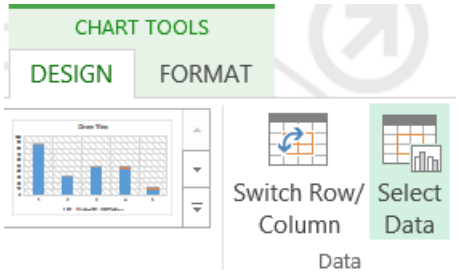
Step	Windows Instructions + Screen Shot
1. Create chart base for box and whisker plot.	Highlight all of the data in Q1, Q1 median, and Q3 median and insert stacked column chart. You will edit the layout of this chart to make box and whisker charts. 

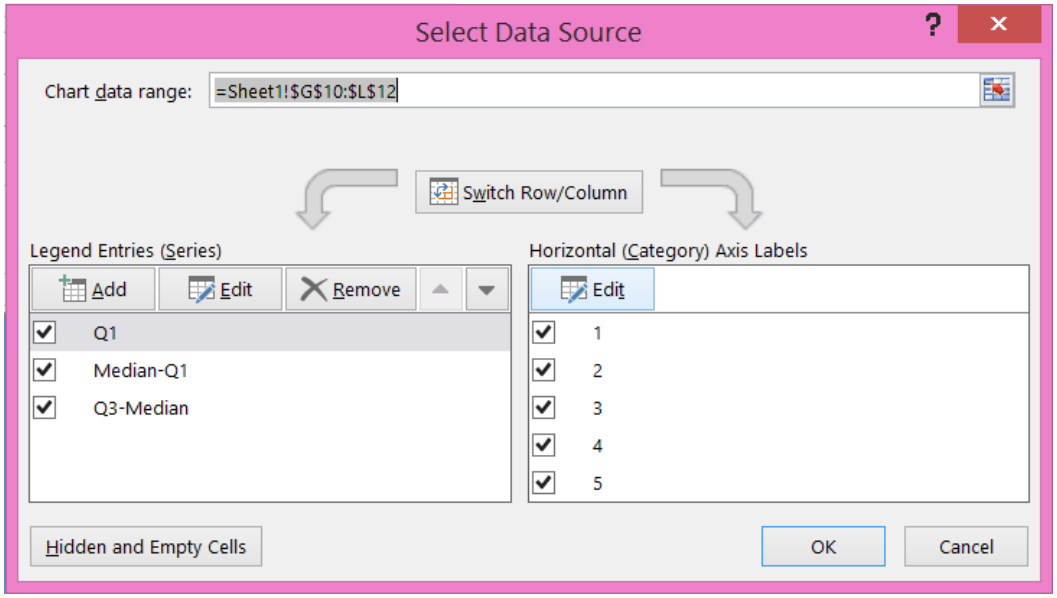
	G	H	I	J	K
1		POLICY A	POLICY B	POLICY C	POLICY D
2	Max	116.2	120.2	78.6	81.1
3	3rd Quartile	88.3	34.7	50.6	49.6
4	Median	87.9	33.1	50.4	47.9
5	1stQuartile	87.5	31.5	50.3	46.1
6	Min	87.1	30.1	50.1	44.5
7					
8		POLICY A	POLICY B	POLICY C	POLICY D
9	Q1-min	0.4	1.5	0.2	1.6
10	Q1	87.453106	31.520202	50.272972	46.125309
11	Median-Q1	0.5	1.6	0.2	1.8
12	Q3-Median	0.4	1.6	0.2	1.8
13	Max-Q3	27.9	85.5	28.0	31.5
14					

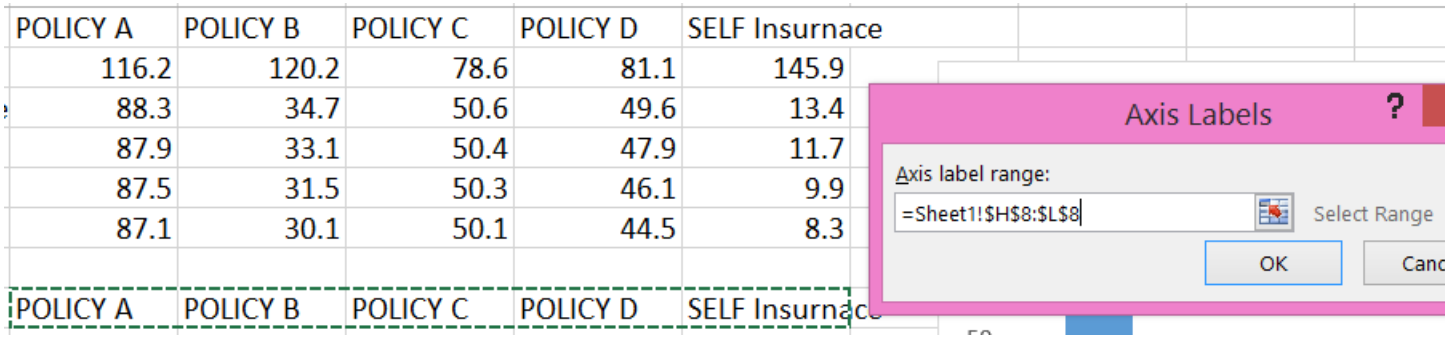
2. Change the chart labels.

Under the Design tab, press Select Data. Under the Horizontal Axis Label category, select edit. When prompted for a range of cells, select the cells with your category labels.

The range of cells are taken as A:B where A is the first column and row of the data and B is the last column and row of the data.

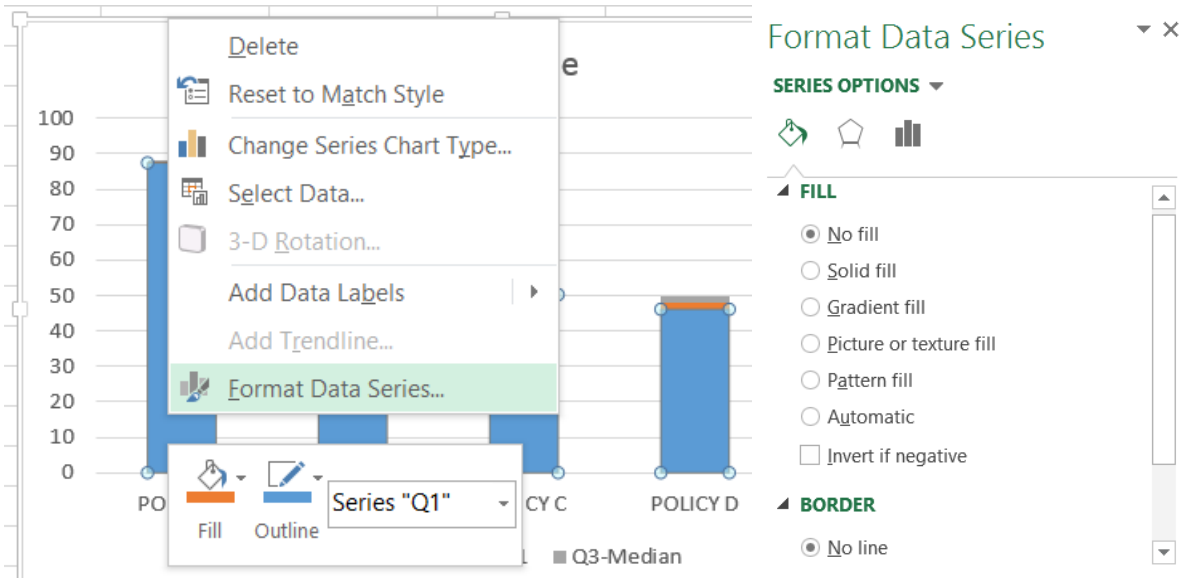






3. Format stacked columns.

Right click on any of the Q1 boxes in the chart and select Format Data Series. Set the fill to be No fill. Set the border to be No line.

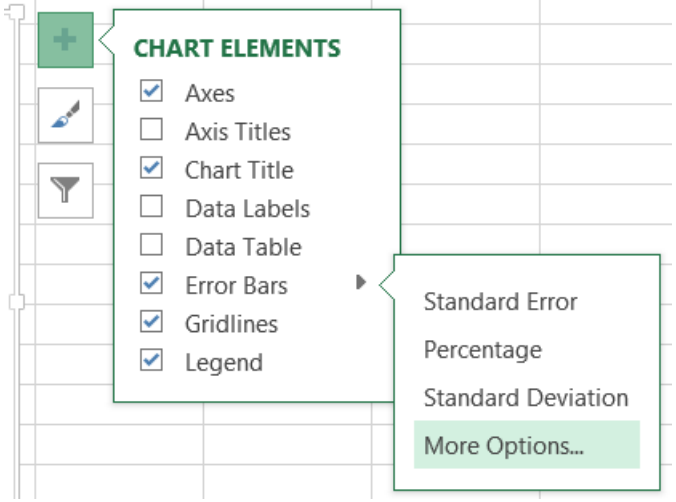


TIP: To quickly move between the stacks:

Use the up and down arrows on keyboard to navigate between the stacks of the column.

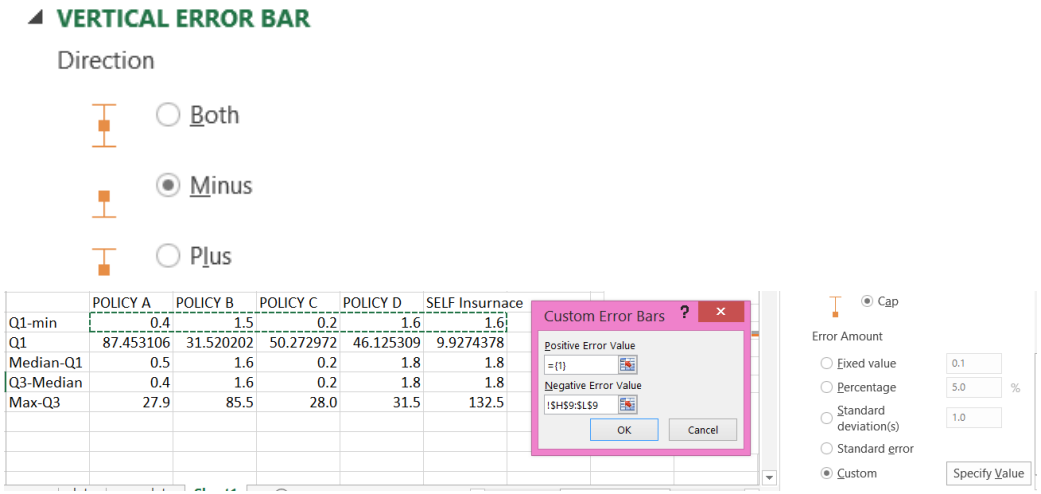
4. Add error bars to the stacked columns to look like bottom whiskers.

Click on the Q1 space on the chart and select the plus button. Click on add error bars. Navigate to the expansion arrow and click More Options.



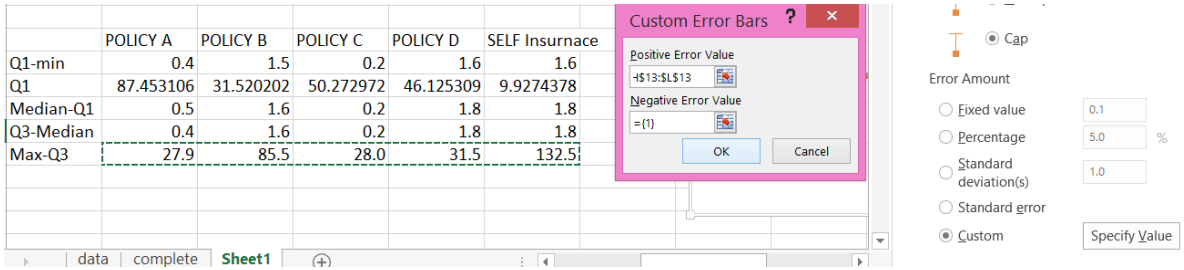
5. Customize the error bars.

Make the direction minus. Make the error amount a custom, specified value. Make the negative error the values of Q1-min by selecting all of the cells in that row of your five number summary.

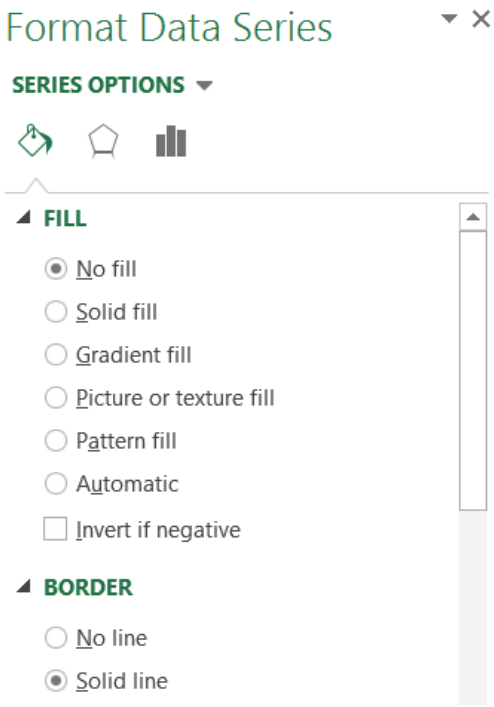


6. Add error bars to look like top whiskers.

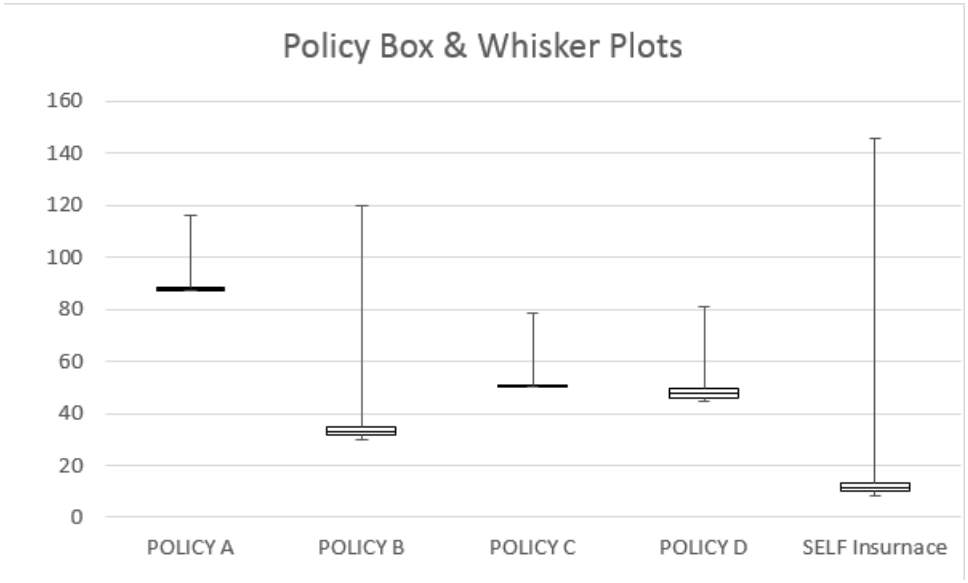
Click on the Q3 space on the chart and select the plus button. Click on add error bars. Navigate to the expansion arrow and click More Options. Make the direction plus. Make the error amount a custom, specified value. Make the positive error the values of max-Q3 by selecting all of the cells in that row of your five number summary.



7. Format the remaining stacks into boxes.	Right click the data series and select Format Data Series. Change the fill to no fill and the border to solid line.



8. Format and label the chart.



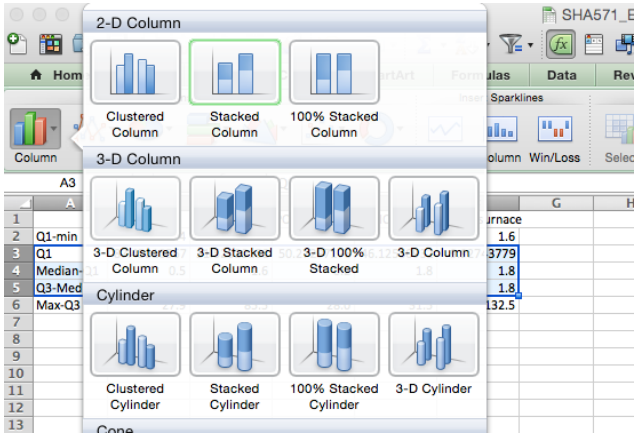
Create a Box and Whisker Plot

Excel Step-by-Step How-to for Mac

Excel for Windows Instructions on page 1

Instructions: Use this guide to create a box and whisker plot.

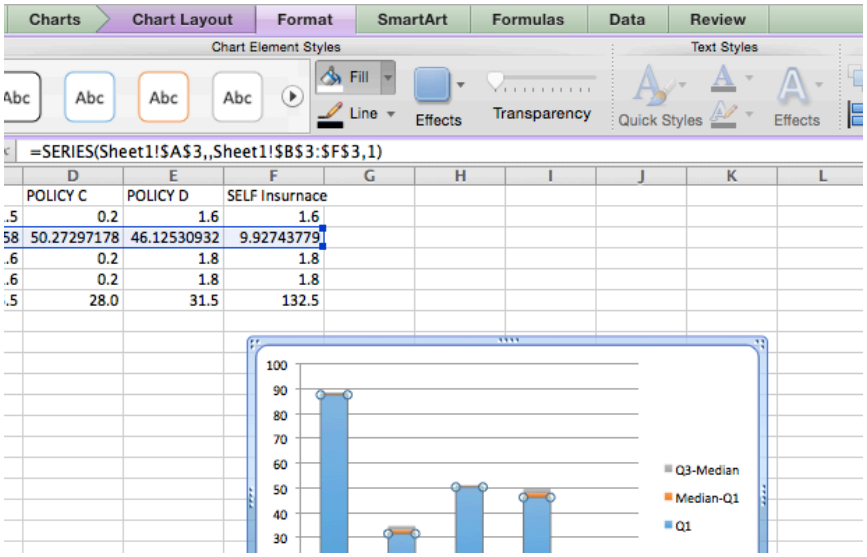
Data requirement: five number summary, quantitative data
Sample Data: multiple insurance policies

Step	Mac Instructions + Screen Shot
1. Create chart base for box and whisker plot.	Highlight all of the data in Q1, Q1 median, and Q3 median and insert stacked column chart. 

2. Change the chart labels.

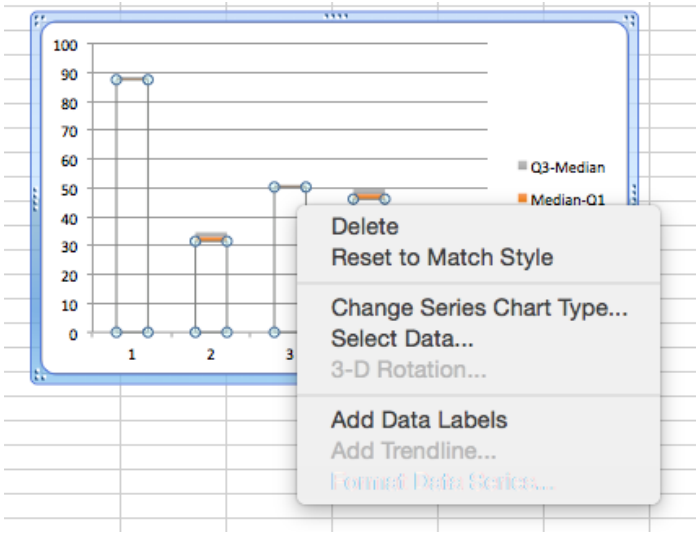
Under the Charts tab, press Select Data. Under the Horizontal Axis Label category, select edit. When prompted for a range of cells, select the cells with your category labels.

The range of cells are taken as A:B where A is the first column and row of the data and B is the last column and row of the data.



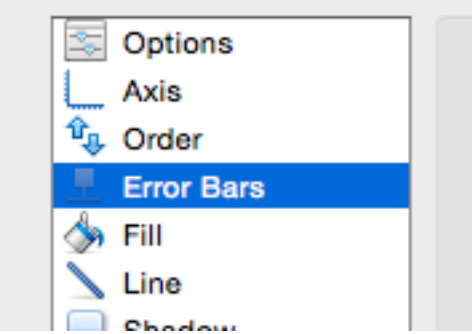
3. Format stacked columns.

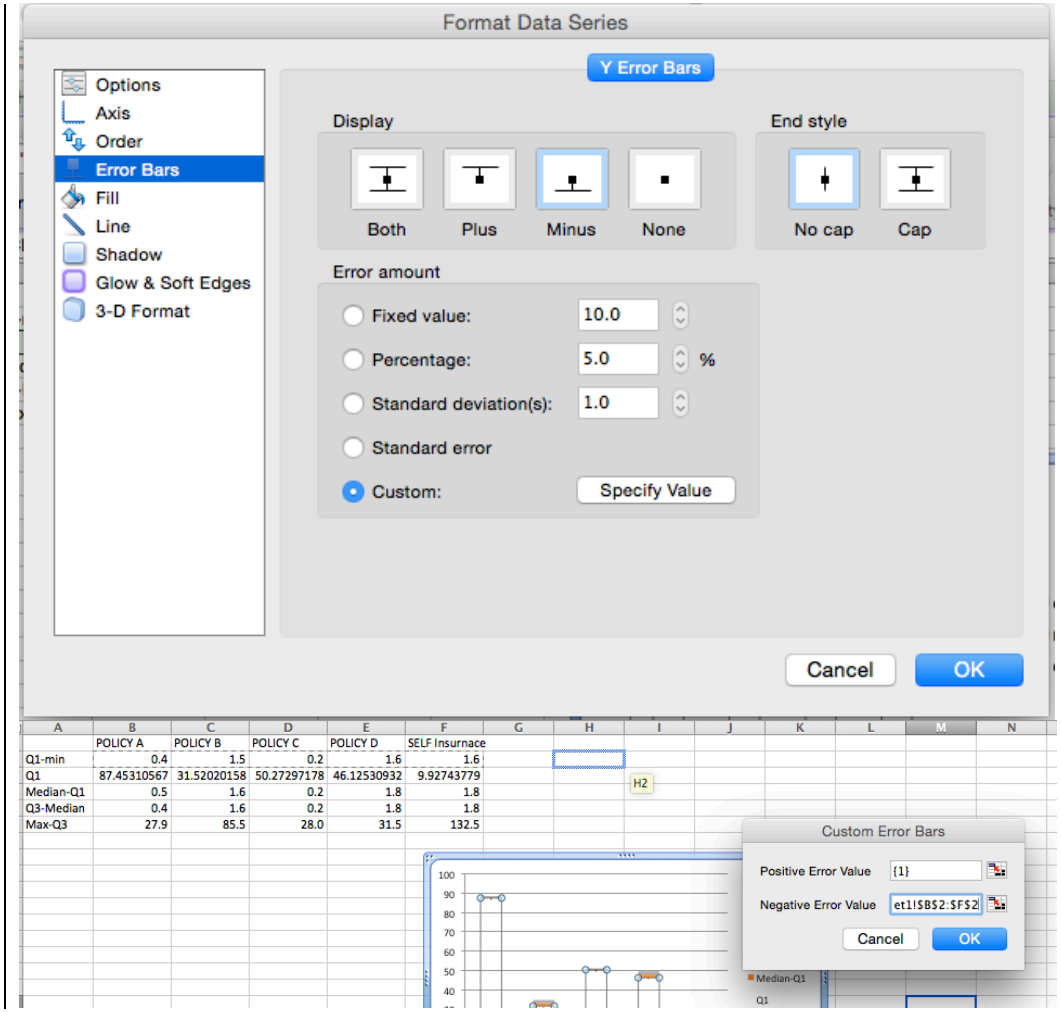
Control-click on any of the Q1 boxes in the chart and select Format Data Series. Set the fill to be No fill. Set the border to be No line.



TIP: To quickly move between the stacks:

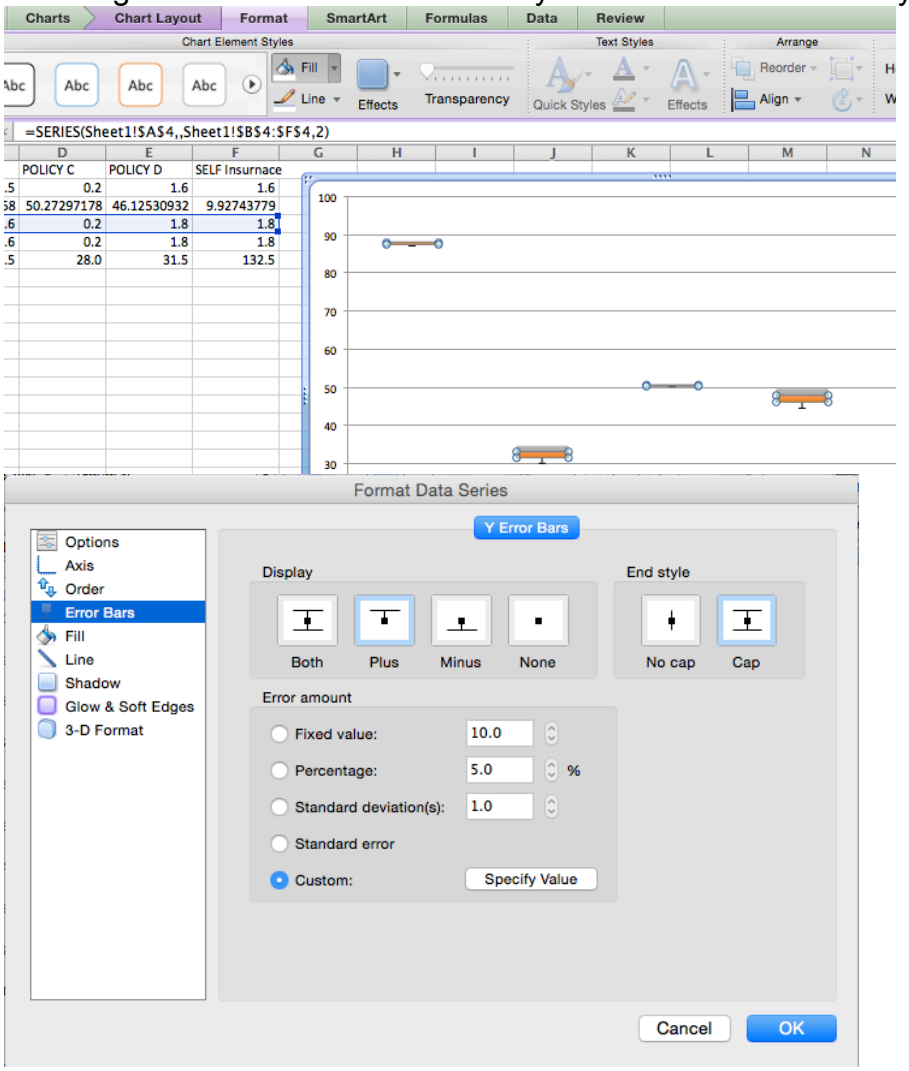
Use the up and down arrows on keyboard to navigate between the stacks of the column.

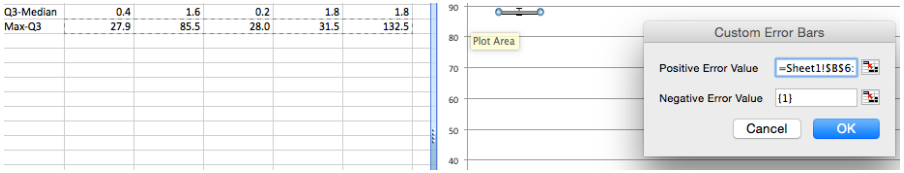
4. Add error bars to the stacked columns to look like bottom whiskers.	Under Format Data Series, select error bars. 
.5. Customize the error bars.	Make the direction minus. Make the error amount a custom, specified value. Make the negative error the values of Q1-min by selecting all of the cells in that row of your five number summary.



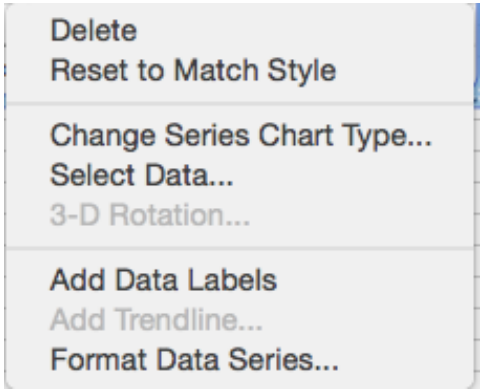
6. Add error bars to the stacked columns to look like top whiskers.

Control-click on the Q3 space on the chart and navigate to error bars. Make the direction plus. Make the error amount a custom, specified value. Make the positive error the values of max-Q3 by selecting all of the cells in that row of your five number summary.

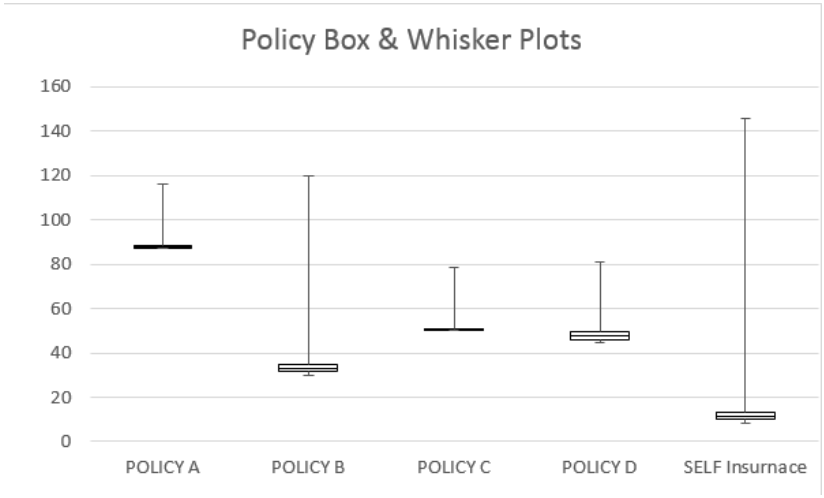




7. Format the remaining stacks into boxes.	Control-click the data series and select Format Data Series. Change the fill to no fill and the border to solid line.



8. Format and label the chart.



Create a Representation of Categorical Data

Excel Step-by-Step How-to for Windows

Excel for Mac Instructions on page 6

Instructions: Use this guide to create meaningful visualizations of categorical data.

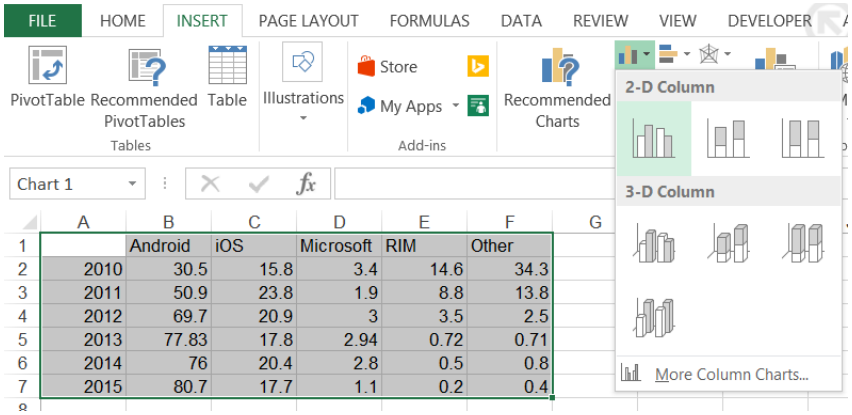
Data requirement: multiple variables, qualitative data
Sample Data: annual and quarterly mobile OS market share

Step	Windows Instructions + Screen Shot
1. Arrange the data you want to use so that each unique variable has its own column.	

	A	B	C	D	E	F
1		Android	iOS	Microsoft	RIM	Other
2	2010	30.5	15.8	3.4	14.6	34.3
3	2011	50.9	23.8	1.9	8.8	13.8
4	2012	69.7	20.9	3	3.5	2.5
5	2013	77.83	17.8	2.94	0.72	0.71
6	2014	76	20.4	2.8	0.5	0.8
7	2015	80.7	17.7	1.1	0.2	0.4
8						

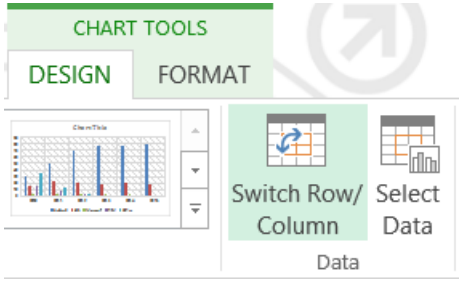
2. Select your data and insert a column chart.

After highlighting the data you want to use, click the insert tab. Navigate to the column chart icon and select a cluster column chart.



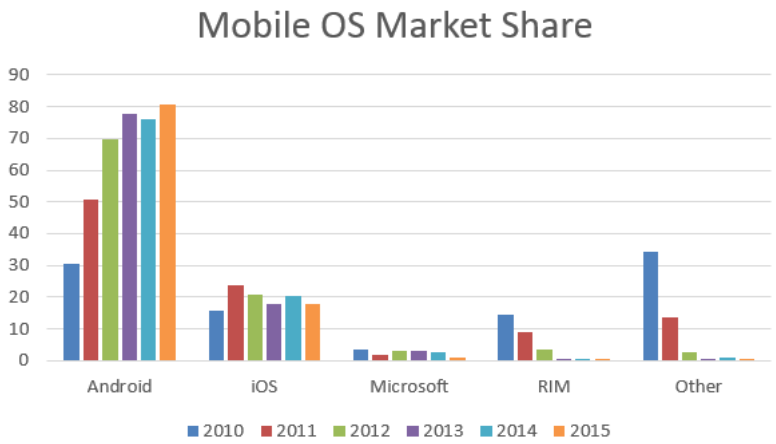
TIP: To switch the column/row in chart:

Navigate to the Data category of the Design tab and select Switch Column/Row.



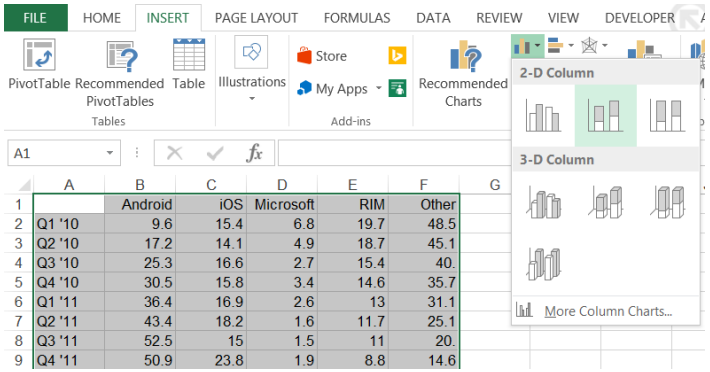
3. Format and label your chart.

You can edit the titles and labels of the chart by selecting them and editing the text.



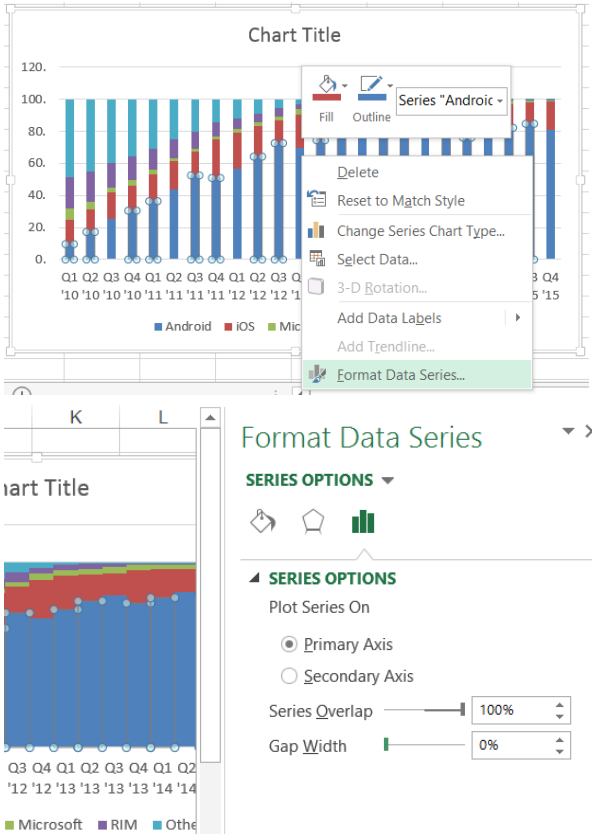
4. Select the data that you want to use and create a stacked column chart.

After highlighting the data that you want to use, click the insert tab. Navigate to the column chart icon and select a stacked column chart.



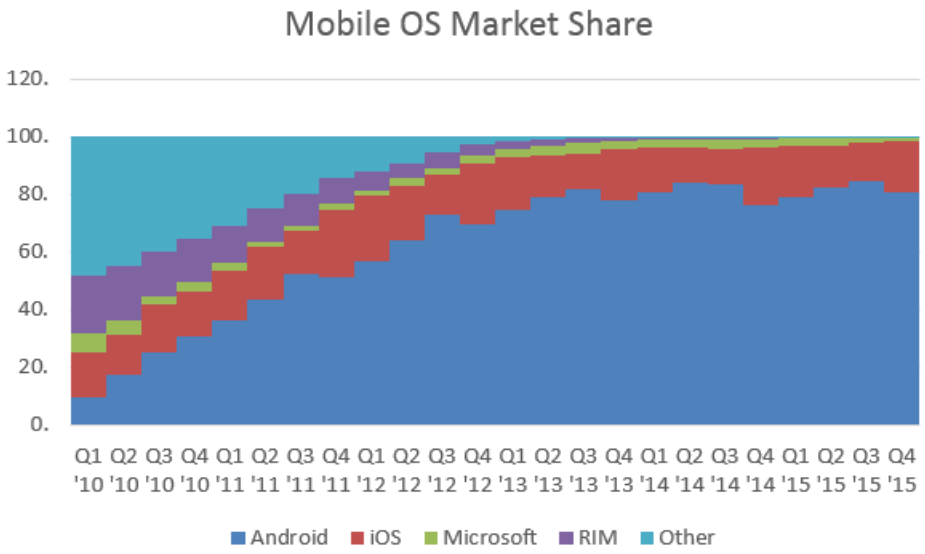
5. Format the data series to remove the gap between columns.

Right-click any column in the chart and select Format Data Series. Then bring the Gap Width to 0%.



6. Format and label your chart.

You can edit the titles and labels of the chart by selecting them and editing the text.



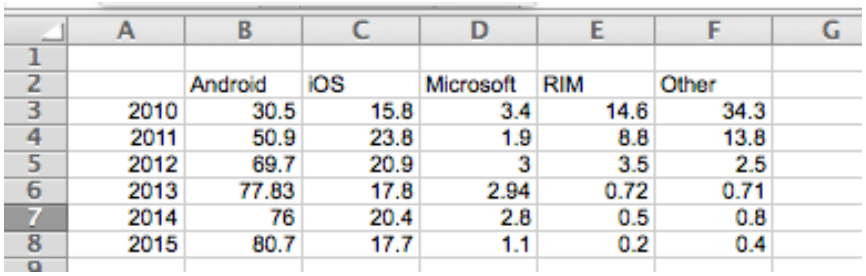
Create a Representation of Categorical Data

Excel Step-by-Step How-to for Mac

Excel for Windows Instructions on page 1

Instructions: Use this guide to create meaningful visualizations of categorical data.

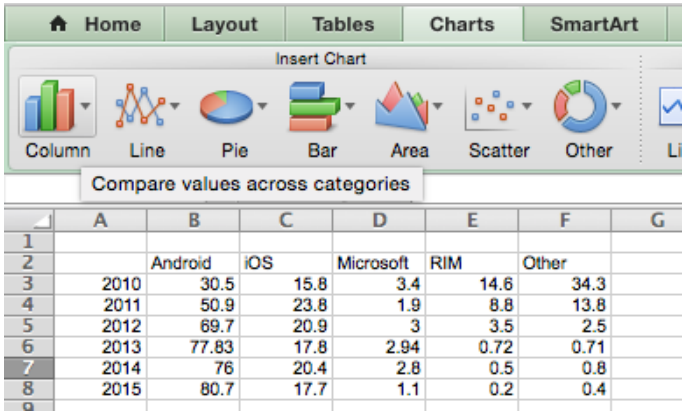
Data requirement: multiple variables, qualitative data
Sample Data: annual and quarterly mobile OS market share

Step	Mac Instructions + Screen Shot
1. Arrange the data you want to use so that each unique variable has its own column.	

	A	B	C	D	E	F	G
1							
2		Android	iOS	Microsoft	RIM	Other	
3	2010	30.5	15.8	3.4	14.6	34.3	
4	2011	50.9	23.8	1.9	8.8	13.8	
5	2012	69.7	20.9	3	3.5	2.5	
6	2013	77.83	17.8	2.94	0.72	0.71	
7	2014	76	20.4	2.8	0.5	0.8	
8	2015	80.7	17.7	1.1	0.2	0.4	
9							

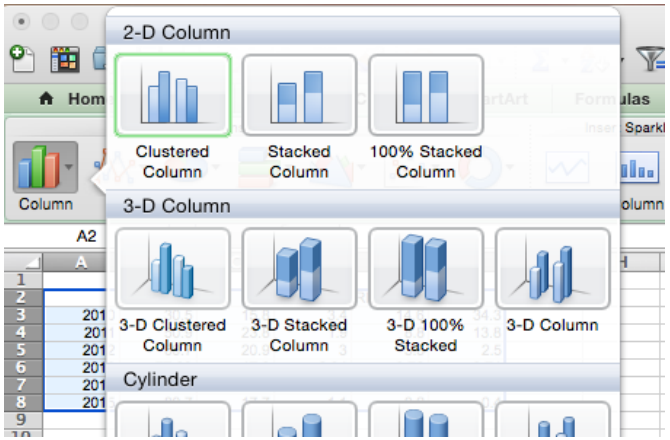
2. Select your data and insert a column chart.

After highlighting the data you want to use, click the Charts tab.



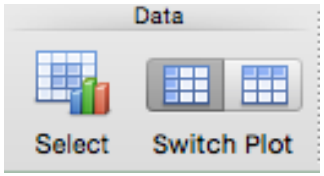
	A	B	C	D	E	F
1						
2		Android	iOS	Microsoft	RIM	Other
3	2010	30.5	15.8	3.4	14.6	34.3
4	2011	50.9	23.8	1.9	8.8	13.8
5	2012	69.7	20.9	3	3.5	2.5
6	2013	77.83	17.8	2.94	0.72	0.71
7	2014	76	20.4	2.8	0.5	0.8
8	2015	80.7	17.7	1.1	0.2	0.4

Navigate to the column chart icon and select a clustered column chart.



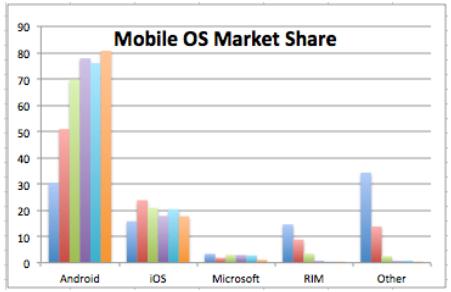
TIP: To switch the column/row in chart:

Navigate to the Data category of the Charts tab and select which plot you would like.



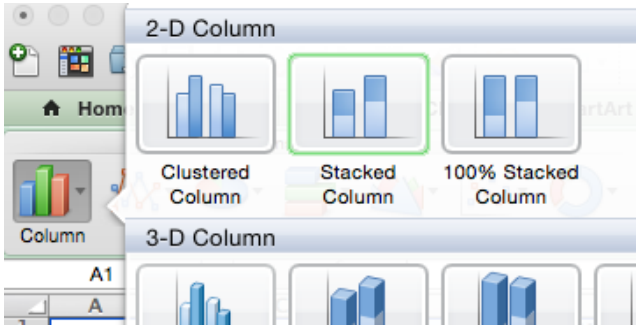
3. Format and label your chart.

You can edit the titles and labels of the chart by selecting them and editing the text.



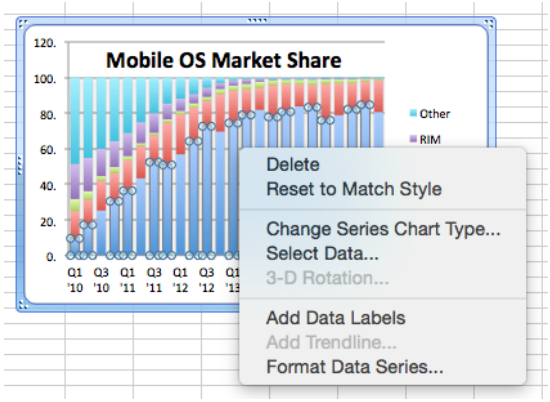
4. Select the data that you want to use and create a stacked column chart.

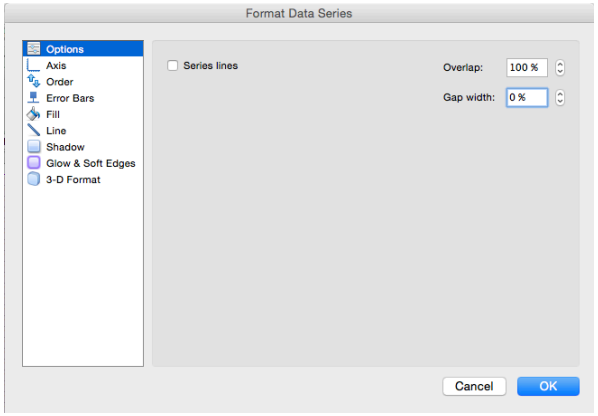
After highlighting the data that you want to use, click the Charts tab. Navigate to the column chart icon and select a stacked column chart.



5. Format the data series to remove the gap between columns.

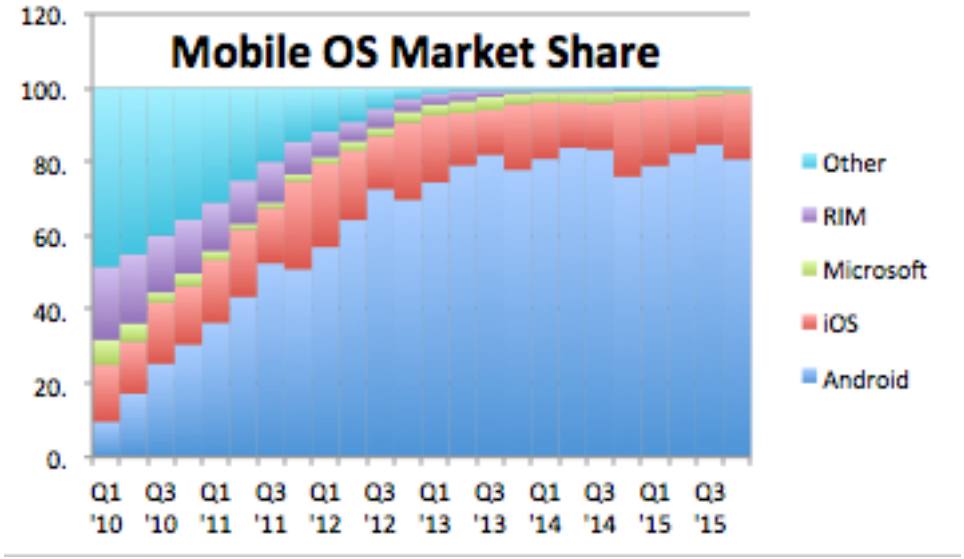
Control-click any column in the chart and select Format Data Series. Then bring the Gap Width to 0%.





6. Format and label your chart.

You can edit the titles and labels of the chart by selecting them and editing the text.



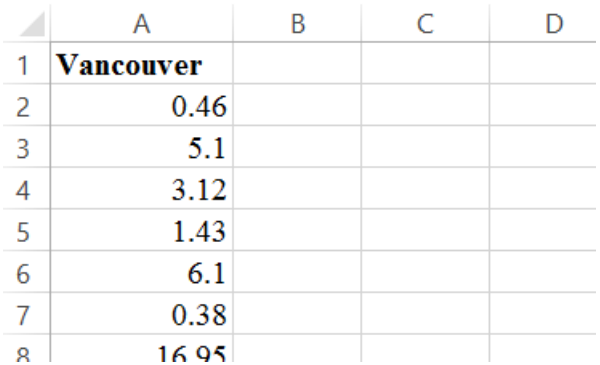
Create a Dot Plot

Excel Step-by-Step How-to for Windows

Excel for Mac Instructions on page 11

Instructions: Use this guide to create a dot plot using Excel.

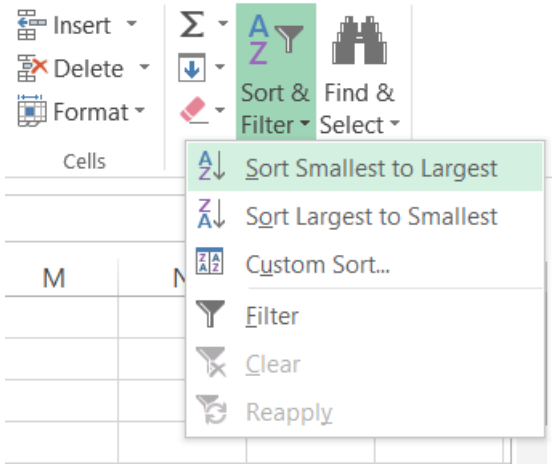
Data requirement: one variable, quantitative data
Sample Data: yearly snowfall in Vancouver

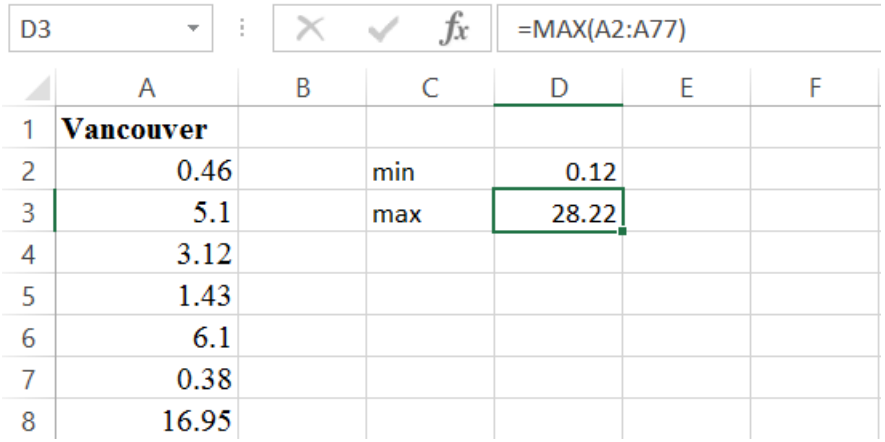
Step	Windows Instructions + Screen Shot
1. Arrange the data you want to use into a column. If you have multiple variables, ensure that each column is a unique variable.	

2. Determine the minimum and maximum values of your data set.

If you have a large data set, you may want to use Excel to find the smallest and largest point in your data. These values make it easier to determine the starting and ending values for your bins.

Select the column with your data in it and then use the “Sort” button. Another method you can use is to type =MIN(A:B) (where A and B are the first and last cell in your column of data) into a blank cell in a new column to find the smallest number and then type =MAX(A:B) to get the biggest number.



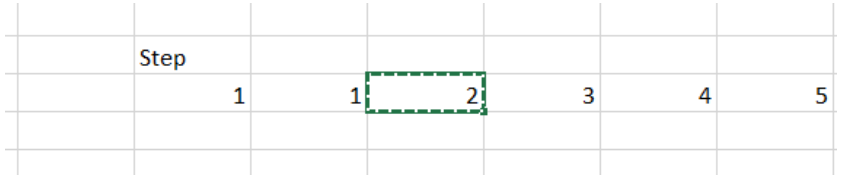


3. Based on the minimum and maximum values, choose an appropriate bin size for your dot plot.

A bin is the interval by which you want to represent your data. A dot plot displays a dot or symbol for every value in your data set that falls into a given bin. It is important to choose a size that is not too small or too large. You want the bin to be wide enough to show a pattern of distribution.

In a cell to the right of your data set, enter your bin size.

4. Set up your range of bins by typing each bin number into a column.



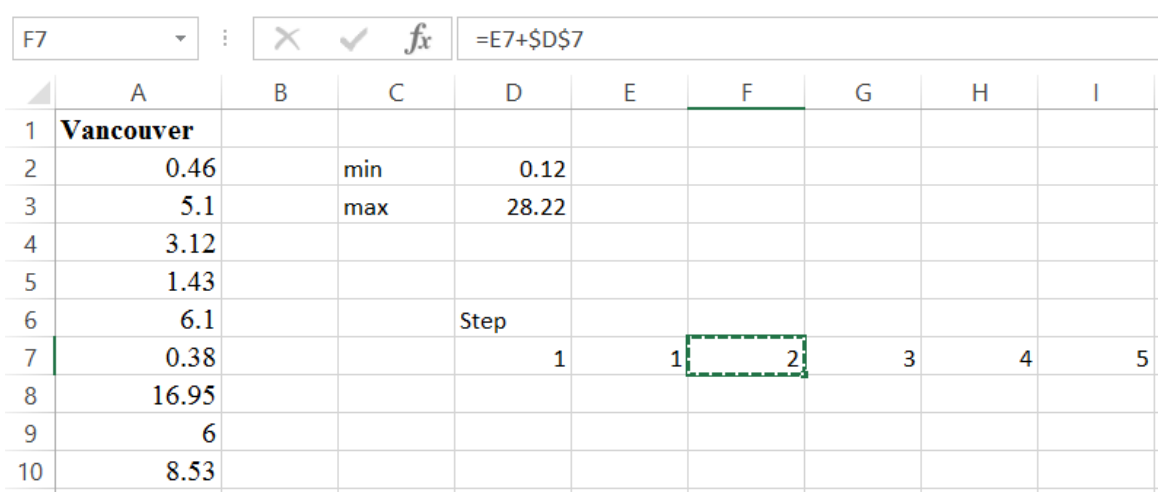
TIP: To quickly auto fill your bin values:

Rather than typing every individual number in your range you can automate the process by using Excel expressions.

In an empty cell type your bin size. In this case, the bin size is 1. In the cell beside the bin size you should use the =INT() function on the cell that holds your minimum data set value. Then add the bin size.

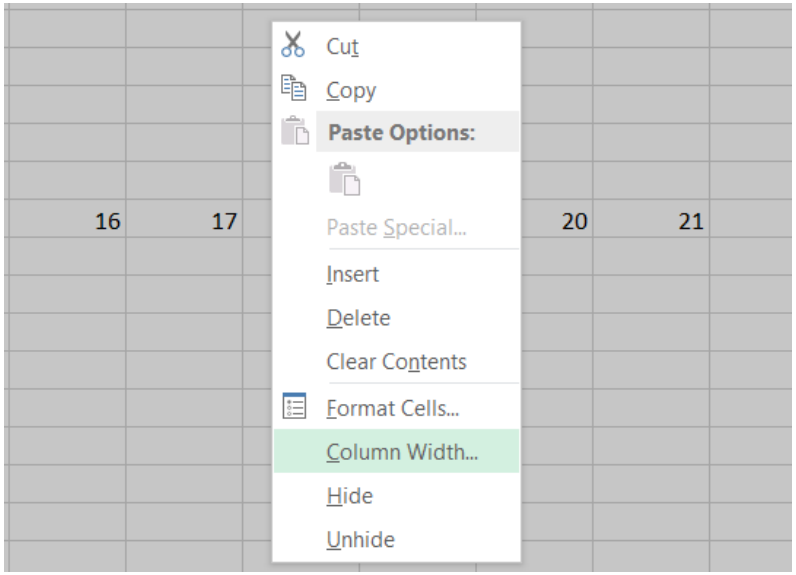
You can repetitively add the bin size to the previous cell by typing = in the blank cell to the right and then selecting the most recently populated cell, typing +, and selecting the cell with the bin size stored. Lock the cell (put $ before the letter and number or press F4) that has the bin size value stored in it and copy and paste this expression to the right until you have as many cells as it takes to reach the max value of your data set.

Another benefit of using this method is that you can dynamically change your bin size by changing the value in your “step” cell.



5. Resize columns to an appropriate width.

Select all of the columns and right click to select resize column. For a dot plot you will only need your column to be as wide as the number stored in it.

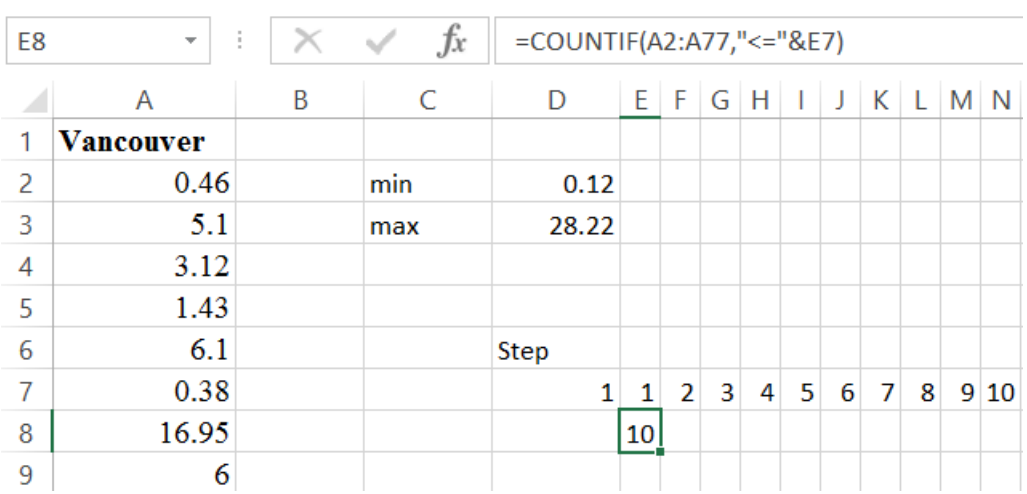


6. Fill in the frequency of the data that falls into each bin.

Select the cell below your first bin. Type =COUNTIF().

=COUNTIF(range, criteria)

- **Range**: all the values of your data set
- **Criteria**: "<="&*firstbinnumber* (this criteria sub-formula indicates that you want to count all the numbers that are less than or equal to your first bin range number. In this example, the completed formula would look like this:



So in this example there are 10 items in the data that are less than or equal to 1.

7. Calculate the difference between the output for each bin and re-assign that value to the cell.

Because you do not want the values that fall in the less than 1 category to also fall under the between 1 and 2 category, you have to calculate the difference between each bin value. Lock the data range by selecting those numbers in the expression and hitting F4 so that our expression references the right cells no matter where we copy and paste the formula. Next you subtract the sum of all of the previous bin values from your count. You can do this by typing =SUM().

=SUM(number1, $number2)

- **number1**: the most recently populated cell in your bin set.
- **$number2**: the locked first output you calculated.

This will give us the number of items in our data that fall exclusively into that given bin.

F8 | =COUNTIF($A2:$A77,"<="&F7)-SUM(E8:E8)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	**Vancouver**																		
2	0.46		min	0.12															
3	5.1		max	28.22															
4	3.12																		
5	1.43																		
6	6.1			Step															
7	0.38			1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
8	16.95				10	6													
9	6																		

8. Display a symbol or dot for the total in each bin.

In the cell above the first bin type =REPT().
=REPT(text, num_times)

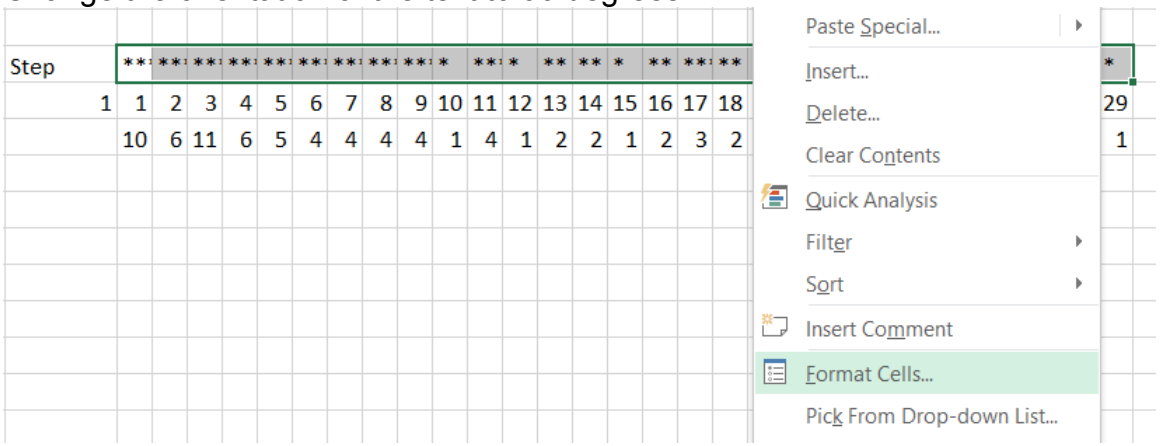
- **text**: the symbol we would like to repeat for the dot plot
- **num_times**: the number of times the text will repeat. For the dot plot you should select the cell that holds the numbers of items from your data set that fit into that bin.

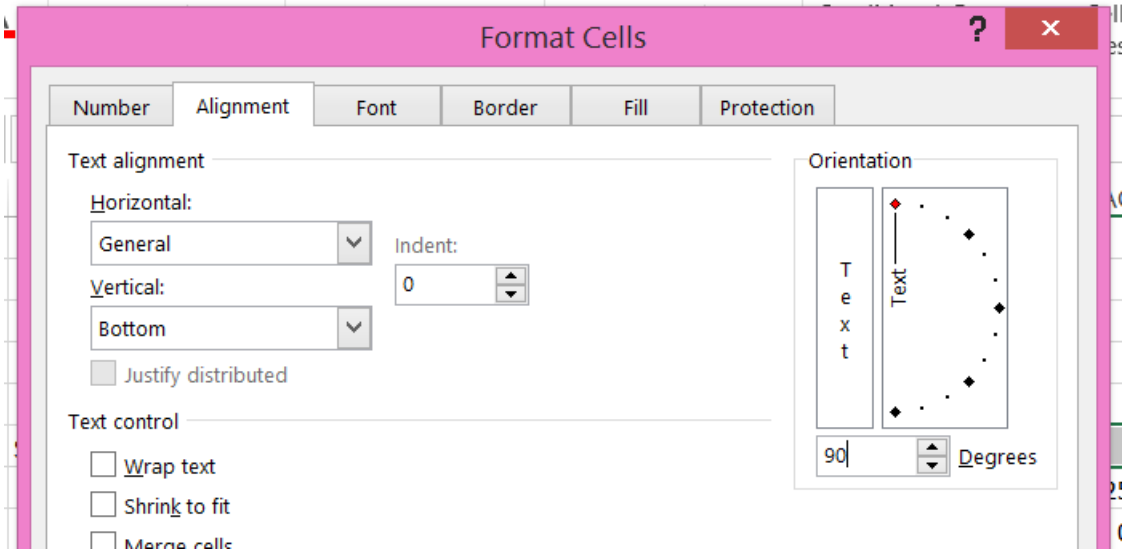
=REPT("*",E8)

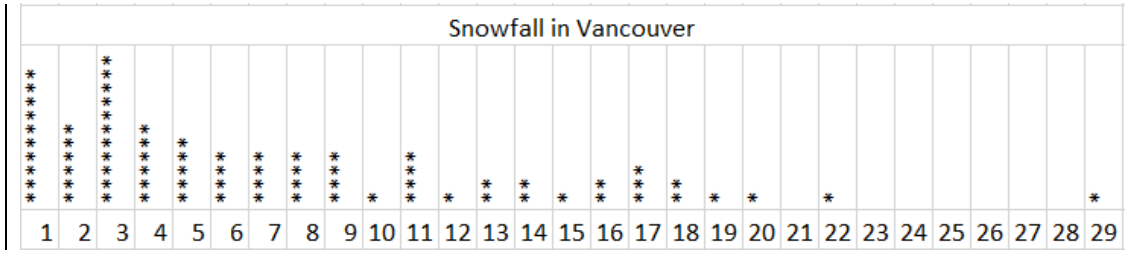
	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
		min	0.12															
		max	28.22															
			Step	**********														
			1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
				10	6	11	6	5	4	4	4	4	1	4	1	2	2	1

9. Format dot plot.

Align the text of the dots to be vertical by selecting format cell and adjusting the alignment. Change the orientation of the text to 90 degrees.







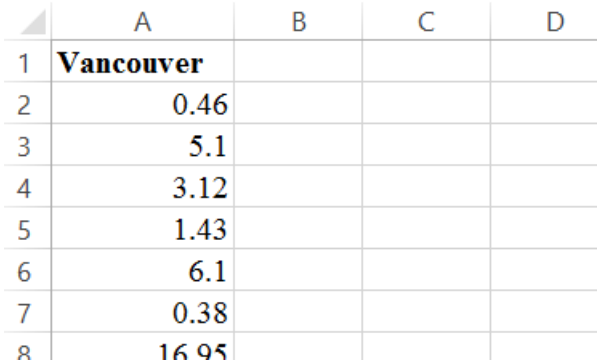
Create a Dot Plot

Excel Step-by-Step How-to for Mac

Excel for Windows Instructions on page 1

Instructions: Use this guide to create a dot plot using Excel.

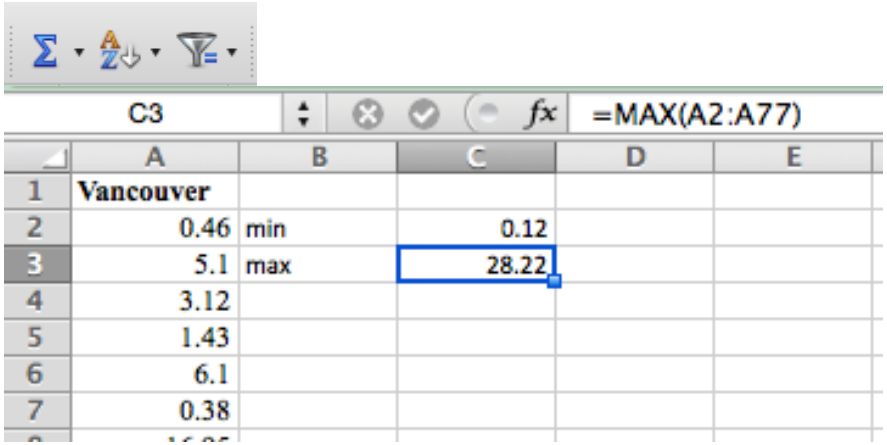
Data requirement: one variable, quantitative data
Sample Data: yearly snowfall in Vancouver

Step	Mac Instructions + Screen Shot
1. Arrange the data you want to use into a column. If you have multiple variables, ensure that each column is a unique variable.	

2. Determine the minimum and maximum values of your data set.

If you have a large data set, you may want to use Excel to find the smallest and largest point in your data. These values make it easier to determine the starting and ending values for your bins.

Select the column with your data in it and then use the “Sort” button. Another method you can use to type =MIN(A:B) (where A and B are the first and last cell in your column of data) into a blank cell in a new column to find the smallest number and then type –MAX(A:B) to get the biggest number.

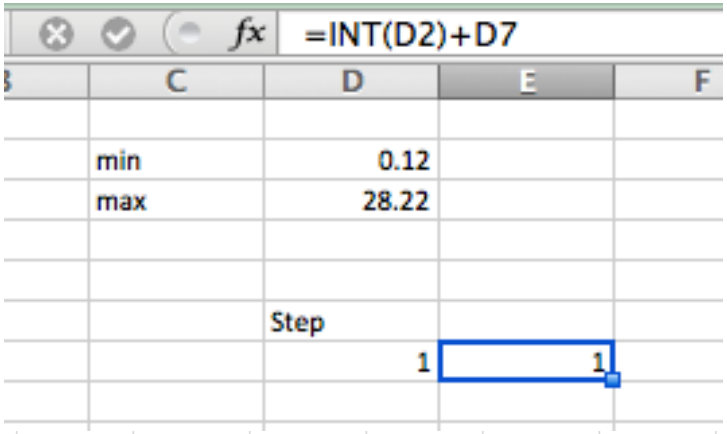


	A	B	C	D	E
1	Vancouver				
2	0.46	min	0.12		
3	5.1	max	28.22		
4	3.12				
5	1.43				
6	6.1				
7	0.38				

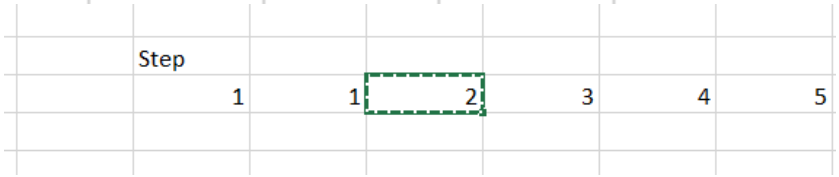
3. Based on the minimum and maximum values, choose an appropriate bin size for your dot plot.

A bin is the interval by which you want to represent your data. A dot plot displays a dot or symbol for every value in your data set that falls into a given bin. It is important to choose a size that is not too small or too large. You want the bin to be wide enough to show a pattern of distribution.

In a cell to the right of your data set, enter your bin size.



4. Set up your range of bins by typing each bin number into a column.



TIP: To quickly auto fill your bin values:

Rather than typing every individual number in your range you can automate the process by using Excel expressions.

In an empty cell type your bin size. In this case, the bin size is 1. In the cell beside the bin size you should use the =INT() function on the cell that holds your minimum data set value. Then add the bin size.

You can repetitively add the bin size to the previous cell by typing = in the blank cell to the right and then selecting the most recently populated cell, typing +, and selecting the cell with the bin size stored. Lock the cell (put $ before the letter and number or highlight and press F4) that has the bin size value stored in it. Now you can copy and paste this expression to the right until you have as many cells as it takes to reach the max value of your data set.

Another benefit of using this method is that you can dynamically change your bin size by changing the value in your “step” cell.

5. Resize columns to an appropriate width.

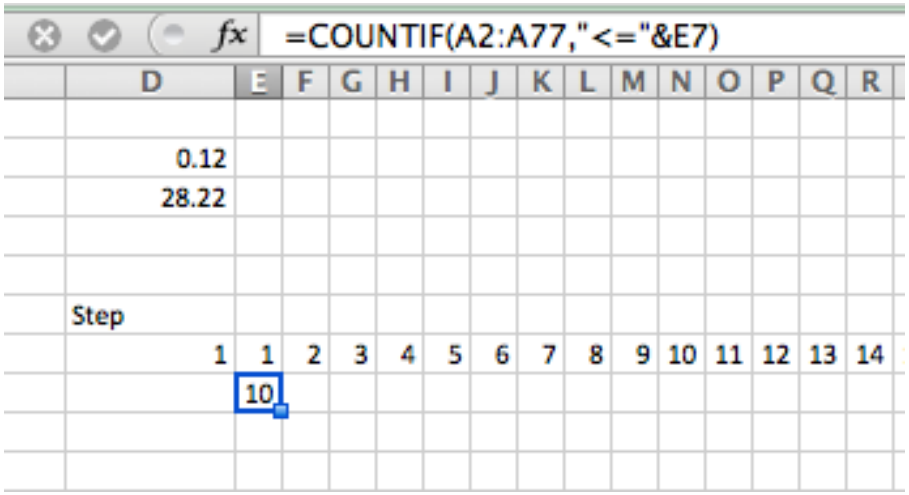
Select all of the columns and control-click to select column width. For a dot plot you will only need your column to be as wide as the number stored in it.

6. Fill in the frequency of the data that falls into each bin.

Select the cell below your first bin. Type =COUNTIF().

=COUNTIF(range, criteria)

- **Range**: all the values of your data set
- **Criteria**: "<="&*firstbinnumber* (this criteria sub-formula indicates that you want to count all the numbers that are less than or equal to your first bin range number. In this example, the completed formula would look like this:



So in this example there are 10 items in the data that are less than or equal to 1.

7. Calculate the difference between the output for each bin and re-assign that value to the cell.

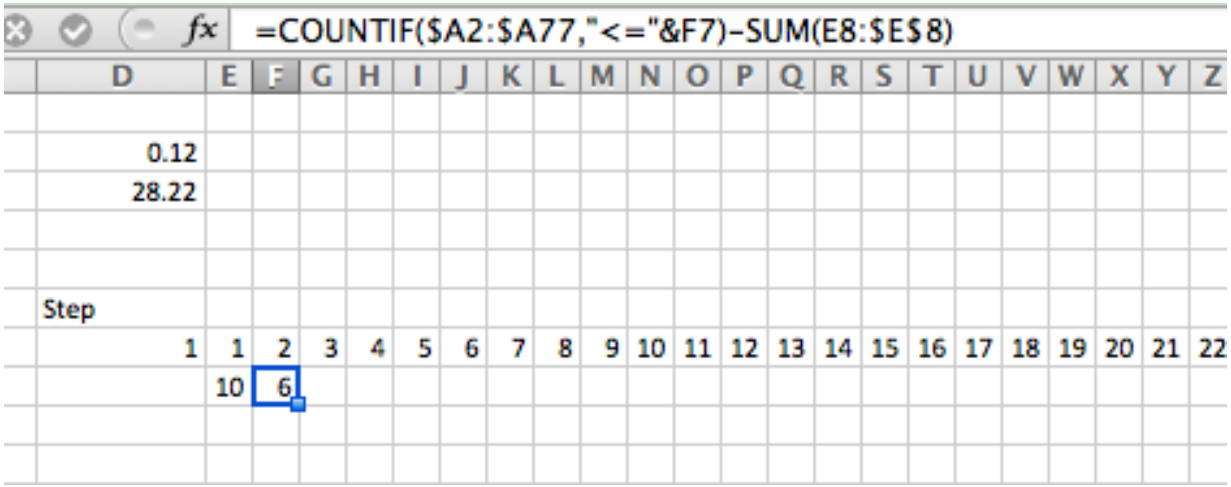
Because you do not want the values that fall in the less than 1 category to also fall under the between 1 and 2 category, you have to calculate the difference between each bin value.
Lock the data range by selecting those numbers in the expression and hitting F4 so that our expression references the right cells no matter where we copy and paste the expression.
Next you subtract the sum of all of the previous bin values from your count. You can do this by typing –SUM().

=SUM(number1, $number2)

- **number1**: the most recently populated cell in your bin set.
- **$number2**: the locked first output you calculated.

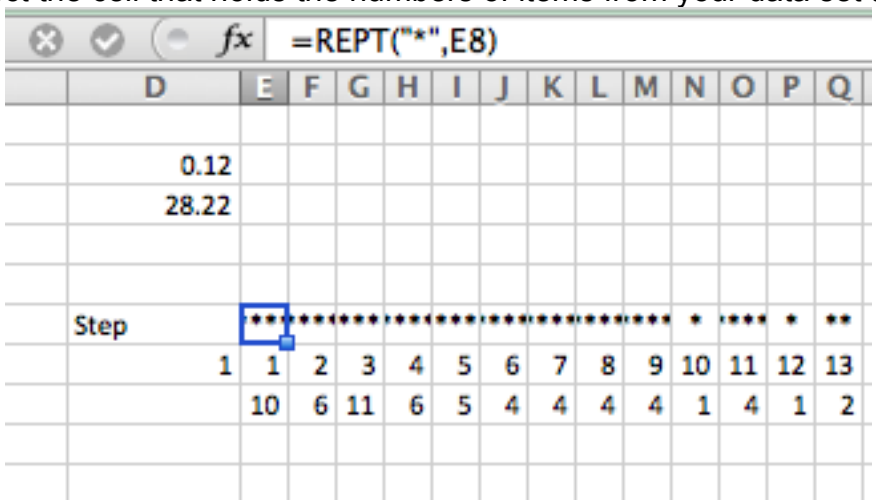
This will give us the number of items in our data that fall exclusively into that given bin.

8. Display a symbol or dot for the total in each bin.



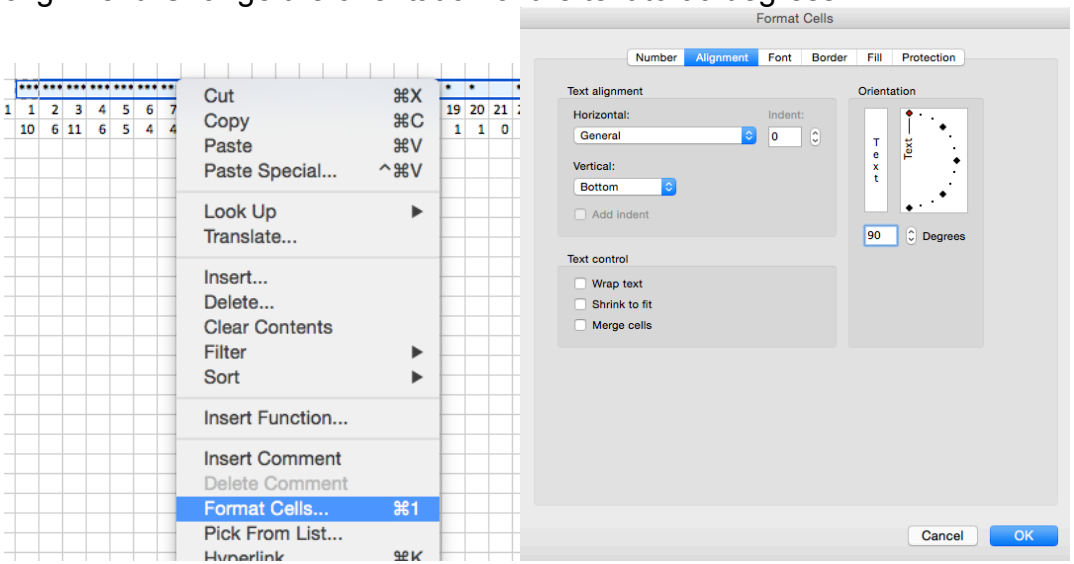
In the cell above the first bin type =REPT().
=REPT(text, num_times)

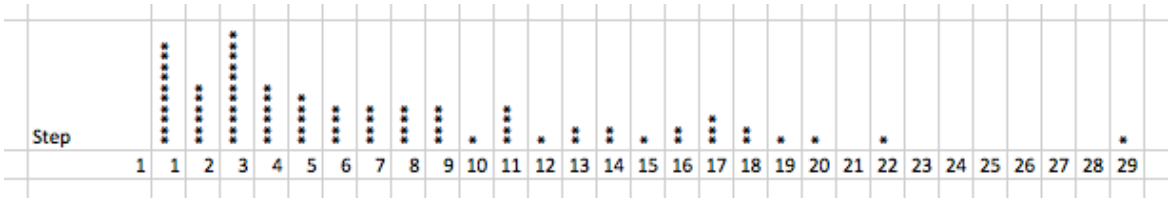
- **text**: the symbol we would like to repeat for the dot plot
- **num_times**: the number of times the text will repeat. For the dot plot you should select the cell that holds the numbers of items from your data set that fit into that bin.



9. Format dot plot.

Align the text of the dots to be vertical by selecting format cell and adjusting the alignment. Change the orientation of the text to 90 degress.





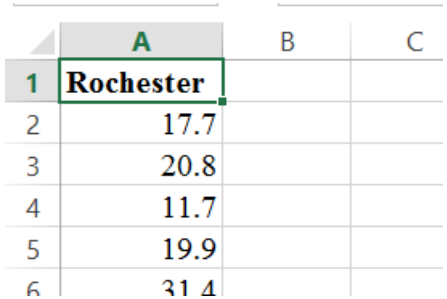
Create a Histogram

Excel Step-by-Step How-to for Windows

Excel for Mac Instructions on page 12

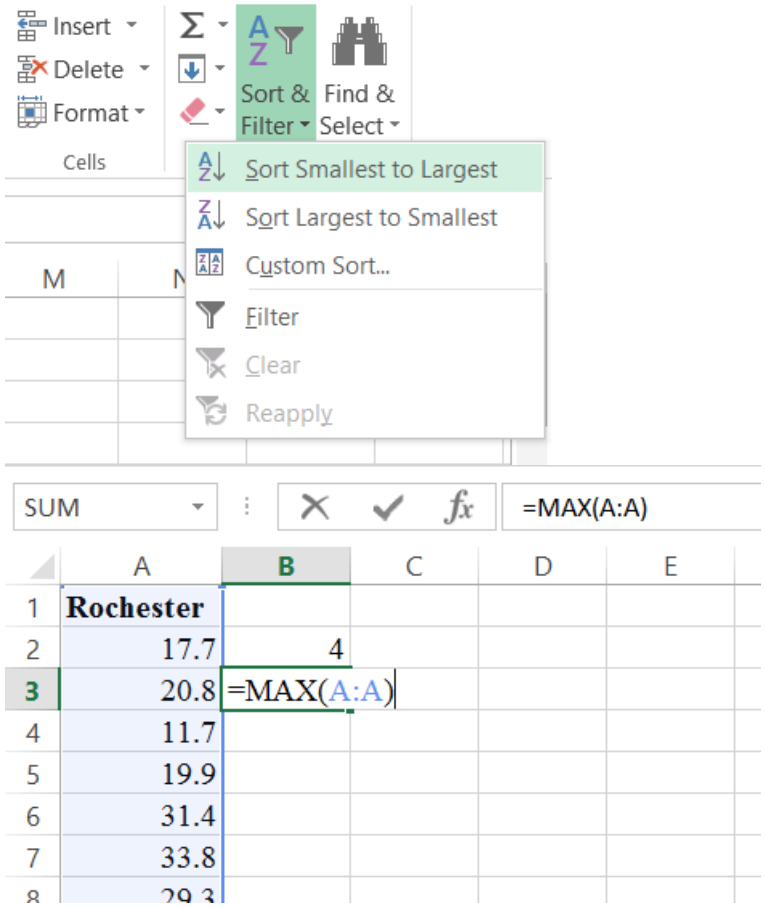
Instructions: Use this guide to create a histogram using Excel.

Data requirement: one variable with quantitative data.
Sample Data: yearly snowfall in Rochester.

Step	Windows Instructions + Screen Shot
1. Arrange the data you want to use into a column. If you have multiple variables, ensure that each column is a unique variable.	

2. Determine the minimum and maximum values of your data set.

If you have a large data set, you may want to use Excel to find the smallest and largest point. These values make it easier to determine the starting and ending values for your histogram. Select the column with your data and then use the “Sort” function or type =MIN(A:A) in a blank cell in a different column (i.e. column B) and then type =MAX(A:A) to get the biggest number. The data range A:A will select everything in column A.



3. Based on your minimum and maximum values, choose an appropriate bin size for your histogram.

A bin is the interval by which you want to sort your data. A histogram displays how many values from your data set fall into each bin (this is known as the frequency of the bin).

TIP: Selecting the right bin sizes for your histogram:

It is important to choose a bin size that is not too small or too large. You want the bin to be wide enough to show a pattern of distribution. If your bin is too small, it will be hard to make sense of the patterns.

4. In a separate area of your Excel sheet, type-in your bin ranges.

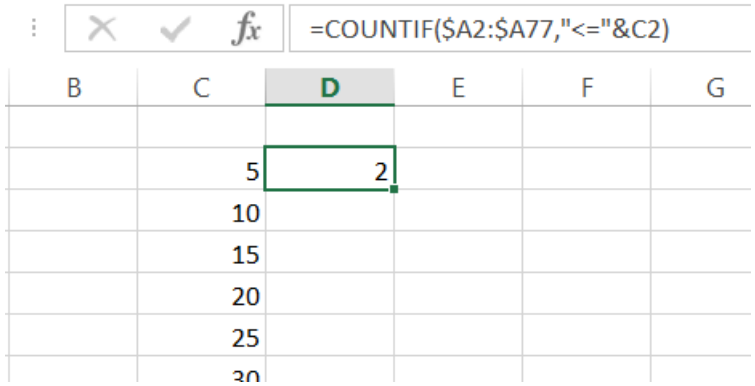
In the Rochester snowfall example, the first bin would start at 5, the next one at 10, the next at 15, etc. Each bin should have the same width, or range.

	A	B	C	D	E	F	G
1	**Rochester**						
2	17.7		5				
3	20.8		10				
4	11.7		15				
5	19.9		20				
6	31.4		25				
7	33.8		30				
8	29.3		35				
9	15.2		40				
10	29.8		45				
11	14		50				
12	49.7		55				
13	61.3		60				
14	43.4		65				
15	11.9						

5. In the empty column next to your bin ranges, use an Excel formula to count how many values are less than or equal to each bin range.

=COUNTIF(range, criteria)

- **Range**: all the values of your data set
- **Criteria**: "<="&*firstbinnumber* (this criteria sub-formula indicates that you want to count all the numbers that are less than or equal to your first bin range number. In this example, the completed formula would look like this:



	A	B	C	D	E	F	G
1	**Rochester**						
2	17.7		5	2			
3	20.8		10	7			
4	11.7		15	19			
5	19.9		20	30			
6	31.4		25	41			
7	33.8		30	52			
8	29.3		35	56			
9	15.2		40	60			
10	29.8		45	63			
11	14		50	65			
12	49.7		55	64			
13	61.3		60	63			
14	43.4		65	64			

=COUNTIF(B2:B77, “<=”&E9), where “E9” is populated with the number 5, which is the first bin range number. Note that your formula may call to another cell in your sheet that is not “E9.” The key here is to reference the appropriate bin range number. Repeat this for each bin range.

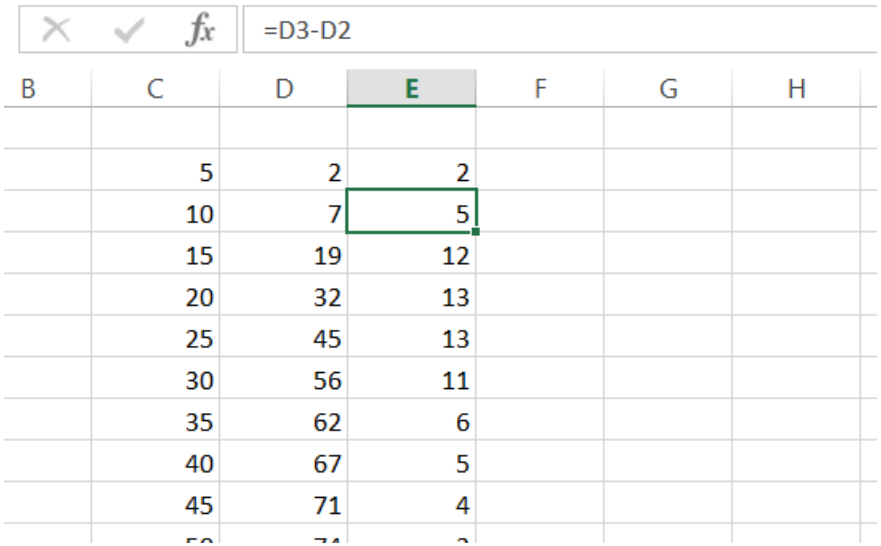
TIP: To quickly replicate formulas across cells:

First, it is important to lock the appropriate values in your formula. Locking a values makes the formula always reference the locked cell instead of changing based on location. You can lock the values in each formula by placing a "$" in front of each coordinate. For example, by representing the formula as B$24+C$35, the value will always calculate using rows 24-35 no matter where that formula is placed. Note that if you also place a “$” in front of the column letter, that will also lock the column.

Then, highlight the cell with the formula you want to copy, and drag the cursor across or down to the next cell.

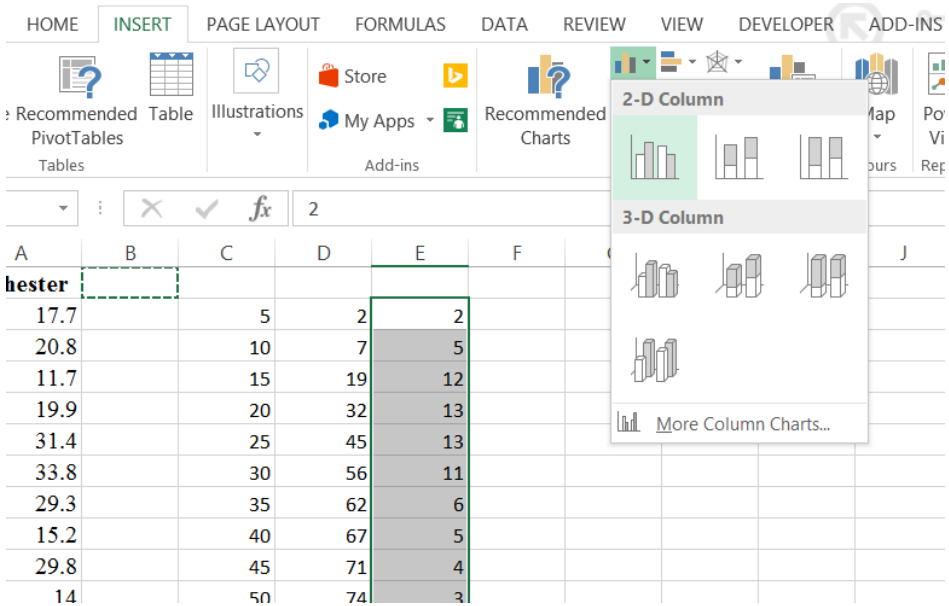
6. Generate frequency values that count how many values fall into each bin range.

In a third column next to your results from step 4, calculate the difference between each pair of bin count totals. This step is necessary because you don't want any values to be counted as falling into multiple bins. If there are 7 values less than 10, and 2 values less than 5, then there must be 5 values in the 5-10 bin. Note that the frequency of your first bin should equal whatever result you got for that bin in step 4. These values will be the histogram bar heights.



7. Insert your histogram.

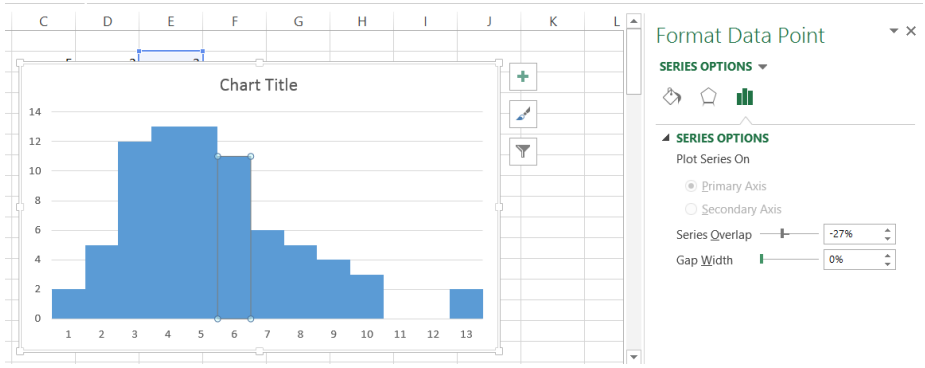
Highlight all the bin frequency values you calculated in step 5. Then, from the “insert” tab in the top toolbar, click to insert a simple 2-D bar chart.



8. Adjust the bar gap width to “0.”

To do this, double click on any of the bars of the chart to open the "Format data series" menu. Change "gap width" to “0%”.

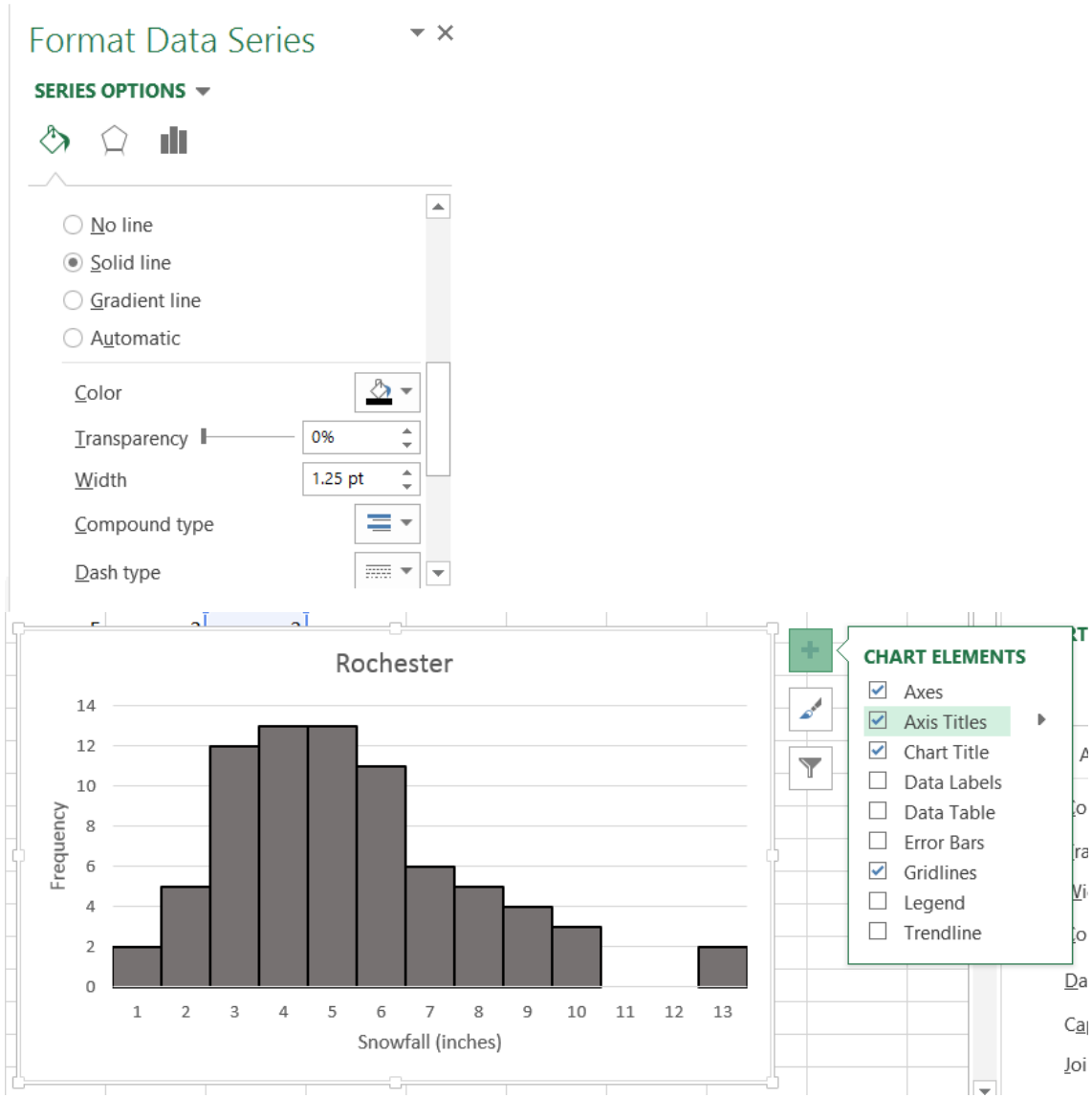
Note that histograms should never have gaps between bars since bins are continuous.



9. Format the histogram's display settings for legibility.

To do this, double click on any of the bars of the chart to open the "Format data series" menu.

- Change the "fill" color to gray
- Change the "border" to a solid black line
- Click on the "+" sign to add "axis titles" to the graph. In this example, the horizontal axis title should be "Snowfall (inches)" and the vertical axis title should be "Frequency"



TIP: To hide all the field buttons on a chart:

Right click on each field button and select "Hide all fields buttons on chart."

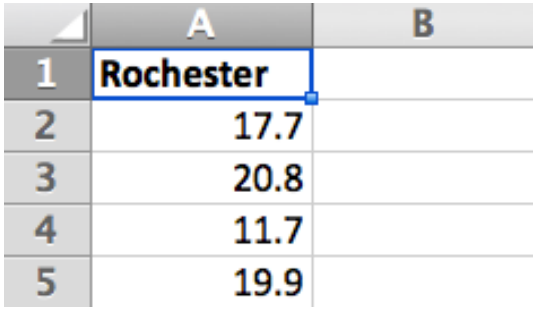
Create a Histogram

Excel Step-by-Step How-to for Mac

Excel for Windows Instructions on page 1

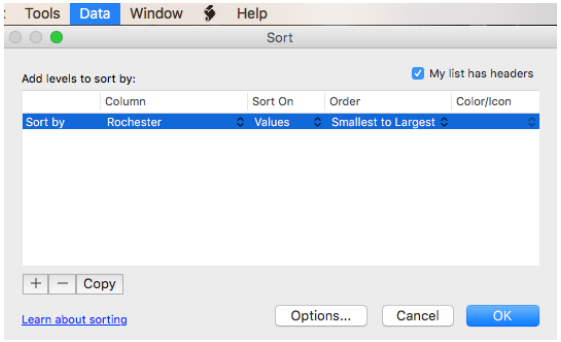
Instructions: Use this guide to create a histogram using Excel.

Data requirement: one variable with quantitative data.
Sample Data: yearly snowfall in Rochester.

Step	Mac Instructions + Screen Shot
1. Arrange the data you want to use into a column. If you have multiple variables, ensure that each column is a unique variable.	

2. Determine the minimum and maximum values of your data set.

If you have a large data set, you may want to use Excel to find the smallest point and the largest point to make it easier to determine the starting and ending values for your histogram. Type your data into a single column and then use the “Sort” function or type =MIN(A:A) in a blank cell in a different column (i.e. column B) and then type =MAX(A:A) to get the biggest number.
*For Macs: The “Sort” function is found within the Data tab. The ‘Order’ is where ‘Smallest to Largest’ can be selected.



	A	B
1	Rochester	
2	17.7	4
3	20.8	=MAX(A:A)
4	11.7	
5	19.9	

3. Based on your minimum and maximum values, choose an appropriate bin size for your histogram.

A bin is the interval by which you want to sort your data. A histogram displays how many values from your data set fall into each bin (this is known as the frequency of the bin). It’s important to choose a bin size that is not too small or too large. You want the bin to be wide enough to show a pattern of distribution. If your bin is too small, it will be hard to make sense of the patterns.

TIP: Selecting the right bin sizes for your histogram:

Professor input here… It's easiest to first calculate the range of your data set…

4. In a separate area of your Excel sheet, type-in your bin ranges.

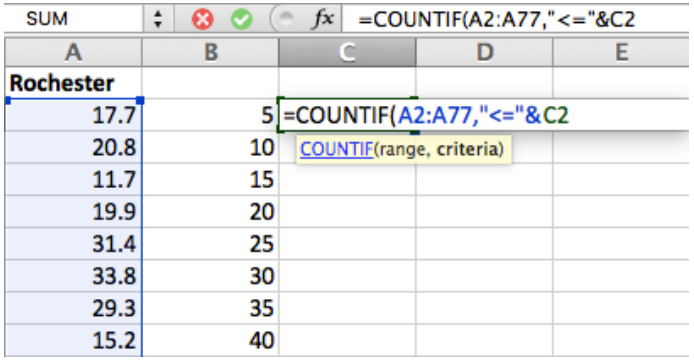
In the Rochester snowfall example, the first bin would start at 5, the next one at 10, the next at 15, etc. Each bin should have the same width, or range.

	A	B
1	**Rochester**	
2	17.7	5
3	20.8	10
4	11.7	15
5	19.9	20
6	31.4	25
7	33.8	30
8	29.3	35
9	15.2	40

5. In the empty column next to your bin ranges, use an Excel formula to count how many values are less than or equal to each bin range.

=COUNTIF(range, criteria)

- **Range**: all the values of your data set
- **Criteria**: "<="&*firstbinnumber* (this criteria sub-formula indicates that you want to count all the numbers that are less than or equal to your first bin range number. In this example, the completed formula would look like this:



=COUNTIF(B2:B77, "<="&E9), where "E9" is populated with the number 5, which is the first bin range number. Note that your formula may call to another cell in your sheet that is not "E9." The key here is to reference the appropriate bin range number. Repeat this for each bin range.

	A	B	C
1	**Rochester**		
2	17.7	5	2
3	20.8	10	7
4	11.7	15	19
5	19.9	20	30
6	31.4	25	41
7	33.8	30	52
8	29.3	35	56
9	15.2	40	60

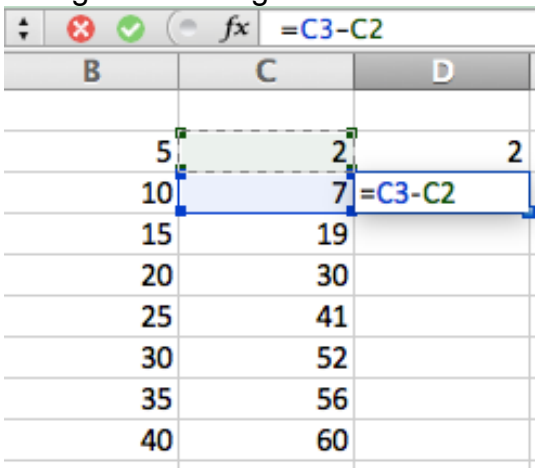
TIP: To quickly replicate formulas across cells:

First, lock the values in each formula by placing a "$" in front of each coordinate. For example, by representing the formula as B$24+C$35, the value will always calculate using rows 24-35 no matter where that formula is placed. Note that if you also place a “$” in front of the column letter, that will also lock the column.

Then, highlight the cell with the formula you want to copy, and drag the cursor across or down to the next cell.

6. Generate frequency values that count how many values fall into each bin range.

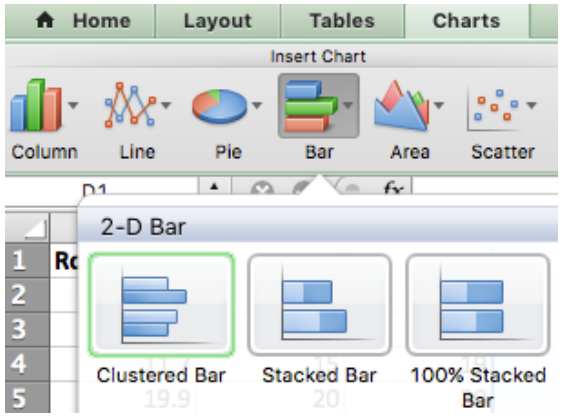
In a third column next to your results from step 4, calculate the difference between each pair of bin count totals. This step is necessary because you don’t want any overlap in the numbers that fall into each bin. If there are 7 values less than 10, and 2 values less than 5, then there must be 5 values in the 5-10 bin. Note that the frequency of your first bin should equal whatever result you got for that bin in step 4. These values will be the histogram bar heights.



7. Insert your histogram.

Highlight all the bin frequency values you calculated in step 5. Then, from the “charts” tab, click the “Bar” menu. Then, click “Clustered Bar” to insert a simple 2-D bar chart. A bar chart will automatically be inserted once the “Clustered Bar” button is clicked.

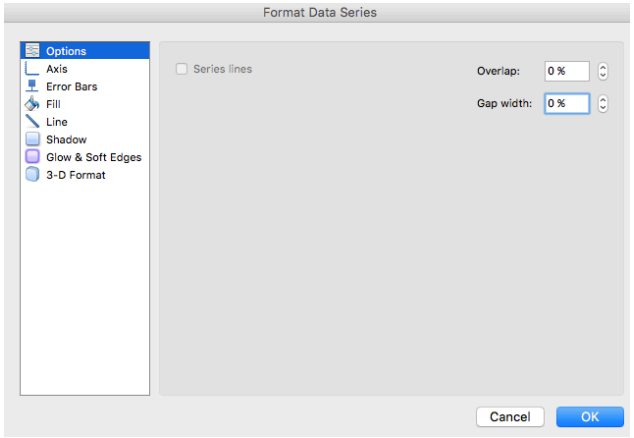
A	B	C	D
Rochester			
17.7	5	2	2
20.8	10	7	5
11.7	15	19	12
19.9	20	30	13
31.4	25	41	13
33.8	30	52	11
29.3	35	56	6
15.2	40	60	5



8. Adjust the bar gap width to “0.”

To do this, double click on any of the bars of the chart to open the "Format data series" menu. Change "Gap width" to “0%”

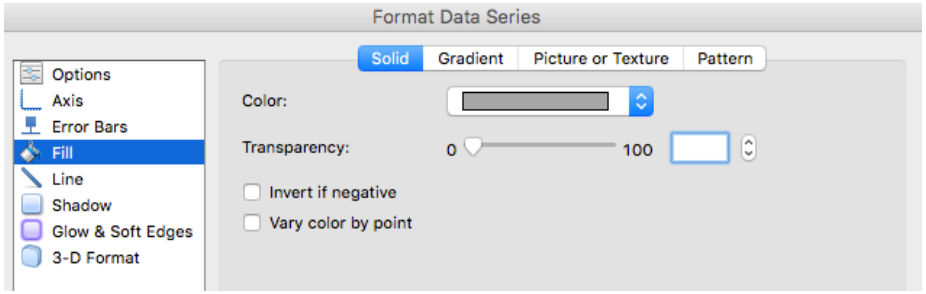
Note that histograms should never have gaps between bars since bins are continuous.

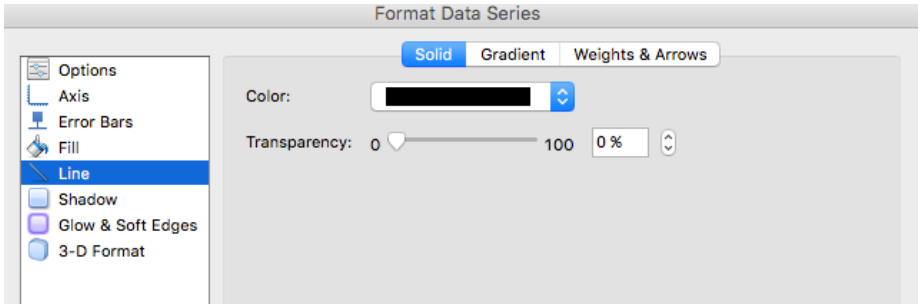


9. Format the histogram's display settings for legibility.

To do this, double click on any of the bars of the chart to open the "Format data series" menu.

- Change the "fill" color to gray
- Change the "border" to a solid black line
- Click on the "+" sign to add "axis titles" to the graph. In this example, the horizontal axis title should be "Snowfall (inches)" and the vertical axis title should be "Frequency"





TIP: To hide all the field buttons on a chart:

Right click on each field button and select "Hide all fields buttons on chart."

Create a Pivot Table

Excel Step-by-Step How-to for Windows

Excel for Mac Instructions on page 7

Instructions: Use this guide to create a pivot table.

Data requirement: three variables (two with quantitative data, one with categorical data)
Sample Data: yearly snowfall in Rochester and Syracuse.

Step	Windows Instructions + Screen Shot
1. Arrange the data you want to use into columns, ensuring each column is a unique variable.	

	A	B	C	D	E
1	**Year**	**Snowfall**	**City**		
2	2015	17.7	Rochester		
3	2014	20.8	Rochester		
4	2013	11.7	Rochester		
5	2012	19.9	Rochester		
6	2011	31.4	Rochester		
7	2010	33.8	Rochester		
8	2009	29.3	Rochester		
9	2008	15.2	Rochester		
10	2007	29.8	Rochester		

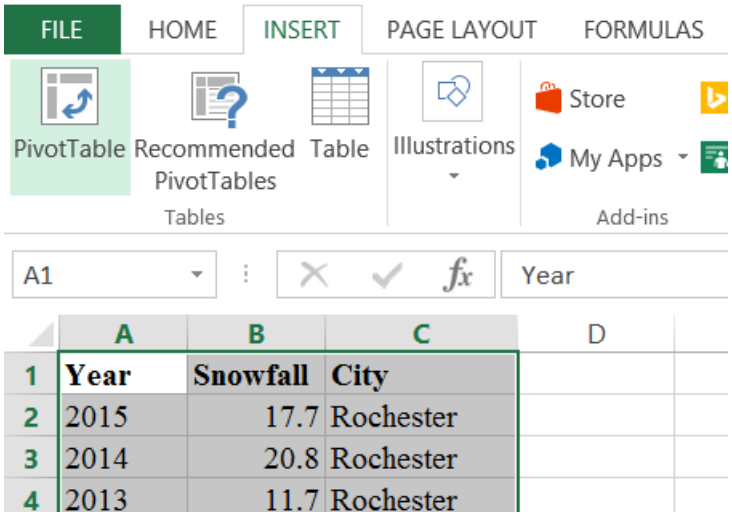
2. Select all data values for all variables.

To do this, click on the first cell in the top left corner of the excel sheet, and drag your cursor to the bottom right cell of the last column in your data.

	A	B	C	D	E
1	**Year**	**Snowfall**	**City**		
2	2015	17.7	Rochester		
3	2014	20.8	Rochester		
4	2013	11.7	Rochester		
5	2012	19.9	Rochester		
6	2011	31.4	Rochester		
7	2010	33.8	Rochester		
8	2009	29.3	Rochester		
9	2008	15.2	Rochester		
10	2007	29.8	Rochester		
11	2006	14	Rochester		
12	2005	49.7	Rochester		

3. Create a pivot table from your data set.

From the “insert” tab, click “Pivot Table.” Click “ok” to insert the pivot table in a new sheet. Label the new sheet “Descriptives.”

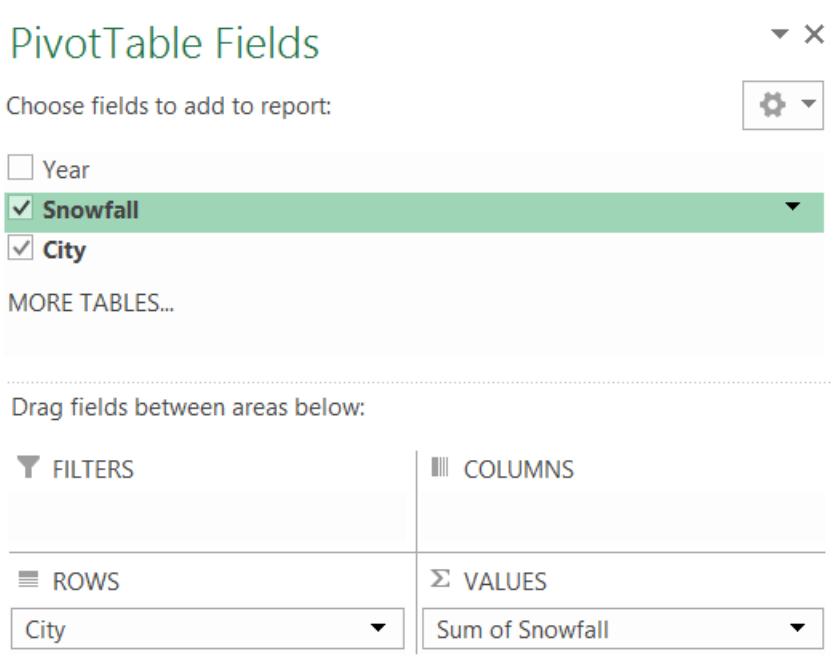


4. Specify data that you want represented in pivot table.

From the new sheet that has the pivot table, click on any part of the pivot table to make the "Pivot Table Fields" menu appear on the right.

You can select any of your variables by checking them off. Whatever category they appear under in the PivotTable Fields menu will be there they are represented in your table.

Check off the "city" field to add it to a report. Make sure it appears under the "Rows" area. Then check off the "snowfall" field to add it to the report. Make sure it appears under the "Values" areas.

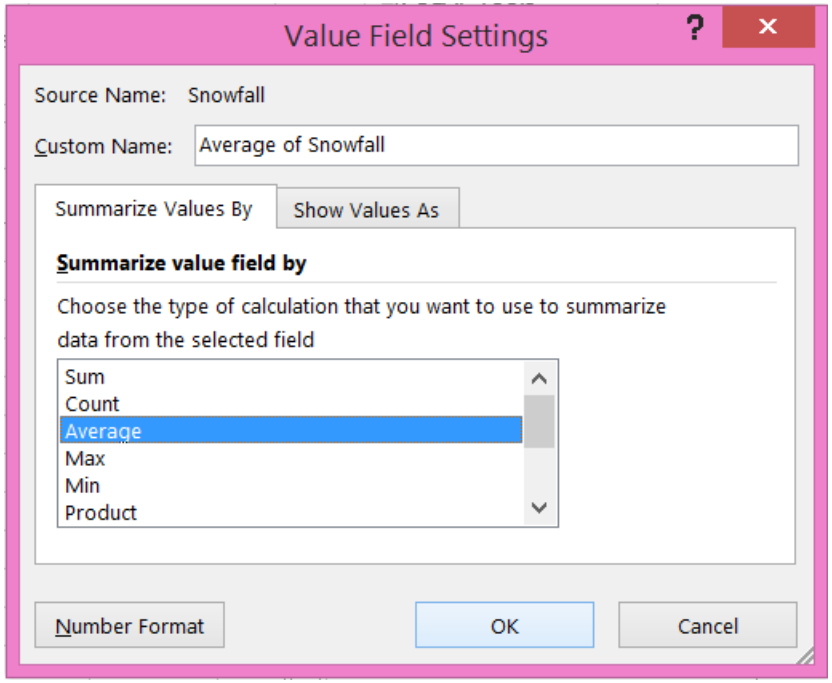


TIP: To easily change fields of your pivot table:

Simply drag and drop fields between the four areas in the pivot table menu.

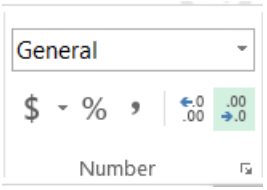
5. Change a value data summary from being a sum to being an average.

To do this, click the dropdown arrow on “Sum of Snowfall” field in the “Values” area, then click “value field settings.” Select “average” from the “Summarize values by” tab on the “value field settings” menu. Click “ok” to apply changes.



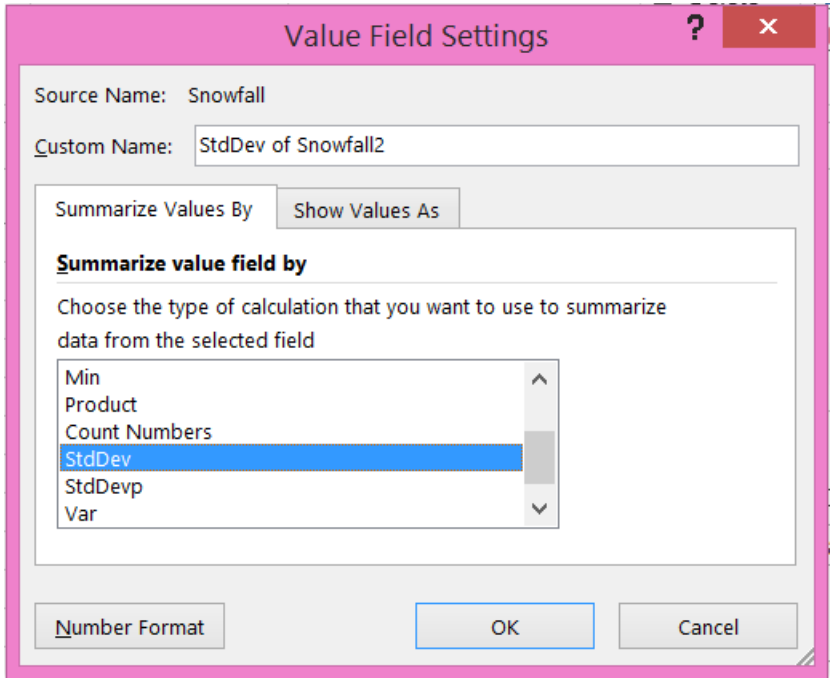
6. Decrease the number of decimal places to one or two for all the values in your pivot table.

Use the decimal button on the top toolbar to decrease the number of decimal places to one or two for all the values in your pivot table.



7. Add another field to values and change the field to be a standard deviation.

Using the “value field settings” menu (as in Step 5), change the field setting of the second "snowfall" field to be a standard deviation by selecting "StdDev" from the "Summarize values by" tab on the "value field settings" menu. Click "ok" to apply changes.



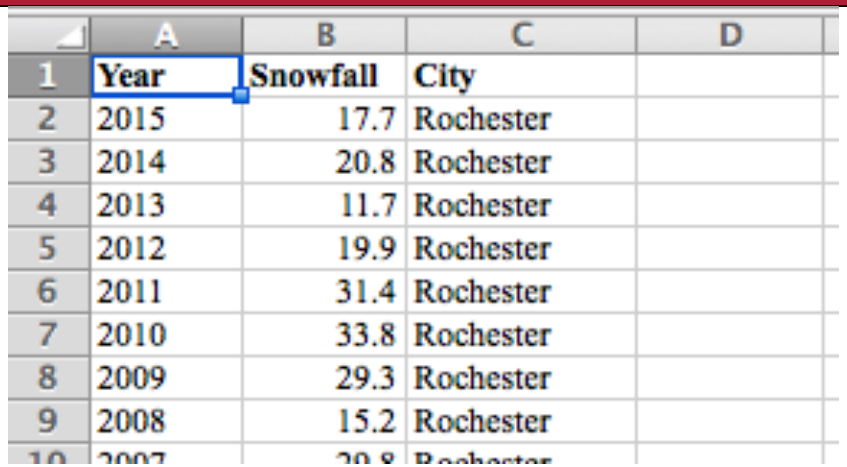
Create a Pivot Table

Excel Step-by-Step How-to for Mac

Excel for Windows Instructions on page 1

Instructions: Use this guide to create a pivot table.

Data requirement: three variables (two with quantitative data, one with categorical data)
Sample Data: yearly snowfall in Rochester and Syracuse.

Step	Mac Instructions + Screen Shot
1. Arrange the data you want to use into columns, ensuring each column is a unique variable.	

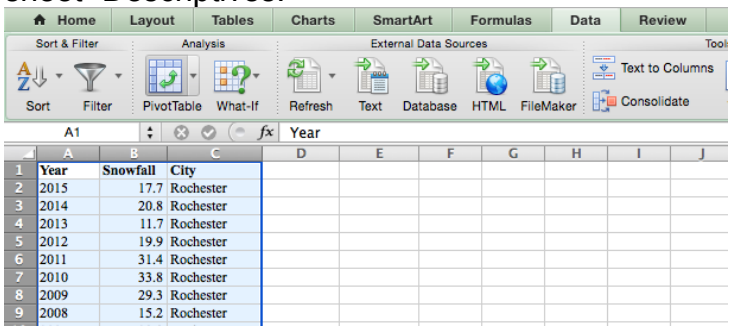
	A	B	C	D
1	Year	Snowfall	City	
2	2015	17.7	Rochester	
3	2014	20.8	Rochester	
4	2013	11.7	Rochester	
5	2012	19.9	Rochester	
6	2011	31.4	Rochester	
7	2010	33.8	Rochester	
8	2009	29.3	Rochester	
9	2008	15.2	Rochester	

2. Select all data values for all variables.

To do this, click on the first cell in the top left corner of the excel sheet, and drag your cursor to the bottom right cell of the last column in your data.

3. Create a pivot table from your data set.

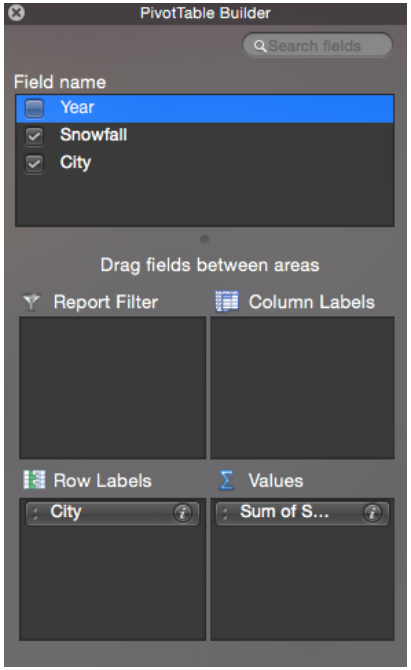
From the “Data” tab, click “PivotTable” to insert the pivot table in a new sheet. Label the new sheet “Descriptives.”



	A	B	C	D	E	F	G	H	I	J
1	Year	Snowfall	City							
2	2015	17.7	Rochester							
3	2014	20.8	Rochester							
4	2013	11.7	Rochester							
5	2012	19.9	Rochester							
6	2011	31.4	Rochester							
7	2010	33.8	Rochester							
8	2009	29.3	Rochester							
9	2008	15.2	Rochester							

4. Customize pivot table.

Select the variable you would like to add to a report. Make sure it appears under the proper area.

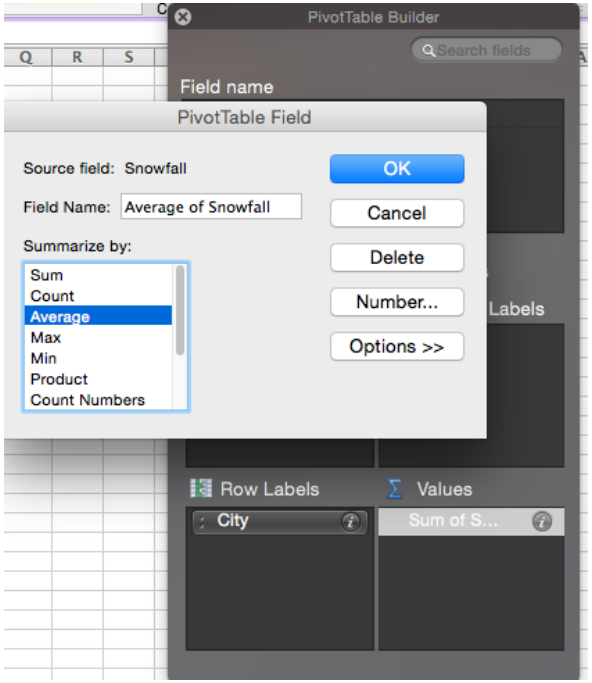


TIP: To easily move fields around in the pivot table menu:

Simply drag and drop fields between the four areas in the pivot table menu.

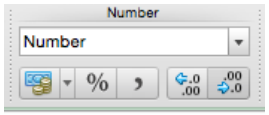
5. Change a value data summary form being a sum to being an average.

To do this, click the “i” icon on “Sum of Snowfall” field in the “Values” area, then click “value field settings.” Select “average” from the “Summarize values by” tab. Click “ok” to apply changes.



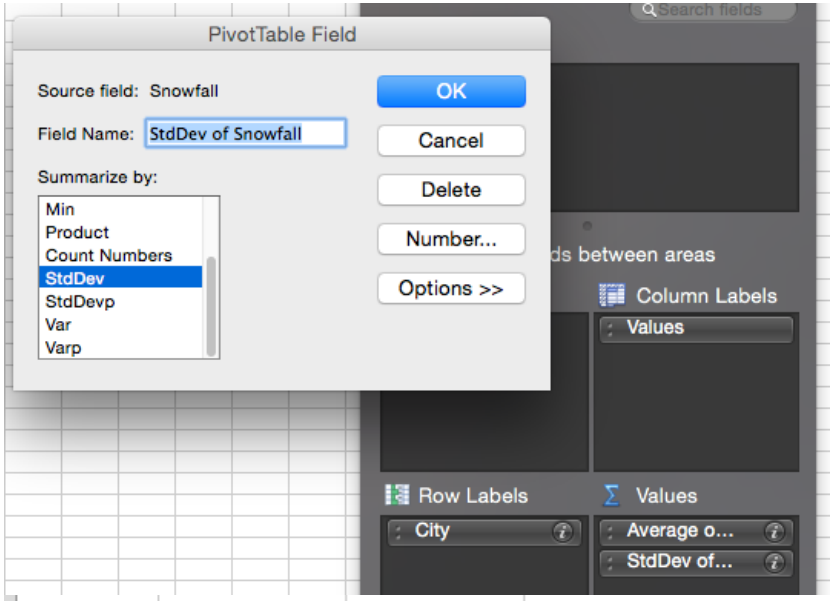
6. Decrease the number of decimal places to one or two for all the values in your pivot table.

Use the decimal button on the top toolbar to decrease the number of decimal places to one or two for all the values in your pivot table.



7. Add another field to values and change the field to be a standard deviation.

Using the “value field settings” menu (as in Step 5), change the field setting of the second "snowfall" field to be a standard deviation by selecting "StdDev" from the "Summarize values by" tab. Click "ok" to apply changes.



	Values	
Row Labels	Average of Snowfall	StdDev of Snowfall
Rochester	24.4	12.3
Vancouver	7.0	6.1
Grand Total	**15.7**	**13.1**

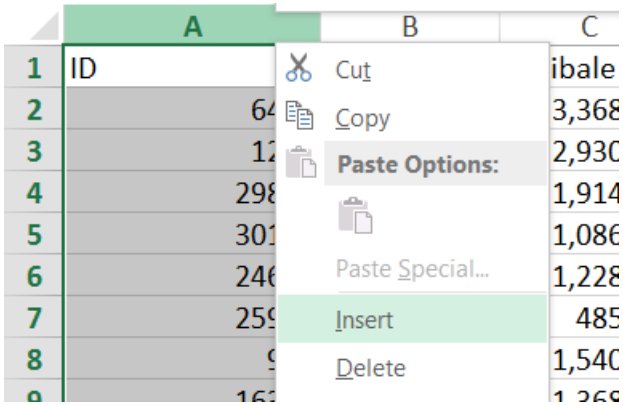
Create a Random Sample

Excel Step-by-Step How-to for Windows

Excel for Mac Instructions on page 7

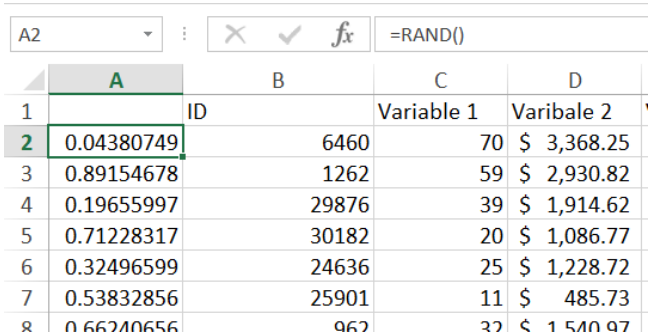
Instructions: Use this guide to create a random sample from a list of a population

Data requirement: data organized as a list of a population
Sample Data: all past consumers

Step	Windows Instructions + Screen Shot
1. Insert a column to the left of your data.	Right click the column and select insert. 

2. Fill this column with random values.

Select the cell beside your first line of data. Enter =RAND() to fill the cell with a random decimal between 0 and 1.



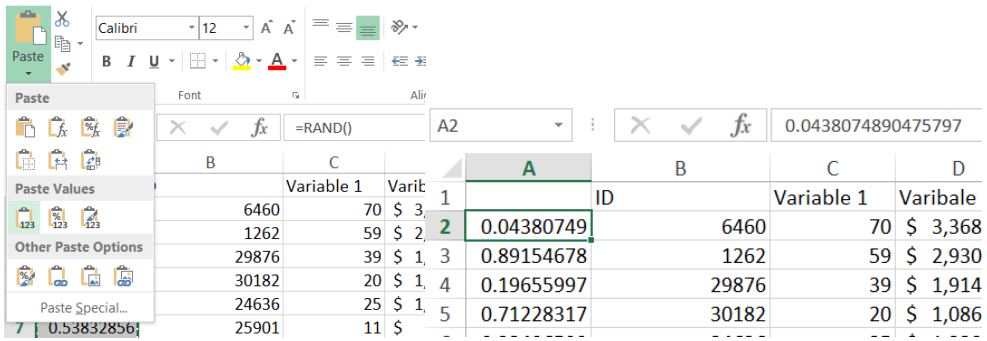
	A	B	C	D
1		ID	Variable 1	Varibale 2
2	0.04380749	6460	70	$ 3,368.25
3	0.89154678	1262	59	$ 2,930.82
4	0.19655997	29876	39	$ 1,914.62
5	0.71228317	30182	20	$ 1,086.77
6	0.32496599	24636	25	$ 1,228.72
7	0.53832856	25901	11	$ 485.73
8	0.66240656	962	32	$ 1,540.97

TIP: To quickly fill a column with numbers:

Double click the bottom right corner of the top cell to autofill the column with that information.

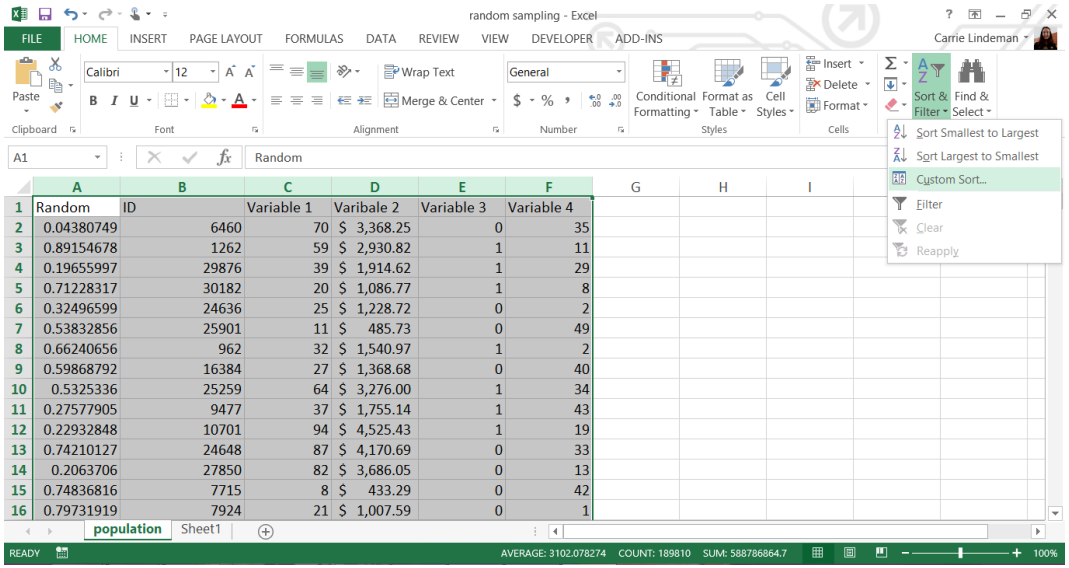
3. Replace the column of functions with a column of values.

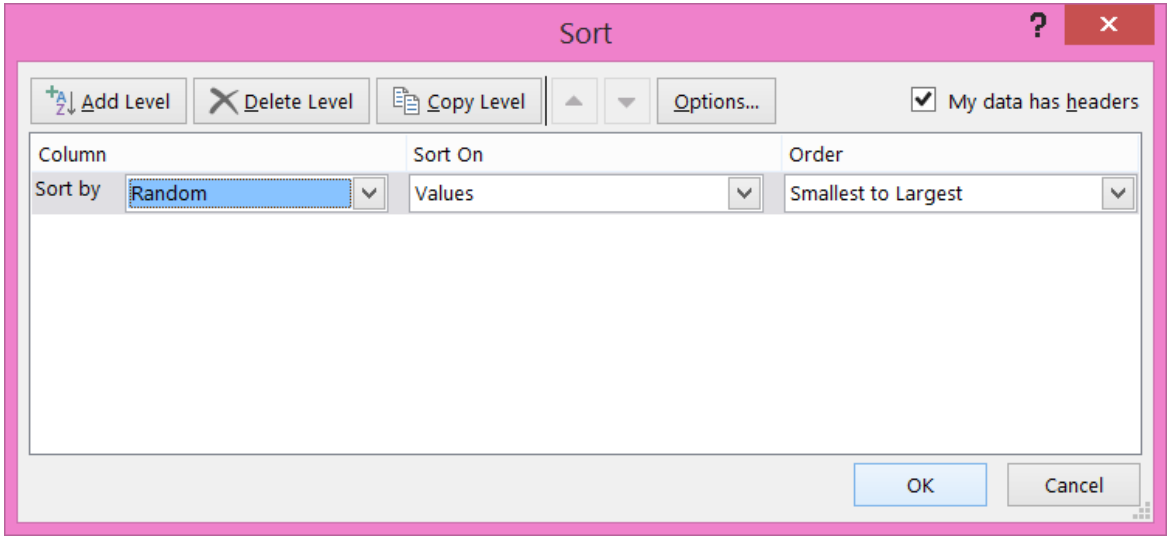
Select the column you just created and copy it. Navigate to Paste and select Special Paste Values.



4. Sort the data by the random column.

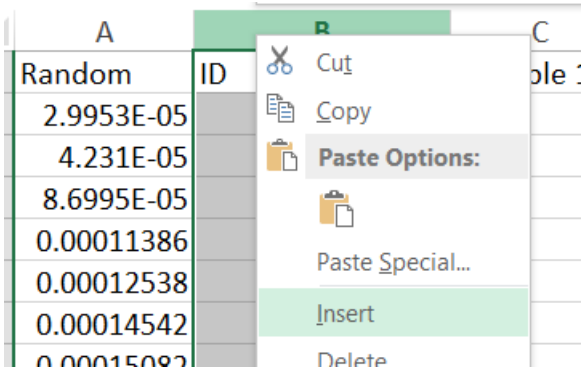
Select your random column and all of your data. Navigate to Sort & Filter and select Custom Sort. In the Sort by field, select your random column.



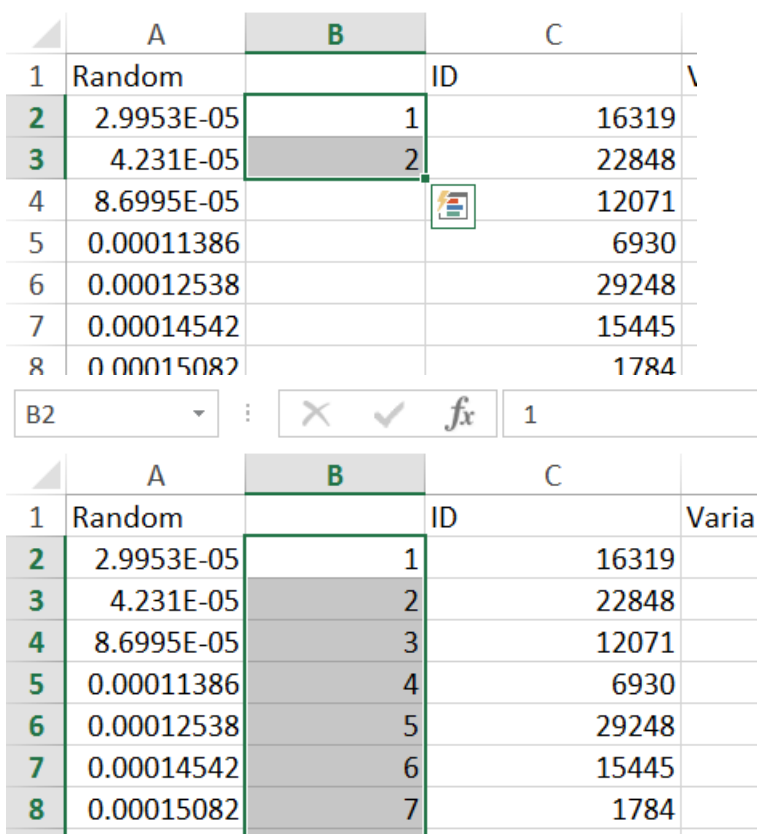


5. Insert another column to beside the random column and fill it with a counter.

Right click the first column of your data and select insert.
Enter 1 in the first cell and 2 in the second cell. Highlight both of these cells and double click the bottom right corner of the selection to autofill the whole column.



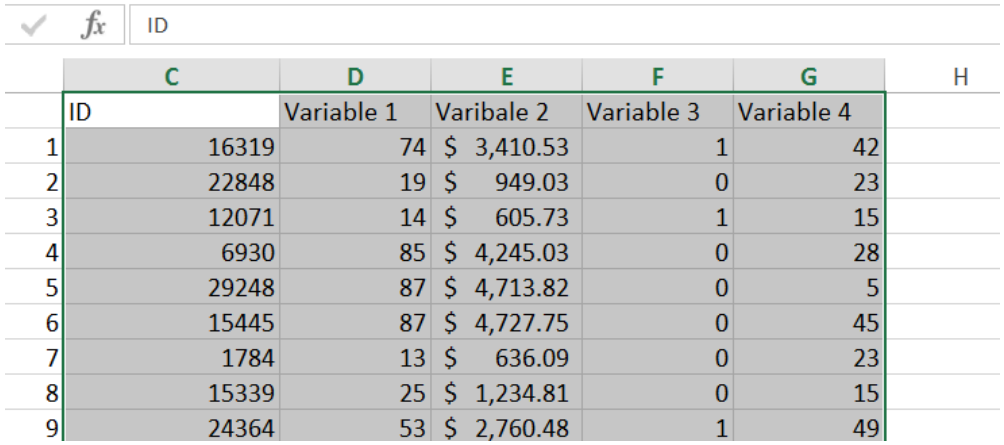
6. Select from row 1 to the size of the sampling you want.	Scroll to the row number of the sampling you want and select all of the rows from there to the top.



	A	B	C
1	Random		ID
2	2.9953E-05	1	16319
3	4.231E-05	2	22848
4	8.6995E-05		12071
5	0.00011386		6930
6	0.00012538		29248
7	0.00014542		15445
8	0.00015082		1784

B2 | 1

	A	B	C
1	Random		ID
2	2.9953E-05	1	16319
3	4.231E-05	2	22848
4	8.6995E-05	3	12071
5	0.00011386	4	6930
6	0.00012538	5	29248
7	0.00014542	6	15445
8	0.00015082	7	1784



ID

	C	D	E	F	G	H
	ID	Variable 1	Varibale 2	Variable 3	Variable 4	
1	16319	74	$ 3,410.53	1	42	
2	22848	19	$ 949.03	0	23	
3	12071	14	$ 605.73	1	15	
4	6930	85	$ 4,245.03	0	28	
5	29248	87	$ 4,713.82	0	5	
6	15445	87	$ 4,727.75	0	45	
7	1784	13	$ 636.09	0	23	
8	15339	25	$ 1,234.81	0	15	
9	24364	53	$ 2,760.48	1	49	

Create a Random Sample

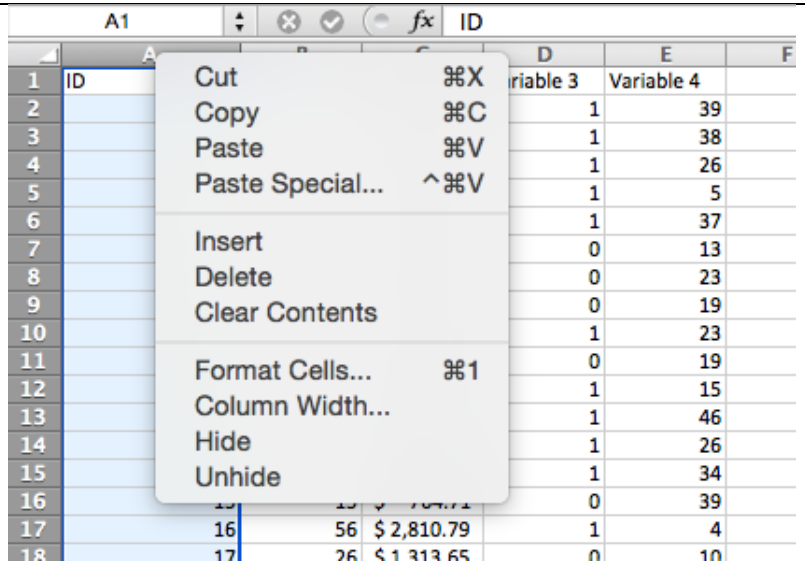
Excel Step-by-Step How-to for Mac

Excel for Windows Instructions on page 1

Instructions: Use this guide to create a random sample from a list of a population

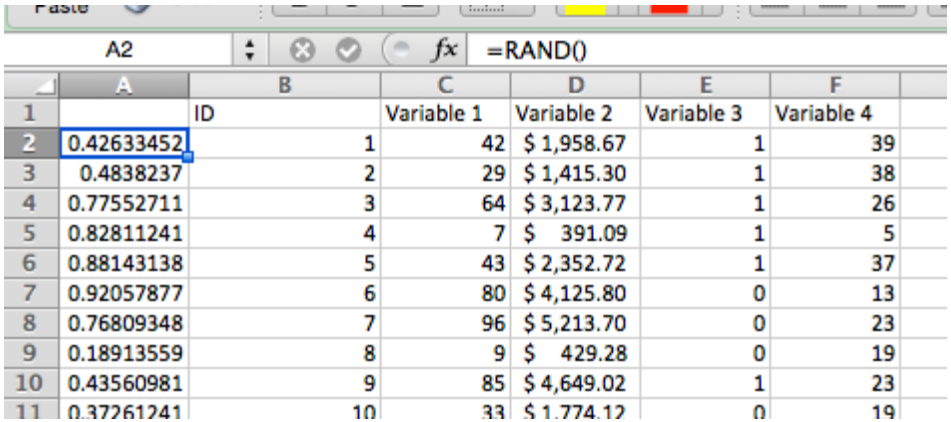
Data requirement: data organized as a list of a population
Sample Data: all past consumers

Step	Mac Instructions + Screen Shot
1. Insert a column to the left of your data.	Control-click the column and select insert.



2. Fill this column with random values.

Select the cell beside your first line of data. Enter =RAND() to fill the cell with a random decimal between 0 and 1.



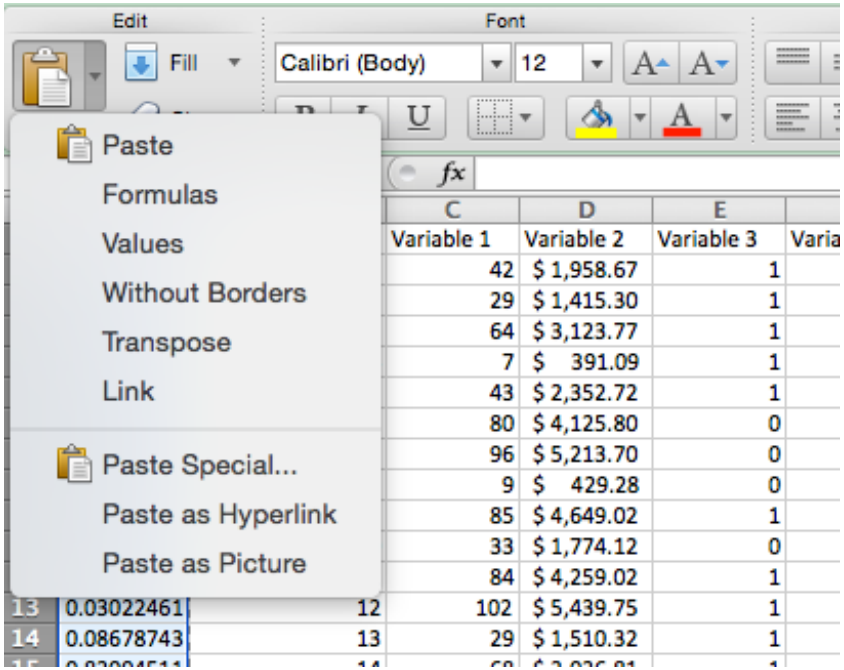
	A	B	C	D	E	F
1		ID	Variable 1	Variable 2	Variable 3	Variable 4
2	0.42633452	1	42	$ 1,958.67	1	39
3	0.4838237	2	29	$ 1,415.30	1	38
4	0.77552711	3	64	$ 3,123.77	1	26
5	0.82811241	4	7	$ 391.09	1	5
6	0.88143138	5	43	$ 2,352.72	1	37
7	0.92057877	6	80	$ 4,125.80	0	13
8	0.76809348	7	96	$ 5,213.70	0	23
9	0.18913559	8	9	$ 429.28	0	19
10	0.43560981	9	85	$ 4,649.02	1	23
11	0.37261241	10	33	$ 1,774.12	0	19

TIP: To quickly fill a column with numbers:

Double click the bottom right corner of the top cell to autofill the column with that information.

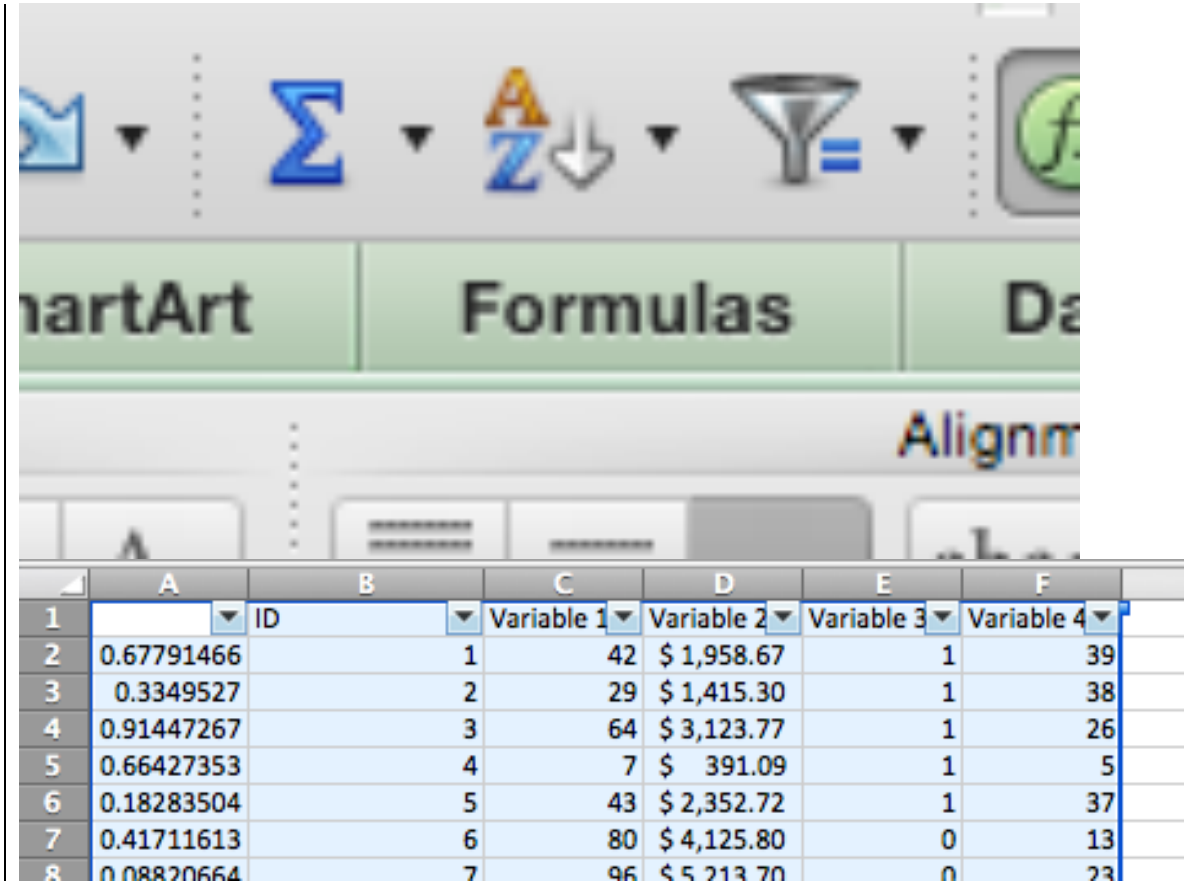
3. Replace the column of functions with a column of values.

Select the column you just created and copy it (command C). Navigate to Paste and select Values.

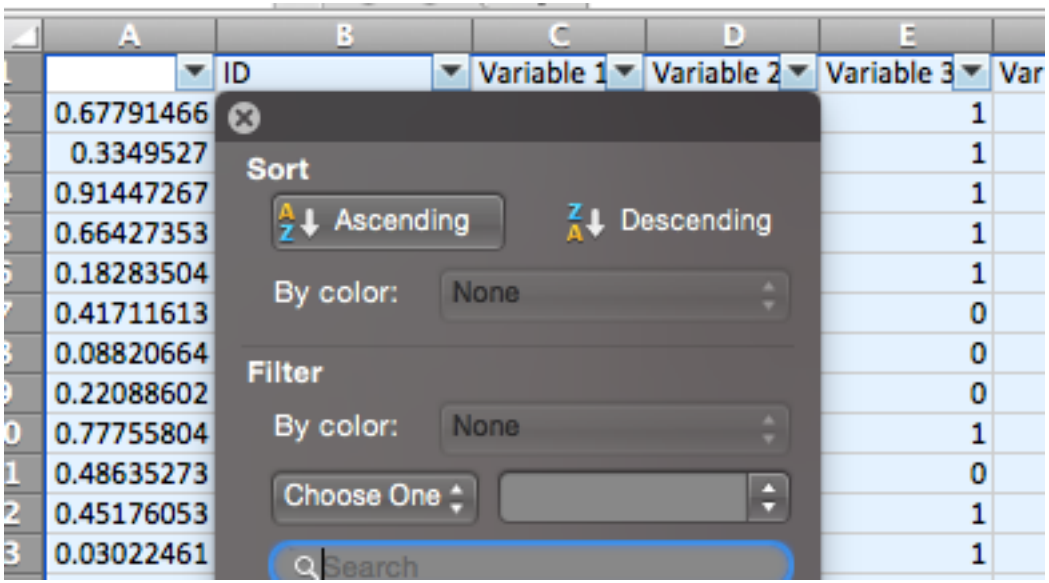


4. Sort the data by the random column.

Select your random column and all of your data. Navigate to Sort & Filter icon and click it. Drop down arrows should appear beside each variable of your data. Select the arrow beside your random variable and sort it by ascending.

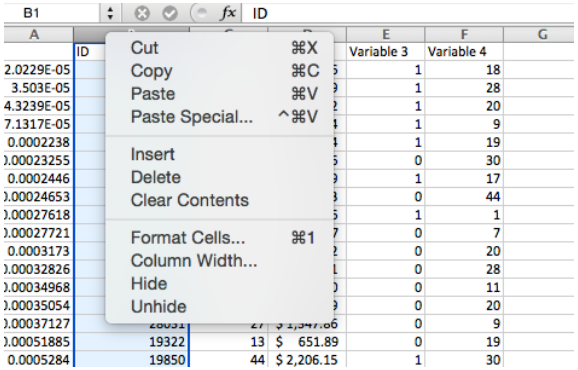


	A	B	C	D	E	F
1		ID	Variable 1	Variable 2	Variable 3	Variable 4
2	0.67791466	1	42	$ 1,958.67	1	39
3	0.3349527	2	29	$ 1,415.30	1	38
4	0.91447267	3	64	$ 3,123.77	1	26
5	0.66427353	4	7	$ 391.09	1	5
6	0.18283504	5	43	$ 2,352.72	1	37
7	0.41711613	6	80	$ 4,125.80	0	13
8	0.08820664	7	96	$ 5,213.70	0	23



5. Insert another column to beside the random column and fill it with a counter.

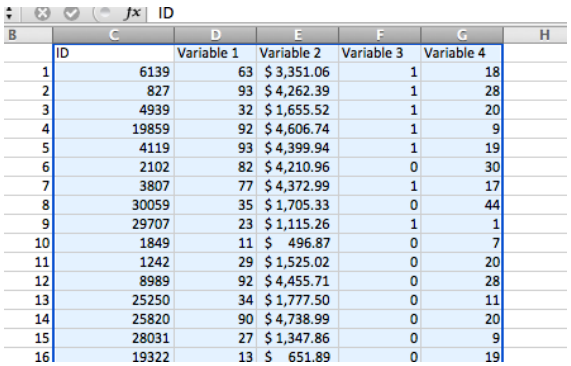
Control-click the first column of your data and select insert.
Enter 1 in the first cell and 2 in the second cell. Highlight both of these cells and double click the bottom right corner of the selection to autofill the whole column.



	B	C	D	E	F	G	H
		ID	Variable 1	Variable 2	Variable 3	Variable 4	
-05	1	6139	63	$ 3,351.06	1	18	
-05	2	827	93	$ 4,262.39	1	28	
-05	3	4939	32	$ 1,655.52	1	20	
-05	4	19859	92	$ 4,606.74	1	9	
238	5	4119	93	$ 4,399.94	1	19	
255	6	2102	82	$ 4,210.96	0	30	
146	7	3807	77	$ 4,372.99	1	17	
553	8	30059	35	$ 1,705.33	0	44	
518	9	29707	23	$ 1,115.26	1	1	
721	10	1849	11	$ 496.87	0	7	
173	11	1242	29	$ 1,525.02	0	20	
326	12	8989	92	$ 4,455.71	0	28	
968	13	25250	34	$ 1,777.50	0	11	
054	14	25820	90	$ 4,738.99	0	20	
127	15	28031	27	$ 1,347.86	0	9	
385	16	19322	13	$ 651.89	0	19	
284	17	19850	44	$ 2,206.15	1	30	
121	18	12081	93	$ 4,198.58	1	17	
192	19	3521	104	$ 4,899.48	0	35	
725	20	18668	11	$ 468.14	0	35	
147	21	21320	38	$ 1,890.28	0	1	
574	22	24631	8	$ 419.14	0	18	

6. Select from row 1 to the size of the sampling you want.

Scroll to the row number of the sampling you want and select all of the rows from there to the top.

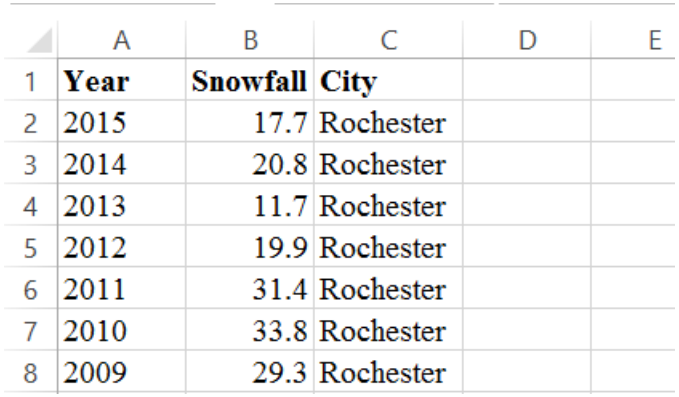


Create a Slicer

Excel Step-by-Step How-to for Windows

Instructions: Use this guide to create a slicer in an Excel spreadsheet.

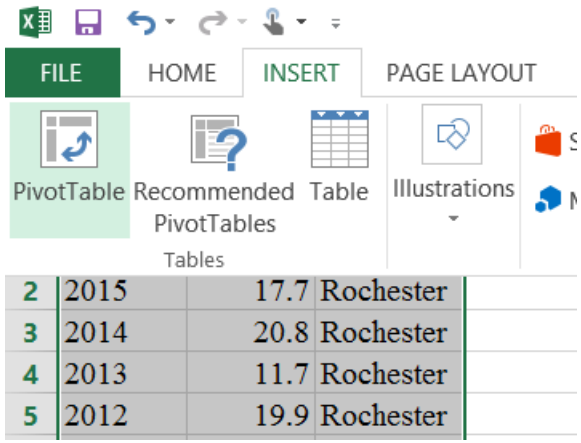
Data requirement: multiple variables, quantitative and/or categorical data
Sample Data: yearly snowfall in Rochester and Vancouver

Step	Windows Instructions + Screen Shot
1. Arrange your data so that each unique variable has its own column.	

	A	B	C	D	E
1	**Year**	**Snowfall**	**City**		
2	2015	17.7	Rochester		
3	2014	20.8	Rochester		
4	2013	11.7	Rochester		
5	2012	19.9	Rochester		
6	2011	31.4	Rochester		
7	2010	33.8	Rochester		
8	2009	29.3	Rochester		

2. Create a pivot table for you data.

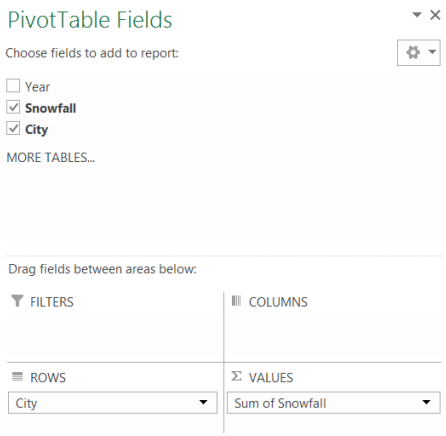
Highlight all of your data and select Pivot Table under the Insert tab.



2	2015	17.7	Rochester
3	2014	20.8	Rochester
4	2013	11.7	Rochester
5	2012	19.9	Rochester

3. Define your pivot table fields.

Select City to populate your rows by the data that have the same city and select snowfall as the values to display.

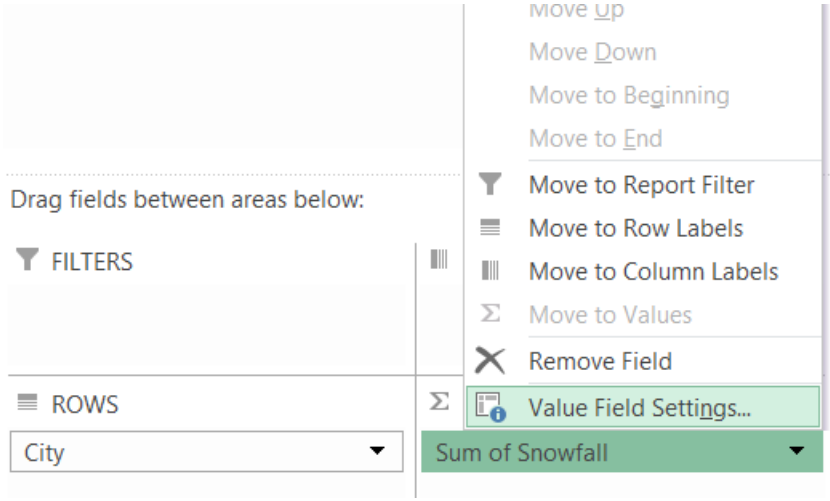


TIP: To quickly add fields to pivot table categories:

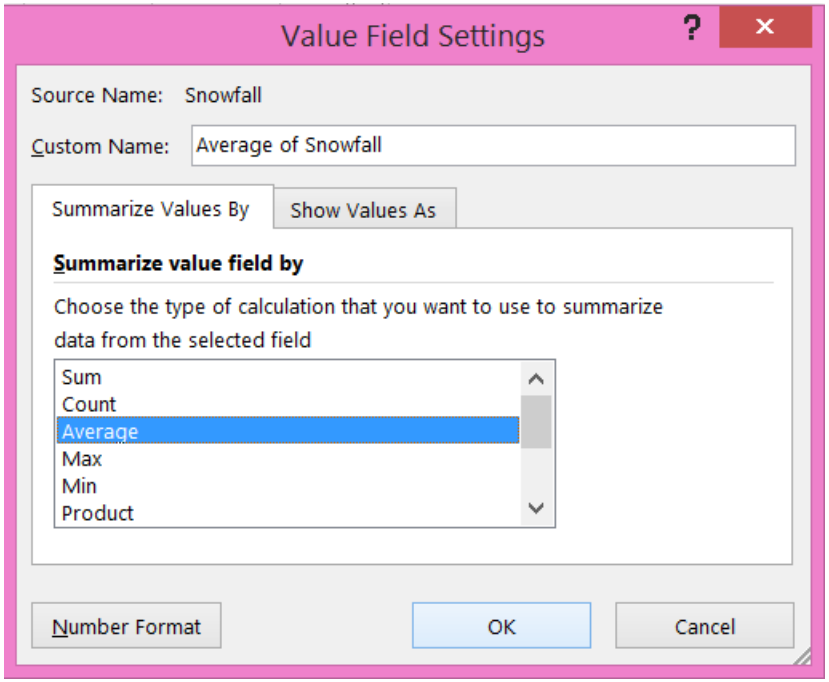
In order to add a field to a pivot table category you can select the variable (for example, snowfall or city) and drag and drop it to the pivot table category you want.

4. Customize your pivot table fields.

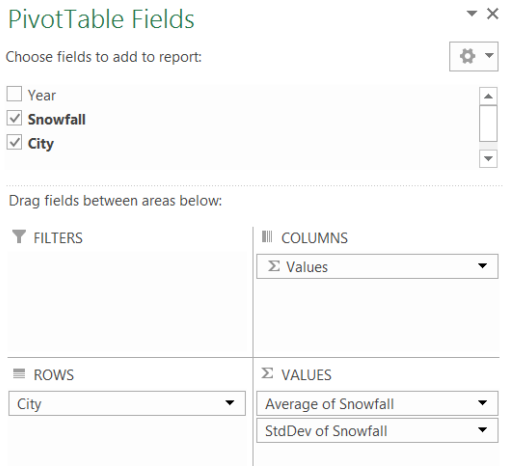
Click the dropdown arrow on your field and select Value Field Settings.



Change the value to your desired type of calculation by selecting it and clicking Okay.

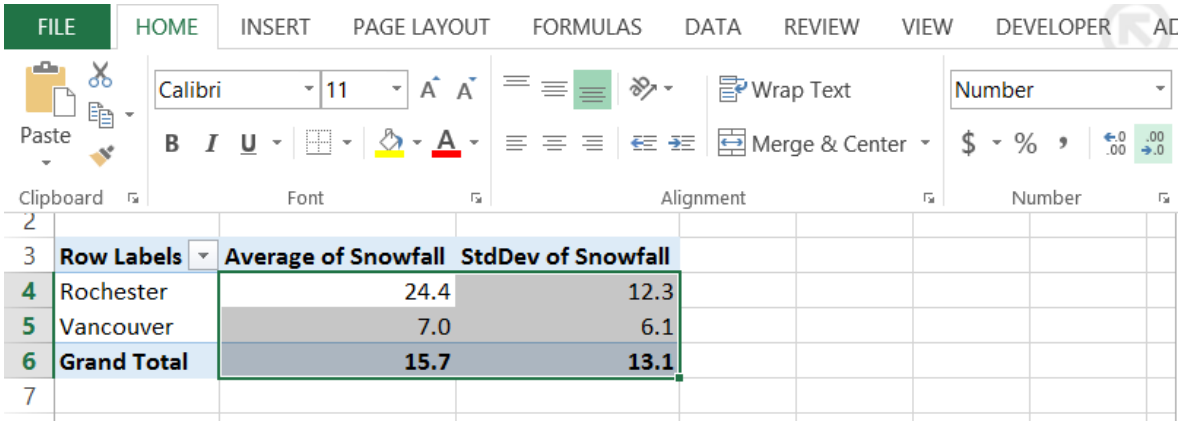


5. Repeat the process of adding more fields for as many data summaries you would like to make.

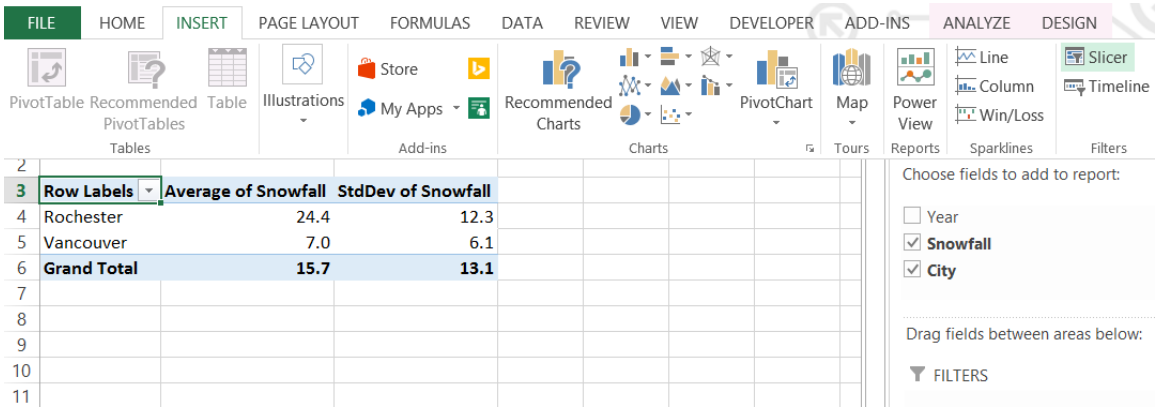


6. Insert a slicer.

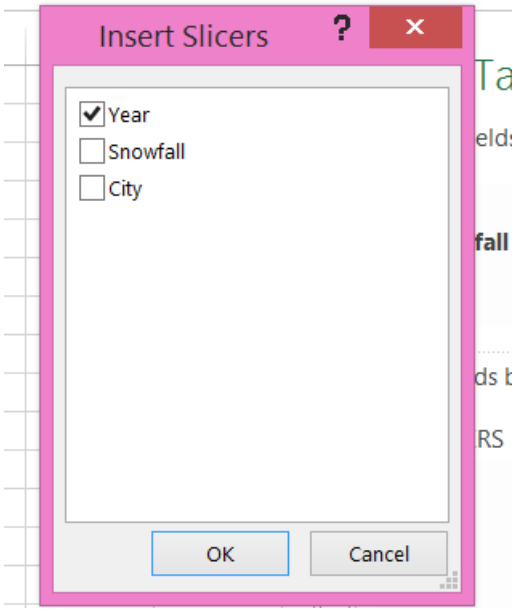
Select your pivot table.



Navigate to the insert tab and select Slicer.



When prompted, select the variable that you would like to slice by.

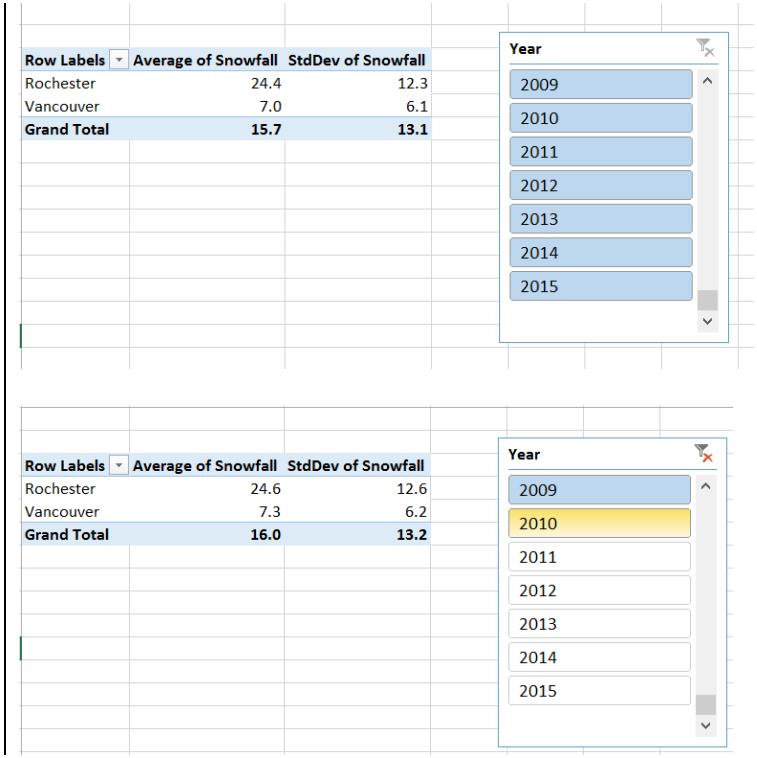


TIP: To choose what variable to slice by:

When choosing the variable to slice by think about what information you would like to quickly remove or add from your data set in order to demonstrate change in your summary.

7. Manipulate the slicer.

In order to unselect values of the slicer you must control click on the value. In order to highlight a large range of values in the slicer you can control-shift click a range.



Row Labels	Average of Snowfall	StdDev of Snowfall
Rochester	24.4	12.3
Vancouver	7.0	6.1
Grand Total	**15.7**	**13.1**

Year
2009
2010
2011
2012
2013
2014
2015

Row Labels	Average of Snowfall	StdDev of Snowfall
Rochester	24.6	12.6
Vancouver	7.3	6.2
Grand Total	**16.0**	**13.2**

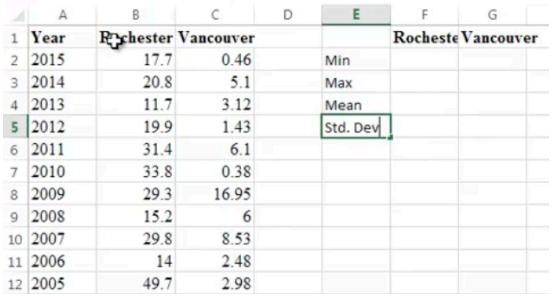
Year
2009
2010
2011
2012
2013
2014
2015

Create a Summary Statistics Table (Descriptive Summary)

Excel Step-by-Step How-to

Instructions: Use this guide to create a table of summary statistics.

Data requirement: one variable, quantitative data
Sample Data: yearly snowfall in Rochester and Vancouver

Step	Instructions + Screen Shot	MAC Variations
1. Type-in descriptions of the summary statistics you want to include in your table.	Typically, these values include: – minimum value – maximum value – mean value – standard deviation 	

	A	B	C	D	E	F	G
1	Year	Rochester	Vancouver			Rochester	Vancouver
2	2015	17.7	0.46		Min		
3	2014	20.8	5.1		Max		
4	2013	11.7	3.12		Mean		
5	2012	19.9	1.43		Std. Dev		
6	2011	31.4	6.1				
7	2010	33.8	0.38				
8	2009	29.3	16.95				
9	2008	15.2	6				
10	2007	29.8	8.53				
11	2006	14	2.48				
12	2005	49.7	2.98				

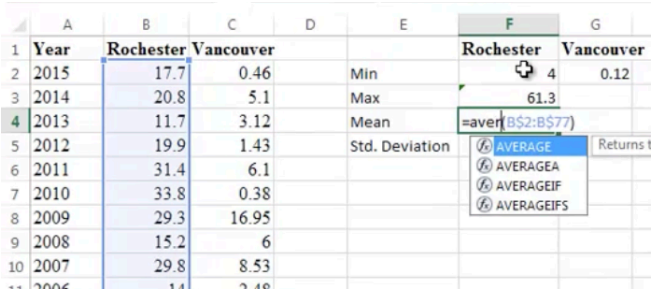
2. Use an Excel formula to calculate the summary statistic for each variable.

For each summary statistic, enter the corresponding Excel formula and then select all data values (i.e. a range of cells) for each variable (in this example, Rochester and Vancouver are the variables).

Common Excel formulas:

=MIN(range…)
=MAX(range…)
=AVERAGE(range…)
=STDEV(range…)

A range is written simply as A:B such that A is the starting row/column, and B is the finishing row and column. In Excel you can simply select the cell to have an expression fill with the correct row and column.



TIP: To quickly select a range of cells in a column:

Replace the range with the column letter. For example, if you want all of the data in column B, your range can be B:B.

Select the first cell in the column you would like to select and press CMD+Shift+Down

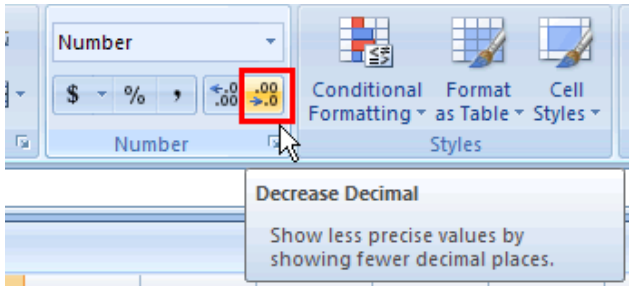
TIP: To quickly replicate formulas across cells:

First, it is important to lock the appropriate values in your formula. Locking a value makes the formula always reference the locked cell instead of changing based on location. You can lock he values in each formula by placing a "$" in front of each coordinate. For example, by representing the formula as B$24+C$35, the value will always calculate using rows 24-35 no matter where that formula is placed. Note that if you also place a "$" in front of the column letter, that will also lock the column.

Then, highlight the cell with the formula you want to copy, and drag the cursor across to the next cell.

3. Decrease the number of decimal places for all the values in your summary statistics table to one or two; this makes it easier to read.

To do this quickly, use the decimal button on the top toolbar.



4. Format your finished summary statistics table as desired.

Simple things like making the column and row titles bold can help you and others more easily read and interpret the data.

	Rochester	Vancouver
Min	4	0.12
Max	61.3	28.2
Mean	24.4	7.0
Std. Dev	12.3	6.1

Use these guides to help you complete data analysis tasks in Excel. Instructions are provided for Excel for Windows and Excel for Mac. Note that some of these tasks are not supported in Excel for Mac.

Back to Table of Contents