Statistics

Lab Book & Project

Peter Keep

Statistics Lab Book and Project

Lab assignments and project details for a project-based Statistics for Business class

Statistics Lab Book and Project

Lab assignments and project details for a project-based Statistics for Business class

Peter Keep
Moraine Valley Community College

February 1, 2026

Contents

Chapter 1

Software: jamovi

Welcome to the class! This initial chapter will serve as a guide to getting started with the software that we'll be using for these labs, and eventually for your project.

1.1 About jamovi

Jamovi is the name of the statistical software that we'll be using and that these labs and project are written for. It's a good choice for a lot of reasons, including:

- *Price:* Statistical software can get expensive! Software built to be used by large companies can come with a huge cost associated with it, and the software that is often used in education is bundled up with textbooks and online homework systems. Jamovi is free for anyone to use.
- *Ease of Use:* Jamovi isn't the only free statistical software out there. In fact, most statistics majors will end up using at least one of many different programming languages: R, python, julia, stan, etc. These are all, in some way or another, built to be extremely useful tools for doing statistical analysis. Jamovi is what we call "point-and-click" software: you'll navigate through some menus and click on options to do the work we're going to do. No programming knowledge required!

  Jamovi is very intuitive to use, with clear menus and pretty intuitive options. We
- *Flexibility:* I think one of the best aspects of this software is that there is just enough to use for an introductory statistics class, and much more to use if needed. Some of that is hidden in "modules" that are installed separately, and some of that is in the options we can select when we are setting up our statistical analysis. It's not overwhelming, and it will cover everything we need to cover.

  I have also had students in the past decide to extend their projects slightly beyond the content we cover in class. In those cases, jamovi has been able to provide access to some more advanced statistical tools.

The instructions for this lab are written to reflect the jamovi 2.7 release. They likely will not become out of date very quickly, but note that some menu options described could change.

1.2 Download and Installation

There are two main ways to access and use jamovi: installing the software, or using the cloud version. There are a couple of things to note if you're planning on using the cloud version.

Cloud Version

The cloud version of jamovi can be accessed at https://cloud.jamovi.org/. You can access the software using a free account, but there are some limitations.

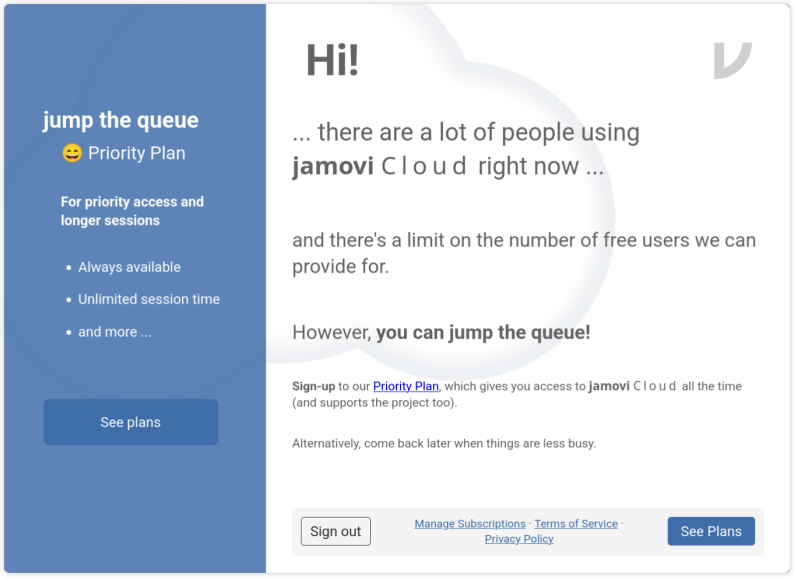

Figure 1.2.1 jamovi limits access for free users when things are busy.

Some of the limitations can involve:

- Limitations on the size of the data sets you can use.
- Time limitations on the session, forcing users to save regularly.
- Limitations on access during busy sessions.

Overall, I recommend a local installation, and it is useful to have this software installed in a computer lab on campus for use by the class.

Local Installation

To find a local installation file, visit https://www.jamovi.org/download.html. Here, you'll find installation files for Windows, macOS, Linux, and also ChromeOS (for use on Chromebooks).

For more instructions on platform-specific installations, visit https://www.jamovi.org/user-manual.html#installation.

> **Note 1.2.2**
>
> For users with Chromebooks, there are some specific instructions for installation that require the user to change a few settings in order to access the linux installation more directly. Visit https://flatpak.org/setup/Chrome%20OS for setup instructions, and then follow the jamovi

> installation instructions in the manual.

Chapter 2

Data Sets

We start with the data documentation. This will serve as a helpful reference as you work with different data sets, but also as a way of deciding which data set you would like to use for your project.

Each section will include:

- A link to download the data set in multiple formats.
- A brief description of what the data set is about and where it came from.
- A data dictionary, with details about each variable (or column) in the data set.

> **Note 2.0.1**
>
> The AirBnB data set will be used for the labs, while the remaining data sets are options to select for your project.

2.1 AirBnB Data Set (Lab Data Set)

This is a reference page to quickly find the data set we'll be using for the labs this semester. As you work through the different labs, you should be saving your jamovi file to reflect the changes we'll make to this data set as we work through it, but you can always download the original data here. Similarly, you can find the data dictionary to get a basic review of what the data set is and what the variables measure.

This data set includes information about different AirBnB listings in New York City. It has 18 different variables collected from over 38,000 listings. Some variables are identifiers for things like the host or the specific listing, but we have information about pricing, location, type of listing, etc.

Download

- AirBnB.csv
- Alternatively, you can download a jamovi-specific file: AirBnB.omv

Data Dictionary

Table 2.1.1 General Information

Rows:	38199
Columns:	18
Source:	InsideAirBnB.com
Year:	2024

Table 2.1.2 Variable Descriptions

`id`	AirBnB's unique identifier for the listing. Often, this corresponds with the URL of the listing: `https://www.airbnb/rooms/[ID Goes Here]`
`name`	Name of the listing, as set by the host.
`host_id`	AirBnB's unique identifier for the host. Often, this corresponds with the URL of the host's profile: `https://www.airbnb/user/show/[ID Goes Here]`
`host_name`	Name of the host (typically their first name)
`neighbourhood_group`	Name of the borough where the listing is located, based on the longitude and latitude.
`neighbourhood`	Name of the neighbourhood where the listing is located.
`latitude`	The latitude of the listing, a part of the coordinates to give a precise location.
`longitude`	The longitude of the listing, a part of the coordinates to give a precise location.
`room_type`	The type of listing: Entire home/apt, Hotel room, Private room, or Shared room.
`price`	The daily price (in USD).
`minimum_night`	The minimum number of nights required to book the listing, as set by the host.
`number_of_reviews`	The total number of reviews that the listing has received.
`last_review`	The date of the latest review, where date is in the format YYYY-MM-DD.
`reviewers_per_month`	The total number of reviews that the listing has received divided by the total number of months the listing has been posted.
`calculated_host_listings_count`	The total number of listings the host has in New York City.
`availability_365`	The number of days that the listing is available in the next year.
`number_of_reviews_ltm`	The number of reviews that the listing has received in the last 12 months.
`license`	The license, permit, or registration number.

2.2 Steam Games Data Set

This data set includes information about video games on the Steam Store, a popular online store and platform for computer games. The data was collected in May, 2024. It has 21 variables collected from over 27,000 games. There is a lot of identification and classification information in here, but we get some other metrics like age ratings, estimated number of owners, average play time, price, etc. There are a decent number of true/false categories that could be fun to look through (like whether or not mac is a supported operating system), and there is some information about reviews in here as well.

Download

- SteamGames.csv
- Alternatively, you can download a jamovi-specific file: SteamGames.omv

Data Dictionary

Table 2.2.1 General Information

Rows:	27075
Columns:	21
Source:	Steam API
Year:	2024

Table 2.2.2 Variable Descriptions

appid	A unique app ID for the game listing.
name	Name of the game in the Steam Store.
release_date	Release date, formatted as YYYY-MM-DD
english	Binary variable describing whether English is one of the supported language.
developer	Name, or names, of the game developer(s).
publisher	Name, or names, of the game publisher(s).
windows	Binary variable describing if the game has a version developed for Windows.
mac	Binary variable describing if the game has a version developed for MacOS.
linux	Binary variable describing if the game has a version developed for Linux and SteamOS.
platform_num	Total number of supported operating systems (between Windows, Mac, and Linux).
required_age	Minimum required age according to PGEI UK standards. Some of the 0 age ratings could represent games that are unrated or that are missing this age requirement.
categories	Game categories.
genre	Game genres.
steamspy_tags	Community-voted tags that are listed for each game.
achievements	The total number of in-game achievements that are available.
positive_ratings	Total number of positive ratings.
negative_ratings	Total number of negative ratings.
average_playtime	Average user playtime, measured in hours.
median_playtime	Median user playtime, measured in hours.
owners	Estimated range of the number of Steam users that own the game.
price	Price of the game, measured in USD.

Chapter 3

Lab Assignments

Each of the lab assignments for this course are listed below. These each have detailed instructions to follow, as well as "Checkpoint" questions to complete and submit.

These labs should showcase some of the different statistical tests and techniques that will be useful for developing and exploring research questions in an independent project.

3.1 Introduction to jamovi

Our first jamovi lab! In this lab, we'll prioritize just getting a handle on what jamovi is, how to navigate it, and do a small amount of data exploration: nothing too much for now, but enough to start getting used to the software.

What You'll Need.

1. *jamovi:* This is the software we'll be using for the whole semester. It's free and open source and available on (almost) any setup you've got. Visit https://www.jamovi.org/download.html to get started on downloading and installing this. You'll need to do this on a computer (laptop, desktop, or chromebook is fine), and please make sure this is something you'll be able to use for the semester.
2. *Data:* We're going to mostly work through the same data set all semester in these labs (we might mix it up once or twice). The dataset that we'll use for the labs is a collection of AirBnB data from AirBnB listings in New York City. More information on the data and a link to download can be found in the data dictionary: AirBnB Data Set (Lab Data Set).

Tasks

Starting jamovi

Once you've downloaded and installed it on whatever computer you are using, go ahead and launch jamovi. You should see a screen similar to the screenshot below.

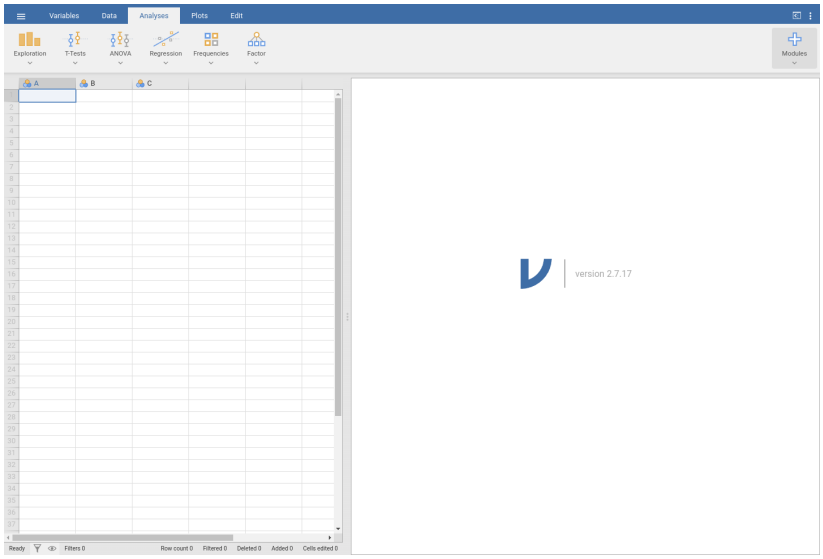

Figure 3.1.1 Screenshot of jamovi, after it opens.

Jamovi is a mix of spreadsheet (to hold, organize, and view our data) and analysis (which will appear on the right-most panel as we perform it). For us to use it, we'll need to load up the data set we are working with.

Save the `AirBnB.csv` file from the AirBnB Data Set (Lab Data Set) page somewhere that is easy to find. In jamovi, open the menu in the top left (the three horizontal bars) and select Open, and then Browse. Locate the `.csv` file that you downloaded and open it.

> **Note 3.1.2**
>
> Alternatively, you can download the data in the jamovi `.omv` format, and use that. Whenever we reference the `.csv` file, you can simply use the `.omv` file that you downloaded.

Once this is opened, you should see the data in jamovi.

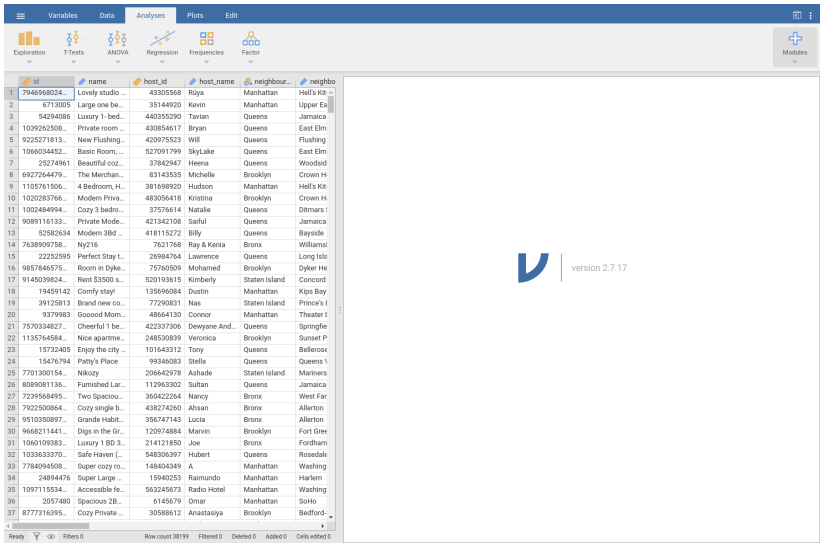

Figure 3.1.3 Screenshot of jamovi, with the `AirBnB` data set opened.

You can see that the spreadsheet portion has been filled in with different columns, each representing a variable in the dataset. Each row represents a

single observational unit. Each cell (the intersection of a row and column) is the measurement recorded from that unit for that variable.

Exploring Variables

Something to notice is the icon on each variable name in the columns of the spreadsheet. These icons correspond with the types of variables. Sometimes jamovi misinterprets what the variables are, so it's always good to check these. Click on the Variables tab on the top of the screen. You should see a list of all 18 variables. You can add a description to each, which can be helpful for long-term projects, so you don't have to keep looking up what each variable represents. If you double-click on any variable, a menu will pop up on the top with some options. The main thing we'll be concerned with, for now, is "Measure type."

For instance, the variable `id` is labeled as a "Continuous," with the "Data type" listed as "Integer." This is clearly a mistake, due to the fact that every observational unit was assigned an ID number. Jamovi has a specific measure type for IDs, so change it from "Continuous" to "ID." As you look through the rest of the variables, you'll see a mix of ID variables, nominal variables, and continuous variables. For categorical variables (nominal or ordinal), you can also see the levels. For an ordinal variable, you can arrange the order of the levels.

Checkpoint 3.1.4

Do any other variable types need to be fixed? You'll need to know what these variables actually represent, so now is a great time to read through the data documentation below as well as view the data itself in the spreadsheet view, by clicking on the Data tab.

Checkpoint 3.1.5

Pick a variable that jamovi identifies (or that you fix, so that jamovi identifies) as continuous. Explain why this is a continuous variable.

Checkpoint 3.1.6

Pick a variable that jamovi identifies (or that you fix, so that jamovi identifies) as nominal. Explain why this is a nominal variable.

You should have a decent idea of what these variables look like and what they are measuring. Let's investigate them a bit more carefully.

Click on the Analyses tab on the top of the screen, and then click on Exploration and then Descriptives. You now have a list of the variables on the left: You can add one or more to the Variables pane to display some summaries and descriptions of the variables. You can change what statistics are displayed in the Statistics menu underneath the list of variables, and you can change the orientation of the table that is displayed by switching the dropdown menu back and forth between "Variables across rows" and "Variables across columns."

Checkpoint 3.1.7

How many observations are recorded for the variable `number_of_reviews`? How many are missing? What about the number of observations for `reviews_per_month`? How many are missing? Explain.

You will also notice that in the bottom Plots menu, you can summarize variables visually with some different options. Feel free to try some of these, although you are not required to create any specific plots yet.

New Variable

Something that you might have noticed is that there's not a comfortable way to handle the `last_review` variable. This variable is a date, but jamovi labels it as an "ID" measurement, and there's not really a nice alternative...it's not a continuous measurement: it's ordinal! But we can't just drag and drop every single date to be put in order. Different statistical software will include a "date" variable type, but jamovi doesn't have this. This isn't a problem for us, and we'll extract some of the information about the date of the most recent review by creating a new variable.

In the Variables tab (at the top of your screen), you'll notice that we have the option to add a new variable. You can either click the Add button on the top, or use the + button near the bottom. Select "Computed Variable," and you can either insert it in the list of variables where you've selected or append it to the end (it doesn't really make a difference).

Name your computed variable `last_review_year`, and in the formula box we'll use the following code.

```
INT(SPLIT(last_review, "-", 1))
```

Here's what it does:

- The `SPLIT()` function does exactly what it says: it splits up some text.
- The first input to this function is the name of the variable that we will split. So `SPLIT(last_review,` is saying that we will divide up the responses to the `last_review` variable.
- The second input to this function is the symbol that we will use to split up these dates. We're going to direct jamovi to split up the dates based on the hyphens, since the dates are in the format YYYY-MM-DD.
- The third input tells jamovi which "piece" to report. We have split these dates into three sections: YYYY, MM, and DD. We are directing jamovi to report the first piece, the year.
- Lastly, we wrap this up in the `INT()` function, which tells jamovi to treat this as if it were an integer. Otherwise, jamovi would treat this as a nominal variable.

Checkpoint 3.1.8

What is the median year for the last review? What is the minimum year for the last review?

If, later on, we would like to have information about the month of the most recent review, we can tinker with the code to create a new variable that pulls the 2nd piece from the split up date.

Another New Variable

Let's add one more variable. This one won't be as useful immediately, but will hopefully model some useful ideas for your own project.

We have two variables that we'll look at: `price` and `minimum_nights`. Wouldn't it be nice to look at a minimum price? Since the price variable is a measurement of price/night, we might just multiply `price` and `minimum_nights` together.

Create a new calculated variable called `minimum_price` where the variable is defined by the following formula.

```
price * minimum_nights
```

Checkpoint 3.1.9

What is the average (mean) value of `minimum_price`? What is the maximum value?

Hopefully you're feeling pretty confident and familiar with this data set. We'll work with it some more throughout the rest of the lab assignments. As we go, we'll learn a lot about this data set, but we'll also learn some tricks and tools to use on your own data sets as you construct the pieces of your project for the semester.

Save Your Progress. You have made some small changes to this data set (fixing variable types and creating new variables), and it will be helpful to use this in future labs. In the main menu in the top left (the three horizontal bars), you can either Save As a `.omv` file, or you can Export a `.csv` file. If you export the `.csv` file, you will be able to use this data with many other applications (this is a common file storage type for datasets), but you will lose the changes you made to the variable types and descriptions to the variables. The `.omv` file will retain this information, but is not as common (since it's specific to jamovi). It will definitely be useful, though, as you work on your project.

Choose to Save As a `.omv` file, and remember where you save it: we'll access this to work on the rest of the labs.

Note 3.1.10

It will likely be useful to set up a folder to organize these labs and your project. Something like:

```
/StatsClass
  - /Project
  - /Labs
      - AirBnB.csv
      - LabAirBnB.omv
```

3.2 Summarizing Variables

In this second jamovi lab, we'll work on building some basic summaries (plots and calculations) of variables (both numerical and categorical). This is really one of the first things we should do with a new data set after learning what the variables represent. We'll explore the data by getting to know not only what each variable represents, but also what the different observations look like and how they act as a group.

What You'll Need.

Data: we're going to use the same AirBnB data set from Lab 1, and it might be nice to use the version you saved afterwards. Load that data up by either opening the .csv file or the .omv file that you saved from the Introduction to jamovi lab.

Tasks

We're going to start off small with just getting an idea of how to visualize histograms, boxplots, and different bar graphs, and how to calculate the different summary statistics we might use to talk about the center and variability of a variable. Afterwards, we'll do a bit more data manipulation to get some more interesting plots.

Summarizing Numerical Variables

You've actually already seen how to get some summary statistics, but let's remind you: head over to the Analyses tab and select Exploration and then Descriptives. First, move the `price` variable into the "Variables" box, and then under "Plots", select "Histogram."

Histograms are a great way to visualize and understand the distribution of a numerical variable, but we've seen how the shape can be impacted by the number of bins used to group our data: jamovi tries to choose a reasonable number of bins, but this isn't customizable (unfortunately). That said, these plots are great for exploration, even if we might want to be able to customize them before putting them into an official report.

Anyways, you now have a table of different summary stats (and you can toggle different ones off or on in the "Statistics" menu) as well as a histogram for the `price` variable.

Checkpoint 3.2.1

Describe the distribution of prices by describing the shape, interpreting a reasonable measure of center as well as a reasonable measure of variability. Notice that the choice of measure of center and variability should depend on the shape.

Note that there are a bunch of missing prices: these represent listings that aren't actually available to book currently, for some reason.

Checkpoint 3.2.2

Add the variables `latitude` and `longitude` to the "Variables" box. Why do these histograms look different from the one describing prices? Explain the differences.

In the table of summary statistics, we can see that the mean and median are reported with values rounded. This is an option we can change! In the menu in the top-right, with three dots, you can change the general number format of results. The default is to use 3 significant figures. Change this to 3 decimal places.

Checkpoint 3.2.3

You can search coordinates on google maps by writing them in the form latitude, longitude: a search of "41.692, -87.840" on google maps will find Moraine Valley Community College's campus. Use the mean latitude and longitude to find the average location of an AirBnB on google maps. Now use the median latitude and longitude. What are the differences in the types of locations you find? Which one do you think would better represent the average location of an AirBnB in New York City? Is there a meaningful difference?

You'll likely have noticed that the histogram isn't actually that good: we can see that the prices are highly skewed right, but that means that we don't really get much information about what's happening, since most of the prices are grouped into a single "bin." Let's look briefly at some more plotting options. Find and open the Plots menu in the top bar, and select Histogram. This will bring us a more detailed menu where we can create higher quality plots.

Add `price` to the "Variable" pane. The histogram you'll find is slightly different: the bins or bars are smaller, and so there are more of them. It still has the general shape, but we can see now that there really isn't much data beyond 25000: there are not many listings that have a cost of $25,000/night or more. This *should* make sense, except for the fact that there *are* some listings with high prices like that. We'll investigate this more fully later.

For now, let's experiment with some options. Let's first change the range of prices we'll look at, just so we can see some detail. In the Axes menu, change the option for the Range on the X-Axis to "Manual", and view the histogram for just prices between $0 and $1000 per night.

Note 3.2.4

Careful with doing this! We should note that we're just doing this for exploratory reasons: we don't want to remove the other listings from our data set or from our consideration here. Removing "outliers" is typically only something we'll do for observations that we believe are not actually part of our population. For now, we're just zooming in on one section of the prices to get some more detail.

Now, in the General Options, experiment with the different options: change the Bin Width from "Auto" to "Manual" and find a bin width that you think is pretty good for seeing some details in the histogram. Remember, this value is just the size of the intervals of prices that will be grouped together! Play with some of the options with the display to get a feel for what they do. We won't worry about the Density option for now.

Using the different options, replicate this plot (without the "Example" watermark, of course).

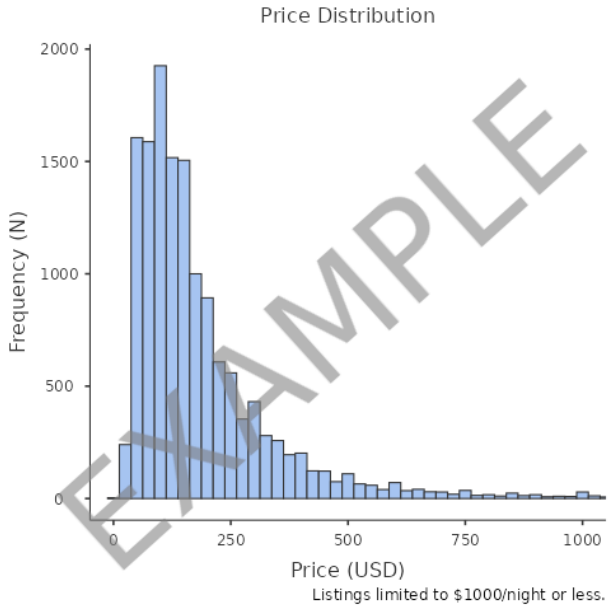

Figure 3.2.5 Example plot to replicate.

You can save your plots in a couple of different ways: if you right click on the picture of it, there is an option to copy the image. In my experience, this is a bit buggy, and the image doesn't always paste (I think it depends on what program you're pasting it into, but I don't know). You can also export the image, where you get different options to save it as a `.pdf` file or a `.png` image file. Save your graph as a `.png` file.

Checkpoint 3.2.6

Upload your histogram in the Lab Submission assignment!

Summarizing Categorical Variables

Let's swap out these numerical variables for some categorical ones. Change the variables we're looking at to `room_type`. Notice that in the table of descriptive/summary statistics, there's nothing calculated: of course not, since the responses aren't numbers! Feel free to deselect all of the different calculations in the "Statistics" menu, although it's still sometimes nice to have the sample size and the number of missing observations in there, just for reference.

Now, select the "Frequency tables" option (right under the "Variables" and "Split by" windows). This option is only available to nominal and ordinal variables (you can see the little icons for them), and should give you a quick display of the frequency (counts) and proportion (% of Total) for each level of the `room_type` variable. In the "Plots" menu, you can select "Bar plot" to visualize this distribution.

Checkpoint 3.2.7

Which room types are most common? Describe the distribution of different room types in words.

Add `neighborhood_group` to the "Variables" window.

Checkpoint 3.2.8

Which neighborhoods are most popular for hosting an AirBnB? Describe the proportions for each borough.

The default bar plots are very basic, but there are more options available in jamovi by installing a Module. Click on the Modules button in the top right of your screen, and select jamovi library. You should see the following popup.

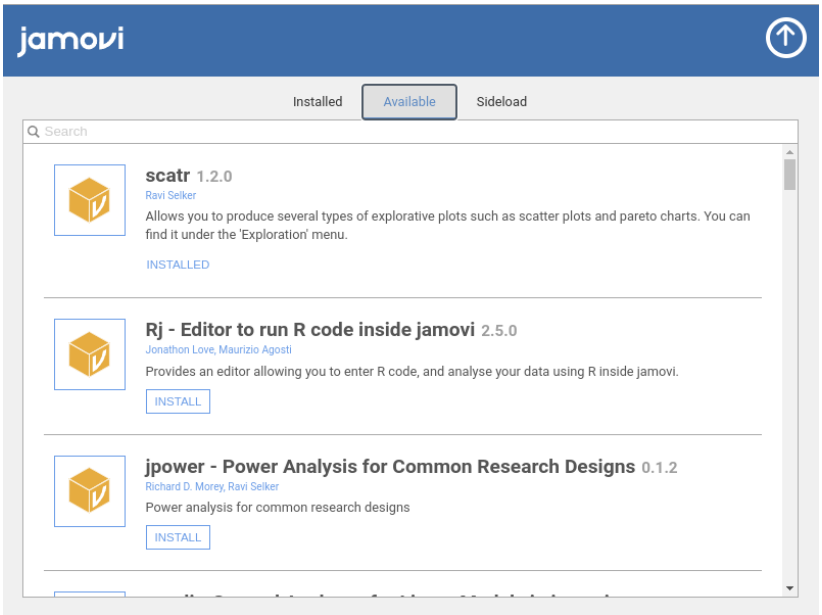

Figure 3.2.9 The jamovi Modules menu.

This is where we can manage different add-ons for jamovi to accomplish specific tasks. In the search bar, search for "surveymv", a module that "Generates summary plots for your survey data." Click Install, and then cluck the up arrow in the top right to close the menu. You'll now have a new option to select in the Exploration menu called "Survey Plots". Select it now.

This should look familiar: you can drag your variables into a "Variables" window, and there are two menus with different plot options (one menu for categorical variables and one for continuous variables). Using the different options, replicate this bar plot (without the "Example" watermark, of course).

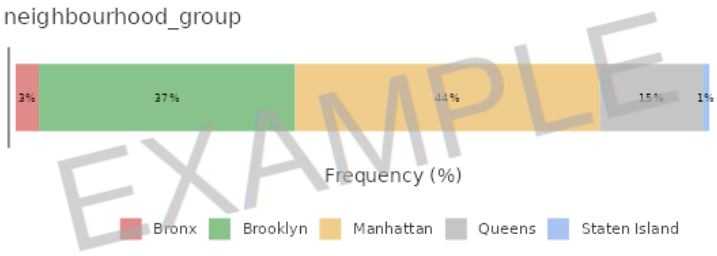

Figure 3.2.10 Example bar plot to replicate.

Checkpoint 3.2.11

Upload your bar plot in the Lab Submission assignment!

Adding a Filter

Let's say that we only want to look at "active" AirBnB listings: ones that have been reviewed in 2023 or 2024 (the most recent year in this dataset). Luckily you have already created a `last_review_year` variable! (If you need to re-create it, go back to the instructions from the Introduction to jamovi lab).

Click on the Data tab on the top of your screen, and you'll notice that one of the buttons near the top is for Filters. We'll create a filter, where the rule is `last_review_year >= 2023`. This will only keep the rows in our data set where the value for the `last_review_year` variable is greater than, or equal to, 2023. So really this is just selecting the listings that were reviewed in 2023 or 2024.

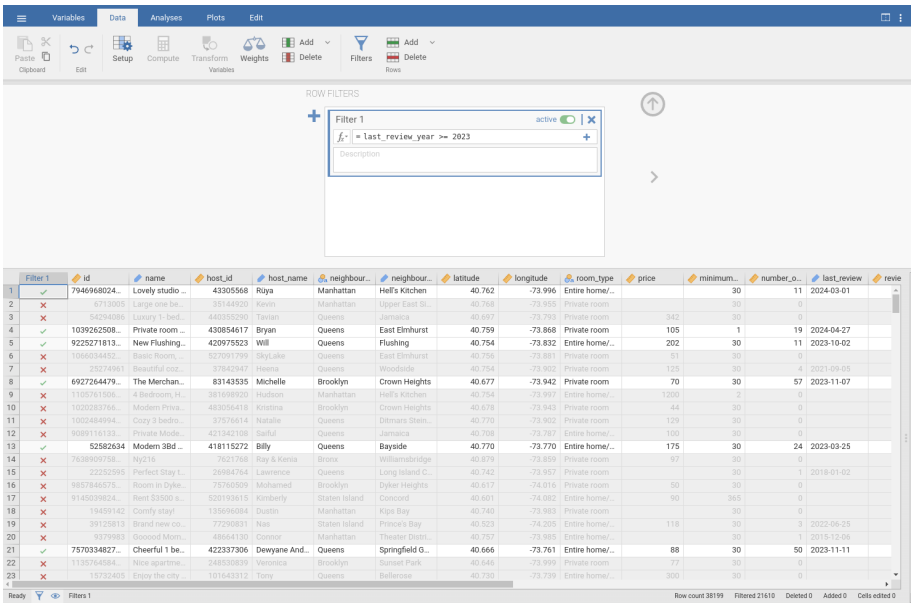

Figure 3.2.12 Active filter, showing only the listings with recent reviews.

You can see a couple of things to note: the rows that have their most current review prior to 2023 are greyed out, you can see how many rows were filtered in the bottom of the screen, and there's a toggle on the filter itself to turn it off and on.

Let's investigate one more numerical variable in this new context (with only the filtered data). Find the five number summary (minimum, 25th percentile, median, 75th percentile, and maximum) for the `availability_365` variable. Then, display a boxplot to see this summary visually.

Checkpoint 3.2.13

Calculate the IQR for the filtered `availability_365` variable (either by hand, or by selecting it as a descriptive statistic to show). Interpret this value, noting that it might be helpful to also report Q1 and Q3 in this interpretation.

Save Your Progress. It will still be helpful to save your progress, here! We'll likely return to this, whether to use the actual Filter that we set up or to use as a reference for some other data set for a project. If you don't want to keep all of the plots and tables up, you can right-click and select "Remove" from the "All" menu.

3.3 More Data Visualization

In this lab assignment, we're going to build a couple of more data visualizations, where the focus is less on summarizing a single variable and more on exploring connections between variables. We aren't doing any formal tests, but we are doing some preliminary work towards answering a research question.

What You'll Need.

1. *Data:* we're going to continue to use the AirBnB data set. Hopefully you have an up-to-date version of this, with the changes we've made in the previous labs.
2. *Questions:* before getting into this lab, see if you can come up with some questions to consider, specifically about relationships between variables. I'll add a couple of mine, and we'll look at those, but this is an important process to get comfortable with, since you'll be doing this on your own for your project. We'll come up with questions about:
   - a pair of two categorical variables.
   - a pair of two numerical variables.
   - a categorical variable and a numerical variable.

Tasks

We're going to begin investigating three questions in this lab.

1. Is the distribution of AirBnB listings different across different boroughs?
2. Is there a connection between the price of a listing and the minimum number of nights required to book it?
3. Are the different types of listings available in the same way for the next year?

Listing Type by Borough

We'll start by opening the Analyses tab and selecting Descriptives from the Exploration menu, as normal. Since we want to explore the distribution of the listing types, we'll add `room_type` to the "Variables" box. But now we can add a variable to the "Split by" window as well: let's put `neighbourhood_group` into that. Since these are both categorical variables, we'll care about seeing a frequency table: select that. You might want to remove the selections from the "Statistics" menu, although I think seeing the overall sample size of the different boroughs is useful.

Notice that this does not produce a two-way table (a contingency table): instead, we have a table that is a bit hard to read, although the same information is there. Rows are grouped into the types of listing, and then each room type is broken up into the 5 boroughs. We'll see later on this semester how to build a contingency table easily in jamovi, but let's move on for now and start visualizing.

The default (and only) option in the "Plots" menu is the Bar plot: check it and look at the bar plot. Now go back to the setup with the variables selected and swap them: put `room_type` in the "Split by" window and `neighbourhood_group` in the "Variables" pane. You can compare the bar plots, and see which one helps you get a handle on the distributions of listing types across different boroughs. Does it look like there are some differences in the distribution of listing types across boroughs? What about this plot makes things easy to see, or difficult to see?

Let's try to create a better bar plot. First, we might want to not visualize the raw counts: with Brooklyn and Manhattan being so much more popular of a listing location than the other 3 boroughs, it can be hard to compare the distributions of listing types. Second, we want to try to condense this plot a bit: making it smaller will hopefully make it easier to compare the different boroughs.

Find the Survey Plots option in the Exploration menu, add `room_type` to the "Variables" window and then add `neighbourhood_group` to the "Grouping Variable" option.

Checkpoint 3.3.1

Create a useful and readable plot that shows the different (proportional) distributions of room types grouped by boroughs. Save your plot as either a `.pdf` or a `.png`, and upload to the Lab Submission assignment.

Checkpoint 3.3.2

Summarize your plot: are there some notable differences in the distribution of listing types across the different boroughs?

Price Compared to Minimum Number of Nights

In order for us to look at the relationship between two numerical variables, we'll think about a scatterplot. Scatterplots aren't included as a default plot type in the Descriptives menu. Instead, you'll find it under Exploration, where you can click "Scatterplot".

Note 3.3.3

If you don't have the "Scatterplot" option in the Exploration mennu by default, that's fine! You can just install and add the `scatr` module. This is normally included as a default, but if that wasn't in your installation of jamovi, it's no problem! You can just add it yourself. Otherwise, an identical scatterplot menu should be found in the Plots menu up top, in the Scatter Plot menu.

Now that we have our scatterplot menu open, we can add `minimum_nights` to the "X-Axis" box, and we can add `price` to the "Y-Axis". You should notice that there are two observations with an *enormous* price. You can investigate them a bit!

Ok, I did, because I was curious! Here's the first one: https://www.airbnb.com/rooms/605115521796576121. And here's the second one: https://www.airbnb.com/rooms/17160286. You can see that

> they're both listed by the same person and both are listed with a minimum stay of 30 days for $100,000.00 per night. I'm assuming that these are artificially high prices so that the person listing these doesn't have to book it out for now! This (clearly) doesn't reflect the actual pricing of AirBnB listings.

Let's filter these ones out, so that we don't get these artificially high prices! Add a filter for `price < 100000`.

You'll also notice that there are some absurdly high values for the minimum number of nights. I did a quick check of everything where the minimum number of nights was larger than a year: all of them were listings that hadn't been reviewed in a while, and had been booked for short periods of time in the past. In short, these listings aren't currently available: instead of jacking the price way up, they just have priced themselves out by forcing prospective visitors to stay for over a year!

Change your filter to `price < 100000 and minimum_nights <= 365`.

Now we can look at the scatterplot! How would you explain the association (if any) between these two variables? Do we expect the prices of the listings to behave or be distributed differently for different numbers of minimum nights?

Again, you can fiddle with some of the settings here! You can change the size of the points, the dimensions of the plot itself, titles, labels, etc. Work on creating a plot that is descriptive and clear, with a good title and clear axes labels.

Checkpoint 3.3.4

Save your plot as either a `.pdf` or a `.png`, and upload to the Lab Submission assignment.

Listing Type Availability

We'll keep the filter that we just created on: this will continue to show us only the listings that are presumably available to reserve. But now we'll look at how available these listings are in the next year. You have some good options for this plot, but, regardless of how you plot this, you'll be looking at `availability_365` grouped or split by `room_type`. You can build histograms, boxplots, or density plots (a kind of "smooth" histogram, almost) to compare these. Let's start!

In Descriptives (in the Exploration menu), display the following statistics:

- N
- Mean
- Standard Deviation
- Q1, Q2, Q3 (these are the 25th, 50th, and 75th percentiles, and note that the 50th percentile can also be found using the Median)
- IQR

Before you visualize any plots, think about whether or not there are some differences in the way this variable is distributed across the different groups.

Now plot a histogram and box plot. Do these visuals match what you had thought when you looked at the summary statistics? You can also swap out the histogram for a density plot to get a "smoother" depiction of this histogram.

Try some of the options in the Plots menu up top, since you can change how these boxplots are displayed.

For a different version of these same plots, open up the Survey Plots menu under Exploration. Set up your variables, and look at the different plots. You have options for what to display under "Continuous Plots". For instance, I think it might be useful to get rid of the Data plot, which is just a dot plot where the dots aren't stacked. Since there are so many observations here, it's pretty messy and not very enlightening. Notice also that the "Violin" plot is just the density plot from earlier. Normally violin plots are reflected, to kind of act like a combination of bloxplot and density plot, but these ones aren't.

Play with the different plots and options until you are happy with what they're showing you. Try out different options, and add whatever detail you think is helpful to titles, labels, etc. This isn't always required, because sometimes the automatic labels are fine!

Checkpoint 3.3.5

Pick one of the plots you made that you think best demonstrates the differences or similarities of the availability in the next year for different types of listings available. Save it as a `.pdf` or `.png` file and upload it to the Lab Submission assignment.

Checkpoint 3.3.6

Describe the distribution of the number of days available for booking in the next year for the different room types. In your explanation, reference both the summary statistics you have displayed as well as the plot you uploaded.

Save Your Progress. We added another filter which will certainly be useful, so save your progress. As a reminder, you can right-click on the different plots you've created and select "Remove" from either the "Analysis" or "All" menus in order to clean up your output.

3.4 Introduction to Inference

In this lab assignment, we'll introduce some of the basics of statistical inference. Some of the specifics will be a bit different than what we're learning, but overall we're just going to focus on the concept of a confidence interval and hypothesis test.

What You'll Need.

Just the *Data:* we're still working with the AirBnB data. Make sure you're working with the most up-to-date version of the data set as possible (with the new variables and filters we've created). If you need a new version of the data set or a reminder of what the variables mean, visit the AirBnB Data Set (Lab Data Set) page.

Tasks

New Variable

We've had a filter set up (it hasn't been on, just there) to look only at the AirBnBs that have been reviewed "recently." Let's turn this into a categorical variable. In the Variables tab, add a new computed variable named `recent`. We want this to flag our recently reviewed listings (where "recent" means anything in 2023 or 2024). The first thing we need is for the listing to actually be reviewed, so we need `number_of_reviews > 0`. We want to combine this with the listings reviewed in 2023 and 2024, where `last_review_year >= 2023`. So the actual formula to compute the recent variable will be:

```
number_of_reviews > 0 and last_review_year >= 2023
```

This will output a 1 for listings that have been reviewed and have been reviewed since 2023 and a 0 for the other listings.

Before we move on, we should make sure we're all on the same page. We have two filters right now.

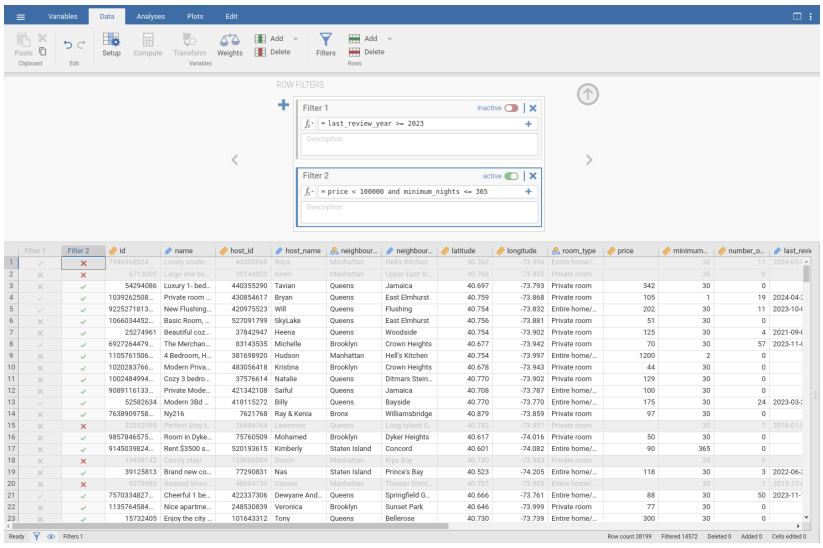

Figure 3.4.1 Two filters.

1. The first is the one we just used to set up a new variable. You can now delete the filter for `last_review_year >= 2023`, since we have a variable measuring it.

2. The second one is the one filtering out the artificially inflated prices, where `price < 100000 and minimum_nights <= 365`. Let's keep that one here, since we have reason to believe that these listings are ones that aren't really there for booking purposes: the price has been artificially inflated to make them unavailable to book.

Confidence Interval to Estimate a Proportion

We're going to calculate the confidence interval to estimate the proportion of listings that have recently been reviewed in two ways:

- We'll calculate the confidence interval by hand, using the formula:

$$\hat{p} \pm z^* SE_{\hat{p}}.$$

- We'll let jamovi calculate the confidence interval using the *Binomial Test*.

> **Note 3.4.2**
>
> These are different from each other! In the one we do by hand, we'll use a normal approximation of the binomial sampling distribution, while jamovi will calculate based on the actual binomial sampling distribution. We'll see how close they are, and how well the normal approximation of the binomial sampling distribution works!

To construct a 95% confidence interval estimation of p, we only need a value for $\hat{p}$ (the sample proportion) and n (the sample size). We can use $z^* = 1.96$, which is the critical value that corresponds to a 95% confidence level, and the standard error of the sampling distribution:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Pull up the Descriptives menu and add `recent` to the "Variables" window. From there, we can de-select all of the statistics other than sample size, and tell

jamovi to display a Frequency Table. This gives us two bits of information that are valuable, and one of them isn't necessarily the % of Total column in the Frequency Table!

This value is rounded, but we can see how it was calculated: it's the number of times the value of recent was 1 divided by the total sample size. Sure, we can change the number of decimal points to get a more accurate percentage displayed, but I think we can just use

$$\hat{p} = \frac{14302}{23627}.$$

Checkpoint 3.4.3

Calculate the 95% confidence interval for the sample proportion, using $\hat{p} = \frac{14302}{23627}$ and $z^* = 1.96$.

We'll compare this with the exact binomial confidence interval to see how well our calculated confidence interval matches.

We're actually going to construct the confidence interval at the same time as doing a hypothesis test, but we'll see that part specifically later.

Find the Frequencies menu in the Analyses tab and select "2 Outcomes (Binomial Test)" from the menu. It should be listed under One Sample Proportion Tests. This menu should look familiar, and you can add the `recent` variable to the window on the right. Then check the box to display a confidence interval with 95% confidence level. You'll get a table on the right with the counts for the two values of `recent` as well as the sample proportions, and, most important for us right now, the lower and upper bounds of a 95% confidence interval.

Checkpoint 3.4.4

Report the lower and upper values of the 95% confidence interval. How much does this confidence interval differ from the one you calculated by hand (using the normal approximation)? Explain what this means.

Checkpoint 3.4.5

Pick one of the confidence intervals and interpret it. What does this tell you about the actual amount of listings on AirBnB that have recent reviews?

Hypothesis Test for a Proportion

Now take some time to look at some of the other options in the Proportion Test. We can add a Test value, which is just the claim about the population proportion that the null hypothesis makes. We can also change the alternative hypothesis. For now, we'll leave things as the default options:

$$H_0 : p = 0.5$$
$$H_A : p \neq 0.5$$

So nothing now has changed in our output: we're displaying a table with a p-value and a confidence interval. Notice, now, that the p-value is small enough that jamovi just displays it as being less than 0.001. This is typical in statistical software: with very small p-values, we might just get a display of a kind of "order of magnitude" or a general benchmark of how small it is.

Let's compare that to a p-value we might calculate using our normal approximation hypothesis test.

1. *Prepare*

   Our parameter of interest is the population proportion of AirBnB listings that have been reviewed "recently" (where we define that to be reviewed in 2023 or 2024). We have our hypotheses, with the null hypothesis

   $$H_0 : p = 0.5$$

   and the alternative hypothesis

   $$H_A : p \neq 0.5.$$

   We'll test this at the typical significance level, with $\alpha = 0.05$. From our sample, we have that $n = 23627$ and the sample proportion is

   $$\hat{p} = \frac{14302}{23627}.$$

2. *Check*

   Our sample size is very large here, so we more than meet our conditions to assume that the sampling distribution of $\hat{p}$ is normal. Since our null hypothesis claims that the population proportion is 0.5, we expect to see 11813.5 listings that have been reviewed recently and 11813.5 listings that haven't. These are both well over 10. We can also assume that each listing being reviewed recently or not is independent of each other, but a deeper examination of this might lead us to think that it is possible that there could be some small amount of dependence: maybe some specific location or host is more or less likely to be reviewed after being visited, or more or less likely to be visited at all. For now, let's not worry too much about this.

3. *Calculate*

   We need a calculation for standard error, a Z-score, and then a p-value.

   For the standard error, we can use the following:

   $$SE = \sqrt{\frac{p_0(1-p_0)}{n}}$$

   where $p_0 = 0.5$, the claim from the null hypothesis.

   Then we can calculate the Z-score for our sample statistic under the null hypothesis' sampling distribution:

   $$Z = \frac{\hat{p} - p_0}{SE}.$$

   Now look up a probability for the p-value using a normal distribution tool, like the one below.

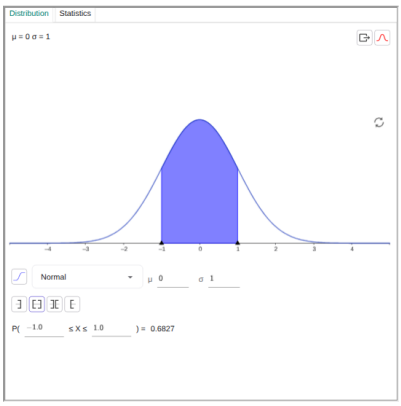

Figure 3.4.6 Normal Distribution Calculator

You'll notice that the Z-score is so big that this application rounds the p-value down to 0. We know the probability corresponding with this Z-score isn't actually 0, it's just extremely small: that's why jamovi gives the output $< .001$.

Let's do this one more time. In the little applet above, there are two tabs at the top: one for "Distribution" and one for "Statistics". Click the "Statistics" one. From the drop down menu at the top, change "T Test of a Mean" to "Z Test of a Proportion". Now we have a spot to enter the claim from the null hypothesis, select our alternative hypothesis, and then put in 14302 for "Successes" and 23627 for n. This should report results for a Z-score (which should match what you calculated) and a p-value (which should be 0, just like what you looked up).

4. *Conclude*

   Compare the p-value to our significance level. What does this mean?

Checkpoint 3.4.7

Interpret the results of the hypothesis test. What does this tell us about the population proportion of recently reviewed AirBnB listings?

Checkpoint 3.4.8

Connect this result or conclusion to the conclusion you got from the confidence interval. Are they related to each other? How? Why might it be useful to pair these kinds of results together when we analyze and test data?

Save Your Progress. We added another variable, so it could be useful to save this. As a reminder, you can right-click on the different test results you've created and select "Remove" from either the "Analysis" or "All" menus in order to clean up your output.

3.5 Inference for Categorical Data

In this lab, we'll apply some different tests and build some different confidence intervals that are concerned with proportions.

What You'll Need.

1. *Data:* make sure you're working with a copy of the AirBnB data set that has the most recent filters in it!
2. *Notes:* it will be helpful to have any notes or resources about hypothesis tests or confidence intervals for categorical data. You'll likely want to be able to look up conditions easily, to remind yourself of what is happening in each calculation.

Tasks

We've already looked at a test of a single proportion, so let's compare proportions a bit more heavily.

Distribution of Room Types

Build a frequency table for the room_type variable in the Descriptives menu. We've already noted that the types of listings are dominated by the Entire home/apt and Private room categories. Let's compare this distribution to a claim made in a 2017 publication[1]. Below is a relevant table, copied from the paper.

Table 3.5.1 The percentage of entire home, private room, and shared room listings in Airbnb

	Total cumulative listings			Active cumulative listings[2]		
Year	Entire home	Private room	Shared room	Entire home	Private room	Shared room
2008	50%	50%	0.0%	66.6%	33.3%	0.0%
2009	57.6%	42.3%	0.0%	58.1%	41.8%	0.0%
2010	53.2%	46.7%	0.0%	50.4%	49.6%	0.0%
2011	47.4%	50.1%	1.3%	46.8%	51.5%	1.6%
2012	50.2%	48.1%	1.7%	50.3%	48.1%	1.5%
2013	48.8%	49.4%	1.7%	49.2%	49.4%	1.3%
2014	50.2%	47.8%	1.9%	51.3%	47.4%	1.2%
2015	49.5%	47.8%	2.6%	51.8%	46.4%	1.7%
2016	49.5%	47.4%	2.9%	51.1%	46.4%	2.4%
2017	49.3%	47.6%	2.89%	50.4%	46.9%	2.4%

We'll have to fiddle with this a bit to get this claim for the null hypothesis. We'll use the 2017 numbers, and the total cumulative listings. Note that this doesn't have "Hotel" room as one of the options, since that wasn't an option in 2017. It's fine, the percentages don't quite add up to 100% anyways, so let's put the remaining bit in the hotel room option. Here's our null hypothesis claim!

[1] Dogru-Dr. True, Tarik & Mody, Makarand & Suess, Courtney. (2017). The hotel industry's Achilles Heel? Quantifying the negative impacts of Airbnb on Boston's hotel performance. Boston Hospitality Review. 2017.

[2] Listings with at least one booking within the past 12 months as of June 2017.

Table 3.5.2 Null Hypothesis Distribution

`room_type`	Proportion
Entire home/apt	0.493
Hotel room	_____
Private room	0.476
Shared room	0.0289
Total	*1.00*

Checkpoint 3.5.3

What is the percentage that we'll put in for hotel room to make the total add up to 1.00?

Now we need to do a Goodness of Fit Test! In the Analyses tab, select Frequencies and then find the "N Outcomes, χ^2 Goodness of fit" option. We should have a pretty familiar setup, with all of our variables on the left. Add `room_type` to the "Variable" field. You'll need to add your expected proportions! Open up the "Expected Proportions" menu and fill in your null hypothesis distribution (using your proportion for hotel rooms).

On the right we have a table with the frequency distribution of our variables, as observed (we can add the expected distribution using the Expected counts checkbox) and a table with the test statistic, degree of freedom, and p-value.

Checkpoint 3.5.4

What was the χ^2 test statistic for this test? How do you interpret the results of this test? Does the distribution of AirBnB listing types in New York match what was reported for the overall distribution of AirBnB listings in 2017?

Listing Types by Borough

Let's see if there is some connection with the types of listings and the boroughs of New York City. We'll perform a χ^2 Test for Independence or Association. Under the Frequencies menu, find the "χ^2 Test for Association" option. Here we'll have an option to perform the test for association by building a contingency table or a two way table. We'll add `room_type` and `neighbourhood_group` as Rows and Columns (it doesn't matter which one is which). This will set up our test for association, but notice that we have some options in the "Statistics" menu. If we had a 2x2 table (two groups with a proportion in each), we'd be performing the "Z Test for 2 Proportions", and we have options to change the null hypothesis, output a confidence interval, and show the Z Test statistic. There are also a couple of other nice measurements that get used in different field, especially for the 2x2 case. For our test, we'll leave it as just showing the χ^2 statistic.

Under the "Cells" menu, we can show the expected counts under the null hypothesis that these two variables are independent.

Checkpoint 3.5.5

What is the expected number of Private rooms in the Bronx under the null hypothesis? Where does this number come from, or how was it calculated?

Checkpoint 3.5.6

What is the χ^2 statistic, p-value, and conclusion of this test? Are these two variables independent? What does that mean about the types of listings across different boroughs?

Note also that there's an option under "Plots" to include a bar graph. I think this one isn't that great and is often cramped. We've had better options in the "Survey plots" menu. Go to this menu and create a plot that compares the percentage distribution of `room_type` across different boroughs from the `neighbourhood_group` variable or vice versa (distributions of boroughs for different listing types).

Checkpoint 3.5.7

Generate and submit your plot comparing the distributions of `room_type` across different boroughs from the `neighbourhood_group` variable or vice versa (distributions of boroughs for different listing types).

3.6 Inference for Means

This lab will be a bit longer than the previous one: we'll work on doing some of the inference on means, and, unlike the lab on proportions, we have to start with single means tests and confidence intervals. (We did single proportion stuff on the Introduction to Inference lab, remember.) We also want to be a bit more careful with some of these, since there are some conditions to check outside of just big enough sample size.

What You'll Need.

1. *Data:* make sure you're working with a copy of the AirBnB data set that has the most recent filters in it!
2. *Notes:* it will be helpful to have any notes or resources about hypothesis tests or confidence intervals for categorical data. You'll likely want to be able to look up conditions for things like ANOVA or the t-tests easily.

Tasks

We'll work on building a t-test for one mean, two means, and then think about how to apply ANOVA here.

Single Mean

The website AllTheRooms claims that the average price for an AirBnB in New York City in 2021 was \$143/night[1]. Let's test this claim!

We're going to set up a t-test with the following hypotheses:

$$H_0 : \mu_{\texttt{price}} = 143$$
$$H_A : \mu_{\texttt{price}} \neq 143$$

In the Analysis tab, you'll find a T-Tests menu with a "One Sample T-Test" option. Select that. Add `price` as a Dependent Variable, and set up your hypothesis: You'll need to make sure that we're selecting the "Student's t-Test" (there are other options here!) and include the tested value in the hypothesis. You'll select the appropriate alternative hypothesis, and even get some options on what gets displayed.

For now, you should see a test statistic of 26.4185, which is extremely large! Picture that: we're over 26 standard errors into the tail of a nearly-normal t-distribution. You can see the correspondingly low p-value. The evidence that we have about AirBnB prices does not fit the claim from this website, and we reject the hypothesis that the average nightly price is \$143.

How much is it? Well, for this, we need a confidence interval. There are two quick ways of doing this. One of them is in the t-Test itself: we can select "Mean difference" under "Additional Statistics" and choose a 95% confidence interval. This will actually display a confidence interval estimating the difference of the true mean to the one in the null hypothesis. So the bounds of our interval need to be added to the claim from the null hypothesis to get the actual confidence interval on the mean.

[1] Average Airbnb Prices By City: How Much Should You Charge For Your Airbnb? [2022]

Note 3.6.1

Again, this doesn't show us the confidence interval to estimate the mean price: just the confidence interval estimating the *difference* in the mean price from the one we tested against: \$143/night.

Another way of doing this is to find it in the Descriptives option under the Exploration menu. If we add the `price` variable, we can choose to display the mean under the statistics, but we also have some options for the confidence interval of the mean below. A 95% confidence interval will match what we found in the t-Test itself.

Checkpoint 3.6.2

Report the actual confidence interval estimating the average price per night. Which way of finding it do you prefer? Why? (They should be the same results, just a different way of getting jamovi to display it.)

Checkpoint 3.6.3

In the same source that we used, they say that the average per night price of an AirBnB in USA and also in New York *state* is \$216/night. Without doing a hypothesis test, use your confidence interval to explain whether we would reject this null hypothesis or not.

While we're doing this, we should be concerned with our assumptions: we have seen that price is very clearly not normally distributed: this is fine, since our sample size is so big! The Central Limit Theorem says that if the sample size isn't too small, the actual distribution of the values we're sampling does not matter. We have also done our best to think about the outliers on this variable, and were able to filter our variable to remove some of them that clearly did not belong to the population. Good work, us!

Note 3.6.4

Again, notice that we removed data points that we found that *did not belong in the population we were examining.* We're not removing outliers because they were outliers. We examined them and found that they were not real AirBnB listings. This is, in general, a pretty hard thing to do without a lot of background knowledge for the type of data we're working with. We should, in general, be very reluctant to remove outlying data, since we don't want to remove some of the extreme values that actually are representative the variability in our population.

Price More Recently

We have already explored how the prices in this data set might not be representative of the most "current" AirBnB listings: we tried to filter out the listings that were artificially high in price or were seemingly not actually on the market already. We also have the new variable from the Introduction to Inference lab, `recent`. Let's see if the average price is different based on whether or not the listing has been reviewed recently. Really, we're testing these hypotheses:

$$H_0 : \mu_{\texttt{recent}} = \mu_{\texttt{non-recent}}$$
$$H_A : \mu_{\texttt{recent}} \neq \mu_{\texttt{non-recent}}$$

In the same T-Tests menu, select "Independent Samples T-Test". Now we can add `price` to the "Dependent Variables" pane, and add the new variable `recent` as a "Grouping Variable." We again have a ton of options for how we get results. A thing that you might notice is a little footnote under the results of your test: "Levene's test is significant ($p < 0.05$), suggesting a violation of the assumption of equal variances." One of the assumptions that we briefly talked about how sometimes we assume equal variances for the groups, but in our case we don't want to. A couple of points:

- The version where we *don't* assume equal variances is called Welch's t-Test and is an option here. You can select it instead of Student's t-Test.
- Levene's Test is another hypothesis test that looks at whether or not two variances are different: a low p-value rejects the null hypothesis that the variances are the same. With large samples, this sometimes happens even if the sample variances are *very* close.

Let's look at the variances: it might be easiest to just think about standard deviations. Check the "Descriptives" box under "Additional Statistics." This pulls up a table with sample size, mean, median, standard deviation, and standard error of the mean. We can compare standard deviations, which will also compare variances then, to see if they look like they might be close.

Checkpoint 3.6.5

What are the standard deviations for the price of recently reviewed and non-recently reviewed listings? Do they seem close to each other? Is there evidence that we actually could have equal variances here, or not?

Let's go back to our t-Test and use Welch's t-Test, not trying to assume equal variances. We can, again, get a confidence interval for the mean difference under "Additional Statistics". I sometimes think it's hard to keep track of this, since it relies on knowing which way we did the subtraction to find the difference. I like to use the "Descriptives" table to see which one has the higher mean. The confidence interval for the mean difference estimates how much bigger the mean is in the group with the higher mean.

Checkpoint 3.6.6

Explain the difference in prices for the two groups based on the results of the hypothesis test and the confidence interval.

More Groups

Now we'll think about a question that we've been hinting at in these labs: is there a difference in the prices of these listings based on what borough they're located in?

This is a classic ANOVA question: we are looking to see if our means are all consistent across the different groupings or if there is some difference in some of those means somewhere. There are a lot of cool options to use in ANOVA in general, but since we're looking at a very introductory level here, we're going to stick to the "One-Way ANOVA" found in the ANOVA menu. Add `price` to the "Dependent Variables" pane, and include `neighbourhood_group` as the "Grouping Variable." You'll get to select that we won't assume equal variances. The ANOVA table here is a bit different than what we might normally see,

since it doesn't report the Mean Squares: it just gives us the F-Statistic, the degrees of freedom, and the p-value.

Here we can see that there is definitely a difference in between the average prices depending on which borough you're in!

The obvious follow-up question is "what are these differences?" One nice option is to add the "Descriptives plots" under the "Additional Statistics." This will display the confidence intervals of the means of each group for us to compare. You can also get these confidence intervals by using the Mean and Standard Error for each group in the Descriptives table, and then building it by hand (by just using the almost 2 standard errors for margin of error, like normal).

Checkpoint 3.6.7

Build and submit the picture of the confidence intervals for the mean price for each borough.

We can also do a post-hoc test, where we look pair-wise comparisons: we'll use the Games-Howell test, and see which boroughs are significantly different in price. It looks like, as a summary:

- Manhattan's prices are REALLY high comparatively to every other borough.
- Brooklyn's pricing model is different from every other borough as well.
- Staten Island, Queens, and the Bronx may not have any significant differences in the mean price.

Chapter 4

Project Parts and Details

This is the chapter with all of the details about your project! Each section describes a single part or submission for your project. Your submissions will include some typed out write-ups, written in the style of an essay or paper, serving as a report on your data set and your progress. Details for what you should include in your writeups are listed in each section.

Some of these submissions will also include a gallery submission. These are meant to showcase some aspect of your progress to the group. We'll share snippets of our work with each other as we go.

By the end of this project, you will submit a more formal and complete report describing your work. Similarly, the gallery component will include a short presentation.

4.1 Data Description

For this first part of the project, you'll be deciding and specifying what, exactly, you want to study this semester. You're going to sift through data sets, select one, and then describe it in some specific ways. You'll get used to exploring the data in the software (jamovi) that we'll use and overall orient yourself to the process of working on and submitting parts of the project.

Data Selection

The first thing you'll need to do is look through the data sets and pick which one you would like to analyze. Take some time to read through the basic descriptions of each, check out the data dictionaries for different options, and ultimately select an option to work with. You can look through the data sets as much as you'd like while you select them, but I hope you go into this project with a pretty good understanding of what information is included in each data set.

Each of the data sets have different applications, so you can prioritize your selection based on the types of applications you can see for these data sets or the topics of the data themselves.

Find the data sets in Chapter 2. Note that the AirBnB Data Set (Lab Data Set) is used for the labs and is not eligible to use for your project.

In your submission, explain which data set you selected, and give some explanation. Here are some ideas or examples of things to think about:

- Did you pick the data set because the data was describing something you are interested in? Tell me about it!
- Did you pick the data set because you had some questions about it that you wanted to answer? What are your questions?
- Were you interested in something about the types of applications this data could have? Describe them!
- Is there something else about this data set that was intriguing to you?

Data Details

I want you to think about the data you've selected and familiarize yourself with it.

- What is the population that you're data set represents? Be specific, and explain why you think this. Do you think this data is a representative sample?
- We don't get much information about how this data was sampled or collected. Do you think this data was collected from an experiment or an observational study?
- Pick 10 variables from your data set. For each one, describe:
  - Whether it is categorical or numerical.
  - If it is categorical: is it ordinal or nominal? Describe the different levels of the categorical variable briefly.
  - If it is numerical: is it continuous or discrete? Is there a natural maximum/minimum for the values (for example: does it only make sense to have values greater than or equal to 0, or something similar)?

    > In your write-up, you can just include these 10 variables in a list, almost like in the data dictionary for your data set.

- Pick a pair of variables that you think could be associated. Why do you think these variables are associated? What about their description makes you think they could be associated?
- Pick a pair of variables that you think could be independent. Why do you think these variables are independent? What about their description makes you think they could be independent? If you don't think any pairs of variables are independent, explain why you think that.

Application and Decision Making

We want to go into this with some idea of why the data we're selecting could be useful.

- Why is your data set useful? What kinds of business-specific applications might it have? Note that these don't have to be profit-generating applications for a company: there are tons of business applications that aren't centered around making money off of a product.
- What kinds of business decisions could this data set help with?

- What other information or data would be helpful to have alongside this data? If you could collect more information, especially more information about the observational or experimental units, what would you collect? Why would this be helpful?

Write-Up

You should collect all of your answers to these questions in a short report. It should look like a short essay or paper. You can write this in Word, Google Docs, or any normal word processor. Feel free to use the headings we have on this page in your report (Data Selection, Data Details, Application and Decision Making) to help organize your answers. Make sure you answer everything, and write clearly and coherently. I won't be grading you on grammar or spelling, but in any report like this, clarity is most important.

Gallery

For your gallery post, you should explain three things:

1. What data set did you choose?
2. Why did you choose it? Give a short explanation, similar (or the same!) as what you submitted in your write-up.
3. Why might your data set be useful? Give a short (2-4 sentences) summary of the ways that this data set could be used in some business application or decision making process.

This will hopefully give people some more ideas as you move forward: maybe something someone says about your data set will spur on a question or connection that you think would be interesting to explore. Once these have been submitted, try to give a brief glance around to other submissions

4.2 Exploring Data

In this second part of the project, you'll start doing some more exploration and summary of the different variables in your data set. In Part 1 of this project, Data Description, you identified some possible associations between variables as well as some possible applications of this data. I want you to try to think about these overall goals, and think about how we'll build some research questions to think about. While we're preparing to do this, we should think about what variables will be important to these questions or applications. From there, we want to explore them and summarize them.

You might want/need to do some data cleaning here: make sure your variables are classified correctly in jamovi, and, if you have an ordinal variable, you should confirm the order of the levels. You might need to set up some filters or build a new variable here, so be prepared to do that.

Variable Selection

Pick 5 variables from your data. These should:

- Include a mixture of numerical and categorical variables.
- Be variables that you think are central to some application, decision, or question that you could ask (they don't need to all be connected with the same application, decision, or question). These don't need to be the same things you identified in Part 1 of this project. Sometimes we think of these variables as "demographic" variables, but yours don't all have to be exactly this.

Describe your 5 variables and give a brief explanation of why you picked them.

Numerical Variables

For each numerical variable:

- Build both a histogram and box plot of the variable's distribution.
- Describe the general shape of the distribution: Is it symmetric or skewed? Does it look like it is unimodal or not? Explain.
- Report an appropriate measure of center. Interpret what this value means in the context of your variable.
- Report an appropriate measure of variability. Interpret what this value means in the context of your variable.

Spend some time looking more at each one: look at the minimum and maximum values, and generally investigating the values. Does anything look out of place? You'll have more time to look at your variables, but I want you to be aware of any issues that come up and fix them later. You might need to add a filter or create a new variable to describe the thing that you actually care about. You'll have more time to do this, but I want you to start thinking about it as we keep moving through this project.

If you do change anything by adding a filter or creating a new variable, make sure you include this information in your write-up so I know what you're up to!

Categorical Variables

For each categorical variable:

- Include a frequency table for the distribution of the variable. This should include both the raw counts and the proportional frequency/percentage frequency. You don't need to include a cumulative frequency if you don't want to.

  > You might be able to copy/paste this from jamovi, but sometimes that feature doesn't work well with specific text editors, so you might need to manually type in this table using the table creator in your text editor.

- Build a bar graph (whatever kind is most relevant) of the variable's distribution.
- Give a general summary of the variable: what is the mode, and describe what levels of your category have high or low frequencies. Interpret these in the context of the actual variable.

Again, spend some more time looking at each variable. Does anything look out of place? You'll have more time to look at your variables, but I want you to be aware of any issues that come up and fix them later. You might need to add a filter or create a new variable to describe the thing that you actually care about. You'll have more time to do this, but I want you to start thinking about it as we keep moving through this project.

If you do change anything by adding a filter or creating a new variable, make sure you include this information in your write-up so I know what you're up to!

Write-Up

You'll write up your answers in a short essay-style report again. Make sure to include the relevant plots and tables in the actual document. You might be able to copy/paste your plots, but you might have to save these as images to include in your document. Make sure you give yourself some time to figure this out.

Again, feel free to use the headings (Variable Selection, Numerical Variables, and Categorical Variables) in your document to organize your answers. You should write most of this up using complete sentences and paragraphs, since this is meant to mimic an actual report.

Gallery

Here's our second gallery! This one will be a bit more visual. By the end of it, we'll have a nice collage of different plots from different variables from different data sets.

For your gallery post, you should post two plots and an accompanying explanation for each.

1. Pick your favorite plot summarizing a numerical variable and include it in your post. Give a brief summary of what this plot shows you about the distribution of the variable: can you see the shape, center, and variability of the data? Include the calculated measures of center and variability that

you included in your report, and try to make sure your explanation uses these numbers in the actual context of the variable.

2. Pick your favorite plot summarizing a categorical variable and include it in your post. What does this show us about the distribution of that variable? Briefly explain anything noteworthy about the distribution in the context of your variable.

When enough things have been posted, look around at other people's plots. Try to think about which ones are most effective at summarizing the variables they picked. Do any of these plots make you think of specific kinds of questions you could ask about your variables?

Reflect on your own plots: would it be more effective to use a specific kind of plot? Does it make more sense to include percentage frequency instead of raw counts? Try to think about how, moving forward in this project, you can most effectively summarize your variables.

4.3 Association and Questions

Hopefully by now you are all very familiar with the different variables in your data set as well as some of the overall context and applicability of your data. You have hopefully been asking good questions about the connections between variables. In this part of the project, we want to formally state our research questions, hypothesize (give a reasonable guess) about the answers to those questions, and then begin exploring them.

The order here is very important: we really should have formed these questions *before* we even collected the data. Of course, data collection takes so much time that it's not reasonable to start a semester by asking some questions and then collecting the relevant data from there. We are starting with data already collected, but we still want to form our questions before analyzing too much. If we let our sample direct our research questions, we are more likely to find and focus on aspects of our sample that aren't necessarily reflective of our population.

Research Questions/Hypotheses

Think back to some of the questions you've been forming this whole time. Pick two questions.

1. A question about whether or not two categorical variables are associated.
2. A question about whether a numerical variable and a categorical variable are associated.

> **Note 4.3.1**
>
> If you want some help thinking about questions, you can re-visit the More Data Visualization lab for some examples.

For each of these questions, you should say what you think the answer is: this might refer back to things you've already submitted in parts 1 and 2 of this project, but does not need to. You should explain your thinking, but these are just meant to be thoughtful guesses at an answer. You'll be graded on your thoughtfulness, not on how well you predicted a correct answer.

You should also explain which of your questions you think are most important or most interesting, especially in terms of some business application or decision.

Exploration

You're going to start exploring your questions more deeply. Before we begin, it might be helpful to "translate" your questions into ones that describe statistics. For instance, here are two of the three questions we asked in the More Data Visualization lab:

1. Is the distribution of AirBnB listings different across different boroughs?
2. Are the different types of listings available in the same way for the next year?

Here is a way that we could translate those:

1. Is `room_type` independent of `neighbourhood_group`? Does the proportion of different values of `room_type` change for different values of `neighbourhood_group`?

2. Does the mean of `availability_365` differ across different values of `room_type`?

For each of the two questions you ask about your data set, you should:

- Calculate any relevant summary statistics. For example, if you're looking at proportions or means across different groups, report those. These statistics should be directly relevant to your question.
- Construct relevant plots and/or tables. You might want to try some different options and then select what you think best represents the relationship you are exploring in your question.

Summary

For each of your questions, give an explanation of what your different statistics tell you in the context of the variables you're looking at. Describe your plots and tables tell you about your question. In general, discuss your question, your initial thoughts on what you thought might be true, and what this initial exploration has shown you.

Write-Up

In this write-up, the structure will change. Instead of using the headings above, I want you to break up the report into sections dedicated to each individual research question. In each section (for each question), you can then use the headings above (Hypothesis, Exploration, and Summary) in your document. Include all of the relevant tables and plots in the report itself, and, as always, write this up coherently. You should write most of this up using complete sentences and paragraphs, since this is meant to mimic an actual report.

Gallery

This will be our last gallery for the semester, until the final project presentations. Pick one of the plots or tables you built that you think demonstrates a relationship or lack of relationship. Post it to this discussion with a brief (2-4 sentence) summary of what this graph is actually showing. Describe the variables represented and the general relationship or lack of relationship you're showing. After your summary, give a brief explanation of why this useful to know: how can this be applied or what kind of decision might this help with?

Once you've posted, you should try to reflect on your research questions for your data set: how can you use the ideas that others have posted to help you visualize your relationships more clearly? Think about the relationships you're seeing. Do any of these differences in proportions or means or anything else seem like they could be differences in the population, or might they just be differences due to the randomness of building the sample? Which of these relationships seem most important to find, based on what they might mean for some kind of decision-making process?

4.4 Analysis

This is the part of our project where we'll transition into doing some inference! You've had a whole semester so far to get familiar with your data set, build some good research questions, and even explore them preliminarily. Now what we'll do is try to match these questions with a relevant test or statistical technique, apply it, and explain what we are doing.

You're going to have to be familiar with the types of tests and techniques, the assumptions or conditions behind them, and how to interpret the outcomes of those tests and techniques in the context of your data set.

Research Questions

You have two research questions from the Association and Questions submission, but you can come up with others if you'd like. Some other ideas for questions:

- Are two variables measuring the same thing, or are they from different populations? (This is useful when you have historical values of something and new values measuring the same thing, or other instances like this.)
- Is a categorical variable distributed in the way that we have thought it was? (For this, you'll need some reasonable claim on how it was "meant" to be distributed, which might require outside research.)
- Is the average value of some numerical measurement the same as what it was claimed to be? (Again, you'll need strong research to get a realistic claim for this average.)

If you choose to think about new questions, you should give a short explanation of your change, including a description of why you think your new question is important.

For each of your questions, include the following.

1. *Statistical Test*

   Match your question with some test or technique we've learned this semester, including (but not limited to):

   - Z Test (and Estimate) for a Single Proportion
   - Z Test (and Estimate) for Two Proportions
   - χ^2 Test for Goodness of Fit
   - χ^2 Test for Independence/Association
   - T Test (and Estimate) for a Single Mean
   - T Test for Paired Means
   - T Test (and Estimate) for Two Means
   - ANOVA

   Explain why this test is a match for your question.

2. *Prepare*

   Identify and describe the parameter(s) of interest in your test. List the hypotheses of the test, and explain in words what these hypotheses represent in your data. Decide on a significance level: if you use something other than $\alpha = 0.05$, you should explain your reasoning. Finally, report the relevant sample statistics from that you will be using from your data to perform the test (for example, a sample mean and a sample size).

3. *Check*

   List the conditions for the test you've picked, and explain why your data meets these conditions. If there are any issues where they might not, explain why not.

4. *Calculate*

   Apply your test. Include the results in your writeup. You might choose to include a table, similar to the ones that jamovi outputs, with test statistics and p-values. Explain what the test statistic is measuring, and what the p-value means.

5. *Conclude*

   Report your conclusions from the hypothesis test, making sure to interpret them in the context of your data. This is a good time for you to discuss the implications of what you've found and think about how this result might be used. It is absolutely fine if you do not reject your null hypothesis: explain what this means for any possible stakeholders!

Be sure to include any relevant and descriptive plots.

Similarly, include any relevant confidence intervals. You can include these as plots or report them in a table or something similar. Make sure to explain and interpret what these confidence intervals mean.

Write-Up

Again, I think you should structure this in two sections, one for each research question. Then, in each of these sections/questions, you can include each step of our inferential process. Feel free to use some headings to break these up, but you do not need to. Include all of the relevant tables and plots in the report itself, and, as always, write this up coherently. You should write most of this up using complete sentences and paragraphs, since this is meant to mimic an actual report.

4.5 Final Report

We are here! The final project submission! Hopefully you'll find that the different submissions along the way will really help build out this report, so fee free to use what you've already created this semester. You can definitely edit it to make sure the report "reads" well, but much of the work you've done will be able to be used or at least replicated.

In general, for this final project submission, you're going to analyze one research question from start to finish. We're going to follow a pretty traditional structure that you'll find in other statistical reports.

Begin your report with a title page, including:

- A title! You get to create this, but it's helpful to give an informative title. For instance, if we were looking at a connection between the types of AirBnB listings in different boroughs of NYC, we might title our report something like, "Where to Stay in the City that Never Sleeps: AirBnB Listing Types Across the Boroughs of NYC."
- Your name.
- Where you submitted the report. For us, that's this class.
- The semester/year that it was submitted.

In the main body of your report, use the following structure.

Introduction & Background

This is where you'll introduce your research question. Give some explanation on the background here:

- Give some background explanation about your data set. This might mean giving some explanation on the context, some of the terminology, etc.
- State your research question. Explain the significance of your question, or why it might be important to look at. Be specific here, with real, actionable outcomes. Make sure to also include a note about the actual population that this question pertains to.
- As a part of the background work for your question, you should explain the distributions of the variables you mention. Include a plot or table of the distribution (pick whatever is best at showing a clear picture of the distribution of the variable) and describe the distribution using shape, center, and variability.
- Include any other summary statistics or plots that you think are relevant to understanding the background of your question or data. Note that these should be limited to plots describing single variables at a time. You can include plots showing the connections between multiple variables later. For instance, in my fictional report about the types of listings in different boroughs of NYC, I might show just a distribution of the types of listings, as well as a plot of how many listings are in each borough.

Methods

Here we'll explain a bit about the collection of the data, any work you've done to create or filter variables, the analysis you applied etc.

- Describe the source of your data.
- Describe any filters you applied to the data. Why did you use them? (If you didn't use any, that's ok!)
- Describe any changes to variables you made or any new variables you created. How did you create them? (If you didn't create any, that's ok!)
- Include some preliminary analysis: you should have some motivating graphs and maybe some computed statistics. Include them here with a description of what they show, and why they are pointing towards the specific test you used.
- Introduce the test you used. For this test, make sure to include some justification that the conditions of the test have been met. Explain what the hypotheses are, or what the intent of the test or technique is.

Sometimes people split the Methods section into two sections: a kind of exploratory data analysis, and then the formal statistical test. You don't have to use subsections to do this if you don't want to, but you should talk about both.

Results

This is the section where you'll introduce the main results of your test. You'll need to include:

- The outputs of the statistical test you performed. Maybe this is just a test statistic and p-value, but it might include a table with some relevant output from the statistical software.
- Explanations of these results. You should explain the results in statistical terms, as well as in terms that a non-statistician could understand. Make sure to use your context to explain and interpret the results.
- Include any relevant confidence intervals, as well as an explanation of what they represent in the context of your data.
- If you have any follow-up analysis (confidence intervals, other tests, etc.), then report them here as well. Make sure to give motivation for why these were useful.

In general, the results section should be able to be read as a stand-alone section, without the explanation of your methods. You don't have to repeat any explanation of your methods here, but you should explain the results in a way that makes sense without knowing all of the details in your Methods section.

Discussion & Conclusions

There are a couple of uses of the Discussion section:

- Introduce and explain any limitations you can think of: if you have reservations about the data itself, or want to warn about over-applying the results, or if you have concerns about the conditions of your tests, this is the place to be clear about that. We don't want to mislead the readers, and it's a sign of a good and ethical statistician to voice any concerns.
- A really important thing to touch on is the difference between statistical significance and practical significance. For your result, do they actually matter? It's totally ok if your answer is "No!" We want to stress that these results are meant to be used to make business decisions, in our case, and if you have a statistically significant result, but the result itself isn't that useful, it's good to note that.
- You should relate your conclusions back to the original question: were you able to answer the research question? Were you only able to answer part of it?
- You can also include some ideas about future work that makes sense as an extension of your results. What follow-up questions do you have that would be interesting to think about? If you could do things differently (especially when it comes to the data collection, which you didn't have a hand in), what would you do?
- Does this result apply to the population? Is there a way that you'd rather re-do this project and change a bit of what the population was, in order to answer a "better" question?

Written Submission

This is a full length report, and I hope that your submission reflects that. It doesn't necessarily have to be a specific length, but I want you to touch on all of this. Structure and organization is really important for these types of reports. We want to make sure that the sections are consistent and predictable, and we want to be able to easily find the stuff we care about. Often times, people will read an introduction and skip to the results and discussion before heading back to the methodology. Other times people will be really interested in replicating a study or performing a similar one, and so will not care as much about the results, but will really care about the methodology and discussion sections.

You should have high quality visuals that are included, but not distracting. Try to make sure your plots and tables support your writing instead of distract from it. This comes into play when you're considering silly things like sizing: you want things to be large enough to read, but not so large that they take over the report. As always, write this up coherently. You should write most of this up using complete sentences and paragraphs, since this is meant to mimic an actual report.

Presentation

You'll also include a short (up to 5 minutes) recorded video presentation. The presentation will not include everything from this report: just a basic introduction of the research question, overview of the statistical analysis, the important results, and some brief discussion.

Presentation Details

Your presentation should be short! This is meant to be 5 minutes long or less. Depending on how you balance the slides vs your own talking, you can accomplish this presentation in as few as 3 good slides, but definitely don't include more than 8 slides. Your presentation should include:

- A title slide with your project's name and your name.
- A brief introduction to your research question and the data set, including an explanation of why the question is interesting or worth answering.
- Some preliminary analysis: mostly you'll want at least some visual aid here.
- An explanation of the test you used, and why it is relevant for your question.
- A brief summary of the results of your test.
- A conclusion that includes at least one of the things from your Discussion & Conclusions section that you think is important to note or interesting to think about.

If you are worried about length, you could use each bullet point above as the guide to create a single slide (so your presentation would have a title slide and then 5 more slides).

How To Record a Presentation

There are a lot of options here, and you might know of them already, but I think the easiest thing to do is to just record yourself giving the presentation in an empty Zoom call. Here are some instructions for how to accomplish this:

- Download Zoom: you *may* be able to record in the web browser version of Zoom, but I'm not sure. You can download the zoom client at zoom.us/download.
- Once you've downloaded and opened Zoom, start a new meeting. You might need to choose your audio/camera and allow Zoom to use it, and then eventually "Join with computer audio."
- In the Host Settings, you might need to enable recording (either to your computer or to the cloud—I prefer to record directly to my computer, since you'll need to file eventually anyways).
- Click the green "Share Screen" button in the Zoom window, and select your presentation window (or the entire screen), and click share.
- Start the recording (this might be under the "More" option) either to the cloud or to your computer, whichever you've selected.
- Go back to your presentation window, start the presentation, and do it!
- Once you're done, stop your recording and end the meeting. You can close all of the Zoom windows now.
- If you recorded this to the cloud, you'll get an email later when this is available. Once it is, you can find the link to it in your email and download the video.

Practice Is Important

I really recommend that you practice this! Practice your presentation to make sure that you are clear and within the length guidelines. Practice recording a video to make sure you know how to do it. Give yourself time to get this finished and don't leave all of the tech stuff until the last minute!

Colophon

This book was authored in PreTeXt.