Data Analysis Plan

Prepared by: Matthew E. Vanaman

09-27-2021

Table of Contents

Read This First

In my experience, at least half of the time researchers do not have a clear plan for what they are going to do with their data once they get it. This means confusion, intimidation, frustration, and feeling defeated before they've even begun. And once the dataset is in front of them, they are not quite sure what they are looking for, what their first step should be, or why statistical software is giving them confusing errors. In the worst cases, researchers have to learn new statistics during the already-stressful task of analyzing data, having not been able to foresee what kind of statistics their data would call for. Here are some common situations researchers can find themselves in without a data analysis plan:

- A researcher is looking at the dataset and can't figure out what they need to calculate to answer their research question.
- A researcher can't figure out why the statistical software will not give the calculations they want.
- The calculations are working, but the researcher is confused by the results, which seem impossible given what she knows about the variables (for example, the average age is 105 years old).
- The researcher realizes that because of the way the variables were measured, the researcher must use statistics that she has never learned, adding further confusion, frustration, and worry.

The end goal of a data analysis plan is to prevent these situations by planning ahead. Whenever possible, you should have this plan in place before you try to analyze your data, especially if you are collecting you own data with a survey. Broadly, there are three more specific outcomes you want to strive for in a data analysis plan:

1. Identify the level of measurement of your variables.
2. Create a code book for your variables.
3. Use the level of measurements of your variables to identify what statistics you will need and, if applicable, what statistical test you will use.
4. Based on 2 and 3, make last-minute changes to your survey or choice of dataset.

Step 1: Identify Levels of Measurement

What is a *level of measurement*? The level of measurement refers to the way your variable manifests itself in your dataset. Imagine you have a variable (a column in your dataset) that contains the responses to a particular question on a survey. What form do the responses take? There are two broad categories: categorical and numeric.

Categorical contains these three levels of measurement:

- **Binary**: the variable only has two unique values.
  - Example: "Did you vote in the last presidential election?"
  - Values: "Yes" or "No"
- **Nominal**: the variable has more than two unique values.
  - Example: "Which best describes your gender?"
  - Values: "Female", "Male", or "I identify as something else"
- **Ordinal**: Same as nominal, but values can be ranked (they have a natural ordering). Ordinal variables are also not always *equidistant*, meaning that the difference between say, the highest and second-highest rank is not gauranteed to be equal to the difference between second- and third-highest rank.

  - Example: “What is you education level?”
  - Values: “Associate’s”, “Bachelor’s”, “Master’s”, or “Doctorate’s”
    - In this case, values are not equidistant: the “gap” between Bachelor’s and Master’s is smaller than the gap between Master’s and Doctorate’s. In other words, Doctorate’s are much harder to get than a Master’s, but a Master’s is only slightly harder to get than a Bachelor’s.
  - Note that except in rare cases, you usually treat ordinal like a nominal variable.

Numeric are numbered values, and the values are equidistant. Numeric contains these levels of measurement:

- **Discrete**: values are numbers, but cannot be broken down beyond whole units.
  - Example: Number of children someone has. It is numeric because children are expressed as a number, but a decimal would be impossible (your data can’t have 1.5 children).
- **Continuous**, which includes:
  - Interval: values are numbers, but can be broken down because they fall along a continuum or sliding scale. The value 0, in this case, does not literally mean 0, because 0 does not indicate an absence of the thing being measured - it is not meaningful.
    - Example: temperature measured by degrees Fahrenheit.
    - 0 degrees does not mean “absence of heat”, because that is still warmer than -10 degrees Fahrenheit.
  - Ratio: Same as interval, but value 0 is meaningful.
    - Example: temperature measured by degrees Kelvin.
    - In kelvin, 0 degrees does literally mean “absence of heat”, and the scale does not go below 0 - so 0 is meaningful.

Guide for Identifying Levels of Measurement

If you can already identify the level of measurement of your variable based on the definitions above, then great. If you are still having trouble figuring that out, you can use the flowchart below. For each variable, start at the top of the flowchart below and ask yourself each question about your variable one at a time. Based on your answers to the question, the flowchart should lead you to the level of measurement of your variable.

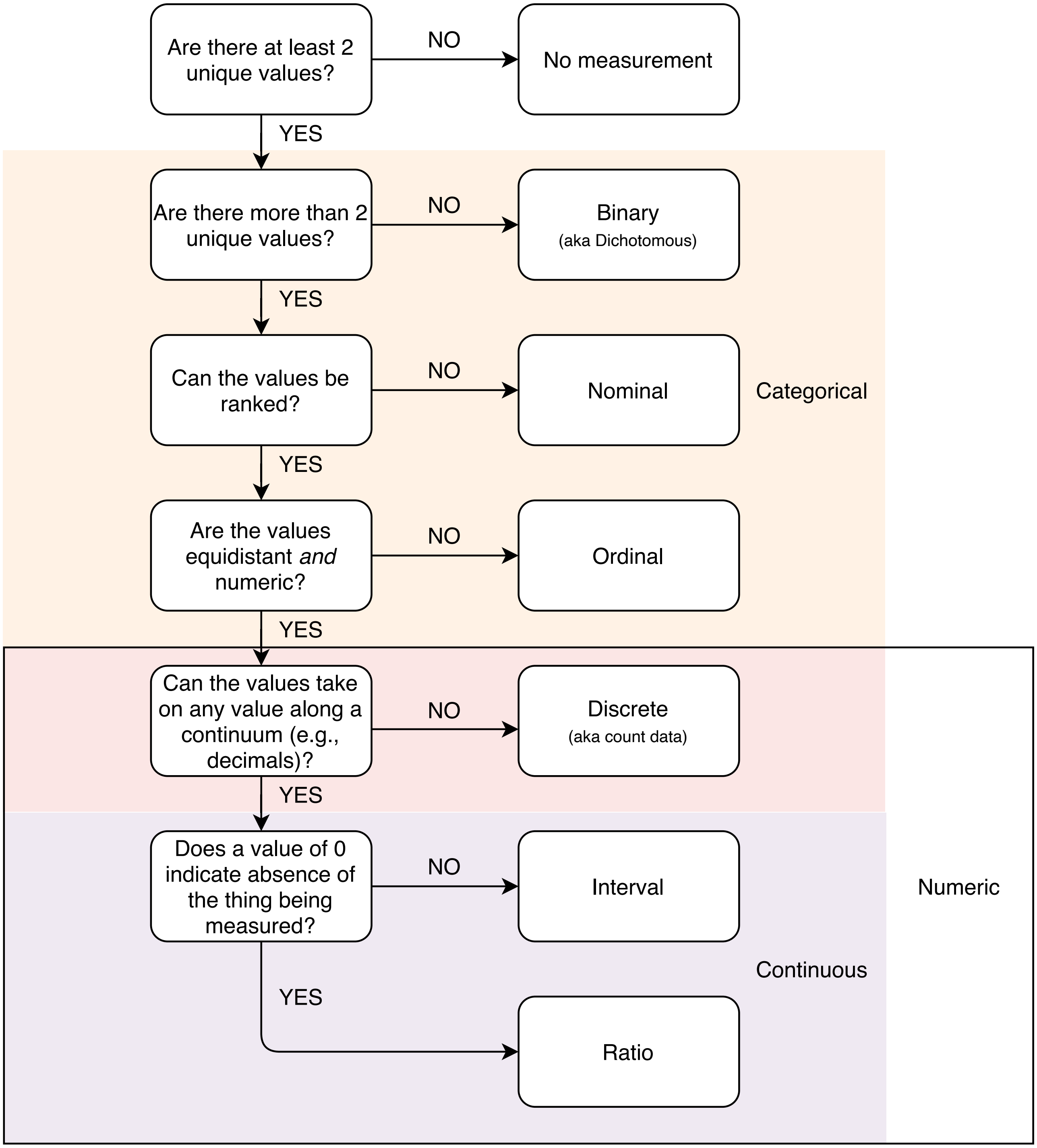
In the boxes below, type the name of your variable and it’s level of measurement:

Dependent Variable (variable you want to explain):

Independent Variable 1 (variable that might explain your dependent variable):

Independent Variable 2 (another variable that might explain your dependent variable, if applicable to your research question):

Determine Levels of Measurement


Are there at least 2 unique values?
NO
No measurement
YES
Are there more than 2 unique values?
NO
Binary
(aka Dichotomous)
YES
Can the values be ranked?
NO
Nominal
Categorical
YES
Are the values equidistant *and* numeric?
NO
Ordinal
YES
Can the values take on any value along a continuum (e.g., decimals)?
NO
Discrete
(aka count data)
YES
Does a value of 0 indicate absence of the thing being measured?
NO
Interval
Numeric
Continuous
YES
Ratio

Step 2: Create a Codebook

At this point, you have identified the levels of measurement for your variables. The next step is to make a code book. The goal of the code book is to:

- Give you a record of any changes your dataset will go through (e.g., recoding, or converting old values to new ones).
- Clarify the values you expect your variables to take on.
- Before you have data (especially if using a survey): help you check whether your survey produces the kind of data that you expect and need.
- After you have data: help you identify problems in your data.

The code book, which you can make in an Excel spreadsheet, might have the following columns:

- Variable: name of your variable. Personally, I like to put the survey wording here so that I do not have to go back to the survey to remind myself of the question that the name refers to.
- New Name: name of the variable in your dataset (you can shorten the name yourself, or sometimes statistical software will shorten the names automatically).
- LOM: level of measurement of the variable. This indicates how you expect the level of measurement to be in your data. But, you can sometimes find another dataset with the same variable measured on a different level (if using databases) or modify your survey to generate data on a different level (if using a survey).
- Scoring: how you will score the variable. This is mostly for variables that are created from other variables. If not applicable, put "as-is".
- Expected Values: unique values your variable can take on. Type the exact values that you expect to see in your dataset. For categorical variables, type each unique value. For numeric variables, you can leave this blank or put NA.
- Value Recode: if you have to convert old values to new ones, type out your recoding scheme.
- Min: the smallest value your numeric variables should be able to take on. Sometimes, this is based on the structure of the survey or dataset. Others, this depends on your expert knowledge of the variable. For example, if the question is "how many children do you have?", the smallest possible value should be zero since you can't have negative children (i.e. Ratio).
- Max: the largest value your numeric variables should be able to take on.
  - Sometimes this number is just a guideline. For example, if the question is "how many children do you have?", the largest possible value is not any particular value, but some values are definitely too extreme. If I see a participant with 11 kids, that might be unusual but still possible. But 110 might indicate a typo or other issue with that data point since no one has 110 kids.
  - In other cases, the value is definite. If my survey has a question that takes responses on a 1 to 5 scale, the highest possible number should be 5; if there is a higher number than that, this is a problem.

The table below is an example, based on the survey in the Appendix. You can add or subtract columns depending on your own circumstances. When you are getting data from an online database, the database almost always has an accompanying code book or web page with the information needed to construct one. If they don't have one, you might create one based on your understanding of how the data came to be, if possible. Note also that the next steps might require to go back and change elements of your codebook. The idea is that the code book should be finalized before you analyze your data (if using databases) or release your survey (if using a survey).

Table 1: Example Code Book

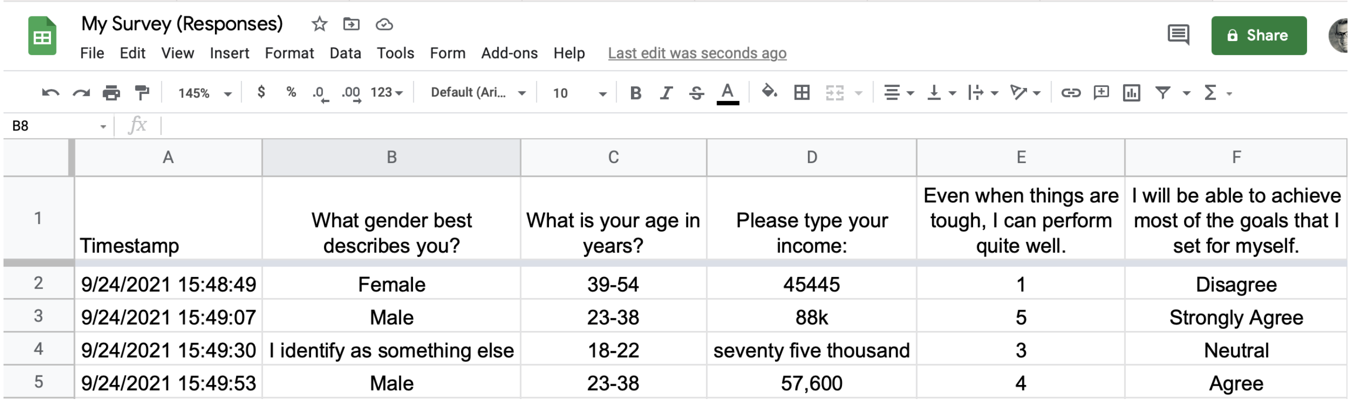
Variable	**New Name**	**LOM**	**Scoring**	**Expected Values**	**Value Recode**	**Min**	**Max**
What gender best describes you?	gender	nominal	as-is	"Female", "Male", "I identify as something else"	NA	NA	NA
What is your age in years?	age	ordinal	asi-is	18-24, 23-38, 39-54, 55-73, 74-91	NA	NA	NA
Please type in your income.	income	ratio	as-is	NA	"88k" = 88000 "seventy five thousand" = 75000 "57,600" = 57600	0	100000
Even when things are tough, I can perform quite well.	SE_perform	discrete	as-is	NA	NA	1	5
I will be able to achieve most of the goals that I set for myself.	SE_achieve	ordinal	as-is	"Strongly Agree", "Agree", "Neutral", "Disagree", "Strongly Disagree"	"Strongly Agree" = 5 "Agree" = 4 "Neutral" = 3 "Disagree" = 2 "Strongly Disagree" = 1	1	5
Compsite scale of self-efficacy items	SE_comp	continuous	mean of self-efficacy 1 and self-efficacy 2	NA	NA	1	5

Using the Code Book

For Those Using a Survey

If you are using a survey, the process of creating a code book can help you identify problems in your survey that create (or allow for) confusing or problematic data. Make sure to do a pilot round where you take your own survey along with a few friends or classmates. Then download the data, take a look at the format, and make sure the data are all at the expected level of measurement. Also check to see whether your dataset allows for odd responses.

For example, when people fill out the survey in the Appendix, it is like to result in the dataset below:


My Survey (Responses)
File Edit View Insert Format Data Tools Form Add-ons Help
Last edit was seconds ago
Share
145%
$ % .0 .00 123
Default (Ari...
10
B I S A
B8
fx

	A	B	C	D	E	F
1	Timestamp	What gender best describes you?	What is your age in years?	Please type your income:	Even when things are tough, I can perform quite well.	I will be able to achieve most of the goals that I set for myself.
2	9/24/2021 15:48:49	Female	39-54	45445	1	Disagree
3	9/24/2021 15:49:07	Male	23-38	88k	5	Strongly Agree
4	9/24/2021 15:49:30	I identify as something else	18-22	seventy five thousand	3	Neutral
5	9/24/2021 15:49:53	Male	23-38	57,600	4	Agree

We can see that our survey allows for some less-than-ideal situations. First, we will have to recode all of the values in the "I will be able to..." column based on the coding scheme in the code book. At this point, I might prefer to change that question to be in the same format as the "Even when things..." question, which produces data that is already numeric. Likewise, I might try to find another way of asking about income, such as using a drop-down menu that lets participants select the income closest to theirs (rounded to the nearest $5,000, for example). This would prevent me from having to go back and recode income data points with the equivalent number (like I've done in the example code book), which could be a lot of work.

Once you've done this, you can edit your code book to reflect any changes you made. After you're confident that your survey will produce data in the way you want, you can consider your code book finalized and release your survey.

For Everyone

Once you have your data in front of you, add an extra column that indicates how much of each variable is missing - both the number of missing responses for that variable, and the percentage.

Then use the code book to check for problems in your data. For example:

- Unlikely values. For example, someone reporting that they have 22 kids - perhaps that person meant to put 2.
- Impossible values. For example, a value of 10 on a variable that should be on a 1-5 scale.
- Redundant values - for example, the values `"Male"` and `"male"` will be read by statistical software as different responses because statistical software is case-sensitive. Also, statistical software picks on spaces too, so `"female"` and `" female"` will be read as different values. Any unique combination of characters will cause statistical software to treat them as unique values, so design your survey to prevent this if you can.
- Always make a note somewhere indicating how you dealt with any problems that you found. If you needed to recode some values, such as by converting `" male"` to `"male"`, record this change in the Value Recode column of the code book. If you decided to drop (delete, convert to missing data) a value that you knew was problematic, take a note somewhere that records which variable it was and what the subject ID was (i.e., which person).

- If you are editing directly in Excel, ALWAYS make a copy of your data first. That way, you retain the original untouched data as a comparison point. This copy is invaluable for if you make a mistake editing your data.

Step 3: Identify Your Statistics

The most important aspect of your code book is identifying what the levels of measurement will be. Once you have this, you can refer to this section to help you identify which statistics you will eventually get for your variables.

I have included inferential statistical tests here which allow you to make inferences back to the larger population that your data came from. In other words: gives you things like confidence intervals (margin of error) and *p*-values (which tell you how well your data fit with a null hypothesis), which help you understand the extent to which your data, and the relationships within it, accurately represent the larger population. They also can help you test hypotheses about your data, particularly with respect to relationships between variables. Inference might not be useful for e.g., country-level data or census data, where you already have population-level data. Inference is particularly useful for survey data, but check with your instructor first - you might not be required to use inference in your study.

For Individual Variables

When analyzing individual variables, here are some statistics you might consider. The list is not exhaustive.

Table 2: Statistics for Individual Variables Based on Level of Measurement

Level of Measurement	**Descriptive Statistics**	**Inferential Statistical Test**
Binary	proportions, percents, odds	z-binomial test
Nominal (> 2 unique values)	proportions, percents, odds	Chi-Squared goodness of fit
Discrete, Interval, or Ratio	mean, median, mode, standard deviation	One sample t-test

For Relationships Between Variables

Often a researcher will want to know how two or more variables relate to each other - as one variable increases, what happens to the other variable? Using the level of measurement, use this table to identify which statistics you might need, what statistics to compare, and which test of inference you might use (where applicable).

Keep these in mind as you use this table:

- If your independent variable is continuous and your dependent variable is binary or nominal, the guide suggests you use logistic regression. However, logistic regression is hard to understand, and it would be just as valid to flip your variables around and use the *t*-test instead. This will be much easier to unpack, and just as valid.
- If the level of measurement leads you to statistics you don't want, you can go back and choose a different database that has the same variable on a different level of measurement (if using a database), or edit your survey so that the variable comes out as the level of measurement that you want (if using a survey).

Table 3: Identifying Statistics

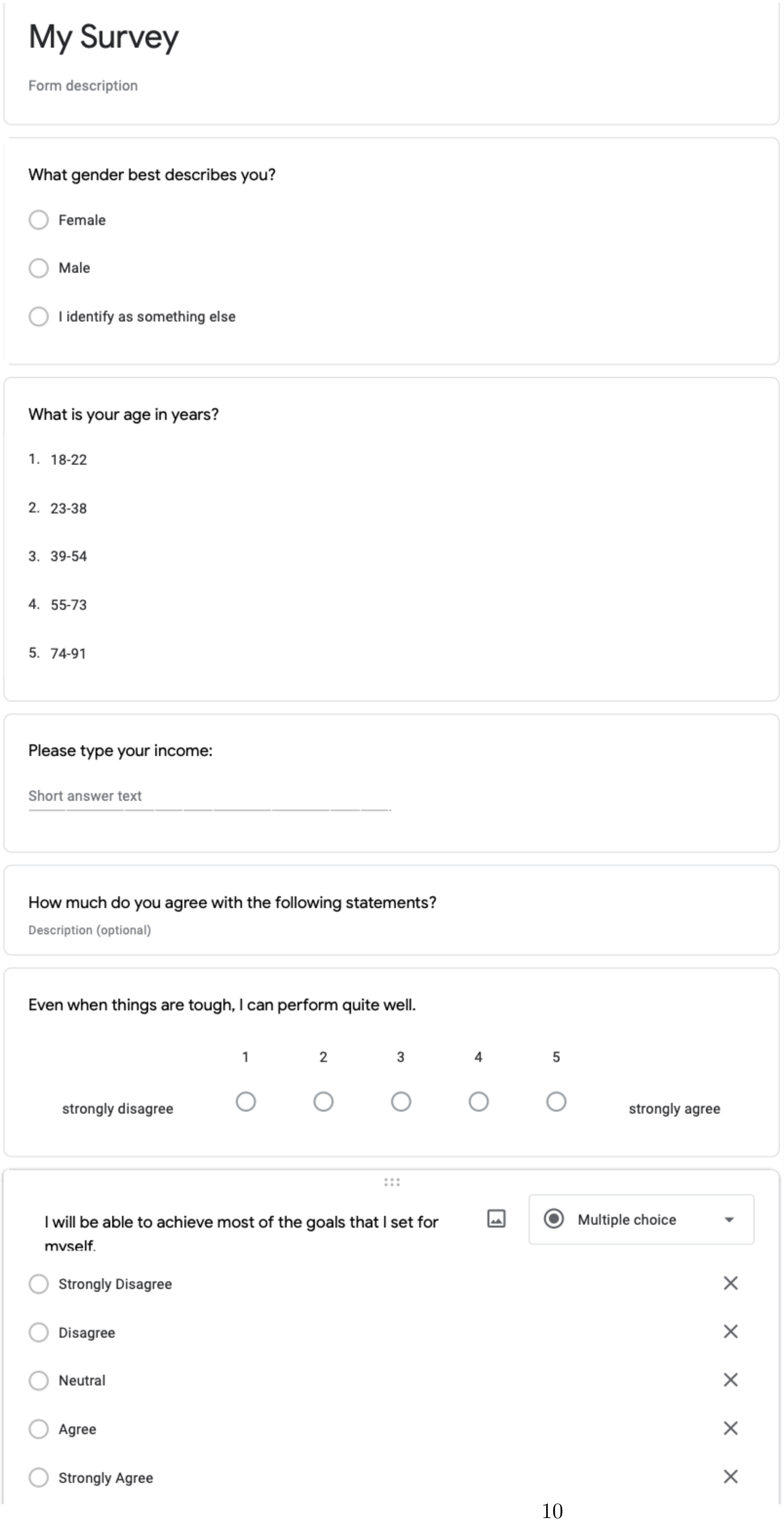
Number of IVs	IV LOM	DV LOM	Descriptive Statistics	Statistical Comparison	Inferential Statistical Test
1	Discrete, Continuous	Discrete, Continuous	correlation coefficient	r coefficient (r = zero indicates no correlation)	Pearson's r correlation
1	Binary	Discrete, Continuous	difference in means	mean difference (mean diff = 0 indicates no difference)	(In)dependent samples t-test
1	Discrete, Continuous	Binary	flip and use difference in means	(log)odds of one of two values (0 indicates the odds are the same for both values)	Logistic regression, or flip IV with DV for convenience and use t-test (much easier)
1	Binary, Nominal	Binary, Nominal	counts, percentages, or proportions of each combination of IV and DV. Odds and risk ratios, and aboslute risk difference	Cross-tab table, compare percentages to your hypothesis	Chi-squared test of independence
1	Nominal	Discrete, Continuous	differences in means across different combinations of IV	mean difference between mean of each nominal category (mean differences of 0 indicates no differences among groups)	One -way Analysis of Variance (ANOVA)
2+	Anything	Anything	Ask instructor or me :)		

Time to Collect Data and Analyze!

At this point, you are ready to dive into your data, or release your survey to your participants. If you are using a survey, make sure to delete the responses from your pilot round. Once you have data in hand, you can refer again to the Using the Code Book section of this guide. Next, you can jump straight into getting the statistics you need, because by the time you have your data, you should already know what you'll need to calculate.

Good luck!

Appendix



My Survey

What gender best describes you?

○ Female

○ Male

○ I identify as something else

What is your age in years?

1. 18-22
2. 23-38
3. 39-54
4. 55-73
5. 74-91

Please type your income:

How much do you agree with the following statements?

Even when things are tough, I can perform quite well.

	1	2	3	4	5	
strongly disagree	○	○	○	○	○	strongly agree

I will be able to achieve most of the goals that I set for myself.

○ Strongly Disagree

○ Disagree

○ Neutral

○ Agree

○ Strongly Agree