Guidelines: Project report.

Stat 152

Read “Writing Lab Reports” from the book *Stat Labs* by Deb Nolan and Terry Speed. I have uploaded it to bcourses. It gives general guidelines for writing statistical reports, and is very well written. You can read some of the chapters in their book as well to get an idea of how to write a report, though yours may not be as long - you are not writing a book.

The following is mostly taken from Prof. Purdom’s report writing guidelines from when she taught the course, and I think gives a nice checklist and description of how you should write up your report.

General comments:

- The report should be *readable* paragraphs and sentences. You should make tables and figures, but you should refer to them in the text. You should **not** be giving any R code in your text.
- You should not have an outline or list in your report. You can have section headers to distinguish sections. Please don’t submit the entire report in bullet points. There should be some text that makes it readable.
- You should also not give lists of things in your report; if you want to list something, such as the levels of a factor, you can instead write full sentences that give that information, or you can create a table with that information. Again, I have used lists to outline components you should include, but you should not; you should give them in paragraphs.
- Tables and figures should be given numbers so that you can clearly refer to them in the text.
- **Run a spell check and grammar check.**
- Everything you write should be completely comprehensible to someone else who has the background at the level at 152 but who has never seen the survey you are consider. You can assume that the reader has basic knowledge of the United States regarding what is a state, a county, the Census Bureau, and so forth (if your survey is not from the US, you should not even assume this). However, technical or structural information about the government, governmental agencies, or the census should not be assumed and should be explained to the best of your abilities.

Any guidelines about length are just to give an idea, and are not requirements or guidelines. They are rather to help you gauge how detailed you should be in different parts. There are no specific page guidelines.

The Structure

The report should consist of

1. Introduction – give a birds-eye overview of what you have done: brief description of the survey you are using, the questions you are looking at, and a summary of the results of your analysis, all in broad strokes. You should generally not have to refer to tables, plots or other details for this material. Non-statisticians should be able to understand the introduction. It should not be terribly long – somewhere between 2 paragraphs and a page.

2. Survey Design – describe the survey in detail. This should be based on documentation that comes with the data and is written by the organization that creates or distributes the data. Make sure you cite the information you are using, including a url. See Prof. Purdom's write-up:'How to write background' (it is on bcourses) to make sure that you are using your own words appropriately and not plagiarizing the online material.

   - **The basics of the survey:** Give broad information about the survey. This should include things like: who generates the survey, how often is it run, what kind of survey is it (telephone, in person, computer assisted, etc), how long has it been running, what is the purpose of the survey generally. You should not take that to be a comprehensive list, but just to give you an idea of the kinds of things that you should include.
   - **The design elements of the survey:** Describe how the survey was designed. Depending on what detail is given you should specify the stages of the survey, what stratification there was at each stage, what PSU there was at each stage, and how the probability of selection was determined at each stage. You will likely have to include definitions as to what the PSUs are. If 'households' were the PSU at some stage, you should specify what is meant by a household. You do not have to define obvious terms like 'state' or 'county', but more technical terms like 'census tracts' should be described.
   - **The public release data:** Describe what is changed in the public release data versus the raw data. This will probably include weight adjustments that you should describe. It may involve not releasing all observations. It may involve reporting stratum or psu variables that were not the true stratum or psus but that give similar estimates. You should also note any special instructions that are given in how to use the design correctly.
   - **Exploration of the design elements:** You should describe the design elements based on the public data you downloaded. This could include boxplots of the weights relative to the variables you are interested in, a summary of how the variables differ across strata, what is the range of the weights, are there large weights, etc. For this section of the report, it should reflect your exploration of the *design*, and not be the analysis of the data – that you will describe below in the Results section. You can use the information released about the design to aid you in thinking about how to explore the design. For example, if the survey description says that a particular ethnic group is over-represented in the sample, you should be able to see it in the weights for that ethnic group.

3. Methodology – this section should describe *your* methodology, i.e. what you did to the data and how you analyzed it. The idea is to describe important steps that you did so that if someone else wanted to copy your work, they would know what to do. This section should only describe what you did, *not* what was already done by the makers of the survey.

   Here are some things that should be included, but this is not exhaustive. You can address this issues in any order.

   - Information about how you got the data and any preprocessing you did. This should include the URL of where the public data can be obtained. It should describe what you had to do to get the data ready for analysis. For example, if you had to combine two datasets together to have all the necessary information, describe what you did. If you had to change variables from '99' to missing, or recode some variables, note that here.
   - Describe the specific variables you used in your analysis (i.e. their name in the code book)
   - Describe how you dealt with any non-response in your variables, if relevant (i.e. item non-response – not what was done by the survey designers before you got it). You should in this part also give summary information / plots about the extend of non response for the

variables you are using, and anything you can gather about the non-responders based on the variables they did respond to. And you should generally comment on what possible problems the non-response might cause.

- Describe how you created your survey object in the survey package and the options you used. This should include what is the variable name (from the code book) for the PSU, the strata and the weight that you give to `svydesign`. For this part ONLY, you should give the R command you used to make the survey object (as well as the written description) in the text.

4. Results – this should discuss the results of analyzing the data with respect to the question you were interested in to begin with, as well as any possible problems that you encountered. This section should restate the question(s), describe the analysis that you did to address that question, and what conclusions you would draw.

   I expect you to explore a variable of interest in depth and it's relationship to other variables. You are not expected to use all data in the dataset, but I would expect that there will be roughly 3-5 other variables (beyond the design variables like weights/PSU/Stratum) that you will draw upon to better understand your variable. I expect that you will evaluate relationships between variables with either regression or contingency table analysis (as appropriate) as well as plots and tables. I also expect you to analyze/summarize the variable(s) own its own via estimation, like the mean/median, and via plots like histograms or boxplots. As stated above, your analysis should address non-response.

   If you have any doubts about the appropriateness of any assumptions or the effect of something you do, you should state it up front and weigh the pros and cons. For example, you might do something to correct for non-response, but then you might want to weigh whether the assumptions make sense in your context, and if not whether that affects whether you trust the results of the analysis. Statistical reports should be nuanced, not black and white assertions of the truth.

   This text should form a logical narrative, and should not be just a series of plots or statistics. Your text does not need to follow the order in which you did the analysis but written in the best order to make a coherent text. You should give plots and tables to supplement your narrative, and again they should be a logical part of the narrative of your results and appropriate to the analysis. Results of an analysis (e.g. contingency tables, regression results) should be included as *formatted tables*, not just cut and paste of the output of R.

   Furthermore, your text should be readable without looking at the plots or the tables; the plots and tables *support* your statements in your text, but the text should be able to stand alone. You should not say "Figure 2 shows that the assumption of normality is satisfied." because I don't know what Figure 2 is a plot of nor what it is suppose to demonstrate. A figure shows a pattern, it does not, on its own, give any conclusions you should draw from those points. A better way to write this would be "As we can see in a qq-plot of the values (Figure 2), the data appear to follow a straight line, so we conclude that the assumption of normality is satisfied."

5. Discussion/Conclusion – you should draw a "big-picture" view of the results, as well as any further questions your results suggests. It should bring together all the details that you've brought up. This material is like the introduction only now you can be more specific because you assume the reader has read everything. So you can draw more specific conclusions, as well as refer to previous discussions from your earlier text. However, you should generally not have to refer to specific tables, plots or other fine details for this material. The discussion is generally longer than the introduction.

You should also include an appendix of R code, organized and commented so that it is possible to quickly find the code that goes with the discussion. For example,

```
###############
#Reading in and formatting data
###############
...
###############
#stripplot (Figure 1 in text)
###############
...
###############
#ANOVA (Table 2 & 3 in text)
###############
...
```

You should not give a dump of every piece of code that you entered, but the final code sufficient to replicate your analysis starting with the public data downloaded from the internet. More or less, the code should link up with the analysis in the text, though if there are some additional components that is fine.