STAT 332 Exercises: Chapter 4

Here are a few exercises in addition to those in the Course Notes. I'd recommend trying as many of these as possible without reference to any notes!

Some general advice:

- Sometimes it will be correct to state that there is insufficient information provided in a question to answer it (sometimes this will be deliberate). However, in such cases you should explain very carefully what information would be required. (In other words, if you just write 'insufficient information', you're unlikely to get any credit!)
- In exam situations some questions will explicitly state that you must show your working, in which case you *will* lose marks if you do not. However, it is generally a good idea to show your working even if it is not explicitly requested; you may still receive partial credit if your eventual final answer is incorrect.

1. Consider this study description and answer the questions that follow:

A researcher is interested in studying the proportion of news articles on a particular website published during 2016 that include fake information (i.e., inaccuracies or lies). To form a dataset, the researcher constructs a list of the site's 'most popular' 200 news articles from 2016, which is based on the number of comments left by site readers. This resulted in a list of 200 links to news articles, from which the researcher sampled 30 via simple random sampling without replacement. The researcher then read each of the articles and determined whether each story featured any fake information.

a) Briefly describe (i) the target population; (ii) the sample population; (iii) the sampling frame; (iv) the sampling units; (v) the population quantity of interest.

b) Briefly describe whether there are any potential sources of (i) frame error; (ii) sample error; (iii) measurement error. For each of (i), (ii), and (iii), if you do not think there is a potential source of that type of error, you should explain why.

2. Consider this study description and answer the questions that follow:

The Math Faculty at UW wishes to know the average salary of its graduates who find a job within three months of graduation, for students who graduated in the last five years. The faculty has a mailing list of former students (alumni) who have donated money to the university since they have graduated. They chose 100 students from this list using simple random sampling without replacement and sent each a questionnaire by email asking for the information.

a) Briefly describe (i) the target population; (ii) the sample population; (iii) the sampling frame; (iv) the sampling units; (v) the population quantity of interest.

b) Briefly describe whether there are any potential sources of (i) frame error; (ii) sample error; (iii) measurement error. For each of (i), (ii), and (iii), if you do not think there is a potential source of that type of error, you should explain why.

3. Consider this study description and answer the questions that follow:

The Globe Poll is a public opinion survey conducted on Canadians 18 or older in the 10 provinces through telephone interviews. In a recent survey, each selected respondent was asked to choose one of the three answers, **A** = Agree; **D** = Disagree; and **U** = Undecided, to the survey question "*...Unemployment is so bad that the federal government must increase government spendings on programs that create more jobs.*". The goal is to find the population distribution over the three possible opinions. Sample data were taken using separate lists of telephone numbers for each of the provinces.

a) Briefly describe (i) the target population; (ii) the sample population; (iii) the sampling frame; (iv) the sampling units; (v) the population quantity of interest.

b) Briefly describe whether there are any potential sources of (i) frame error; (ii) sample error; (iii) measurement error. For each of (i), (ii), and (iii), if you do not think there is a potential source of that type of error, you should explain why.

STAT 332 Exercises: Chapter 4: SOLUTIONS

Here are my solutions to the first exercise set. **Please** only review these after you've given 100% (or a close approximation) to the original problems! I provide my answers in blue, and provide additional comments in red. **If you think you spot any mistakes, please let me know ASAP! I've (literally) triple-checked these, but as you may find out it's easy for mistakes to creep in!**

Please note that in many cases there are more correct answers than I can provide (especially with questions about things like study error), so if you're unsure whether your answer would be considered correct please ask!

1. Consider this study description and answer the questions that follow:

A researcher is interested in studying the proportion of news articles on a particular website published during 2016 that include fake information (i.e., inaccuracies or lies). To form a dataset, the researcher constructs a list of the site's 'most popular' 200 news articles from 2016, which is based on the number of comments left by site readers. This resulted in a list of 200 links to news articles, from which the researcher sampled 30 via simple random sampling without replacement. The researcher then read each of the articles and determined whether each story featured any fake information.

a) Briefly describe (i) the target population; (ii) the sample population; (iii) the sampling frame; (iv) the sampling units; (v) the population quantity of interest.

- (i) The target population is all news articles on a particular website published during 2016.
- (ii) The sample population is the 200 most popular news articles on the site in 2016.
- (iii) The sampling frame is the list of URLs/links to the news articles.
- (iv) The sampling units are the individual news articles that are sampled.
- (v) The population quantity of interest is the proportion of news articles defined in (i) that contain fake information.

Things to look out for with questions such as these:

- Be as specific as you can with these definitions/interpretations. If you said the target population was just 'news articles', that would not be precise enough. Note also that it's fine to refer to previous definitions (as I've done in part (v)) - this can save you having to write the same thing multiple times.
- With questions like this it's totally fine to use bullet points for each definition - you don't need to write using a formal paragraphs/sentence structure, but you do need to make what you've written clear and unambiguous.

b) Briefly describe whether there are any potential sources of (i) frame error; (ii) sample error; (iii) measurement error. For each of (i), (ii), and (iii), if you do not think there is a potential source of that type of error, you should explain why.

- (i) A potential source of frame error is that the sampling frame is made up of articles with the most comments. It seems reasonable that articles with more comments may (for example) be more controversial, potentially as a result of the articles containing fake information.
- (ii) As the sample was taken via simple random sampling without replacement (SRSWOR), we do not have any reason to be concerned about sample error.
- (iii) A potential source of measurement error is that the researcher may make mistakes in determining whether an article contains fake information.

- As mentioned above, we might think of other potential sources of error. In exams/tests, you will only need to provide 1 example for each unless otherwise stated.
- In part (i) we might have concluded that we require more information to determine whether there is a risk of frame error, as we might not consider 'most comments' to be an indicator that our sample population differs from the target population on the attribute of interest. In this case, an alternative answer could have been "We would require more information to determine whether there is a potential source of frame error. The sample population is composed of the most popular articles on the site, and we would need to determine whether 'most popular' was related to whether an article was more or less likely to contain fake information." - note how I have been very precise about what we would need to know!
- Watch out for the difference between sample error and frame error - this seems like an obvious case of frame error, but the sampling approach was entirely fine!

2. Consider this study description and answer the questions that follow:

The Math Faculty at UW wishes to know the average salary of its graduates who find a job within three months of graduation, for students who graduated in the last five years. The faculty has a mailing list of former students (alumni) who have donated money to the university since they have graduated. They chose 100 students from this list using simple random sampling without replacement and sent each a questionnaire by email asking for the information.

a) Briefly describe (i) the target population; (ii) the sample population; (iii) the sampling frame; (iv) the sampling units; (v) the population quantity of interest.

- (i) The target population is UW graduates who graduated in the last five years and found a job within three months of graduation.
- (ii) The sample population is UW students who have donated money to the university since they have graduated.
- (iii) The sampling frame is the mailing list of the students in (ii).
- (iv) The sampling units are the 100 students sampled from the population in (ii).
- (v) The population quantity of interest is the average salary of individuals defined in (i).

b) Briefly describe whether there are any potential sources of (i) frame error; (ii) sample error; (iii) measurement error. For each of (i), (ii), and (iii), if you do not think there is a potential source of that type of error, you should explain why.

- (i) A potential source of frame error is that the sample population comprises students who have donated money to the university. This might indicate they are wealthier, or have higher incomes, which is the population quantity of interest.
- (ii) As the sample was taken via simple random sampling without replacement (SRSWOR), we do not have any reason to be concerned about sample error. [Alternative answer: we do not know whether there is non-response in this example. If there was non-response this might be a source of sample error.]
- (iii) We might encounter measurement error if individuals lie in response to the questionnaire.

3. Consider this study description and answer the questions that follow:

The Globe Poll is a public opinion survey conducted over Canadians 18 or older in the 10 provinces through telephone interviews. In a recent survey, each selected respondent was asked to choose one of the three answers, **A** = Agree; **D** = Disagree; and **U** = Undecided, to the survey question "*...Unemployment is so bad that the federal government must increase government spendings on programs that create more jobs.*". The goal is to find the population distribution over the three possible opinions. Sample data were taken using separate lists of telephone numbers for each of the provinces.

a) Briefly describe (i) the target population; (ii) the sample population; (iii) the sampling frame; (iv) the sampling units; (v) the population quantity of interest.

- (i) All Canadians aged 18 or older who live in one of the 10 provinces.
- (ii) The sample population is Canadians 18 or older in the 10 provinces who can be contacted with a Canadian telephone number.
- (iii) The sampling frame can be viewed as stratified lists of residential phone numbers for the 10 provinces if each phone number represents a unique person; it is a stratified list of clusters if each phone number represents a house (a group of people).
- (iv) The sampling units are the individuals sampled from the population in (ii).
- (v) The population proportions of who answers A, D or U among all Canadians aged 18 or older.

b) Briefly describe whether there are any potential sources of (i) frame error; (ii) sample error; (iii) measurement error. For each of (i), (ii), and (iii), if you do not think there is a potential source of that type of error, you should explain why.

- (i) A potential source of frame error is that individuals who are unemployed or low-income may not have a telephone or cellphone.
- (ii) As the sample was taken via stratified sampling with simple random sampling without replacement (SRSWOR), we do not have any reason to be concerned about sample error. [Alternative answer: we do not know whether there is non-response in this example. If there was non-response this might be a source of sample error.]
- (iii) The question is biased and seems to be asking two only one concepts the question. One possible change is "*Currently there is low unemployment. Should the federal government increase or decrease government spendings on programs that create more jobs*". Or instead have two questions: question 1 "*Currently, the unemployment rate is low, high or neither high or low.*" and question 2 " When there is low unemployment, should the federal government increase, decrease or maintain current government spendings on programs that create more jobs".