# STAT20 Homework #6

Seth Metcalf

## Table of Contents

Introduction ..... 1
Chapter 19 ..... 1
1. Ch 19 A8: ..... 1
2. Ch 19 A11: ..... 2
3. Ch 19 Rev 5: ..... 2
4. Ch 19 Rev 12: ..... 3
Chapter 20 ..... 3
5. Ch 20 A2: ..... 3
6. Ch 20 C2: ..... 4
7. Ch 20 Rev 6: ..... 4
Chapter 21 ..... 5
8. Ch 21 A7: ..... 5
9. Ch 21 A8 ..... 5
Chapter 23 ..... 6
10.Ch 23 A1: ..... 6

### Introduction

This is Homework #6, which contains questions from Chapters 19, 20, 21, & 23. *Due 19 March 2021.*

## Chapter 19

### 1. Ch 19 A8:

In one study, the Educational Testing Service needed a representative sample of college students. To draw the sample, they first divided up the population of all colleges and universities into relatively homogeneous groups. (One group consisted of all public universities with 25,000 or more students; another group consisted of all private four-year colleges with 1,000 or fewer students; and so on.) Then they used their judgment to choose one representative school from each group. That created a sample of schools. Each school

in the sample then picked a sample of students. Was this a good way to get a representative sample of students? Answer yes or no, and explain briefly.

### Explanation

Since the idea was to choose one preseentative school from each group, the judgement to choose a school never was said to be random and could contain bias. Furthermore, since it was never clarified that they would pick a random student, there is a chance that they pick the "best student" at the school. Overall, this is not a good way to get a representative of the school.

## 2. Ch 19 A11:

The San Francisco Examiner ran a story on September 11, 1988:

3 IN 10 BIOLOGY TEACHERS BACK BIBLICAL CREATIONISM Arlington, Texas. Thirty percent of high school biology teachers polled believe in biblical creation and 19 percent incorrectly think that humans and dinosaurs lived at the same time, according to a nationwide survey published Saturday.
"We're doing something very, very, very wrong in biology education," said Dana Dunn, one of two sociologists at the University of Texas, Arlington. Dunn and Raymond Eve sent questionnaires to 20,000 high school biology teachers selected at random from a list provided by the National Science Teachers Association and received 200 responses … "
The newspaper got it wrong. Dunn and Eve did not send out 20,000 questionnaires: they chose 400 teachers at random from the National Science Teachers Association list, sent questionnaires to these 400 people, and got 200 replies. Why do these corrections matter?

### Explanation

These corrections matter because it changes the validity of the study. With 200 responses out of 20,000, non-response bias is a large issue and could be used in an arguement that the study was flawed. With 200 responses out of 400 questionnaires sent out, then this data becomes more reliable to believe in. However, this is still a small sample size comparatively to the size of the population.

## 3. Ch 19 Rev 5:

(Hypothetical.) A survey is carried out by the finance department to deter-mine the distribution of household size in a certain city. They draw a simple random sample of 1,000 households. After several visits, the interviewers find people at home in only 653 of the sample households. Rather than face such a high non-response rate, the department draws a second batch of households, and uses the first 347 completed interviews in the second batch to bring the sample up to its planned strength of 1,000 households. The department counts 3,087 people in these 1,000 households, and estimates the average household size in the city to be about 3.1 persons. Is this estimate likely to be too low, too high, or about right? Why?

Why?

Since we know that the survey suffered from a non-response bias, as mentioned in the question, we also know that this estimate is probably not accurate enough to be a representative of the entire population. The idea is that if there are more people in a household, there is a higher chance for someone to be home during the day/when they are visiting the homes, meaning that the estimate is probably too high.

### 4. Ch 19 Rev 12:

The San Francisco Chronicle reported on a survey of top high-school students in the U.S. According to the survey, Cheating is pervasive. Nearly 80 percent admitted some dishonesty, such as copying someone's homework or cheating on an exam. The survey was sent last spring to 5,000 of the nearly 700,000 high achievers included in the 1993 edition of Who's Who Among American High School Students. The results were based on the 1,957 completed surveys that were returned. "The survey does not pretend to be representative of all teenagers," said Who's Who spokesman Andrew Weinstein. "Students are listed in Who's Who if they are nominated by their teachers or guidance counselors. Ninety-eight percent of them go on to college."

Why isn't the survey "representative of all teenagers"?

Since this survey was only sent to those who were considered "high achievers" then this is a clear example of selection bias. Therefore this would only be a representative of those that were "high achievers" not all highschoolers.

Is the survey representative "of the nearly 700,000 high achievers included in the 1993 edition of Who's Who Among American High School Students"? Answer yes or no, and explain briefly.

No, since only 1900 or so responded out of the 5000 that were sent, it can imply a non-response bias and should not be trusted to be a representative of a population. This non-response bias could be based off a volunteer bias because it can be assumed that those with a stronger opinion are more likely to respond to the survey.

## Chapter 20

### 5. Ch 20 A2:

A university has 25,000 students, of whom 10,000 are older than 25. The registrar draws a simple random sample of 400 students.

a) Find the expected value and SE for the number of students in the sample who are older than 25.

EV = #draws * avg avg = 10000/25000 = 2/5 draws = 400 EV = 400 * 2/5 = 160 SE = sqrt(#draws) * SD

```
uniStudents = rep(c(0:1),c(10000,15000))
sd(uniStudents)

## [1] 0.4899077
```

SD = 0.4899077 SE = sqrt(400) * 0.4899077 = 9.798154

b) Find the expected value and SE for the percentage of students in the sample who are older than 25.

EV = 160 EV% = EV/#draws * 100% EV% = 160/400 * 100 = 40% SE = 9.798154 SE% = SE/#draws * 100% SE% = 9.798154 * 100% = about 2.45%

c) The percentage of students in the sample who are older than 25 will be around ___ , give or take ___ or so.

The percentage of students in the sample who are older than 25 will be around 40%, give or take 2.45% or so.

**6. Ch 20 C2:**

You have hired a polling organization to take a simple random sample from a box of 100,000 tickets, and estimate the percentage of 1's in the box. Unknown to them, the box contains 50% 0's and 50% 1's. How far off should you expect them to be:

a) If they draw 2,500 tickets?

EV = 50000/100000 = 0.5 SD = (1-0) * sqrt(50000/100000 * 50000/100000) = 0.5 SE = sqrt(2500) * 0.5 = 25 SE% = 25/2500 * 100% * cf cf = sqrt((100000-2500/(100000-1))) SE% = 0.99%

b) if they draw 25,000 tickets?

SE = sqrt(25000) * 0.5 = 79 SE% = 79/25000 * 100% * cf cf = sqrt((100000-25000/(100000-1))) SE% = 0.27%

c) if they draw 100,000 tickets?

SE = sqrt(100000) * 0.5 = 158 SE% = 158/100000 * 100% SE% = 0.16%

**7. Ch 20 Rev 6:**

The Census Bureau is planning to take a sample amounting to 0.1 % of the population in each state in order to estimate the percentage of the population in that state earning over $100,000 a year. Other things being equal (in particular, the rough percentage of the population earning over $100,000 a year in each state), which of the following is most correct, and why? i) The accuracy to be expected in California (population 40 million) is about the same as the accuracy to be expected in Nevada (population 3 million). ii) The accuracy to be expected in California is quite a bit higher than in Nevada. iii) The accuracy to be expected in California is quite a bit lower than in Nevada.

Answer

ii) Is the most correct. Since the absolute size of the sample determines the accuracy, since the population of Nevada is much lower than that of California, the accuracy of California is expected to be a bit higher than in Nevada.

# Chapter 21

### 8. Ch 21 A7:

Suppose there is a box of 100,000 tickets, each marked 0 or 1. Suppose that in fact, 20% of the tickets in the box are 1's. Calculate the standard error for the percentage of 1 's in 400 draws from the box.

SE

draws = 400 SD = (1-0) * sqrt(0.2 * 0.8) = 0.4 SE = sqrt(400) * 0.4 = 8 SE% = 8/400 * 100% = 0.02 * 100% SE% = 2%

### 9. Ch 21 A8

Three different people take simple random samples of size 400 from the box in exercise 7, without knowing its contents. The number of 1's in the first sample is 72. In the second, it is 84. In the third, it is 98. Each person estimates the SE by the bootstrap method.

a) The first person estimates the percentage of 1 's in the box as ___ , and figures this estimate is likely to be off by ___ or so.

P = 72/400 = 0.18 SE% = sqrt((P * (1-P))/n) * 100% = sqrt((0.18 * 0.82)/400) * 100% = 0.0192 * 100% = 1.92% The first person estimates the percentage of 1 's in the box as 18% , and figures this estimate is likely to be off by 1.92% or so.

b) The second person estimates the percentage of 1 's in the box as ___ , and figures this estimate is likely to be off by ___ or so.

P = 84/400 = 0.21 SE% = sqrt((P * (1-P))/n) * 100% = sqrt((0.21 * 0.79)/400) * 100% = 0.0204 * 100% = 2.04% The first person estimates the percentage of 1 's in the box as 21% , and figures this estimate is likely to be off by 2.04% or so.

c) The third person estimates the percentage of 1 's in the box as ___ , and figures this estimate is likely to be off by ___ or so.

P = 98/400 = 0.245 SE% = sqrt((P * (1-P))/n) * 100% = sqrt((0.245 * 0.755)/400) * 100% = 0.0215 * 100% = 2.15% The first person estimates the percentage of 1 's in the box as 24.5% , and figures this estimate is likely to be off by 2.15% or so.

# Chapter 23

### 10.Ch 23 A1:

One hundred draws are made at random with replacement from a box.

a) If the sum of the draws is 7,611, what is their average?

Sum = 7,611 draws = 100 avg = sum/draws = 7611/100 = 76.11

b) If the average of the draws is 73.94, what is their sum?

avg = 73.94 draws = 100 sum = avg * draws = 7394