Statistics Package Exercise: Exploring Simple Random Samples

Learning Objective: Identify the sampling method used in a study and discuss its implications and potential limitations.

The purpose of this activity is to show you how a simple random sample produces a sample that is not subject to any bias and is thus representative of the population from which it was selected. Also, we'll see how a nonrandom sample can produce some sources of bias.

Background

Consider the population of all students at a large university taking introductory statistics courses (1,129 students taking statistics for business, social sciences, or natural sciences).

Suppose we are interested in the values of four specific variables for this population: handedness (right-handed or left-handed), sex, SAT Verbal score, and age. If we were unable to determine the values of those variables for the entire population, we may be able to take a random sample from that population, and use the sample summaries as estimates for population summaries. Would the random sample provide unbiased estimates for the population values?

Next, what if instead of taking a random sample, we sampled the 192 students who happen to be enrolled in the business statistics course? First we will intuit, then check, if they would be a representative sample with respect to each of the four variables: handedness, sex, SAT Verbal score, and age. It may be helpful for you to know that, at this university, all students have comparable options in terms of when they take introductory statistics. You should also know that women, on the whole, tend to do somewhat better than men on the verbal portion of the SAT, and that business is a major that tends to interest males more than females.

To summarize the goals for this activity, we will:

A. Verify that the distributions of the variables handedness, sex, SAT Verbal score, and age are roughly the same for the random sample as they are for the population.

B. Intuit whether the distributions of each of the four variables in the (nonrandom) sample of business students would be roughly the same as those for the population, or whether there is a reason to expect any of them to be biased.

C. Check our intuition by comparing the distributions of each of the four variables for the sample of business students with those for the population, and determine whether they are roughly the same or if the sample values for any of the variables appear to be biased.

Our dataset contains data on the entire population of 1,129 students, which includes students taking introductory statistics who are majoring in the natural and social sciences, as well as for business majors.

- R
- StatCrunch
- TI Calculator
- Minitab
- Excel

R Instructions

To open R with the dataset preloaded, right-click here and choose "Save Target As" to download the file to your computer. Then find the downloaded file and double-click it to open it in R.

The data have been loaded into the data frame "population". Enter the command

```
population
```

to see the data. The variables in the data frame are

```
Course
```

,

```
Handed
```

,

```
Sex
```

,

```
Verbal
```

,

```
Age
```

.

The variables are defined as follows:

- **Course:** natural science, social science, or business
- **Handed:**righthanded or lefthanded
- **Sex:** female or male
- **Verbal:** SAT Verbal scores up to 800
- **Age:** in years

First, we will take a simple random sample of the data. For the sake of consistency, we will make the random sample the same size (192) as the nonrandom sample of business statistics students that will be examined later.

To do this in R, copy the following command:

```
random_sample =
population[sample(length(population$Course),192),];random_sa
mple
```

PART A. Now we will determine whether the four variables' behavior for the random sample is comparable to their behavior for the population.

Step 1. To compare the proportion of right-handed students in the sample to those in the population, create two pie charts.

- R StatCrunch TI Calculator Minitab Excel

R Instructions
If using the R stat package create two side-by-side bar graphs one for handedness in the population and one for handedness in the random sample rather than two pie charts.

To create bar charts in R, copy the commands below:

a.

```
random_sample_percent =
100*prop.table(table(random_sample$Handed));random_sample_pe
rcent
```

b.

```
pop_percent =
100*prop.table(table(population$Handed));pop_percent
```

c.

```
barplot(rbind(pop_percent,random_sample_percent), beside=T,
col=c(0,1),legend.text=T,xlab="Handedness",ylab="Percent in
Group",args.legend=list(x="topleft"))
```

To create pie charts do this in R, copy the entire command below:

```
random_sample_percent =
100*summary(random_sample$Handed)/length(random_sample$Hande
d);random_sample_percent; pop_percent =
100*summary(population$Handed)/length(population
$Handed);pop_percent; par(mfrow=c(1,2));
pie(pop_percent,labels=paste(c("left=","right="),round(pop_p
ercent,0),"%"),main="Population");
pie(random_sample_percent,labels=paste(c("left=","right="),r
ound(random_sample_percent,0),"%"),main="Random Sample")
```

Note: When using R and getting it to display 2 graphs at once requires the "par" command, which tells R to display the next 2 graphs together.

Consider the distributions to be comparable if the sample proportion comes within about 5% of population proportion. Does it? (Use the text box in the first Learn By Doing exercise below to record your answers.)

Consider the distributions to be comparable if the sample proportion comes within about 5% of the population proportion. Does it? (Use the text box in the first Learn By Doing exercise below to record your answer.)

Step 2. To compare the proportion of female students in the sample to the proportion in the population, create two pie charts, one for sex in the population and one for sex in the random sample.

- R
- StatCrunch
- TI Calculator
- Minitab
- Excel

R Instructions

Use the commands above, but replace "

```
$Handed
```

" with "

```
$Sex
```

" and "

```
left
```

" with "

```
female
```

" and "

```
right
```

" with "

```
male
```

".

Below is the code with the changes we suggested above in case you want to check your command.

```
random_sample_percent =
100*summary(random_sample$Sex)/length(random_sample$Sex);ran
dom_sample_percent; pop_percent =
100*summary(population$Sex)/length(population
$Sex);pop_percent; par(mfrow=c(1,2));
pie(pop_percent,labels=paste(c("female=","male="),round(pop_
percent,0),"%"),main="Population");
```

```
pie(random_sample_percent,labels=paste(c("female=","male="),
round(random_sample_percent,0),"%"),main="Random Sample")
```

Consider the distributions to be comparable if the sample proportion comes within about 5% of population proportion. Does it? (Use the text box in the first Learn By Doing exercise below to record your answers.)

Step 3. Create 2 descriptive statistics summary tables—one for SAT Verbal score in the population and one for SAT Verbal score in the sample

- R StatCrunch TI Calculator Minitab Excel

R Instructions
To do this in R, copy the following commands:

```
summary(population$Verbal)
```

```
summary(random_sample$Verbal)
```

Since SAT scores tend to follow a normal (symmetric) distribution, you can focus on means to make a comparison. Consider the distributions to be comparable if the sample mean SAT Verbal score comes within about 10 points of the population mean. Does it? (Use the text box in the first Learn By Doing exercise below to record your answers.)

Step 4. Create 2 more descriptive statistics summary tables—one for age in the population and one for age in the sample.

- R
- StatCrunch
- TI Calculator
- Minitab
- Excel

R Instructions
To do this in R, just substitute "

```
$Age
```

" for "

```
$Verbal
```

" in the two commands above.

Since Age tends to follow a right-skewed distribution, you should focus on medians to make a comparison. Consider the distributions to be comparable if the sample median age comes within about .5 years of the population median. Does it?

Learn By Doing (1/1 point)

Summarize your findings from Part A, exercises 1-4 above.

Your Answer:

No bias in all of them!

Our Answer:
Different random samples will yield different results, but the summaries should be roughly the same for population and sample. Here is what we got: 1. The percentage of right-handed students is 87% in the population and it was 88% in our random sample; close enough to assert that there is no bias. 2. The percentage of females is 62% in the population and 65% in our random sample; there is no clear evidence of bias. 3. The SAT Verbal scores averaged 589 for the population and 581 for our random sample; again, there is no clear evidence of bias. 4. Finally, median age for the population is 19.7 and 19.7 for our random sample; there is no evidence of any bias at all.

Resubmit Reset

Learn By Doing (1/1 point)

PART B. Intuit whether the distributions of each of the four variables in the (nonrandom) sample of business students would be roughly the same as those for the population, or whether there is a reason to expect any of them to be biased. For each of the variables—Handed, Sex, Verbal, and Age—decide whether or not you believe the sample of business statistics students should be fairly representative of the larger population of all students in introductory statistics courses.

Your Answer:

They are both representative, though the minimums and maximums will definitely be subject to change because there will be smaller range in the sample.

Edit: okay i was supposed to think about them myself oops.
I would've (mistakenly) thought that all of them were representative; I didn't notice the "sample of business statistics buttons"

Our Answer:
• Handed—should be representative. Business students shouldn't be any different from other students with respect to this variable. • Sex—could easily be biased. We expect fewer women in business than in the social or natural sciences. • Verbal—could easily be biased. Business students' SAT Verbal scores may tend to be lower, since there are likely to be fewer women, and women tend to do better on the verbal portion of the SAT (as the background to this problem suggests). • Age—should be representative. Regardless of major, students may take introductory statistics at about the same point in their college career (as suggested by the background).

Resubmit Reset

PART C. How representative is the (nonrandom) sample of students in the business statistics course, in actuality? In order to answer this question, we will need to extract this group from the population.

- R StatCrunch TI Calculator Minitab Excel

R Instructions
To do this with R, execute the following command in R:

```
business =
population[population$Course=="Business",];business
```

Next, we explore whether the four variables' behavior for the (nonrandom) sample of business statistics students is comparable to their behavior for the population:

- R StatCrunch TI Calculator Minitab Excel

R Instructions

To do this in R, just use the commands above, but replace the variable "

```
random_sample
```

" with "

```
business
```

". In the pie chart commands, replace "

```
Random Sample
```

" with "

```
Business Students
```

".

1. To compare the proportion of right-handed students in the sample to those in the population, create 2 pie charts, one for handedness in the population and one for handedness in the sample of business statistics students.

Consider the distributions to be comparable if the sample proportion comes within about 5% of the population proportion. Does it? (Use the text box in the Learn By Doing exercise below to record your answers.)

2. To compare the proportion of female students in the sample to the proportion in the population, create 2 pie charts (using the instructions above), one for sex in the population and one for sex in the sample of business statistics students.

Consider the distributions to be comparable if the sample proportion comes within about 5% of the population proportion. Does it? (Use the text box in the Learn By Doing exercise below to record your answers.)

3. Create 2 tables of descriptive statistics (using the instructions above)—one for SAT Verbal score in the population and one for SAT Verbal score in the sample of business statistics students.

Since SAT scores tend to follow a normal (symmetric) distribution, you can focus on means to make a comparison. Consider the distributions to be comparable if the sample mean SAT Verbal score comes within about 10 points of the population mean. Does it? (Use the text box in the Learn By Doing exercise below to record your answers.)

4. Create 2 tables of descriptive statistics (using the instructions above)—one for age in the population and one for age in the sample of business statistics students.

Since age tends to follow a right-skewed distribution, you should focus on medians to make a comparison. Consider the distributions to be comparable if the sample median age comes within about .5 years of the population median. Does it? (Use the text box in the Learn By Doing exercise below to record your answers.)

Learn By Doing (1/1 point)

Summarize your findings from Part C, exercises 1-4 above.

Your Answer:

1. It is! 4% difference.
2. It does not. We can see a 10% instead of 5% difference.
3. They are not within 10 points.
4. They are!

Our Answer:

1. The percentage of right-handed students is 87% in the population and it was 91% in the sample of business students; close enough to assert that there is no clear bias. 2. The percentage of females is 62% in the population and 52% in the sample of business students; apparently biased, in that females are under-represented in business statistics. 3. The SAT Verbal score averaged 589 for the population and 575 for the business students; the sample does seem to be biased, due to lower SAT Verbal scores for business students. 4. Finally, the median age for the population is 19.7 and our sample of business students had a median age of 19.5 years; apparently the business students may be considered an unbiased sample when it comes to age.

Resubmit Reset