Statistics Package Exercise: Creating Similar Treatment Groups

Learning Objective: Explain how the study design impacts the types of conclusions that can be drawn.

The purpose of this activity is to explore the effectiveness of randomization in creating similar treatment groups, in the sense that it balances the groups with respect to other variables that we didn't control for.

Background

A local internet service provider (ISP) created two new versions of its software, with alternative ways of implementing a new feature. To find the product that would lead to the highest satisfaction among customers, the ISP conducted an experiment comparing users' preferences for the two new versions versus the existing software.

The ISP ideally wants to find out which of the three software products causes the highest user satisfaction. It has identified three major potential lurking variables that might affect user satisfaction —gender, age, and hours per week of computer use.

In this activity, we will use adults in a hypothetical city as the population of interest to the ISP. We will:

- create a simple random sample as the basis for the experimental study of the population,
- use randomization to assign individuals to treatment groups, and
- verify that randomization prevented the three treatment groups from being different with respect to the most obvious lurking variables.

R | StatCrunch | TI Calculator | Minitab | Excel

- 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

```
computers
```

. The data frame has three variables:

```
age
```

,

```
gender
```

,

```
comp
```

. This dataset has more than 20,000 entries, so we will not display it. To verify that the data have been loaded into the variable computers, copy the following command to R:

```
summary(computers)
```

You can see information about the

```
age
```

of the population and the number of hours

```
(comp)
```

per week that the subjects use computers. You also see that there are more than 10,000 men and more than 10,000 women in the population

```
(gender)
```

.

Our dataset contains the values of the three possible lurking variables:

- **age:** in years
- **gender:** female or male
- **comp:** hours per week of computer use

- R StatCrunch TI Calculator Minitab Excel

R Instructions

The company must rely upon sampling to study its customers' preferences, since the entire population cannot be assigned to treatments. Therefore, we will first choose a simple random sample (SRS) of 450 people for the subjects in the study.

To choose the sample, copy the following command to R:

```
random_sample = computers[sample(length(computers$age),
450),]
```

Again, we do not wish to view all 450 entries in the random sample. Instead, let's look at a summary of the sample by copying the following command:

```
summary(random_sample)
```

By looking at the numbers of males and females in the sample, we see that the sample indeed has 450 entries.

Now we will randomly assign our SRS of 450 subjects to treatment groups, one for each of the three versions of the ISP's software. Let's denote the versions "1," "2," and "3," and create a categorical variable to identify the treatment for each subject.

To use R to randomly assign the 450 subjects to one of three treatments, copy the following commands (note these are two separate command lines):

```
group = sample(1:3,450,replace=T)
```

```
random_sample = cbind(random_sample,group)
```

Note: Using R—The variable

```
group
```

is a list of 450 randomly chosen values from 1, 2, and 3. You can see them by entering the variable name

```
group
```

into R. These commands add a fourth column to the variable

```
random_sample
```

. The four columns are

```
age
```

,

```
gender
```

,

```
comp
```

and

```
group
```

.

We are finally reaching the goal of this activity. We will now examine whether the randomization was successful in making our three treatment groups similar with respect to the variables age, gender, and comp. In other words, we will now examine whether the distributions of these variables in the three groups are similar or not.

To compare the distribution of age among the three treatment groups, we'll create side-by-side boxplots of age by treatment.

- R | StatCrunch | TI Calculator | Minitab | Excel

R Instructions
Copy the following command to R:

```
boxplot(random_sample$age~random_sample$group, xlab="Group",
ylab="Age (years)")
```

To compare the distribution of gender among the three treatment groups, we'll look at a two-way table of conditional percents:

- R | StatCrunch | TI Calculator | Minitab | Excel

R Instructions
Copy the following commands to R (note these are two separate command lines):

```
two_way_table =
table(random_sample$group,random_sample$gender)
```

```
prop.table(two_way_table,1)*100
```

To compare the distribution of comp (the hours per week of computer use) among the three treatment groups, we'll create side by side boxplots of comp by treatment. Follow the instructions above, making the obvious necessary changes.

Learn By Doing (1/1 point)

Comment on the displays you created. In particular, are the distributions of age, gender, and comp in the three treatment groups similar?

Your Answer:

Mostly the same though gender's is a bit more different in group 3 on mine, more males. Group 1 had more females. Just a bit, though.

That being said, the IQR of of both boxplots can be seen to be very similar to each other.

Our Answer:

Everyone will get slightly different displays here, but they should all "look" about the same. Based upon the side-by-side boxplots, the distribution of ages and hours per week of computer use appears the same in each of the three treatment groups. Similarly, the table of conditional percents suggests that the distribution of the genders is about the same in all three treatment groups.

Resubmit Reset

Learn By Doing (1/1 point)

Based on your answer to the question above, does the randomization allow us to study the differences in user preferences between the three browsers, while eliminating the possible effects of the lurking variables age, gender, and hours per week of computer use? Comment below:

Your Answer:

It should be able to, since there would be no inherent biases in sampling.

Our Answer:

Our results suggest that the distributions of age, gender, and hours per week of computer use among the three treatments are about the same; therefore, the randomization was successful in balancing these three potential lurking variables among the three treatment groups. We can be fairly sure that any difference between the treatment groups that we find on the user tests of the software will be due to differences in the three software versions, rather than the lurking variables of age, gender, and hours per week of computer use.

Resubmit Reset