IPS9 in R: Producing data (Chapter 3)

Shukry Zablah (szablah20@amherst.edu) and Nicholas Horton (nhorton@amherst.edu)

July 18, 2018

Introduction and background

These documents are intended to help describe how to undertake analyses introduced as examples in the Ninth Edition of *Introduction to the Practice of Statistics* (2017) by Moore, McCabe, and Craig.

More information about the book can be found here. The data used in these documents can be found under Data Sets in the Student Site. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at https://nhorton.people.amherst.edu/ips9/.

This work leverages initiatives undertaken by Project MOSAIC (http://www.mosaic-web.org), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the `mosaic` package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignettes (http://cran.r-project.org/web/packages/mosaic). A paper describing the mosaic approach was published in the *R Journal*: https://journal.r-project.org/archive/2017/RJ-2017-024.

Chapter 3: Producing data

This file replicates the analyses from Chapter 3: Producing data.

First, load the packages that will be needed for this document:

```
library(mosaic)
library(readr)
```

Section 3.1: Sources of Data

Section 3.2: Design of experiment

See Example 3.12 on page 178.

We are looking to setup a dataset with randomly assigned a "Control" or "Treatment" group to our dataset.

```
set.seed(1) #guarantees the same random results each time R runs
Randomized <- data.frame(ID = c(1:10), randNum = runif(10))
Randomized
```

```
##    ID    randNum
## 1   1 0.26550866
## 2   2 0.37212390
## 3   3 0.57285336
## 4   4 0.90820779
## 5   5 0.20168193
## 6   6 0.89838968
## 7   7 0.94467527
## 8   8 0.66079779
## 9   9 0.62911404
## 10 10 0.06178627
```

We created a data frame in which the first column corresponds to unique IDs and the second column holds the random numbers that will determine which group will be assigned to which ID.

Our plan is to order the IDs based on their random number and assign half of the observations to the Control group and the other half to the Treatment group. We start by arranging the observations.

```
Randomized <- Randomized %>%
  arrange(randNum)
Randomized
```

```
##    ID    randNum
## 1  10 0.06178627
## 2   5 0.20168193
## 3   1 0.26550866
## 4   2 0.37212390
## 5   3 0.57285336
## 6   9 0.62911404
## 7   8 0.66079779
## 8   6 0.89838968
## 9   4 0.90820779
## 10  7 0.94467527
```

And finally we assign each observation to a group.

```
Randomized <- Randomized %>%
  mutate(Group = c(rep("Treatment", 5), rep("Control", 5)))
Randomized
```

```
##    ID    randNum     Group
## 1  10 0.06178627 Treatment
## 2   5 0.20168193 Treatment
## 3   1 0.26550866 Treatment
## 4   2 0.37212390 Treatment
## 5   3 0.57285336 Treatment
## 6   9 0.62911404   Control
## 7   8 0.66079779   Control
## 8   6 0.89838968   Control
## 9   4 0.90820779   Control
## 10  7 0.94467527   Control
```

Now your observations are randomly assigned to a group! Here we chose to assign the first five ID to the Treatment group and the rest to the Control group, but there are other other ways to do it. To finish you can clean up the dataset by selecting only the ID and the Group columns and order it by ID.

```
Randomized %>%
  select(ID, Group) %>%
  arrange(ID)
```

```
##    ID     Group
## 1   1 Treatment
## 2   2 Treatment
## 3   3 Treatment
## 4   4   Control
## 5   5 Treatment
## 6   6   Control
## 7   7   Control
## 8   8   Control
## 9   9   Control
```

```
## 10 10 Treatment
```

You can do this for many Groups and for as many observations as you want.

Section 3.3: Sampling Design

We want to get a simple random sample (SRS) of 25 countries out of 189.

```
Countries <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter03/EG03-23TTS.csv")
```

In R you just have to use the `sample()` function.

```
Sample <- sample(Countries, size = 25)
Sample
```

```
## # A tibble: 25 x 4
##    CountryName CountryCode  Time orig.id
##    <chr>       <chr>       <int> <chr>
##  1 Costa Rica  CRI            24 39
##  2 Cameroon    CMR            15 34
##  3 Nepal       NPL            17 129
##  4 Croatia     HRV             8 72
##  5 Romania     ROU             9 143
##  6 Korea, Rep. KOR             6 92
##  7 Pakistan    PAK            21 132
##  8 Vietnam     VNM            34 181
##  9 Guyana      GUY            20 69
## 10 Portugal    PRT             3 140
## # ... with 15 more rows
```

Section 3.4: Ethics