

IS5 in R: Sample Surveys (Chapter 10)

Nicholas Horton (nhorton@amherst.edu)

2025-01-15

Table of contents

Introduction and background	1
Chapter 10: Sample Surveys	2
Section 10.1: The Three Big Ideas of Sampling	2
Section 10.2: Populations and Parameters	2
Section 10.3: Simple Random Samples	2
Section 10.4: Other Sampling Designs	5
Section 10.5: From the Population to the Sample: You Can't Always Get What You Want	5
Section 10.6: The Valid Survey	5
Section 10.7: Common Sampling Mistakes, or How to Sample Badly	5

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fifth Edition of *Intro Stats* (2018) by De Veaux, Velleman, and Bock. This file as well as the associated Quarto reproducible analysis source file used to create it can be found at <http://nhorton.people.amherst.edu/is5>.

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 (<https://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>.

We begin by loading packages that will be required for our analyses.

```
library(mosaic)
library(tidyverse)
```

Chapter 10: Sample Surveys

Section 10.1: The Three Big Ideas of Sampling

Section 10.2: Populations and Parameters

Section 10.3: Simple Random Samples

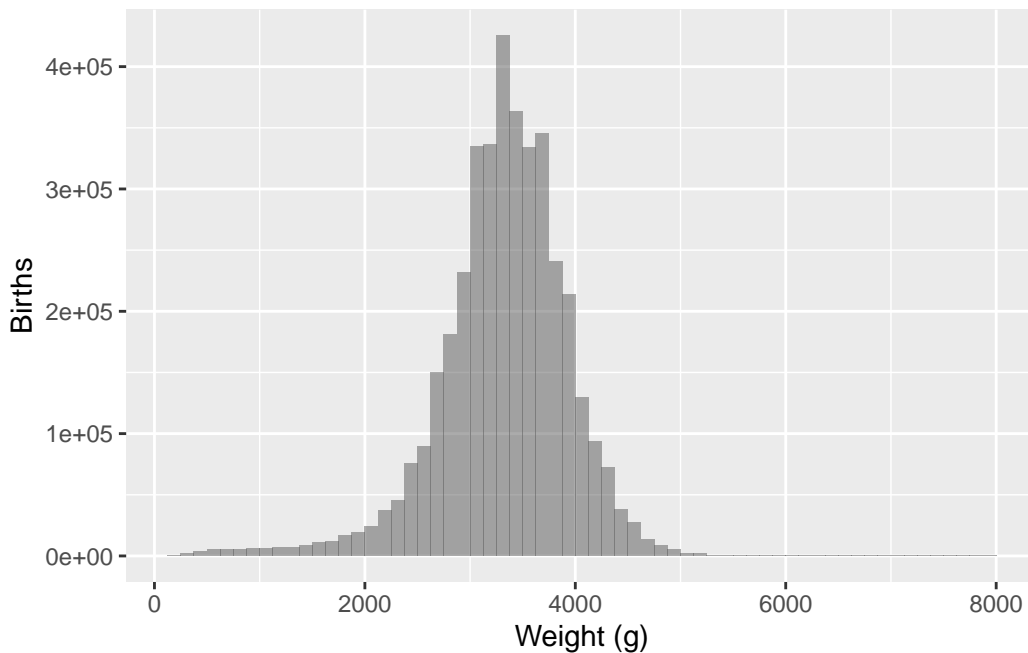
Random Matters

We begin by loading in the data on birthweights.

```
Births <- read_csv("http://nhorton.people.amherst.edu/is5/data/AllBirths1998.csv")
```

By default, `read_csv()` prints the variable names. These messages were suppressed using the `message: false` code chunk option to save space and improve readability.

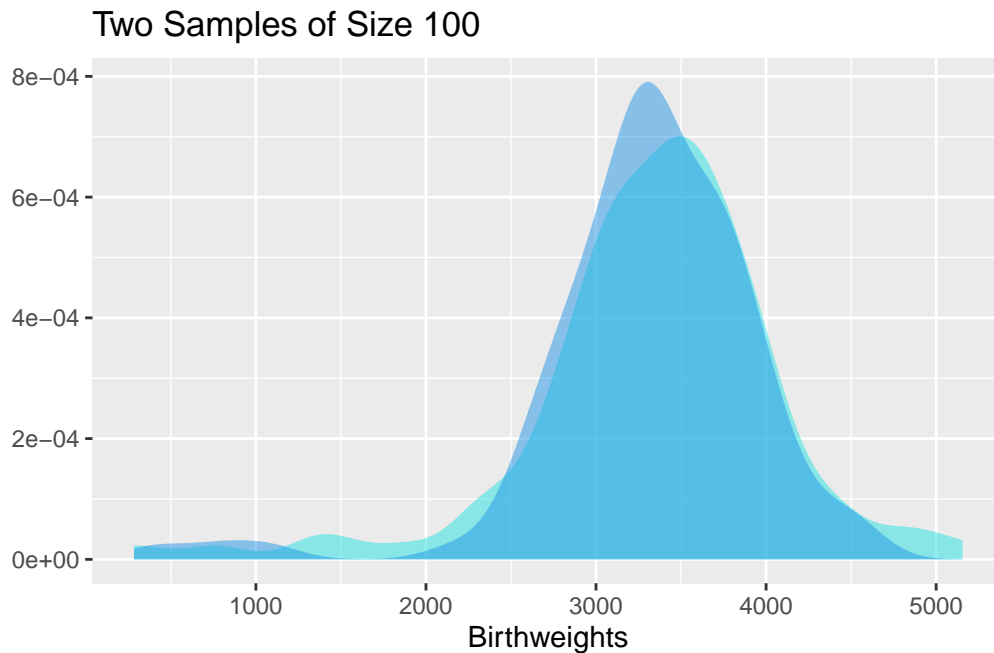
```
# Histogram of known population
gf_histogram(~ birthweight, data = Births, binwidth = 125, center = 62.5) |>
  gf_labs(x = "Weight (g)", y = "Births")
```



The histogram shows the distribution of the population of nearly four million births. Warning messages about missing values were suppressed using the `warnings: false` code chunk option.

Figure 10.2 (page 326):

```
# Samples of 100
set.seed(12452)
gf_density(~ birthweight, data = sample(Births, size = 100), fill = 5) |>
  gf_density(~ birthweight, data = sample(Births, size = 100), fill = 4) |>
  gf_labs(x = "Birthweights", y = "", title = "Two Samples of Size 100")
```

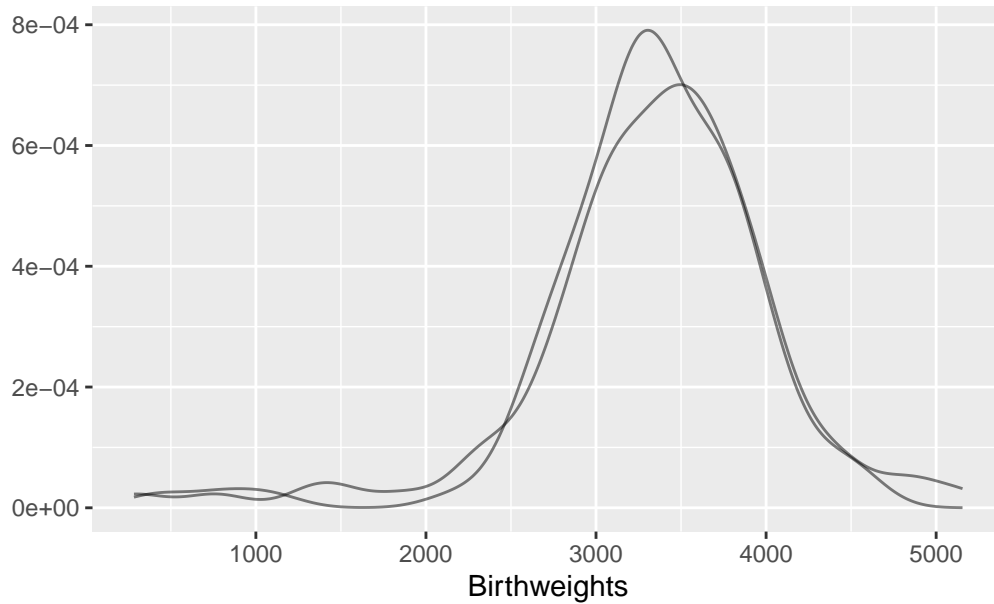


Here we set an (arbitrary) random number seed (12452). This sets our location in a long list of pseudo-random numbers. Use of a seed in this manner allows us to rerun the code and get the same answer.

It's not necessarily a best practice to display overlapping densities (or histograms). Perhaps a better approach would be to use `gf_dens()` to only display the line.

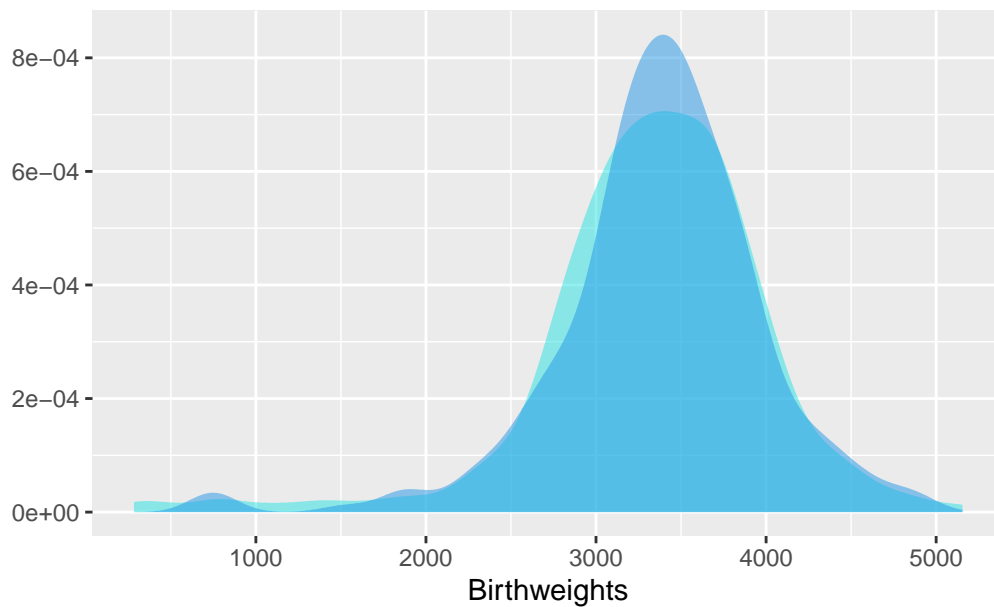
```
set.seed(12452)
gf_dens(~ birthweight, data = sample(Births, size = 100)) |>
  gf_dens(~ birthweight, data = sample(Births, size = 100)) |>
  gf_labs(x = "Birthweights", y = "", title = "Two Samples of Size 100")
```

Two Samples of Size 100

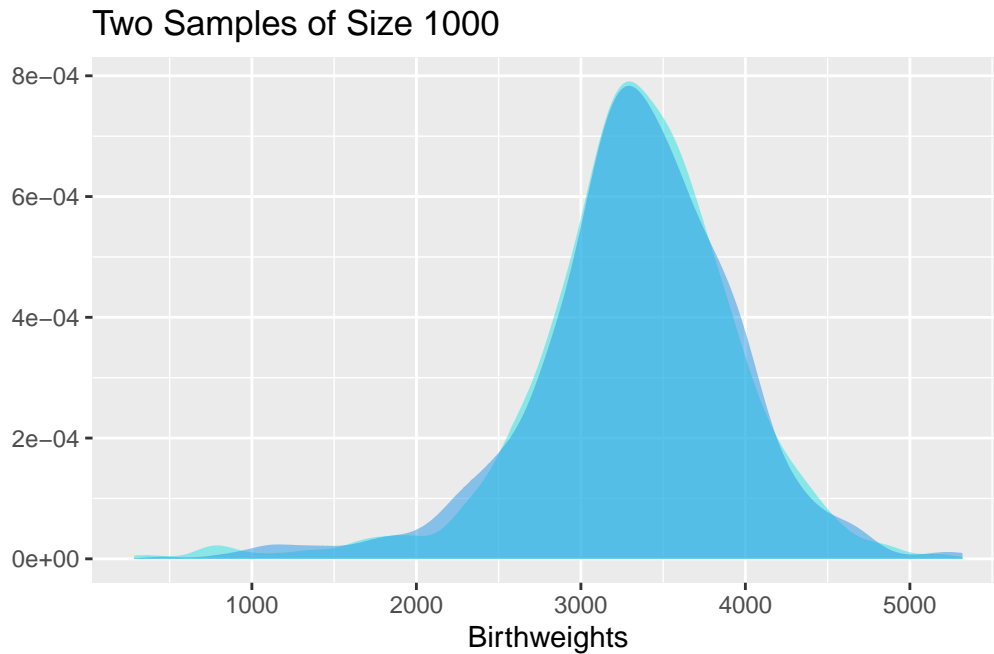


```
# Samples of 250
set.seed(12452)
gf_density(~ birthweight, data = sample(Births, size = 250), fill = 5) |>
  gf_density(~ birthweight, data = sample(Births, size = 250), fill = 4) |>
  gf_labs(x = "Birthweights", y = "", title = "Two Samples of Size 250")
```

Two Samples of Size 250



```
# Samples of 1000
set.seed(12452)
gf_density(~ birthweight, data = sample(Births, size = 1000), fill = 5) |>
  gf_density(~ birthweight, data = sample(Births, size = 1000), fill = 4) |>
  gf_labs(x = "Birthweights", y = "", title = "Two Samples of Size 1000")
```



Section 10.4: Other Sampling Designs

Section 10.5: From the Population to the Sample: You Can't Always Get What You Want

Section 10.6: The Valid Survey

Section 10.7: Common Sampling Mistakes, or How to Sample Badly