

IS5 in R: Sample Surveys (Chapter 10)

Margaret Chien and Nicholas Horton (nhorton@amherst.edu)

July 10, 2018

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. More information about the book can be found at http://wps.aw.com/aw_deveaux_stats_series. This file as well as the associated R Markdown 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 (<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 10: Sample Surveys

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

Section 10.1: The Three Big Ideas of Sampling

Section 10.2: Populations and Parameters

Section 10.3: Simple Random Samples

Random Matters

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

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
```

```
## cols(
```

```
##   X1 = col_integer(),
```

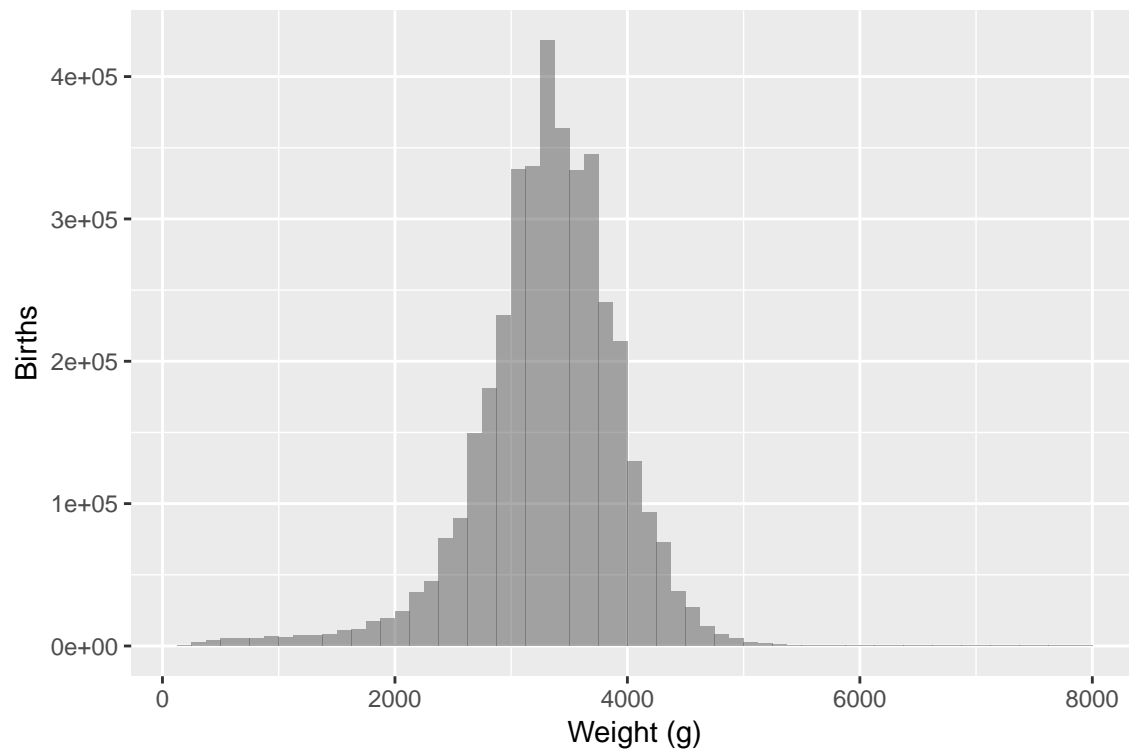
```
##   birthweight = col_integer()
```

```
## )
```

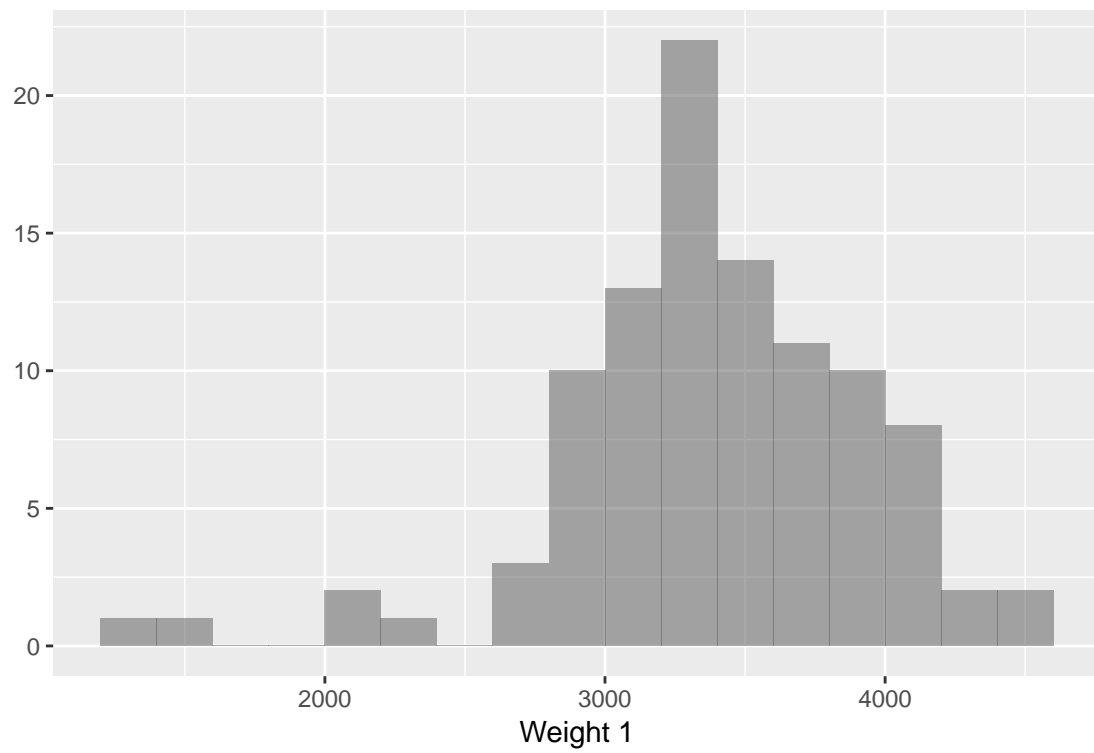
By default, `read_csv()` prints the variable names. These messages can be suppressed using the `message = FALSE` code chunk option to save space and improve readability.

```
gf_histogram(~ birthweight, data = Births, binwidth = 125, center = 62.5) %>%
  gf_labs(x = "Weight (g)", y = "Births")
```

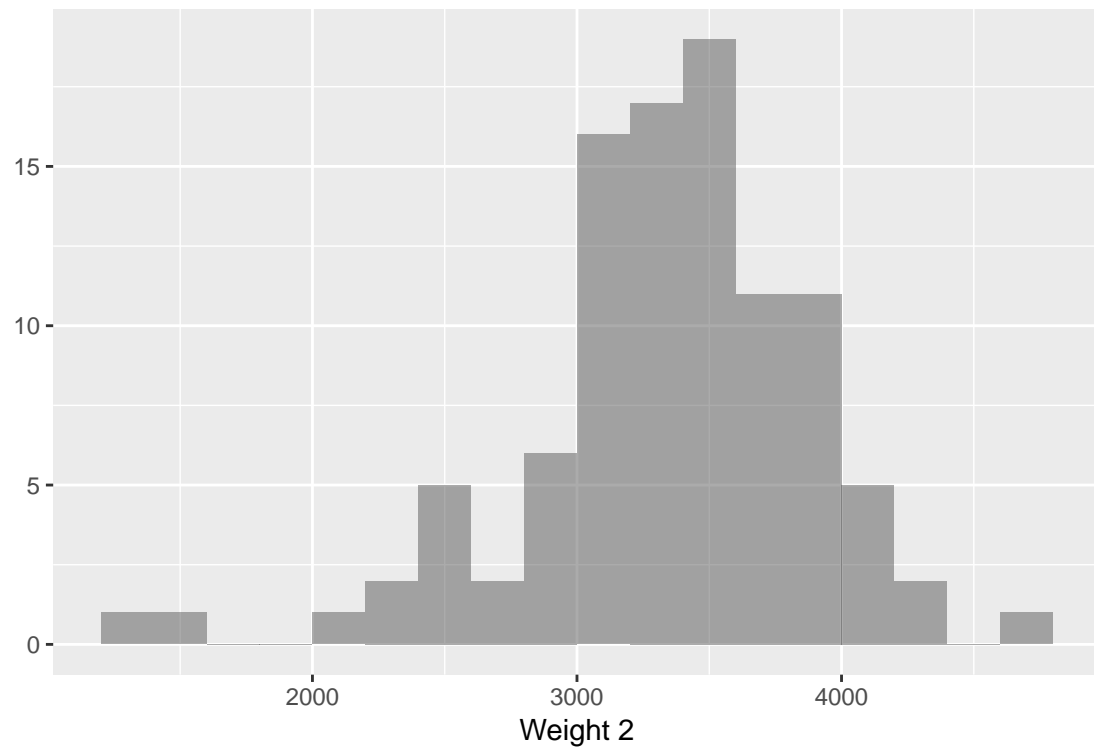
```
## Warning: Removed 4640 rows containing non-finite values (stat_bin).
```



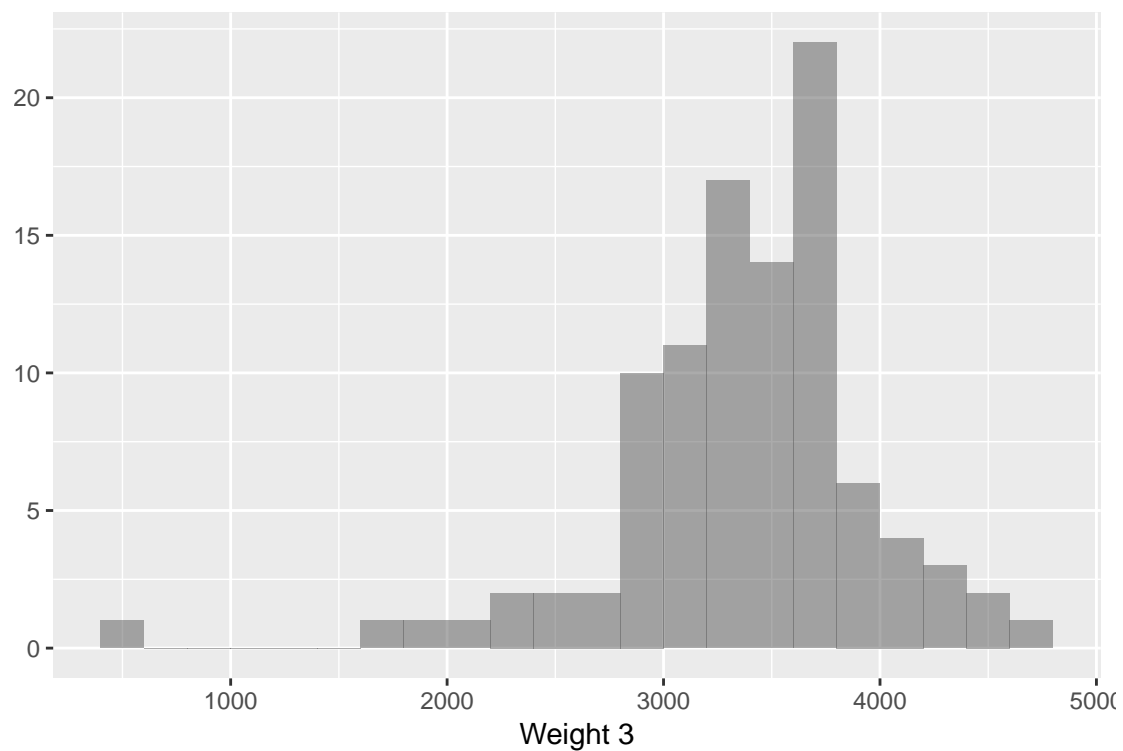
```
# Samples of 100
set.seed(12452)
gf_histogram(~ birthweight, data = sample(Births, size = 100), binwidth = 200, center = 100) %>%
  gf_labs(x = "Weight 1", y = "")
```



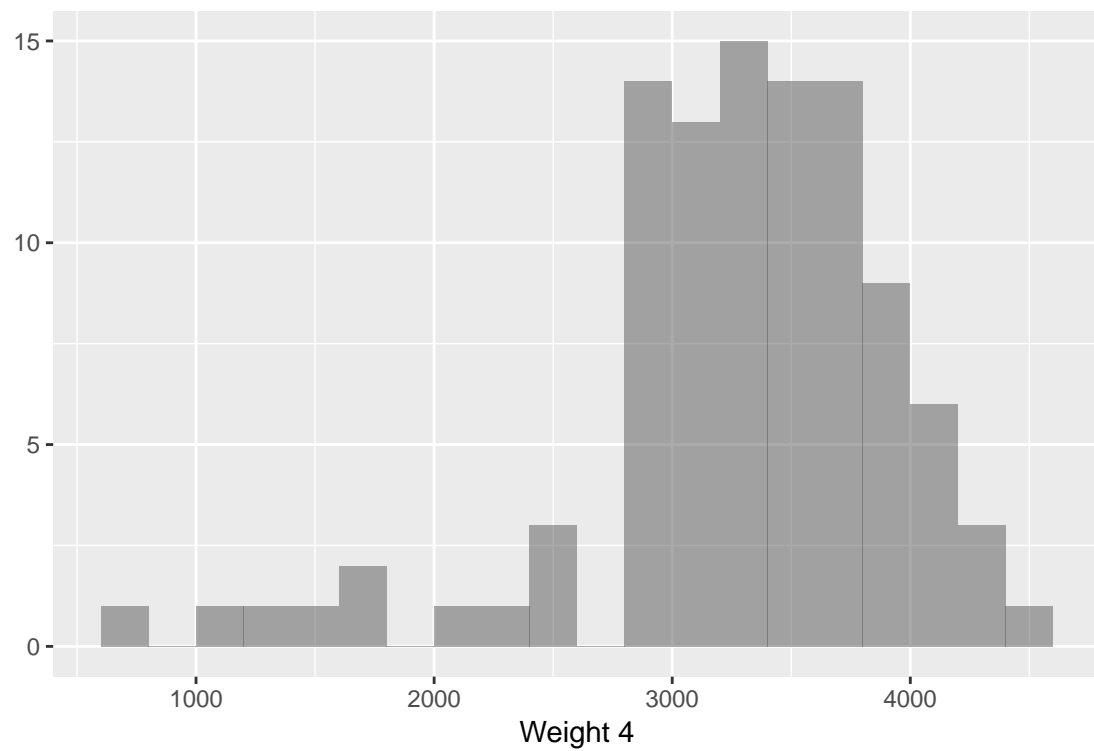
```
gf_histogram(~ birthweight, data = sample(Births, size = 100), binwidth = 200, center = 100) %>%
  gf_labs(x = "Weight 2", y = "")
```



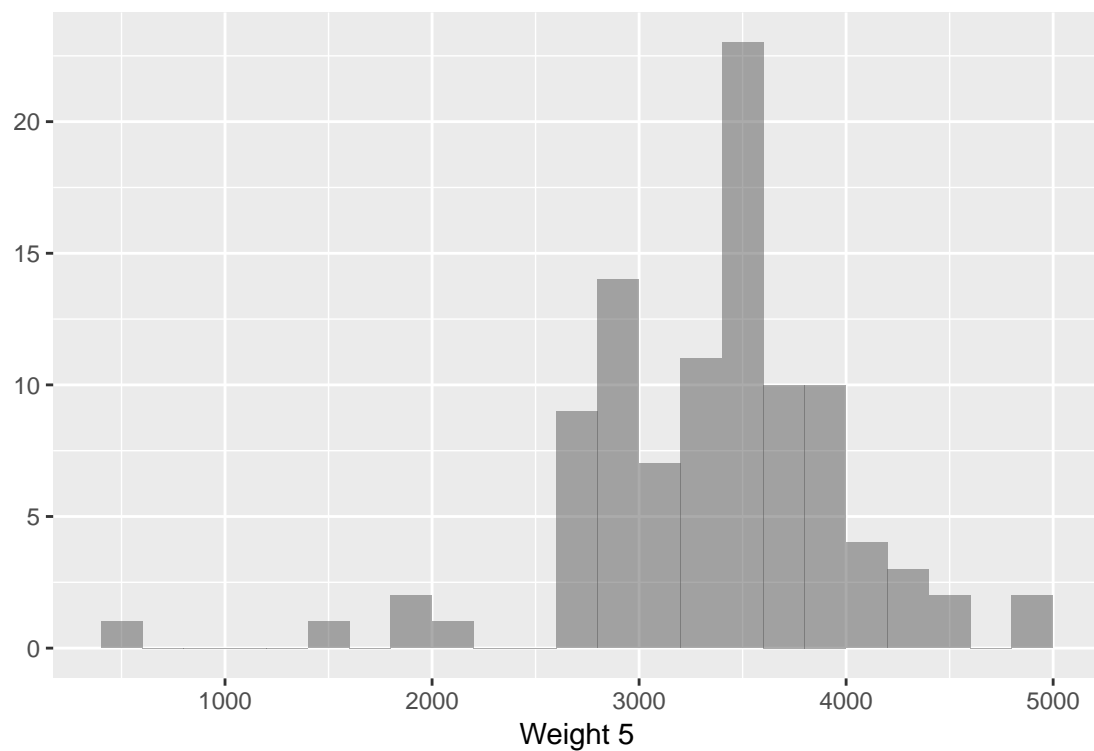
```
gf_histogram(~ birthweight, data = sample(Births, size = 100), binwidth = 200, center = 100) %>%
  gf_labs(x = "Weight 3", y = "")
```



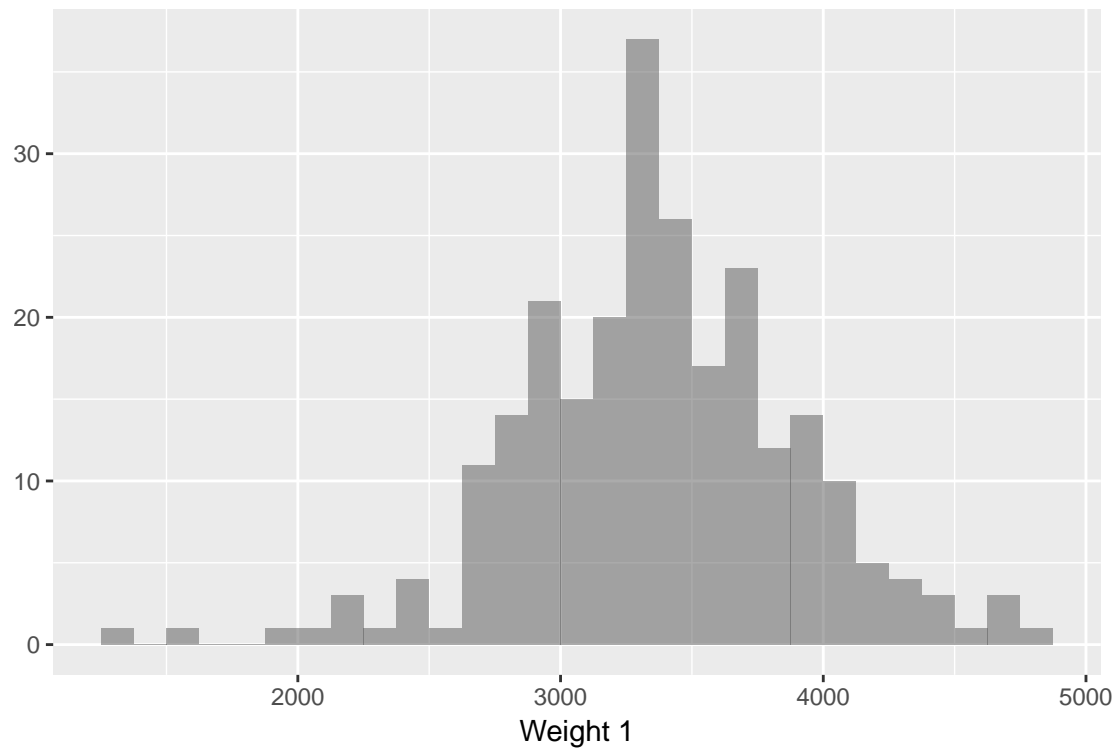
```
gf_histogram(~ birthweight, data = sample(Births, size = 100), binwidth = 200, center = 100) %>%
  gf_labs(x = "Weight 4", y = "")
```



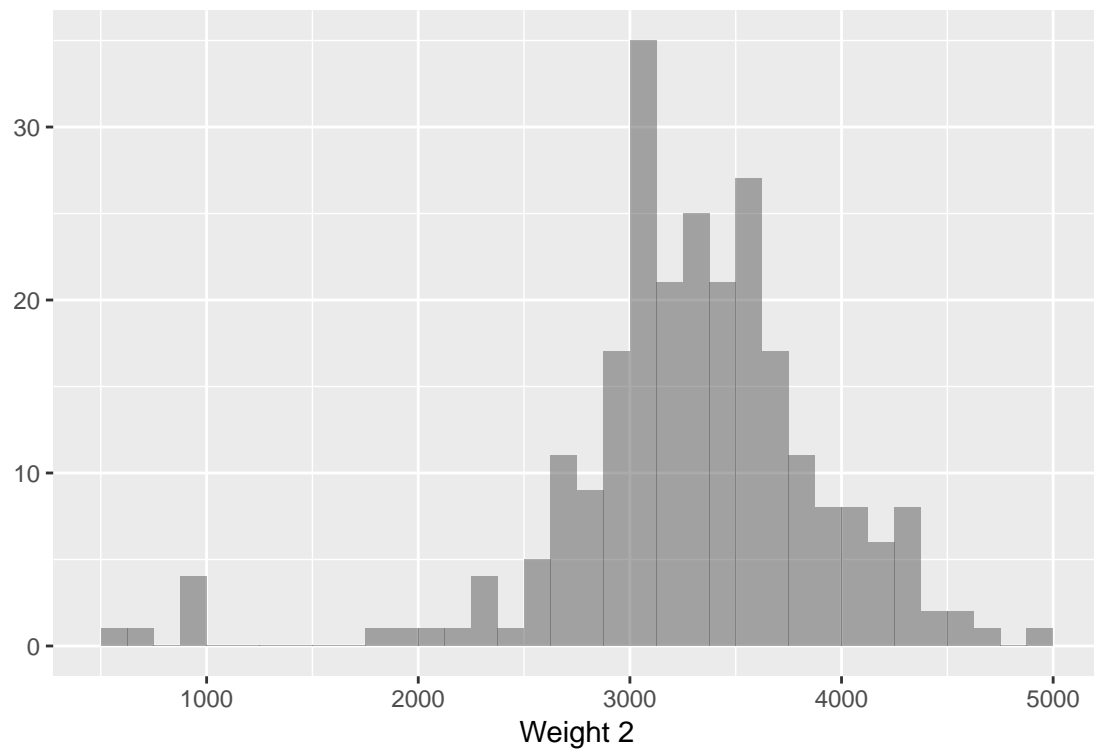
```
gf_histogram(~ birthweight, data = sample(Births, size = 100), binwidth = 200, center = 100) %>%
  gf_labs(x = "Weight 5", y = "")
```



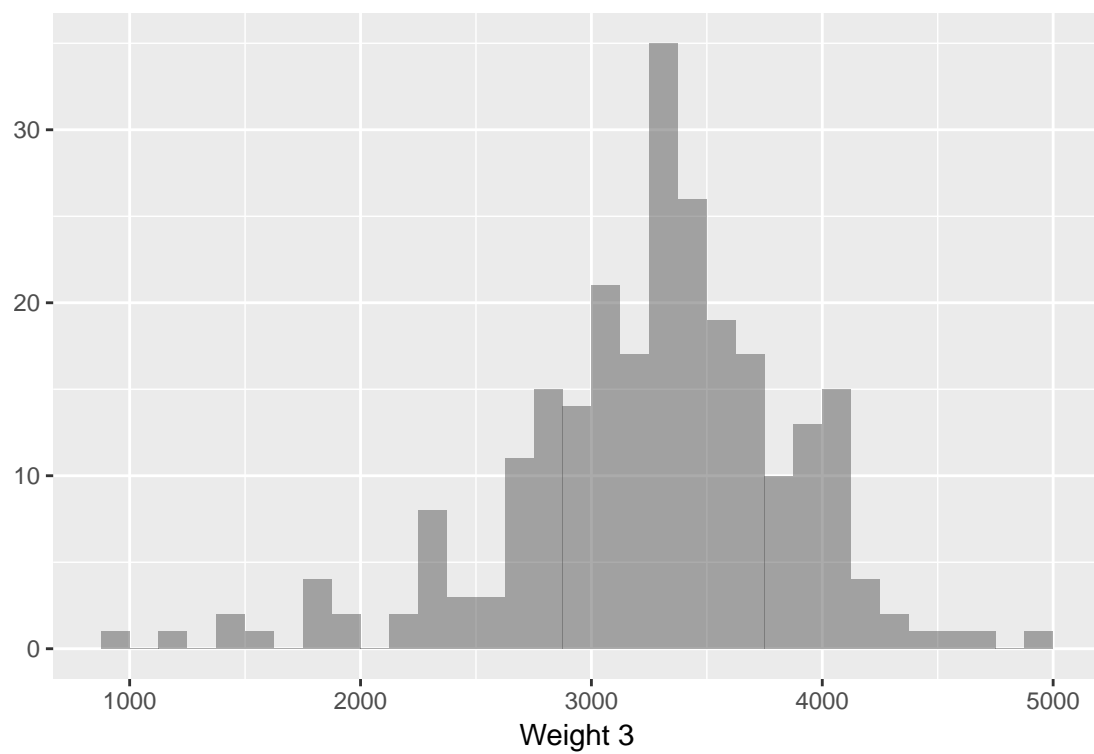
```
# Samples of 250
set.seed(12452)
gf_histogram(~ birthweight, data = sample(Births, size = 250), binwidth = 125, center = 62.5) %>%
  gf_labs(x = "Weight 1", y = "")
```



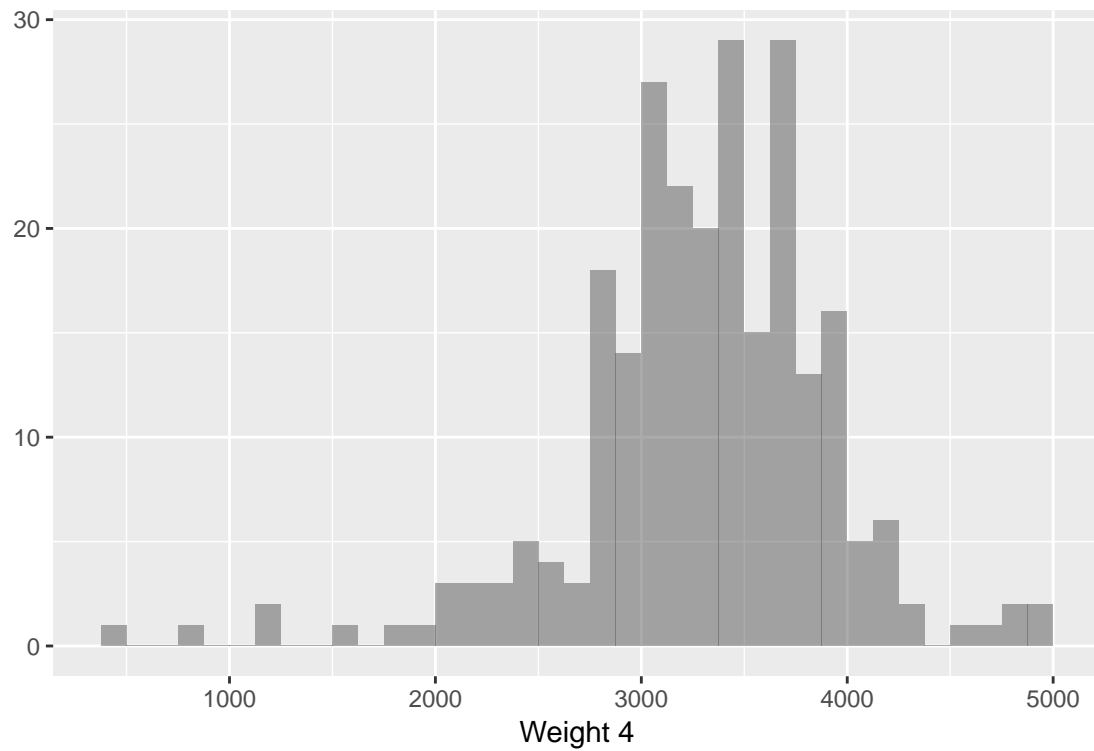
```
gf_histogram(~ birthweight, data = sample(Births, size = 250), binwidth = 125, center = 62.5) %>%
  gf_labs(x = "Weight 2", y = "")
```



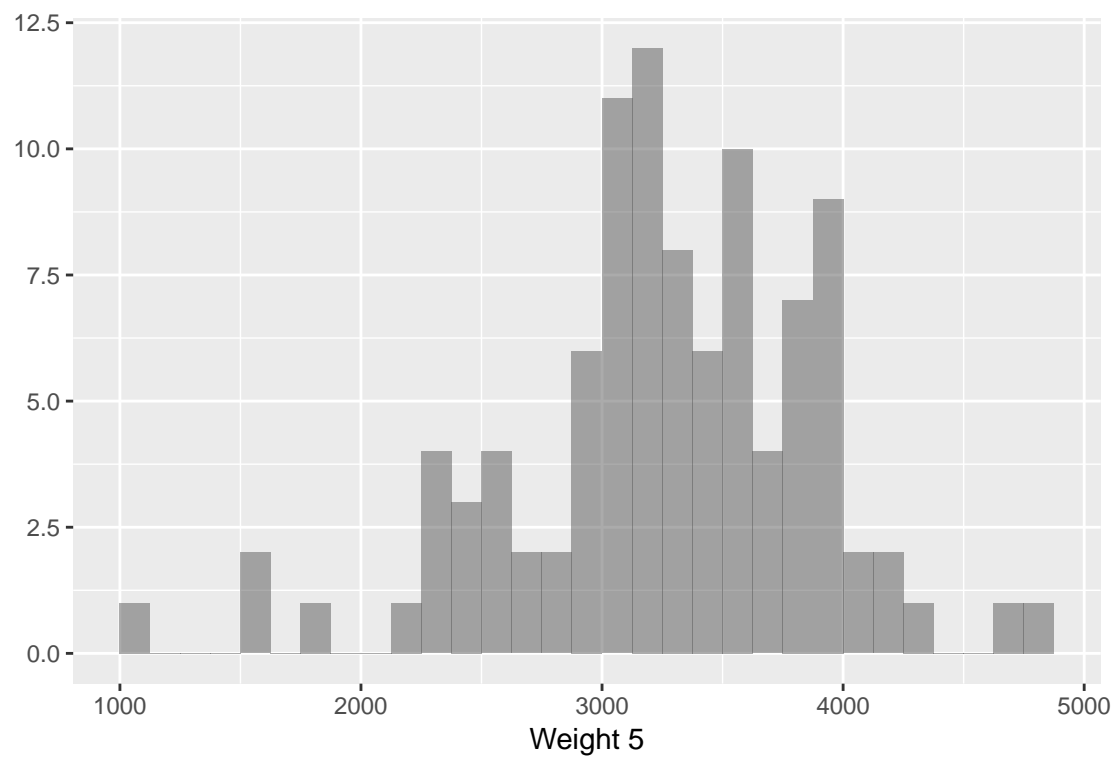
```
gf_histogram(~ birthweight, data = sample(Births, size = 250), binwidth = 125, center = 62.5) %>%
  gf_labs(x = "Weight 3", y = "")
```



```
gf_histogram(~ birthweight, data = sample(Births, size = 250), binwidth = 125, center = 62.5) %>%
  gf_labs(x = "Weight 4", y = "")
```



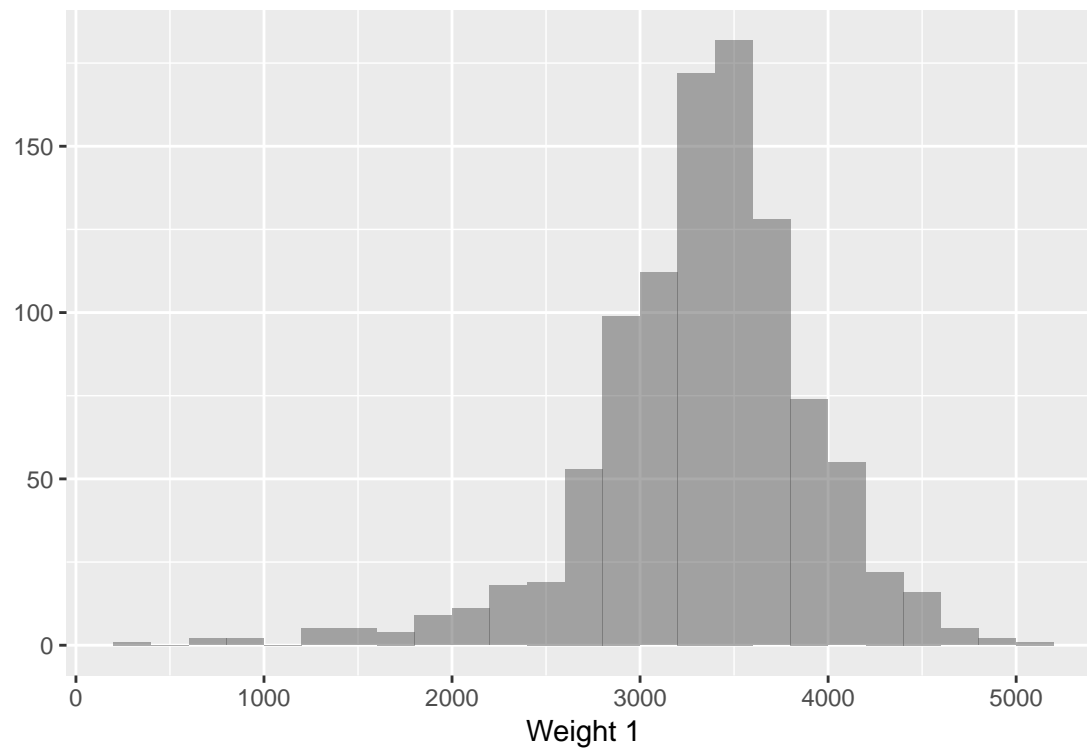
```
gf_histogram(~ birthweight, data = sample(Births, size = 100), binwidth = 125, center = 62.5) %>%
  gf_labs(x = "Weight 5", y = "")
```



```
# Samples of 1000
set.seed(12452)
gf_histogram(~ birthweight, data = sample(Births, size = 1000), binwidth = 200, center = 100) %>%
```

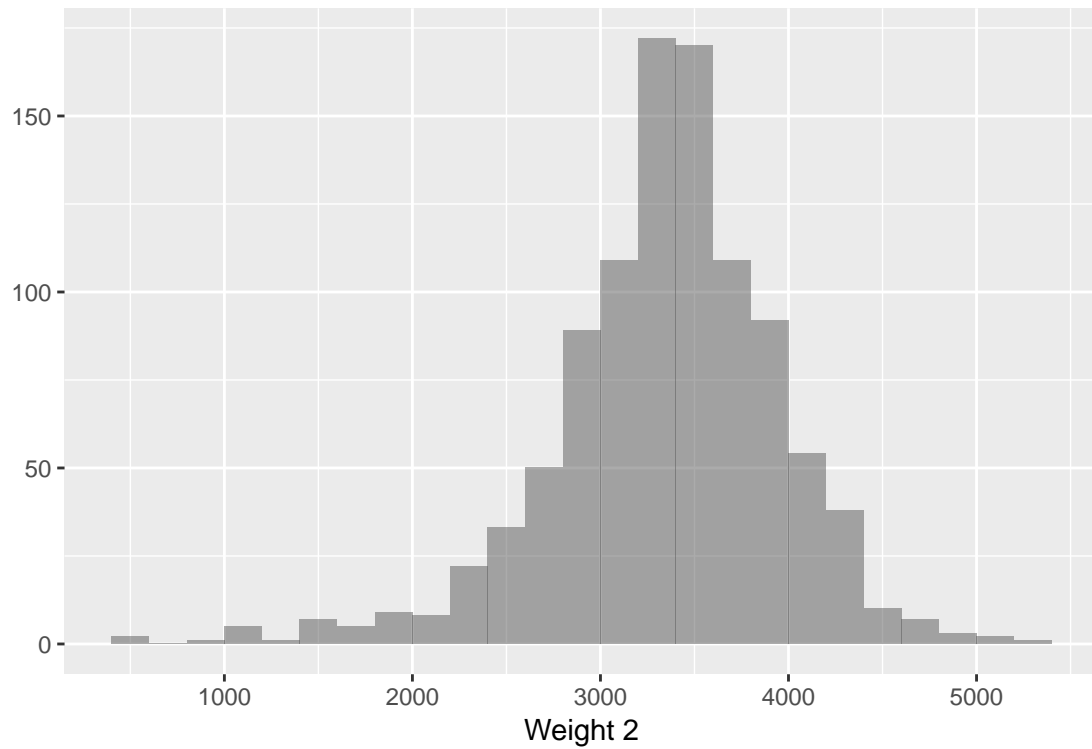
```
gf_labs(x = "Weight 1", y = "")
```

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```



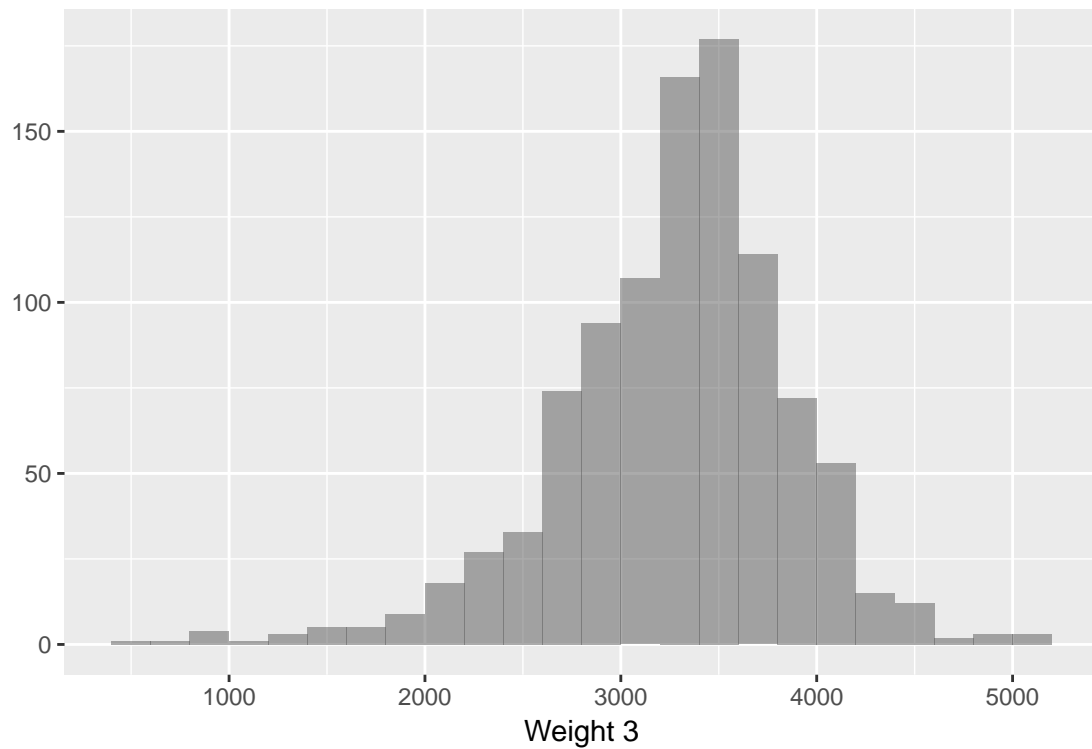
```
gf_histogram(~ birthweight, data = sample(Births, size = 1000), binwidth = 200, center = 100) %>%  
gf_labs(x = "Weight 2", y = "")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

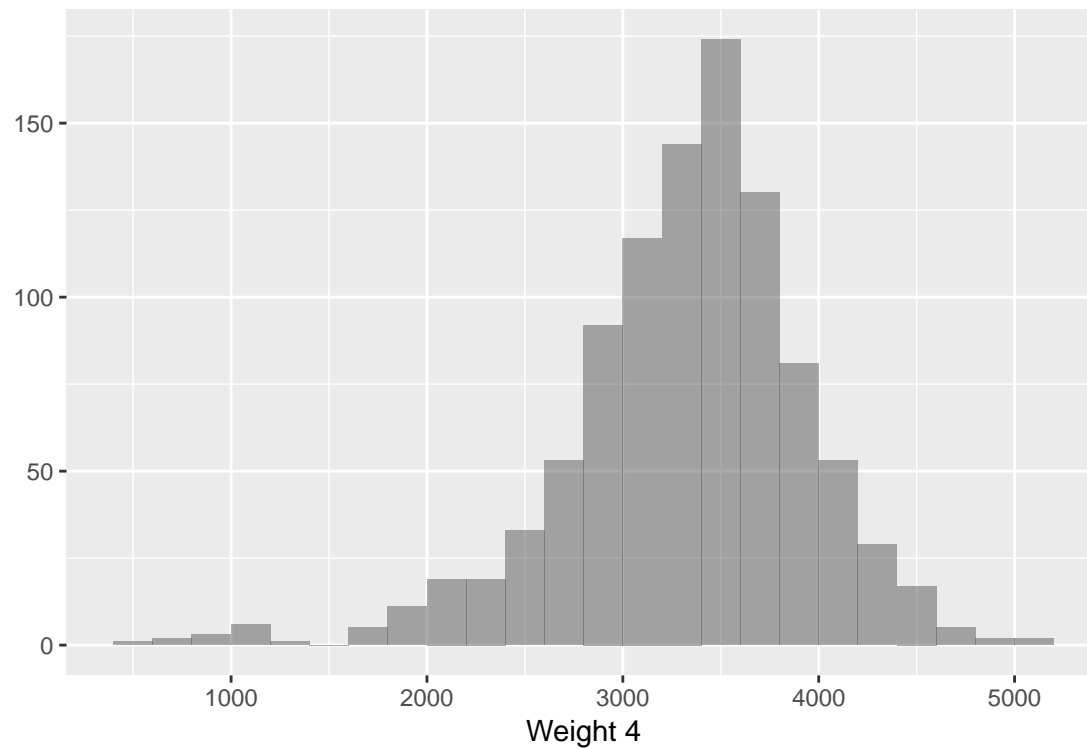
```
gf_histogram(~ birthweight, data = sample(Births, size = 1000), binwidth = 200, center = 100) %>%
  gf_labs(x = "Weight 3", y = "")
```

Warning: Removed 1 rows containing non-finite values (stat_bin).

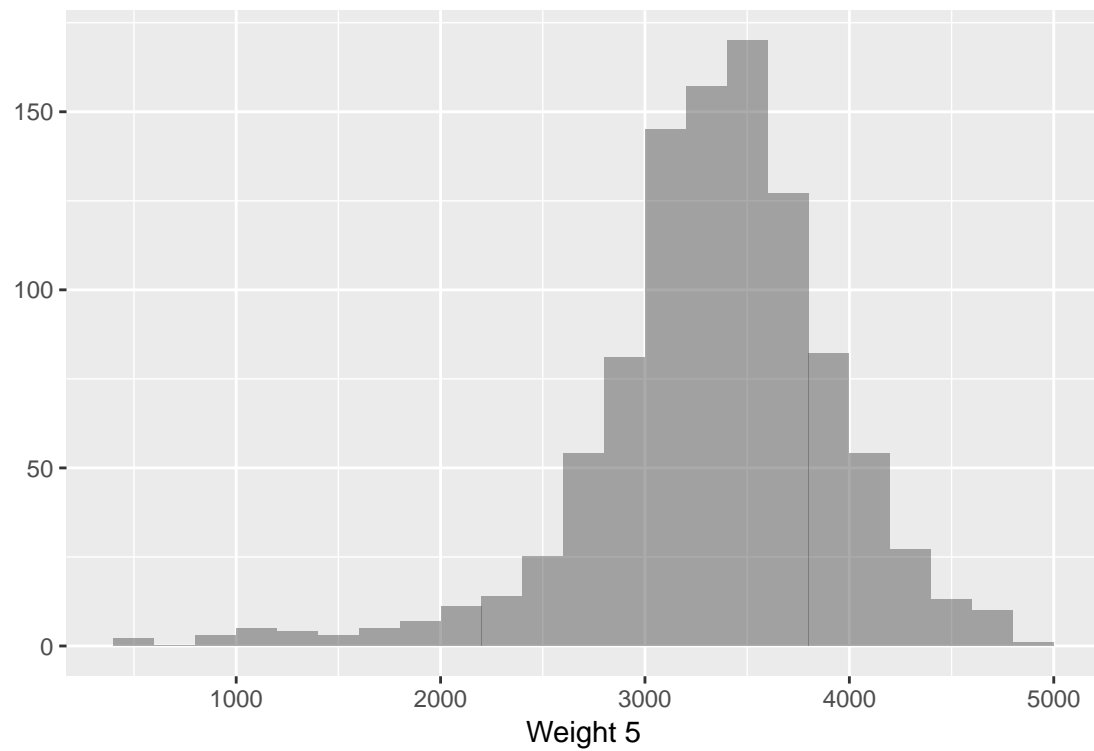


```
gf_histogram(~ birthweight, data = sample(Births, size = 1000), binwidth = 200, center = 100) %>%
  gf_labs(x = "Weight 4", y = "")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



```
gf_histogram(~ birthweight, data = sample(Births, size = 1000), binwidth = 200, center = 100) %>%
  gf_labs(x = "Weight 5", y = "")
```



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