Sampling Distribution Models (Chapter 15)

Patrick Frenett, Vickie Ip, and Nicholas Horton (nhorton@amherst.edu)

July 17, 2017

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fourth Edition of *Intro Stats* (2013) 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 https://nhorton.people.amherst.edu/is4.

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.

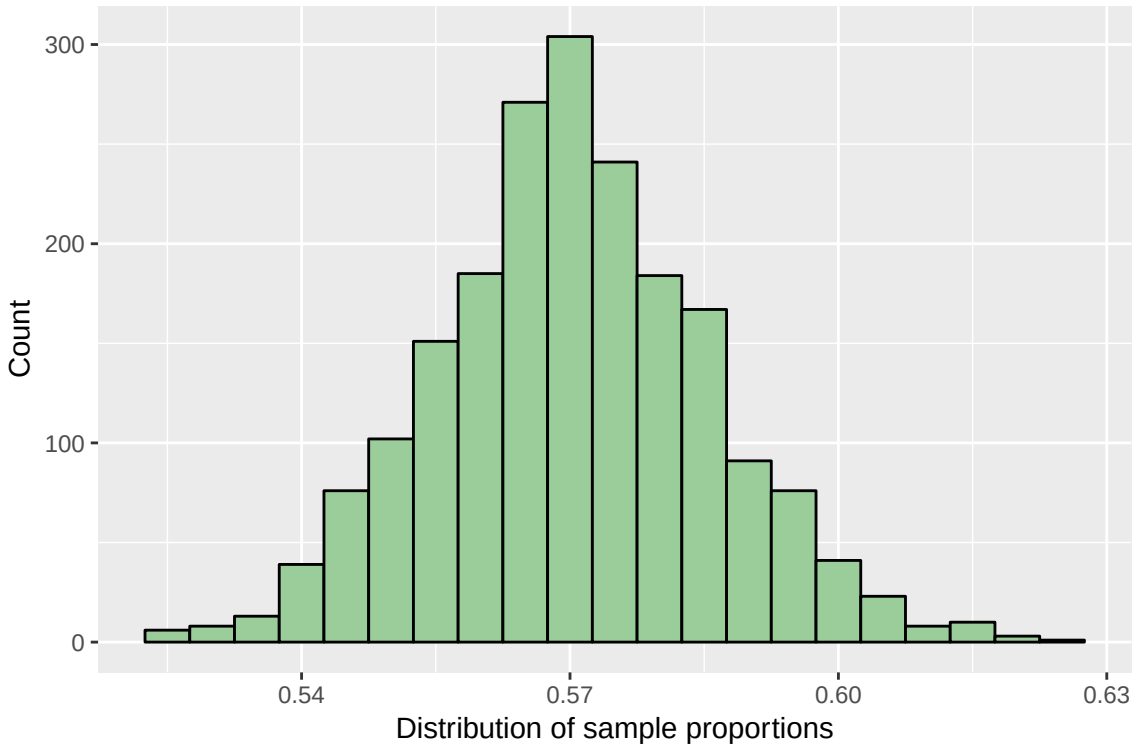
Note that some of the figures in this document may differ slightly from those in the IS4 book due to small differences in datasets. However in all cases the analysis and techniques in R are accurate.

Chapter 15: Sampling Distribution Models

Section 15.1: Sampling distribution of a proportion

Let's regenerate Figure 15.1 (page 400).

```
library(mosaic); options(digits=3)
numsim <- 2000
n <- 1022
p <- 0.57
samples <- rbinom(numsim, size=n, prob=p)/n
gf_histogram(~ samples, binwidth=0.005, center=0.01/2,col=TRUE, fill="darkseagreen3") %>%
  gf_labs(x="Distribution of sample proportions", y="Count")
```



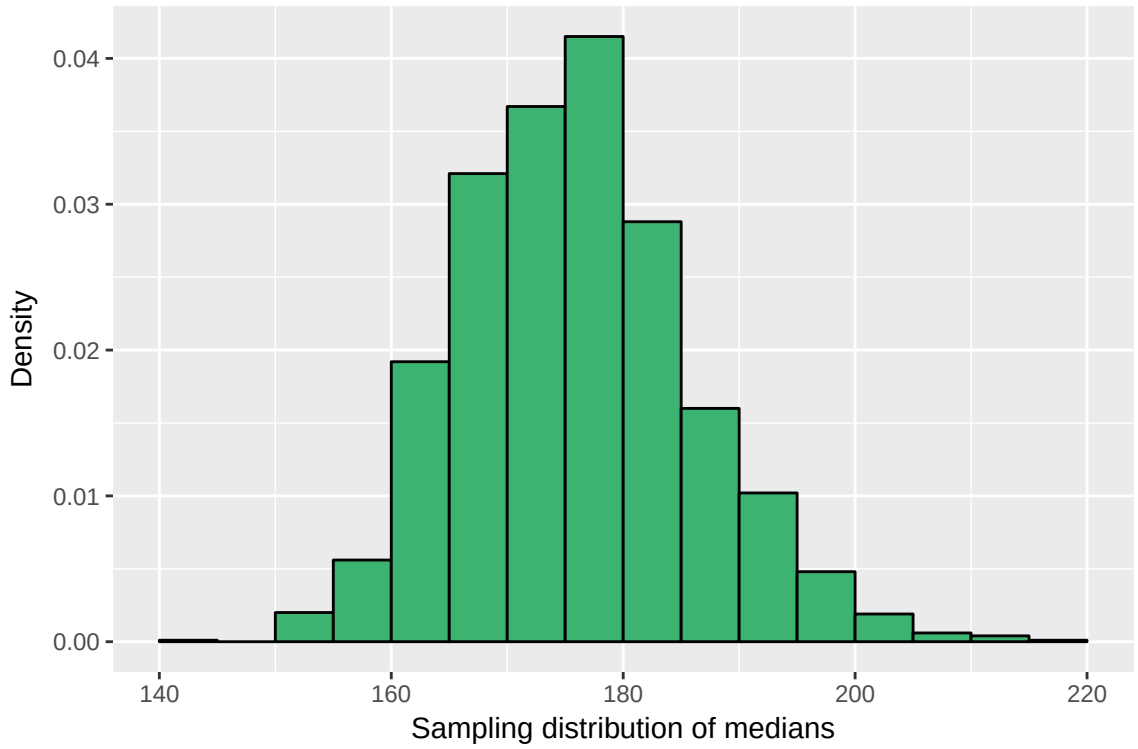
Section 15.2: When does the normal model work? Assumptions and Conditions?

Section 15.3: The sampling distribution of other statistics

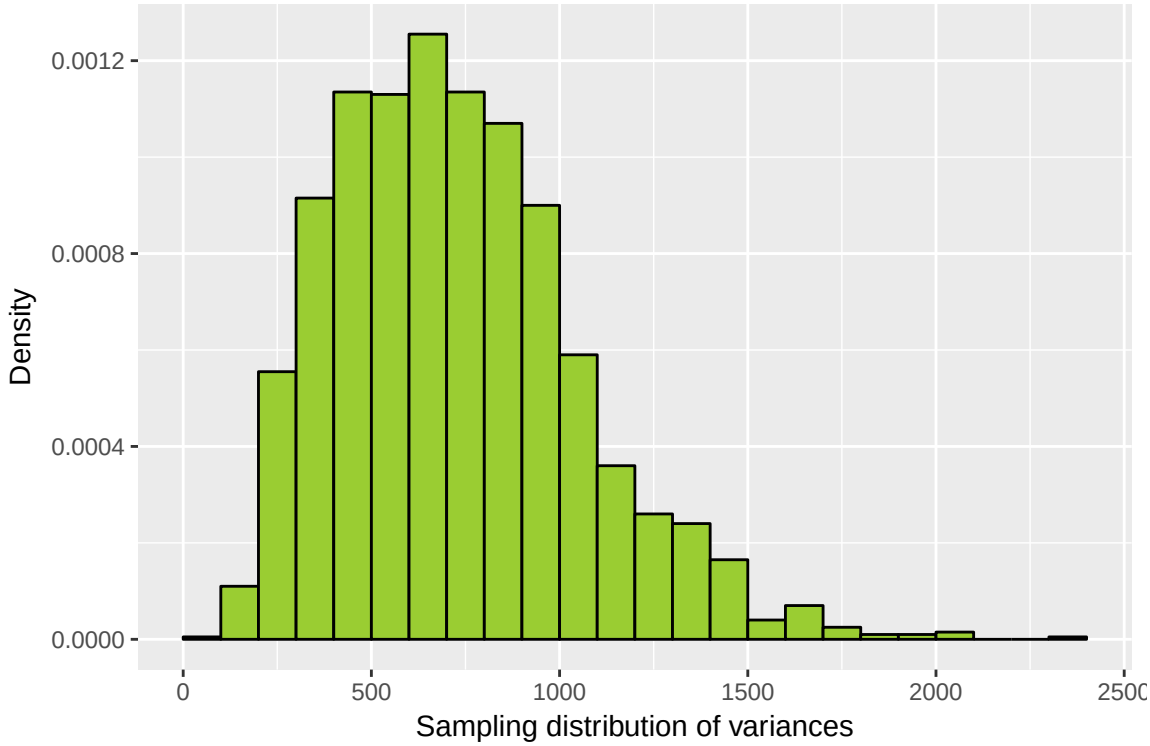
Let's replicate the display on page 407:

```
BodyFat <- read.csv("http://nhorton.people.amherst.edu/sdm4/data/Body_fat_complete.csv")
```

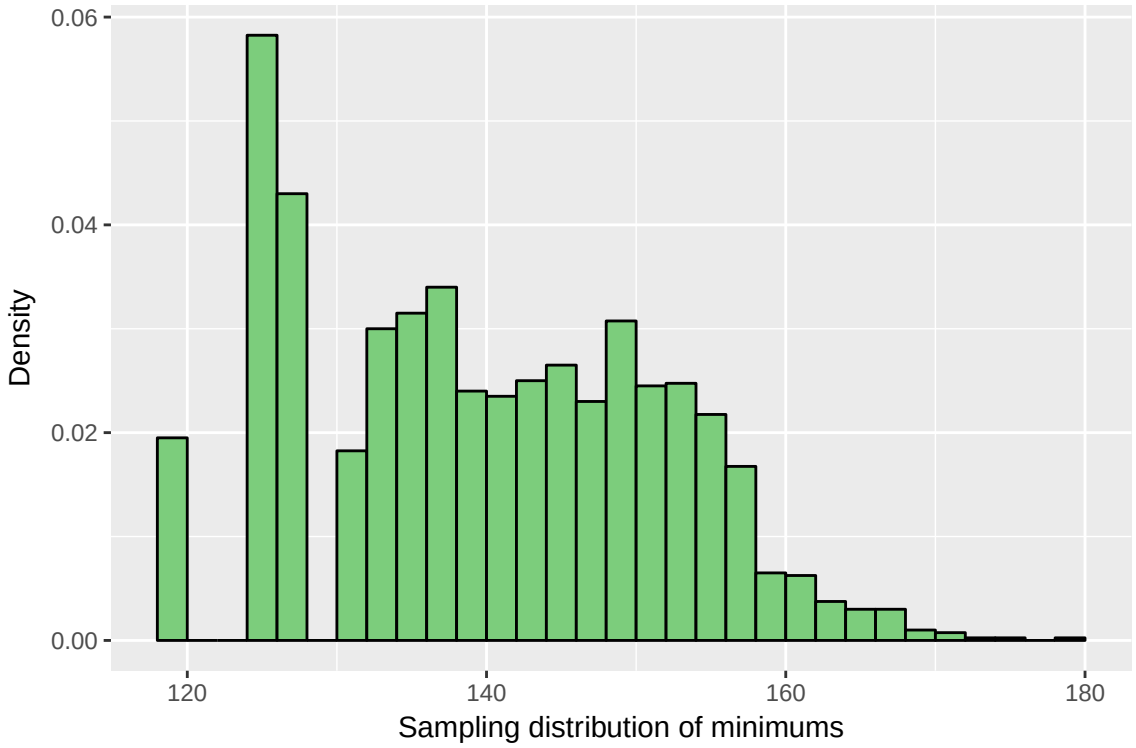
```
medians <- do(2000)*median(~ Weight, data=sample(BodyFat, 10, replace=FALSE))
gf_histogram(..density..~median, binwidth=5, center=5/2,
  data=medians, fill="mediumseagreen",color=TRUE) %>%
  gf_labs(x="Sampling distribution of medians",y="Density")
```




```
variances <- do(2000)*var(~ Weight, data=sample(BodyFat, 10, replace=FALSE))
gf_histogram(..density..~ var, binwidth=100, center=100/2,
  data=variances, fill="olivedrab3",color=TRUE) %>%
  gf_labs(x="Sampling distribution of variances", y="Density")
```



```
minimums <- do(2000)*min(~ Weight, data=sample(BodyFat, 10, replace=FALSE))
gf_histogram(..density..~ min, binwidth=2, center=2/2,
  data=minimums, fill="palegreen3", col=TRUE) %>%
  gf_labs(x="Sampling distribution of minimums", y="Density")
```



Neither of the sampling distributions of the variance or the minimums are normally distributed.

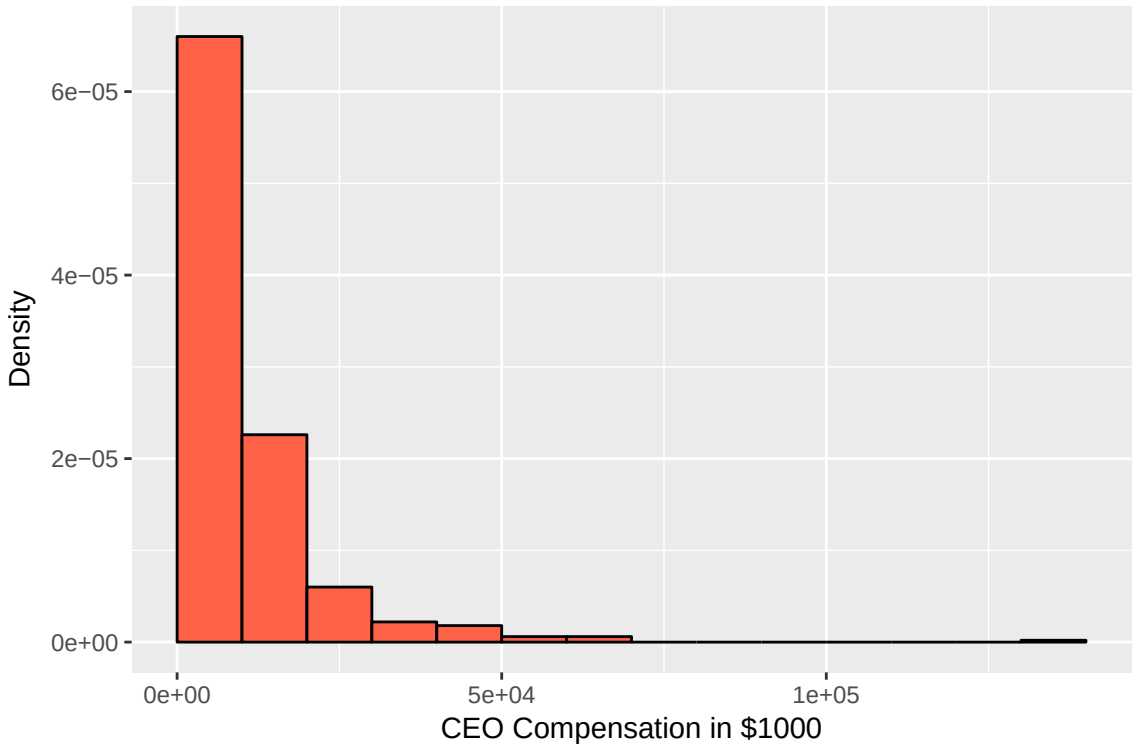
Section 15.4: Central Limit Theorem

Let's replicate the displays on pages 409-410:

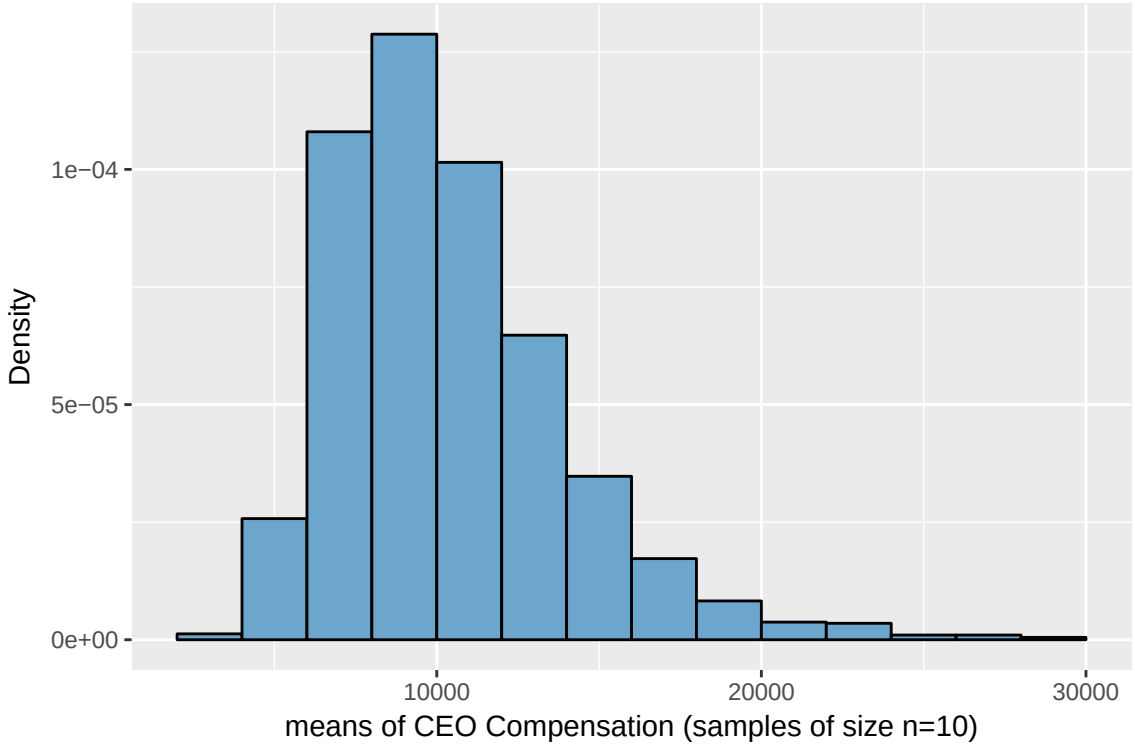
```
require(readr)
CEO <- read_delim("http://nhorton.people.amherst.edu/sdm4/data/CEO_Salary_2012.txt", delim="\t")
CEO <- mutate(CEO, Pay = One_Year_Pay*1000)
```

Note that Figure 15.11 seems to be off by a factor of 10!

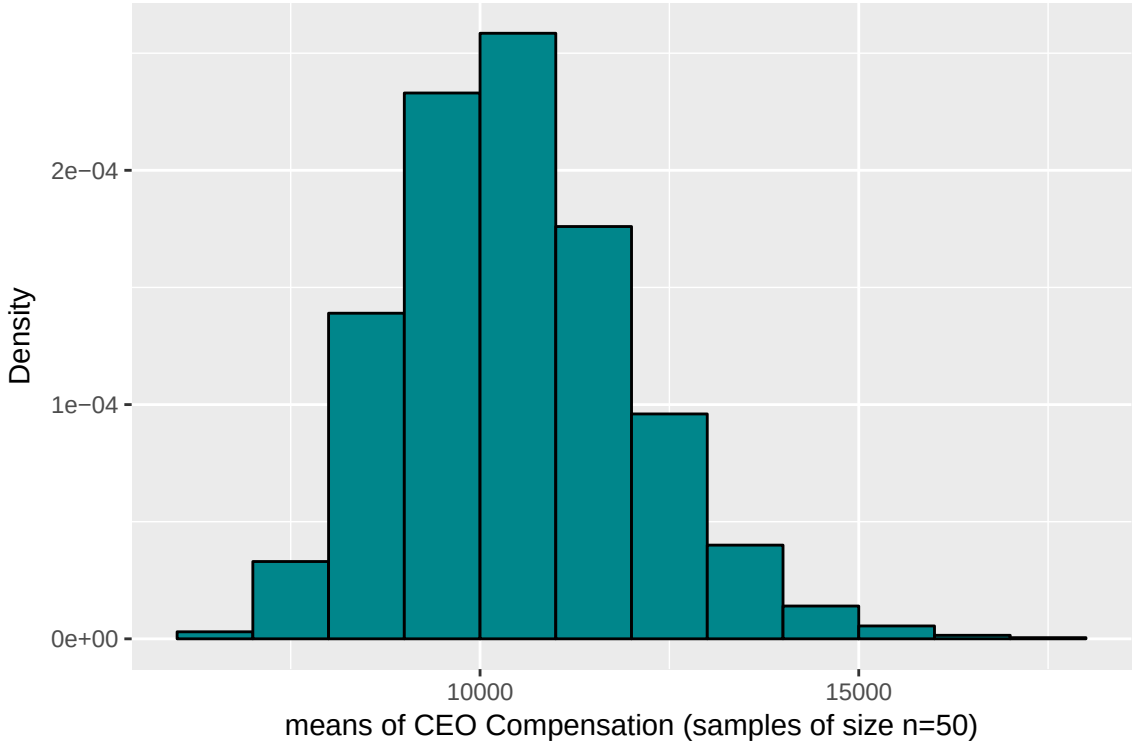
```
gf_histogram(..density..~ Pay, binwidth=10000, center=10000/2-.01,
  data=CEO, fill="tomato",col=TRUE) %>%
  gf_labs(x="CEO Compensation in $1000", y="Density")
```



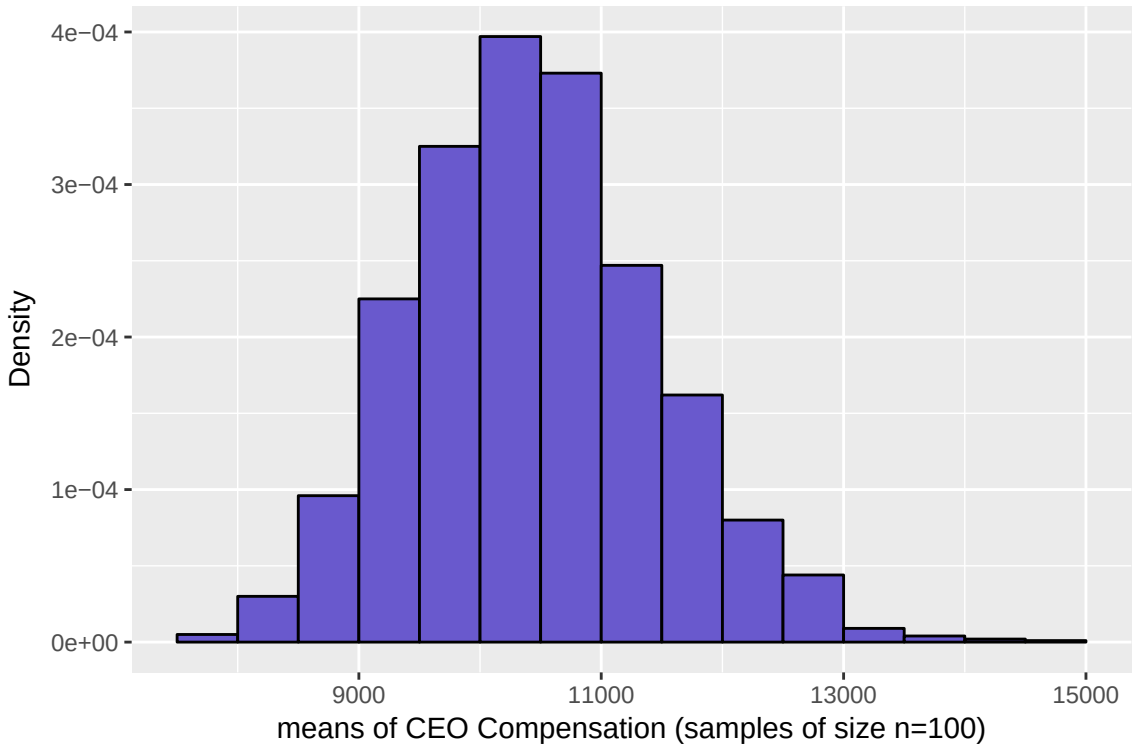

```
samp10 <- do(2000)*mean(~ Pay, data=sample(CEO, 10))
gf_histogram(..density..~ mean, binwidth=2000, center=2000/2,
  data=samp10, fill="skyblue3", col=TRUE) %>%
  gf_labs(x="means of CEO Compensation (samples of size n=10)", y="Density")
```



```
samp50 <- do(2000)*mean(~ Pay, data=sample(CEO, 50))
gf_histogram(..density..~ mean, binwidth=1000, center=1000/2,
  data=samp50, fill="turquoise4",col=TRUE) %>%
  gf_labs(x="means of CEO Compensation (samples of size n=50)", y="Density")
```




```
samp100 <- do(2000)*mean(~ Pay, data=sample(CEO, 100))
gf_histogram(..density..~ mean, binwidth=500, center=500/2,
  data=samp100, fill="slateblue3",col=TRUE) %>%
  gf_labs(x="means of CEO Compensation (samples of size n=100)", y="Density")
```




```
samp200<- do(2000)*mean(~ Pay, data=sample(CEO, 200))
gf_histogram(..density..~ mean, binwidth=500, center=500/2,
  data=samp200, fill="royalblue3", col=TRUE) %>%
  gf_labs( x="means of CEO Compensation (samples of size n=200)", y="Density")
```