Class 7: Bootstrap Distributions

MATH 530-630

- Bootstrapping the sampling distribution of the sample mean
  - Old Faithful Geyser Data
  - Big city dataset

Bootstrapping the sampling distribution of the sample mean

In most real-world analyses, we don't actually know the distribution of our random variable in our population. Instead, we just have our sample- with a measured random variable with a known sample distribution. And most of the time, we only have our one random experiment and corresponding one random sample of data, so we can't assemble a sampling distribution of any one statistic. This is all sounding pretty dire! Luckily, even without knowing the distribution of the rv in the population, we can make some important assumptions about that sample and infer back to our population. So, we take our 1 random sample, then randomly sample from *that* population. So now we are sampling from a finite population rather than an infinite population. This is called **bootstrapping**, where we resample (with replacement) from our sample data.

Here is the general template for doing this in R:

```
Repeat B times {
Draw a resample with replacement from the data. Calculate the resample st
atistic.
Save the resample statistic into a variable.
}
Make a histogram and Normal quantile plot of the B resample statistics.
Calculate the standard deviation of the B statistics.
```

Old Faithful Geyser Data

We'll look at the `eruptions` variable- values are numeric and represent the eruption time in minutes for 272 eruption events.

```
data(faithful)
faithful %>%
  summarise(n = n(), # number of observations
            mean_erup = mean(eruptions), # sample mean
            sd_erup = sd(eruptions), # sample sd
            sem = sd_erup/sqrt(n())) # standard error of mean- formula
```

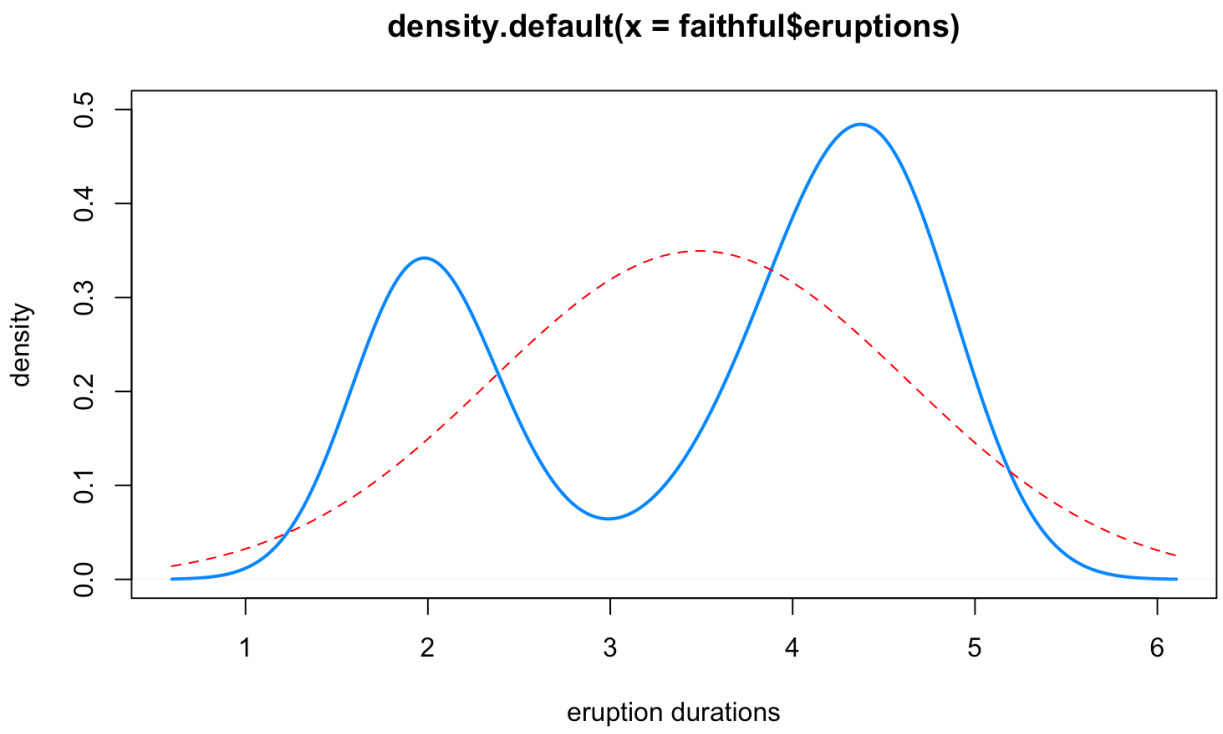
```
    n mean_erup  sd_erup       sem
1 272  3.487783 1.141371 0.0692058
```

```
quantile(faithful$eruptions, c(.025, .975)) # 95% confidence interval for
 the mean
```

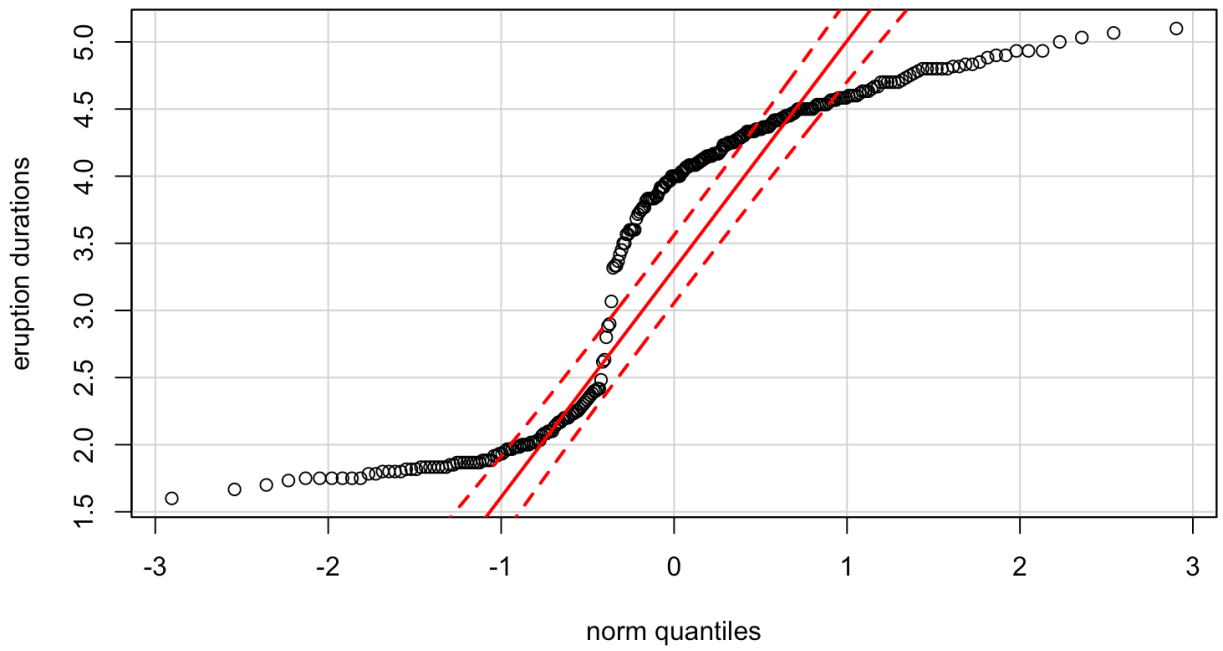
```
    2.5%    97.5%
1.750000 4.907425
```

How would you describe this distribution's shape?

```
# KDE plot
plot(density(faithful$eruptions), col = "dodgerblue", lwd = 2, ylim = c(0
, 0.5),
    xlab = "eruption durations", ylab = "density")
curve(dnorm(x, mean(faithful$eruptions), sd(faithful$eruptions)), lty = 2
, col = "red",
    add = T)
```




```
# library(car)
qqPlot(faithful$eruptions, ylab = "eruption durations")
```

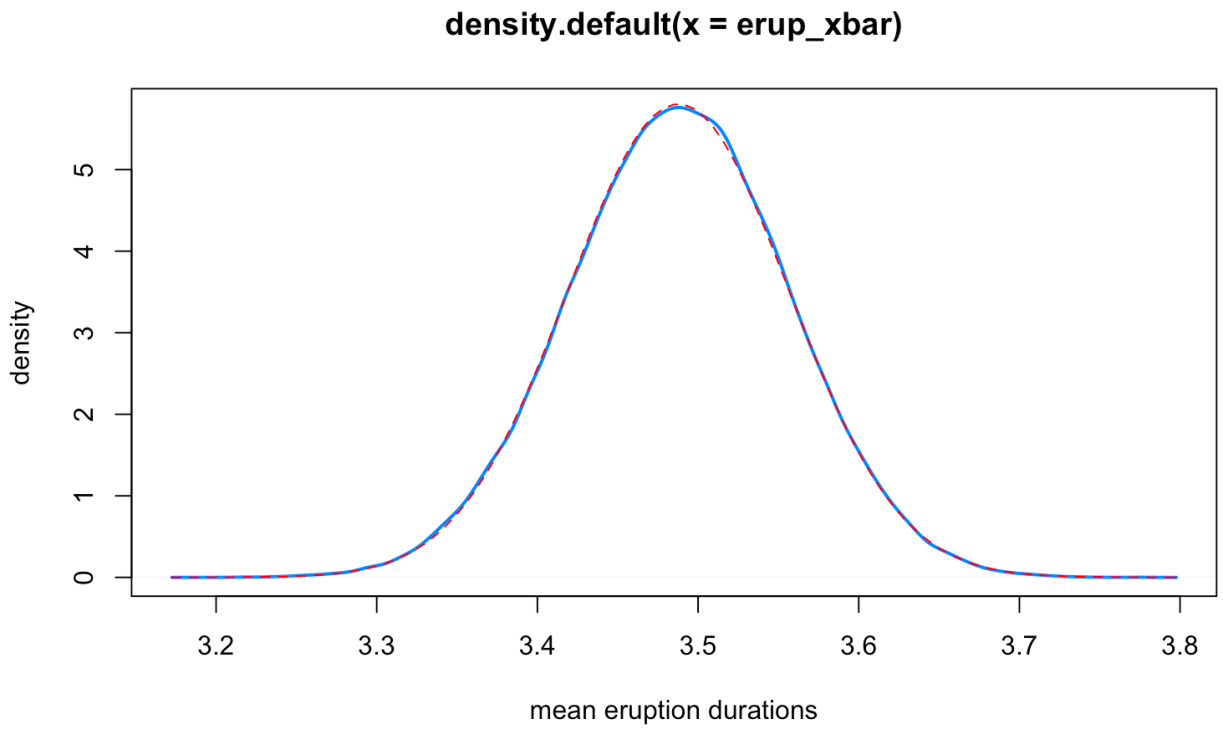


Bootstrap the means using 100,000 random samples of the 272 observations.

```
N <- 10^5  # set the number of random samples to take
erup_xbar <- numeric(N)  # create an empty space to store that many means
 in R
for (i in 1:N) {
    x <- sample(faithful$eruptions, 272, replace = TRUE)  # draw resample
 of same size
    erup_xbar[i] <- mean(x)  # compute mean statistic for each resample
}
```

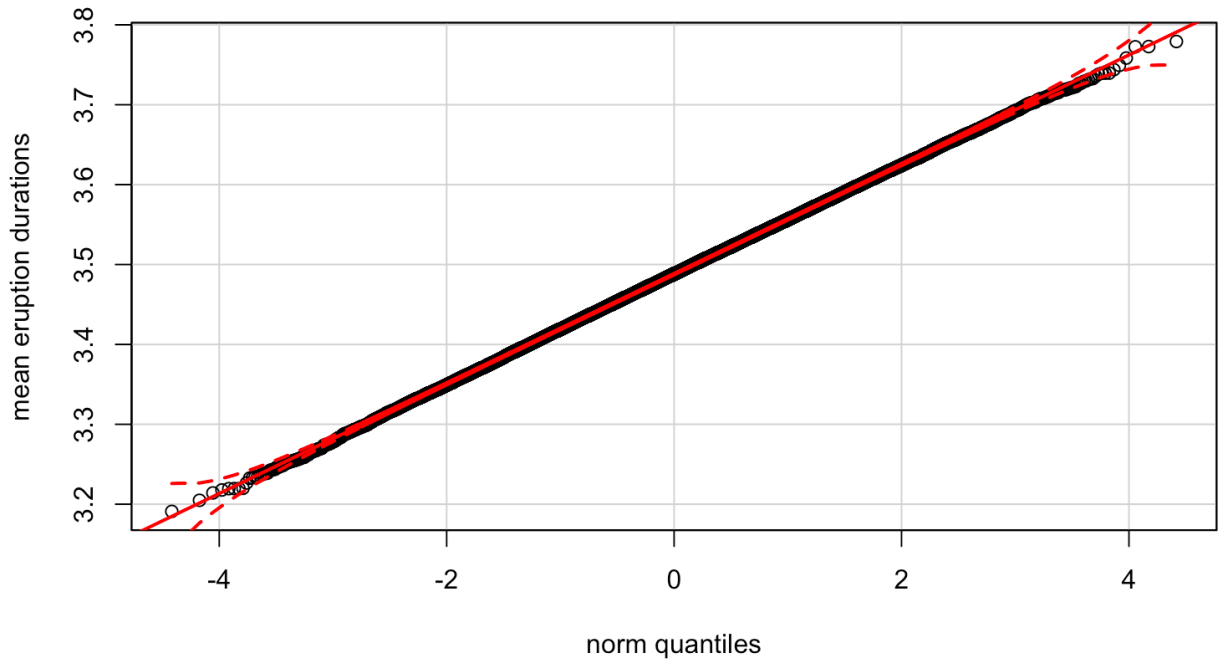
Plot the density of the bootstrap distribution of means, including a curve for the *appropriate* normal distribution:

```
plot(density(erup_xbar), col = "dodgerblue", lwd = 2, xlab = "mean erupti
on durations",
    ylab = "density")
curve(dnorm(x, mean(erup_xbar), sd(erup_xbar)), lty = 2, col = "red", add
 = T)
```



QQplot:

```
# library(car)
qqPlot(erup_xbar, ylab = "mean eruption durations")
```



Calculate the mean and standard deviation of the bootstrap distribution:

```
mean(erup_xbar)  # mean of the bootstrap distribution of means
```

```
[1] 3.487802
```

```
sd(erup_xbar)  # sd of the bootstrap distribution of means (the bootstrap
 SE)
```

```
[1] 0.06873741
```

```
quantile(erup_xbar, c(0.025, 0.975))  # 95% bootstrap confidence interval
 for the mean
```

```
    2.5%    97.5%
3.351882 3.621544
```

The `quantile()` command gives you the 95% bootstrap confidence interval for the mean. This means that we are 95% confident that the true population mean lies within this range.

How does the bootstrap SE compare to our formula?

The "black box" bootstrap via `library(boot)`. A little less intuitive, but gives you four 95% confidence intervals are presented: normal, basic, percentile, and bias-corrected and accelerated. A fifth type, the studentized intervals, requires variances from each bootstrap sample.

```
library(boot)
f <- function(d, i) {
    # second arg here must be 'i', 'f', or 'w'
    d2 <- d[i, ]  # allows boot to select sample
    return(mean(d2$eruptions))
}
erup_boot <- boot(data = faithful, statistic = f, R = 1000, stype = "i")
summary(erup_boot)
```

```
     R original   bootBias   bootSE bootMed
1 1000   3.4878 0.00071331 0.069409  3.4882
```

```
mean(erup_boot$t)
```

```
[1] 3.488496
```

```
sd(erup_boot$t)
```

```
[1] 0.06940928
```

```
boot.ci(erup_boot, type = "all")
```

```
Warning in boot.ci(erup_boot, type = "all"): bootstrap variances needed f
or
studentized intervals
```

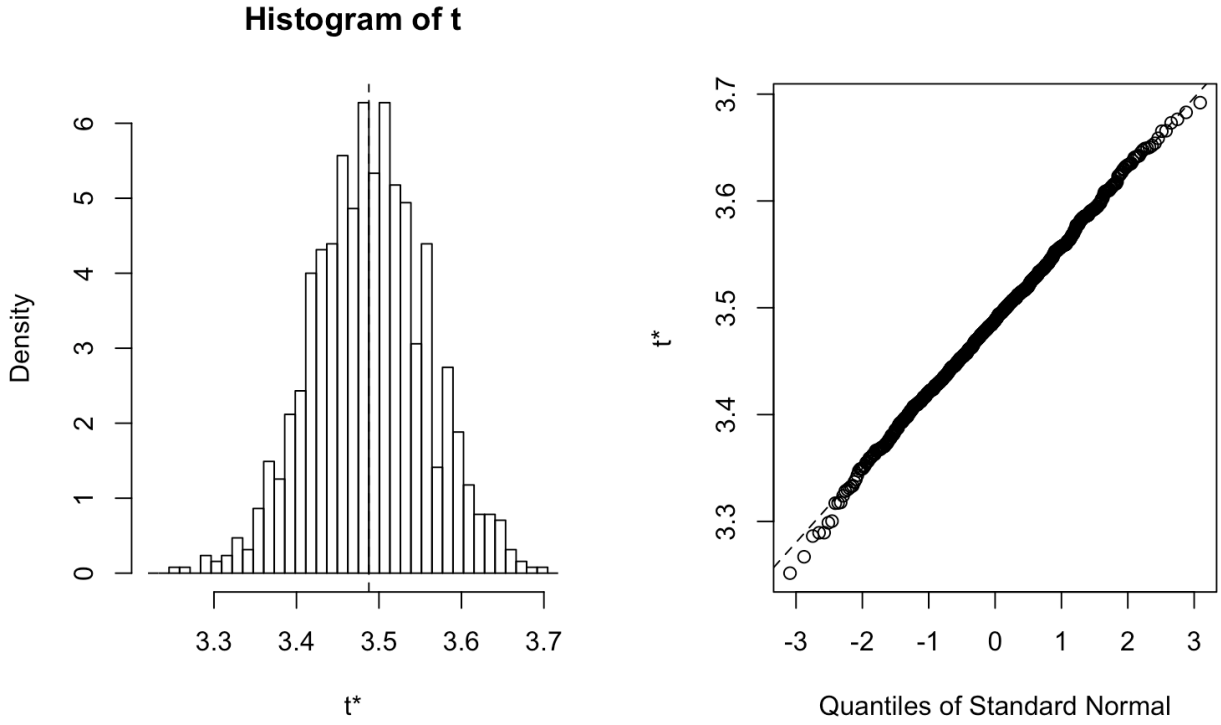
```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = erup_boot, type = "all")

Intervals :
Level      Normal              Basic
95%   ( 3.351,  3.623 )   ( 3.344,  3.623 )

Level     Percentile            BCa
95%   ( 3.352,  3.632 )   ( 3.351,  3.630 )
Calculations and Intervals on Original Scale
```

```
plot(erup_boot)
```



Big city dataset

```
# library(boot)
data("bigcity")
```

The measurements are the population (in 1000â€™s) of 49 U.S. cities in 1920 and 1930. The 49 cities are a random sample taken from the 196 largest cities in 1920. This data includes 2 variables:

- u = the 1920 population
- x = the 1930 population

```
data(bigcity)
bigcity %>%
  summarise(n = n(), # number of observations
            mean_1920pop = mean(u), # sample mean
            sd_1920pop = sd(u), # sample sd
            mean_1930pop = mean(x), # sample mean
            sd_1930pop = sd(x), # sample sd
            ratio = mean_1930pop/mean_1920pop, # ratio
            se_ratio = sd(ratio)/sqrt(n())) # standard error- formula
```

```
   n mean_1920pop sd_1920pop mean_1930pop sd_1930pop    ratio se_ratio
1 49     103.1429   104.4051     127.7959   123.1212 1.239019      Inf
```

We'll use this dataset to highlight that you can bootstrap **any** statistic. Let's do the mean of the ratios of the population in 1930 vs. 1920 for each of the 49 U.S. cities.

Bootstrap the ratio using 100,000 random samples of the 49 observations.

Calculate the mean and standard deviation of the bootstrap distribution. Plot the bootstrapped distribution for the mean ratio- what do you see? How is the distribution different from the geyser data?