Empirical Bootstrap Distribution
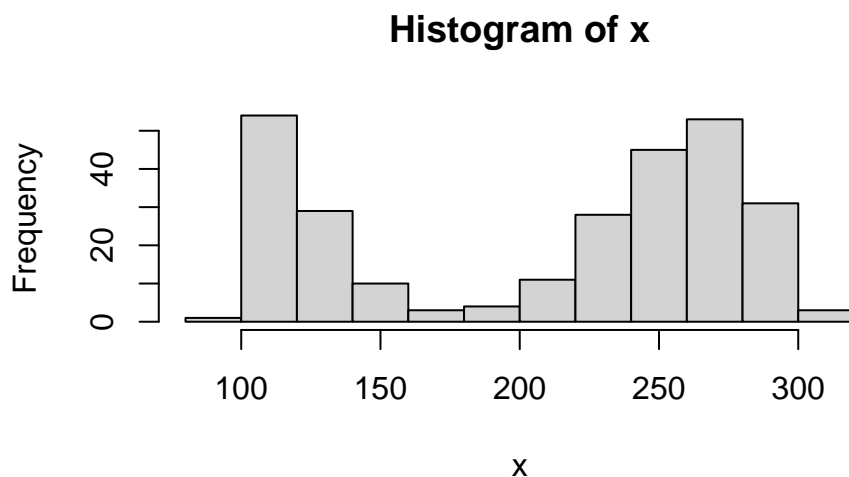
Becky Tang

2024-04-09

Example 1: bootstrapping for a mean

Consider the following sample of eruption times (in seconds) of the Old Faithful geyser:

```r
# obtain data
data("faithful")
x <- faithful$eruptions * 60
hist(x)
```
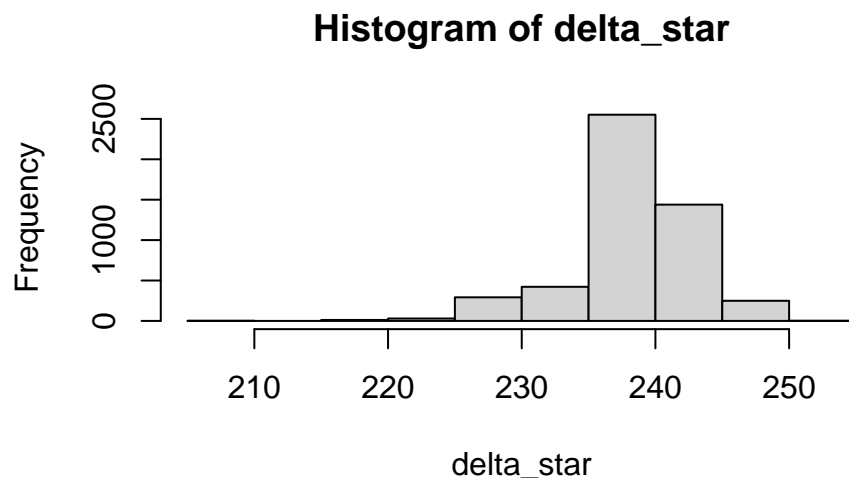
Histogram of x

This data is certainly not Normal!

Perhaps we'd like to obtain a bootstrap distribution for the median eruption time.

```r
n <- length(x)
B <- 5000
delta_star <- rep(NA, B)
for(i in 1:B){
  xstar <- sample(x, size = n, replace = T)
  delta_star[i] <- median(xstar)
}

# bootstrap distribution of medians
hist(delta_star)
```

Histogram of delta_star


```r
# observed sample median
median(x)
```

```
## [1] 240
```

Note that the bootstrap distribution of sample median is centered around the observed sample median.

What can we do with this bootstrap distribution? We can approximate the bias of the sample median for the true median, because now we have a (boostrap) distribution for the sample median to obtain an estimate for the mean of the sample median:

```r
# bootstrap estimate of bias of sample median for true median
mean(delta_star) - median(x)
```

```
## [1] -1.019838
```

Example 2: bootstrapping for MSE

The coefficient of variation of a distribution is the quantity $\frac{\sigma}{\mu}$, where σ and μ are the standard deviation and mean of the distribution, respectively. Let us generate some Poisson data and obtain estimates of the mean squared error (MSE) of the estimator $\frac{s}{\bar{x}}$ where s is the sample standard deviation and $\bar{x}$ is the sample mean.

Notice that in the following, I am setting a seed using the `set.seed()` function. The input doesn't really matter. What this function does is ensure that the random generator in R generates the same sequence of random values. That way when we knit or run the same code that involves random sampling on different laptops, we will get the same result.

```r
# generate some Poisson data
set.seed(309)
lambda_true <- 1
n <- 20
x <- rpois(n, lambda_true)
```
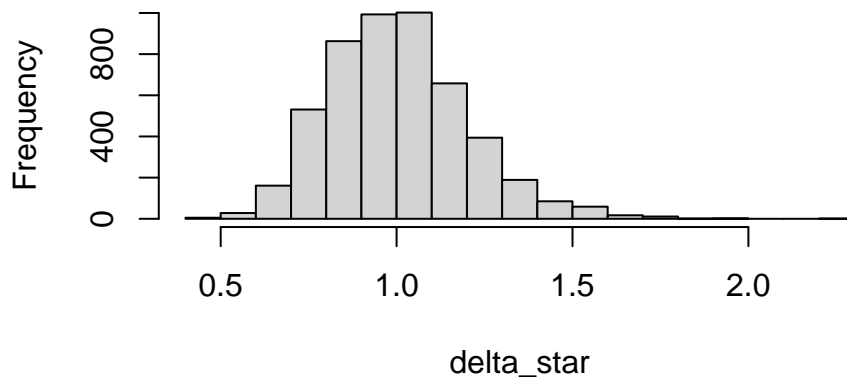
Bootstrap estimate

```r
B <- 5000
delta_star <- rep(NA, B)
for(i in 1:B){
  # if size isn't specified in sample(), defaults to length of x!
  xstar <- sample(x, replace = T)
  delta_star[i] <- sd(xstar)/mean(xstar)
}
hist(delta_star)
```

Histogram of delta_star

We can once again obtain a bootstrap estimate of the bias of this estimate $\frac{\hat{\sigma}_{MLE}}{\hat{\mu}_{MLE}}$ for the true coefficient of variation.

```r
# observed estimate for coefficient of variation
cv_hat <- sd(x)/mean(x)
cv_hat
```

```
## [1] 1.00006
```

```r
# approximate bias of estimate for true coefficient of variation
mean(delta_star) - cv_hat
```

```
## [1] 0.005089808
```

We can also approximate the mean squared error of this estimate for the true quantity:

```r
# bootstrap estimate of MSE
mean((delta_star - cv_hat)^2)
```

```
## [1] 0.03948413
```