Bootstrap confidence intervals

Becky Tang

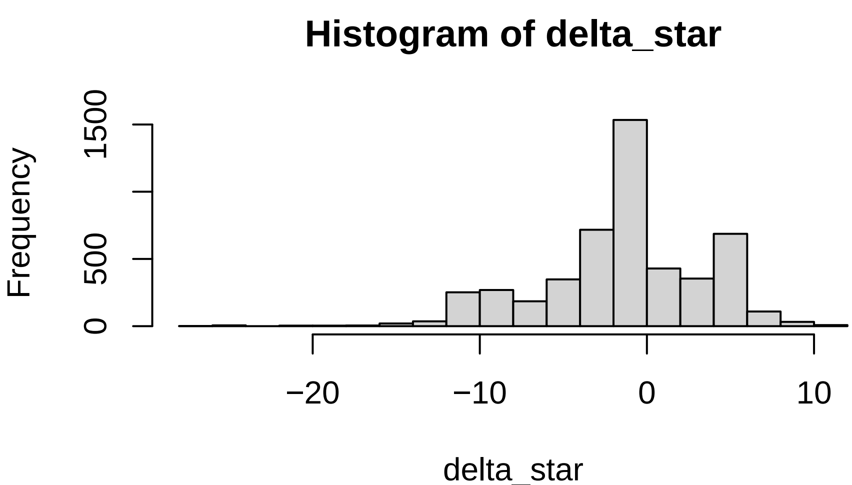
2024-04-12

Non-parametric

Bootstrap CI for median

Suppose we want a symmetric 95% confidence interval for the median eruption time of Old Faithful. Define $\delta = m - M$ where m is the sample median and M is the true median. Then we want quantiles a, b of δ such that $0.95 = P(a \leq \delta \leq b)$. Why? Re-arranging, we have $0.95 = P(m - a \geq M \geq m - b)$! Unfortunately, we don't know the distribution of δ, so we approximate it and its quantiles via bootstrapping. We will assume $\hat{F} = \hat{F}_n$.

```r
set.seed(1)
data("faithful")
x <- faithful$eruptions * 60
n <- length(x)
samp_med <- median(x)
B <- 5000
delta_star <- rep(NA, B)
for(i in 1:B){
  # nonparametric
  xstar <- sample(x, size = n, replace = T)
  delta_star[i] <- median(xstar) - samp_med
}

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



```r
# (approximate) bootstrap CI
a_star <- quantile(delta_star, 0.025)
b_star <- quantile(delta_star, 0.975)
ci <- samp_med - c(b_star, a_star)
names(ci) <- c("2.5", "97.5")
ci
```

```
##     2.5    97.5
## 233.478 250.020
```

The approximate 95% bootstrap CI for the median eruption time of Old Faithul is [233.478, 250.02].

Bootstrap CI for proportion

Recall from Homework 3: in a sample from the Chinese population of Hong Kong in 1937, blood types occurred wit the following frequencies:

- AA: 342
- Aa: 500
- aa: 187

Assuming Hardy-Weinberg, the MLE estimate of the true frequency θ of A is $\hat{\theta} = \frac{2n_{AA}+n_{Aa}}{n} \approx 0.575$. Can we get a 90% confidence interval for θ?

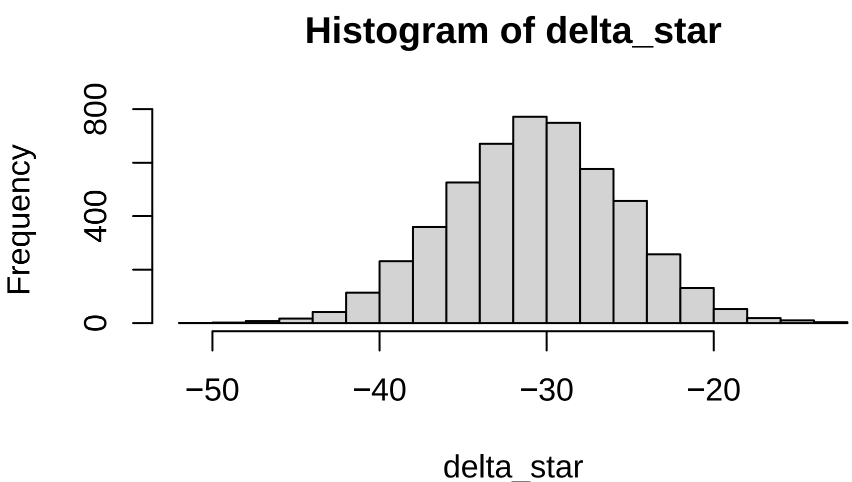
```r
x <- c(rep("AA", 342), rep("Aa", 500), rep("aa", 187))
n <- length(x)
B <- 1000
theta_hat <- (2*sum(x == "AA" ) + sum(x == "Aa"))/(2*n)

delta_star <- rep(NA, B)
for(i in 1:B){
  x_star <- sample(x, size = n, replace = T)
  delta_star[i] <- (2*sum(x_star == "AA" ) + sum(x_star == "Aa"))/(2*n) - theta_hat
}


# (approximate) bootstrap CI
a_star <- quantile(delta_star, 0.05)
b_star <- quantile(delta_star, 0.95)
ci <- theta_hat - c(b_star, a_star)
names(ci) <- c("0.05", "0.95")
ci
```

```
##      0.05      0.95
## 0.5578231 0.5942663
```

Parametric

Returning to the Old Faithful example: suppose we assume that the eruption times are Normal (even though the data clearly show that they are not. . .). For a parametric bootstrap, we will take repeated samples from a Normal distribution with mean and variance estimated from the observed data, then proceed as we did in the nonparametric bootstrap:

```r
x <- faithful$eruptions * 60
n <- length(x)
xbar <- mean(x)
s <- sd(x)
samp_med <- median(x)
B <- 5000
delta_star <- rep(NA, B)
for(i in 1:B){
  # parametric distribution
  xstar <- rnorm(n, xbar, s)
  delta_star[i] <- median(xstar) - samp_med
}
hist(delta_star)
```




```r
# (approximate) bootstrap CI
a_star <- quantile(delta_star, 0.025)
b_star <- quantile(delta_star, 0.975)
ci <- samp_med - c(b_star, a_star)
names(ci) <- c("2.5", "97.5")
ci
```

```
##      2.5     97.5
## 260.8621 280.8380
```