Class 24 in-class problems, 18.05, Spring 2022

Concept questions

Concept question 1. Which stat is easiest

Consider finding bootstrap confidence intervals for

I. *the mean* **II.** *the median* **III.** *47th percentile.*

Which is easiest to find?

(a) I (b) II (c) III (d) I and II

(e) II and III (f) I and III (g) I and II and III

Solution: (g) The program is essentially the same for all three statistics. All that needs to change is the code for computing the specific statistic.

Board questions

Problem 1. Empirical bootstrap

Data: `3 8 1 8 3 3`

Bootstrap samples (each column is one bootstrap trial):

8	8	1	8	3	8	3	1
1	3	3	1	3	8	3	3
3	1	1	8	1	3	3	8
8	1	3	1	3	3	8	8
3	3	1	8	8	3	8	3
3	8	8	3	8	3	1	1

(a) *Compute a bootstrap 80% percentile confidence interval for the mean.*

(b) *Compute a bootstrap 80% percentile confidence interval for the median.*

(a) Solution: $\overline{x} = 4.33$

$\overline{x}^*$: `4.33, 4.00, 2.83, 4.83, 4.33, 4.67, 4.33, 4.00`

Sorted

$\overline{x}^*$: `2.83, 4.00, 4.00, 4.33, 4.33, 4.33, 4.67, 4.83`

So (quantiles), $\overline{x}^*_{0.1} = 3.65$, $\overline{x}^*_{0.9} = 4.72$.

(For $\overline{x}^*_{0.1}$ we interpolated between the bottom two values. Likewise for $\overline{x}^*_{0.9}$. There are other reasonable choices. In R see the `quantile()` function.)

80% percentile bootstrap CI for mean: $[3.65, 4.72]$.

(b) Solution: $m = \text{median}(x) = 3$

m^*: `3.0, 3.0, 2.0, 5.5, 3.0, 3.0, 3.0, 3.0`

Sorted m^*: `2.0, 3.0, 3.0, 3.0, 3.0, 3.0, 3.0, 5.5`

(For $m^*_{0.1}$ we interpolated between the top two values –there are other reasonable choices. In R see the `quantile()` function.)

80% bootstrap CI for median: $[2.7, 3.75]$.

Problem 2. Parametric bootstrap

Data is taken from a Binomial(8, θ) distribution. After 6 trials, the results are

`6 5 5 5 7 4`

(a) *Estimate θ.*

(b) *Write out the R code to generate data of 100 parametric bootstrap samples and compute an 80% confidence interval for θ.*

(Try this without looking at your notes.)

(a) Solution: There are $n = 6$ data points. The MLE for θ is given by

$$\frac{\text{sum of data}}{n \cdot 8} = \frac{32}{48} = \frac{2}{3}.$$

Here are the details done abstractly to verify the formula used above. The likelihood for one trial getting k is

$$P(k \mid \theta) = \binom{8}{k} \theta^k (1-\theta)^{8-k}.$$

So the likelihood over n trials with data $k_1, \ldots, k_n$ is the product of the individual likelihoods

$$L(\theta) = c\theta^{\sum_{i=1}^n k_i} (1-\theta)^{\sum_{i=1}^n (8-k_i)}$$

Here we rolled all the binomial coefficients into one constant called c.

As usual, we look at the log likelihood

$$l(\theta) = \ln(c) + \left(\sum_{i=1}^n k_i\right) \ln(\theta) + \left(\sum_{i=1}^n (8-k_i)\right) \ln(1-\theta).$$

Taking the derivative and setting it equal to zero we get

$$l'(\theta) = \frac{\sum k_i}{\theta} - \frac{\sum (n-k_i)}{1-\theta} = 0 \;\Rightarrow\; \hat{\theta} = \frac{\sum_{i=1}^n k_i}{\sum_{i=1}^n 8} = \frac{\sum k_i}{n \cdot 8}.$$

This is what we claimed at the start of the answer.

(b) Solution: Here's the code with comments

```
data = c(6, 5, 5, 5, 7, 3)
size_binom = 8
n = length(data)
theta_hat = sum(data)/(n*size_binom) # from part a


n = length(sample) # number of sample points
# Generate the bootstrap samples using binom(size_binom, theta_hat)
# Each column is one bootstrap sample (of n resampled values)
n_boot = 100
```

```
x = rbinom(n*n_boot, size_binom, theta_hat)
bootstrap_sample = matrix(x, nrow=n, ncol=n_boot)

# Compute the bootstrap theta_star
theta_star = colSums(bootstrap_sample)/(n*size_binom)

# Compute the differences
delta_star = theta_star - theta_hat

# Find the 0.10 and 0.90 quantiles for delta_star
d = quantile(delta_star, c(0.1, 0.9))

# Calculate the 80% confidence interval for theta
ci = theta_hat - c(d[2], d[1])

s = sprintf("80%% confidence interval for theta: [%.3f, %.3f]", ci[1], ci[2])
cat(s, '\n')
```