STA3030F Module 1

Bootstrapping

Sihle Njonga

Slides credit: Dominique Katshunga

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

1. Overview
2. Bootstrap method
3. Point estimates
4. Confidence Intervals
5. Hypothesis Testing

- STA3030F is mainly a practical course and requires the use of a software.
- In this course, we will be using the statistical software R.
- Some important statistical procedures to be covered: Estimation, Confidence Interval estimates, Hypothesis testing
- Some Statistics: Mean, variance, skewness, kurtosis, median, mode, quantiles, T, F, chi-square tests, SST, SSE, correlation coefficient, regression coefficients, proportions, etc.
- Aim of bootstrapping: Find an estimate of the sampling distribution of a statistic.

1. Start with the observed set of observations (a "random sample"), denoted by $x_1, x_2, \ldots, x_n$ and calculate the sample statistic.
2. Place all observations "in a hat" and "shuffle" them. Then draw n items with replacement to form a new pseudo-sample (the bootstrap sample) denoted by $x_1^*, x_2^*, \ldots, x_n^*$.
3. Recalculate the statistic of interest for this new sample.
4. Repeat steps 1 and 2 as many times as desired.
5. The bootstrapped values form an estimate of the sampling distribution of the statistic.

Original sample		**Bootstrap sample 1**		**Bootstrap sample 2**	
i	x_i	index	x_i^*	index	x_i^*
1	37	7	43	5	39
2	42	5	39	6	35
3	38	4	44	6	35
4	44	8	41	3	38
5	39	2	42	4	44
6	35	3	38	5	39
7	43	1	37	5	39
8	41	8	41	3	38
Means:	39.875		40.625		38.375

Assumptions

1. Original sample is a representative of the population.
2. Bootstrap principle:

$$\hat{\theta}^* - \hat{\theta} \sim \hat{\theta} - \theta$$

   Behaviour of the bootstrapped statistics ($\hat{\theta}^*$) around the original sample statistic ($\hat{\theta}$) reflects the behaviour of the sample statistic ($\hat{\theta}$) around the population parameter (θ).
3. $\hat{\theta} - \theta$ is the sampling error, estimated by $\hat{\theta}^* - \hat{\theta}$

Example in R

Table 1: Delay in payments (in days) for bulk shipments of copper.

37	42	38	44	39	35	43	41	42	38
36	34	37	42	36	38	41	39	39	37
34	34	41	41	40	38	38	46	38	42

```
# Reading the data
pdelays <- c(37,42,38,44,39,35,43,41,42,38,
             36,34,37,42,36,38,41,39,39,37,
             34,34,41,41,40,38,38,46,38,42)

# Calculate the observed sample statistic
obs_mean = mean(pdelays)
obs_mean
# Perform bootstrapping, see R scripts.
```

Estimate of bias and standard errors

1. Bootstrap results can be used to find estimates of bias and standard errors $\hat{\theta}$.
2. Recall: Bias $= \mathrm{E}[\hat{\theta}] - \theta$

$$\text{Bootstrap bias estimate} = \mathrm{E}[\hat{\theta}^*] - \hat{\theta}$$

3. Standard error: standard deviation of the bootstrapped statistics

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2}$$

Bootstrap CI for μ

$$\text{percentiles } c_1 \text{ and } c_2 : \frac{\alpha}{2} \;, \; 1 - \frac{\alpha}{2}$$

$$\Pr[c_1 < \hat{\theta}^* < c_2] = 1 - \alpha$$

$$\Pr[c_1 - \hat{\theta} < \hat{\theta}^* - \hat{\theta} < c_2 - \hat{\theta}] = 1 - \alpha$$

$$\Pr[c_1 - \hat{\theta} < \hat{\theta} - \theta < c_2 - \hat{\theta}] = 1 - \alpha$$

$$\Pr[c_1 - \hat{\theta} - \hat{\theta} < -\theta < c_2 - \hat{\theta} - \hat{\theta}] = 1 - \alpha$$

$$\Pr[c_1 - 2\hat{\theta} < -\theta < c_2 - 2\hat{\theta}] = 1 - \alpha$$

$$\Pr[2\hat{\theta} - c_2 < \theta < 2\hat{\theta} - c_1] = 1 - \alpha$$

$$\left[2\hat{\theta} - c_2 \;;\; 2\hat{\theta} - c_1\right]$$

Bootstrap CI for μ

To obtain the 95% bootstrap confidence interval bounds for the population mean μ, follow the steps:

1. Generate B bootstrap samples from the original sample (in this example $B = 5000$ bootstrap samples were generated).
2. Calculate the bootstrap sample means and sort them from smallest to largest.
3. Identify the bootstrap means ranked $0.025 \times 5000 = 125$ and $0.975 \times 5000 = 4875$.

Bootstrap confidence interval for μ

Rank	1	...	125	...	4875	...	5000
Mean	...	...	37.93	...	40.03	...	...

$$\Pr[37.93 < \overline{X}^* < 40.03] = 0.95$$
$$\Pr[-1.07 < \overline{X}^* - \bar{X} < 1.03] = 0.95$$
$$\Pr[-1.07 < \overline{X} - \mu < 1.03] \approx 0.95$$

95% CI for μ : $(37.97\ ;\ 40.07)$

NB: A 95% confidence interval does not mean a probability of 0.95 that the population parameter is within the stated bounds.

Hypothesis Test

Suppose we want to test the null hypothesis $H_0 : \mu = 38$ against the alternative $H_1 : \mu > 38$.

- If the (population) mean does not actually exceed 38, i.e if H_0 is true, then the observed sample mean corresponds to a sampling error of at least $+1.0$ days
- 168 of the 5000 bootstrapped means (3.36% of the data) exceeded 40.0 (more than 1 day over the original sample mean of 39)
- Conclude that the p-value is 0.034.
- Standard t-test gives $p = 0.0415$