University of BRISTOL

spring semester 2019

Bootstrapping - Lecture 5

APPLIED STATISTICS - EMAT 30007

Nikolai Bode and Ksenia Shalonova

Department Of Engineering Mathematics

Resampling techniques

Resampling techniques are normally used to estimate parameters and confidence intervals from sample data when (1) parametric test assumptions are not met or (2) for small samples from non-normal distributions.

- Non-parametric bootstrap
- Parametric bootstrap
- Jacknife
- Permutation tests
- etc.

Non-parametric bootstrap means that only a random sample is known an no prior knowledge on the population density function.

Matlab resampling functions

- randi
- randperm
- bootci
- bootstrp
- jackknife

Bootstrap and standard error

Random sample from a population (see Monte Carlo simulations in Lecture 1)

The standard error (SE) of a statistic is the standard deviation of the sample statistic. The `standard error` can be calculated as the standard deviation of the sampling distribution

Bootstrap sample

Bootstrap sample is a random sample taken with replacement from the original sample, of the same size as the original sample. A bootstrap statistic is the statistic computed on a bootstrap sample. A bootstrap distribution is the distribution of many bootstrap statistics. The `standard error` of a statistic can be estimated using the standard deviation of the `bootstrap distribution.`

Bootstrap - why to use it? (1)

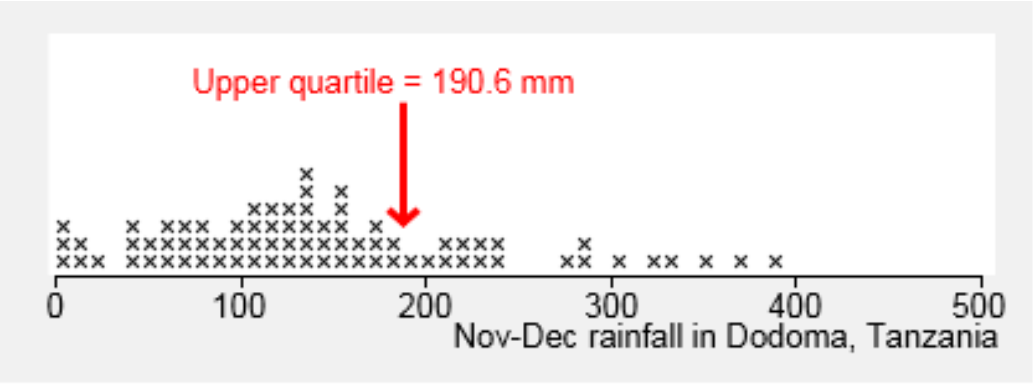
In normal population the mean μ is the parameter that is most often estimated.

- standard deviation
- interquartile range (upper quartile - lower quartile)
- median
- other percentiles (e.g. the upper quartile)

These parameters can be estimated using the corresponding summary statistic from a random sample, but the error distribution may be difficult to obtain theoretically.

Bootstrap - why to use it? (2)

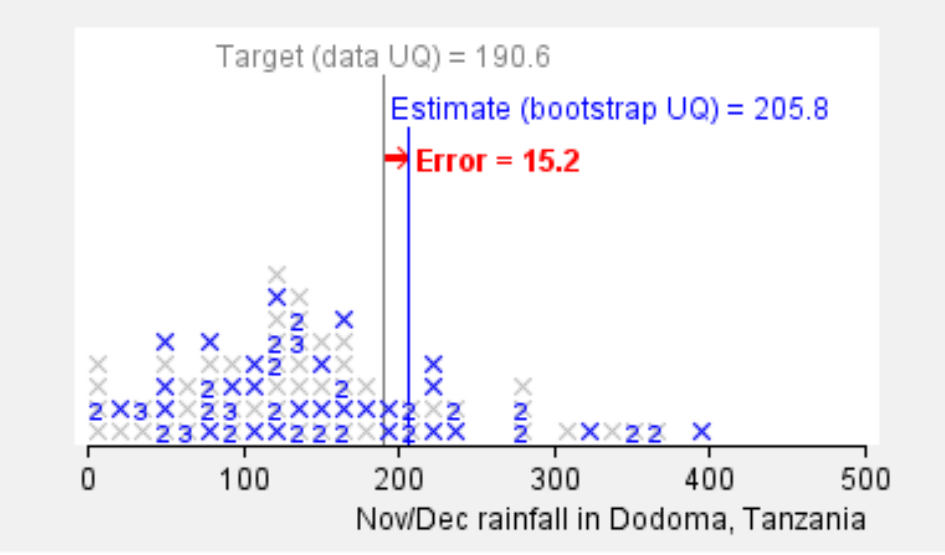
Monthly rainfall in Dodoma, Tanzania has a skew distribution in some months. The distribution of a sample is provided below.

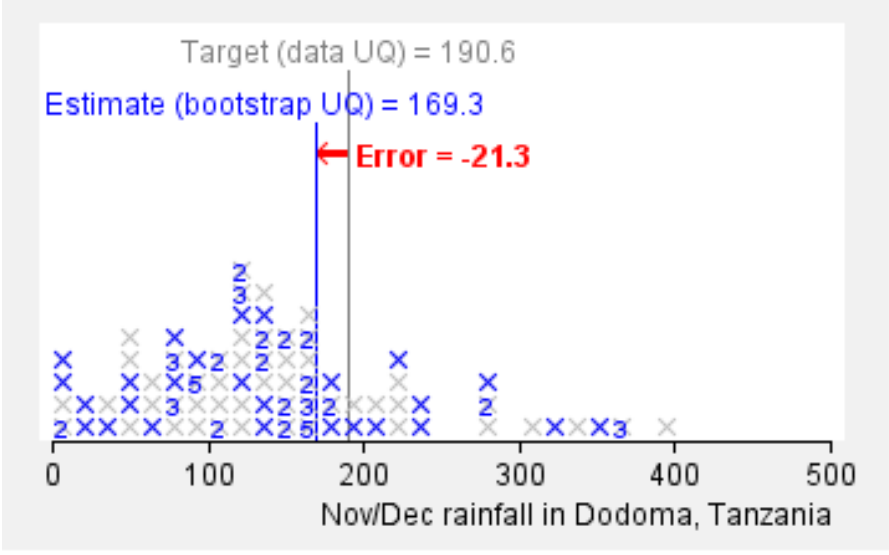


The upper quartile of the rainfall distribution is 190.6 mm, so this is our estimate of the underlying population upper quartile.
If the rainfall distribution remains the same as in the past, we can estimate rainfall of 190.6 mm or more in one out of every four November/December periods.

Bootstrap - why to use it? (3)

If a normal distribution does not seem a reasonable model, an alternative is to treat the actual sample as the 'population' for the simulation and take random samples with replacement from this sample. Such samples are called bootstrap samples. A simulation with these bootstrap samples can again show the error distribution and provide approximate values for the bias and standard error.





Bootstrap distribution

1. Let $\hat{\theta}$ is a `statistic` calculated from a sample ($\hat{\theta} = \bar{x}$).
2. We draw r observations with replacement to create a bootstrap sample and calculate `statistic` $\hat{\theta}^*$ for this sample.

Bootstrap standard error: is the sample standard deviation of the bootstrap distribution: $SE_b = \sqrt{\frac{\sum(\hat{\theta_b^*} - \bar{\theta^*})^2}{B-1}}$ where B is the number of bootstrap replications (usually $B > 10000$).

Bootstrap bias: $\bar{\theta^*} - \hat{\theta}$

Bootstrap confidence intervals: bootstrap percentile interval, t confidence interval with bootstrap standard error, bootstrap t-interval etc.

Bootstrap using t CI - NOT recommended

$\hat{\theta} \pm t_{\alpha/2} SE_b$

Bootstrap `standard error` is the sample standard deviation of the bootstrap distribution: $SE_b = \sqrt{\frac{\sum(\hat{\theta_b^*} - \bar{\theta^*})^2}{B-1}}$ whee B is the number of draws from the sample (usually $B > 10000$).

Bootstrap `bias`: $\bar{\theta^*} - \hat{\theta}$

Can be useful when standard error is difficult to derive

Poor performance if distributions are highly skewed

Bootstrap percentile CI or Efron method (described in the book recommended for this unit)

For a 90% confidence interval keep the middle 90%, leaving 5% in each tail and 5% in the head. The 90% confidence interval boundaries would be 5th percentile and 95th percentile.

In case we have 10000 bootstrap replications: $\theta^*_1 \leq \theta^*_2 \leq \theta^*_3 \leq ... \leq \theta^*_{10000}$
the 90% CI is: $[\theta^*_{500}, \theta^*_{9500}]$

Disadvantages: Can be too narrow for small samples.

Advantages: A very intuitive and easy to implement method. Can also outperform some other bootstrap CI methods for skewed distributions.

Bootstrap percentile CI or Efron method

- Calculate: $\hat{\theta}_b^*$
- Repeat: B times
- Get distribution: $\left\{\hat{\theta}_b^*\right\}_{b=1}^{B}$
- CI interval: $\theta \in \left[q_{\alpha/2}, q_{1-\alpha/2}\right]$

Bootstrap CI - bootstrap t (NOT the same as bootstrap using t confidence intervals)

- Calculate: $\frac{\hat{\theta}_b^* - \hat{\theta}}{SE(\hat{\theta}_b^*)}$
- Repeat: B times
- Get distribution: $\left\{ \frac{\hat{\theta}_b^* - \hat{\theta}}{SE(\hat{\theta}_b^*)} \right\}_{b=1}^{B}$
- CI interval: $\theta \in [\hat{\theta} - SE(\hat{\theta}) * q_{1-\alpha/2}, \hat{\theta} - SE(\hat{\theta}) * q_{\alpha/2}]$

Bootstrap CI symmetric t-percentile - appropriate for hypothesis testing

- Calculate: $\frac{\hat{\theta}_b^* - \hat{\theta}}{SE(\hat{\theta}_b^*)}$
- Repeat: B times
- Get distribution: $\left\{ \frac{\hat{\theta}_b^* - \hat{\theta}}{SE(\hat{\theta}_b^*)} \right\}_{b=1}^{B}$
- CI interval: $\theta \in [\hat{\theta} - SE(\hat{\theta}) * q_{1-\alpha}, \hat{\theta} + SE(\hat{\theta}) * q_{1-\alpha}]$

Bootstrap CI Hall method

- Calculate: $\hat{\theta}_b^* - \hat{\theta}$
- Repeat: B times
- Get distribution: $\left\{\hat{\theta}_b^* - \hat{\theta}\right\}_{b=1}^{B}$
- CI interval: $\theta \in [\hat{\theta} - q_{1-\alpha/2}, \hat{\theta} - q_{\alpha/2}]$

Bootstrap - a word of warning

The limitation of the bootstrap is the assumption that the distribution of the data represented by one sample is an accurate estimate of the population distribution. If the sample does not reflect the population distribution, then the random sampling performed in the bootstrap procedure may add another level of sampling error, resulting in inaccurate statistical estimations.

It is important to get quality data that accurately reflect the population being sampled. The smaller the original sample, the less likely it is to accurately represent the entire population.