

## Terminology

- $\hat{\theta}$  estimates a population parameter  $\theta$ , where  $\hat{\theta}$  is a function of our data

For one example, let our data be  $X_1, \dots, X_n$  iid  $N(\mu, \sigma^2)$ . Let  $\hat{\theta} = \bar{X}$ , which is a function of our data, which we know is an estimator for  $\theta = \mu$ .

For a second example, let our data be  $X_1, \dots, X_n$  iid  $U(0, \theta)$ . Let  $\hat{\theta} = \max(X_1, \dots, X_n)$ , which is a function of our data, which is an estimator for  $\theta$ .

- $\text{var}(\hat{\theta})$  is the variance of this estimator

For example, let our data be  $X_1, \dots, X_n$  iid  $N(\mu, \sigma^2)$ . Let  $\hat{\theta} = \bar{X}$ ,

$$\text{var}(\hat{\theta}) = \text{var}(\bar{X}) = \frac{\sigma^2}{n}$$

- $\text{bias}(\hat{\theta})$  is the bias of this estimator

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

For example, let our data be  $X_1, \dots, X_n$  iid  $N(\mu, \sigma^2)$ . Let  $\hat{\theta} = \bar{X}$ ,

$$\text{bias}(\hat{\theta}) = \text{bias}(\bar{X}) = E(\bar{X}) - \mu = 0$$

- $\text{MSE}(\hat{\theta})$  is the mean-squared error of this estimator

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

$$\text{MSE}(\hat{\theta}) = [\text{bias}(\hat{\theta})]^2 + \text{var}(\hat{\theta})$$

For example, let our data be  $X_1, \dots, X_n$  iid  $N(\mu, \sigma^2)$ . Let  $\hat{\theta} = \bar{X}$ ,

$$\text{MSE}(\hat{\theta}) = \text{MSE}(\bar{X}) = E[(\bar{X} - \mu)^2]$$

$$\text{MSE}(\bar{X}) = [\text{bias}(\bar{X})]^2 + \text{var}(\bar{X})$$

$$\text{MSE}(\bar{X}) = [0]^2 + \frac{\sigma^2}{n}$$

$$\text{MSE}(\bar{X}) = \frac{\sigma^2}{n}$$

1. Use simulation to estimate the bias, variance and MSE when using the sample median to estimate the population **mean** for a uniformly distributed population between 0 and 1. Consider sample sizes 10, 20, 40, and 80.

```

reps <- 1000

bias_med <- NULL
Var_med <- NULL
MSE_med <- NULL

bias_mean <- NULL
Var_mean <- NULL
MSE_mean <- NULL

populationMean <- 1/2

for (n in c(10,20,40,80))
{
  Z <- array(runif(n*reps), c(n,reps))
  Med <- apply(Z, 2, median)
  Mean <- apply(Z, 2, mean)

  bias_med <- c(bias_med, mean(Med) - populationMean)
  Var_med <- c(Var_med, var(Med))
  MSE_med <- c(MSE_med, mean((Med-populationMean)^2))

  bias_mean <- c(bias_mean, mean(Mean) - populationMean)
  Var_mean <- c(Var_mean, var(Mean))
  MSE_mean <- c(MSE_mean, mean((Mean-populationMean)^2))
}

```

## Non-Parametric Bootstrap

Say we only have  $n$  data observations, like the heights of 10 females. We assume that these observations are realizations of  $n$  iid random variables. We could use these 10 observations to compute some statistics, like estimate the mean of the population of female heights, or estimate the maximum of the population of female heights.

Whenever we compute a statistic, like the maximum, we would like to know what the distribution of our statistic is (the sampling distribution), so we can say something meaningful about this particular estimate. If we don't know what the sampling distribution is, we could take advantage of the non-parametric bootstrap to artificially simulate it.

Given *observed* data  $(d_1, \dots, d_n)$ , we form thousands of bootstrap datasets. Each bootstrap dataset will have  $n$  observations, where each of the  $n$  observations was randomly selected from  $(d_1, \dots, d_n)$  with replacement.

1. Let  $X_1, \dots, X_n \sim U(\theta, 10)$ , where  $\theta < 10$ . Use the non-parametric Bootstrap to estimate the relative bias of the estimator  $\hat{\theta}$ :

$$\hat{\theta} = \min(X_1, \dots, X_n)$$

We define relative bias as:

$$RB = \frac{E\hat{\theta} - \theta}{\theta}$$

```

reps = 100
nboot = 1000

theta.POP = 2

SS = c(5,10,20)
Rel.Bias = NULL
for (k in 1:3)
{
  n = SS[k]
  rel.bias = NULL
  for (r in 1:reps)
  {
    ## Generate n observations from Unif(theta, 10)
    X = (10-theta.POP)*runif(n)+theta.POP

    ## Get the "theta" for the RB formula
    theta.X = min(X)

    ## Generate the artificial data from X
    ii = ceiling(n*runif(nboot*n))
    Xboot = X[ii]
    Xboot = array(Xboot, c(nboot, n))

    ## Compute nboot "thetaHat"'s for the RB formula
    ## from this artificial data
    theta.BOOT = apply(Xboot, 1, min)

    ## Compute the Relative Bias formula
    rel.bias[r] = (mean(theta.BOOT)-theta.X)/theta.X
  }
  Rel.Bias[k] = mean(rel.bias)
}

```

## Parametric Bootstrap

Given *observed* data  $(d_1, \dots, d_n)$ , which are realization of iid random variables with a known distribution type. Take these data and estimate the parameters of this known distribution, and then generate artificial data from this distribution with parameters set to these estimates.

Here is the parametric bootstrap estimate of the relative bias from the last problem:

```
reps = 100
nboot = 1000

theta.POP = 2

SS = c(5,10,20)
Rel.Bias = NULL
for (k in 1:3)
{
  n = SS[k]
  rel.bias = NULL
  for (r in 1:reps)
  {
    ## Generate n observations from Unif(theta, 10)
    X = (10-theta.POP)*runif(n)+theta.POP

    ## Get the estimate of "theta"
    theta.X = min(X)

    ## Generate the artificial Uniform data using this estimate
    Xboot = (10-theta.X)*runif(nboot*n)+theta.X
    Xboot = array(Xboot, c(nboot, n))

    ## Compute nboot "thetaHat"'s for the RB formula
    ## from this artificial data
    theta.BOOT = apply(Xboot, 1, min)

    ## Compute the Relative Bias formula
    rel.bias[r] = (mean(theta.BOOT)-theta.X)/theta.X
  }
  Rel.Bias[k] = mean(rel.bias)
}
```