Experimental Statistics for Engineers I

Kamala Dadashova

Problem 1: Let $X_1, X_2, ...X_n$ be a random sample of size n from some distribution with mean μ and variance σ^2.

(a) Show that the sample variance is an unbiased estimator of σ^2.

Solution:

We want to show that the estimator

$$\hat{\sigma}^2 = S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$$

is unbiased for estimation σ^2. Let X be any random variable with $V(X) = E(X^2) - [E(X)]^2$, i.e, $E(X^2) = V(X) + [E(X)]^2$

Firstly,

$$\begin{aligned} E[\sum_i (X_i - \bar{X})^2] =& E[\sum_i (X_i^2 - 2X_i\bar{X} + \bar{X}^2)] = E[\sum_i X_i^2 - \sum_i 2X_i\bar{X} + \sum_i \bar{X}^2] = \\ =& E[\sum_i X_i^2 - 2\bar{X}\sum_i X_i + \sum_i \bar{X}^2] = E[\sum_i X_i^2 - 2n\bar{X}^2 + n\bar{X}^2] = \\ =& E[\sum_i X_i^2 - n\bar{X}^2] = \sum_i E[X_i^2] - E[n\bar{X}^2] \end{aligned}$$

where we used the fact that $\bar{X} = \frac{\sum_i X_i}{n}$. Next,

$$E[X_i]^2 = \sigma^2 + \mu^2$$

$$E[\bar{X}]^2 = \frac{\sigma^2}{n} + \mu^2.$$

By substituting, above expression is equal to

$$= \sum_i (\sigma^2 + \mu^2) - n \cdot (\frac{\sigma^2}{n} + \mu^2) = n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 = \sigma^2(n-1)$$

Hence, we show

$$\frac{1}{n-1}E[\sum_i (X_i - \bar{X})^2] = \sigma^2$$

i.e $E(\hat{\sigma}^2) = \sigma^2$. So, $\hat{\sigma}^2$ is unbiased.

(b) Is $\bar{X}^2$ an unbiased estimator for μ^2? If not, what is the bias of $\bar{X}^2$?

Solution:

Note that

$$E(\bar{X}^2) = V(\bar{X}) + [E(\bar{X})]^2 = \frac{\sigma^2}{n} + \mu^2 \neq \mu^2$$

So, $\bar{X}^2$ an biased estimator for μ^2. Bias term is $E(\bar{X}^2) - \mu^2 = \frac{\sigma^2}{n}$.

(c) How can we obtain an unbiased estimator for μ^2?

Using part (a) and part(b), one unbiased estimator can be

$$\bar{X}^2 - \frac{S^2}{n}$$

Because $E(\bar{X}^2 - \frac{S^2}{n}) = E(\bar{X}^2) - \frac{1}{n}E(S^2) = E(\bar{X}^2) - \frac{\sigma^2}{n} = \mu^2$

Problem 2: Let $X_1, X_2, ...X_7$ be a random sample from the normal model with mean $\mu = 5$ and variance $\sigma^2 = 1$. Consider the following two estimators of μ:

$$\hat{\theta}_1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7}{7}$$

$$\hat{\theta}_2 = \frac{2X_1 - X_2 + X_7}{2}$$

(a) Is either estimator unbiased?

Solution:

We want to check whether $E(\hat{\theta}) = \theta$.

$$\begin{aligned} E(\frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7}{7}) &= \frac{1}{7}E(X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7) = \\ &= \frac{1}{7}(E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) + E(X_6) \\ &+ E(X_7)) = \frac{1}{7} \cdot 7\mu = \mu \end{aligned}$$

Conclusion: θ_1 is is unbiased.

$$E(\frac{2X_1 - X_2 + X_7}{2}) = \frac{1}{2}E(2X_1 - X_2 + X_7) = \frac{1}{2}(2E(X_1) - E(X_2) + E(X_7) = \frac{1}{2} \cdot 2\mu = \mu$$

Conclusion: θ_2 is is unbiased.

(b) Conduct a Bootstrap simulation to compare the estimators (B =1000). Which of the two estimators do you prefer and why?

Solution:

We attach the R code result at the end of this file. According to it, θ_1 is more accurate because the mean and standard deviation estimation for this parameter is much closer to true values.

Problem 3: The pmf for the binomial model is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

for x = 0, 1, . . . , n and 0 otherwise

(a) Assuming that n is known, find the method of moments estimator for p.

Solution:

From binomial distribution, we know that

$$E(X) = np$$

$$V(X) = np(1-p)$$

$$np = \bar{X} = \frac{1}{n} \sum X_i$$

Since n is known, we can find estimator for p,

$$\hat{p} = \frac{\bar{X}}{n}$$

(b) Assuming that both n and p are unknown (i.e., both are parameters), find the method of moments estimator for n and p.

Solution:

From binomial distribution, we know that

$$E(X) = np$$

$$V(X) = np(1-p)$$

$$np = \bar{X} = \frac{1}{n} \sum X_i$$

$$np(1-p) = \frac{1}{n} \sum (X_i - \bar{X})^2$$

We need to solve this system,

$$(1-p) = \frac{\sum (X_i - \bar{X})^2}{\sum X_i}$$

$$\hat{p} = 1 - \frac{\sum (X_i - \bar{X})^2}{\sum X_i}$$

Then,

$$n\hat{p} = \bar{X}$$

$$\hat{n} = \frac{\bar{X}\sum X_i}{\sum X_i - \sum(X_i - \bar{X})^2}$$

Problem 4: Let X be the proportion of allotted time that a randomly selected student spends working on a certain aptitude test. Suppose the pdf of X is

$$f(x;\theta) = (1+\theta)x^{\theta}, \quad 0 \le x \le 1$$

for $\theta > -1$. A random sample of 10 students yields data:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0.92	0.79	0.90	0.65	0.86	0.47	0.73	0.97	0.94	0.77

Table 1

(a) Use the MOM to obtain an estimator of θ; compute an estimate for the given data.

Solution:

Firstly, we find expected value

$$E\left(x\right) = \int_0^1 xf\left(x\right)dx = \int_0^1 x\left(\theta+1\right)x^{\theta}dx = \int_0^1 \left(\theta+1\right)x^{\theta+1}dx = \frac{(\theta+1)}{(\theta+2)}\left[x^{\theta+2}\right]_0^1 = \frac{(\theta+1)}{(\theta+2)}$$

The estimator for θ:

$$\frac{(\theta+1)}{(\theta+2)} = \bar{X}$$

$$\theta + 1 = \bar{X}\left(\theta+2\right)$$

$$\theta\left(1-\bar{X}\right) = 2\bar{X} - 1$$

$$\theta = \frac{2\bar{X}-1}{1-\bar{X}}$$

Next, we find mean value of data given in Table 1,

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{8}{10} = 0.8$$

Plugging it in estimator, we get estimate of parameter θ

$$\hat{\theta} = \frac{2\cdot\bar{X}-1}{1-\bar{X}} = \frac{2\cdot 0.8 - 1}{1-0.8} = 3$$

(b) Obtain the MLE of θ and then compute an estimate for the given data.

Solution:

Likelihood for θ is

$$L = f\left(x_{,1}, x_2, ...x_{10};\ \theta\right)$$

$$L = (\theta+1)\, x_1^{\theta} \times (\theta+1)\, x_2^{\theta} \times (\theta+1)\, x_3^{\theta} \times (\theta+1)\, x_4^{\theta} ... (\theta+1)\, x_{10}^{\theta}$$

$$L = (\theta+1)^{10} \prod_{i=1}^{10} (x_i)^{\theta}$$

$$\log(L) = 10 \ln(\theta+1) + \ln \prod_{i=1}^{10} (x_i)^{\theta}$$

$$\log(L) = 10 \ln(\theta+1) + \theta \sum_{i=1}^{10} \ln(x_i)$$

We take derivative of the last expression and set it equal to zero,

$$\frac{d(\log L)}{d\theta} = 0$$

$$\frac{10}{\theta+1} + \sum_{i=1}^{10} \ln(x_i) = 0$$

$$\theta = \frac{-10 - \sum\limits_{i=1}^{10} \ln(x_i)}{\sum\limits_{i=1}^{10} \ln(x_i)}$$

We plug in each values of x_i in the table into the last expression and get

$$\theta = \frac{-10 + 2.4295}{-2.4295} = 3.12$$

Problem 5: In class we derived the MOM and MLE for an exponential distribution with parameter λ. Conduct a Bootstrap simulation to compare the estimation of λ with sample sizes of n = 10, n = 100, and n = 500. Choose true value $\lambda = 0.2$ and use B = 1000. Calculate and compare the mean and standard error for each set of simulations to each other as well as their theoretical values.
Solution:
The code is provided below. As we observe we get more accurate result when n=500.

CODES

Problem 2 (part(b))

```
set.seed(2)
n <- 7
base.smpl <- rnorm(n,5,1) # create sample of size 7 from N(5,1)

B=1000  # run 1000 bootstrap samples
means <- rep(0,B)
means2<-rep(0,B)
for(i in 1:B){
  rsample <- sample(base.smpl,n,replace=T)
  means[i] <- mean(rsample)
  means2[i]<-(2*rsample[1]-rsample[2]+rsample[7])/2
}
x_bar_bar <- mean(means)  # calculate mean and standard error estimates
x2_bar_bar<-mean(means2)
se_est <- sd(means)
se_est2 <- sd(means2)
se <- 1/sqrt(n) # calculate true standard error
x_bar_bar
```

```
## [1] 5.081006
```

```
x2_bar_bar
```

```
## [1] 5.100792
```

```
se_est
```

```
## [1] 0.3251693
```

```
se_est2
```

```
## [1] 1.021624
```

```
se
```

```
## [1] 0.3779645
```

Problem 5:

```
set.seed(30)
n <- 10
base.smpl <- rexp(n,0.2) # create sample of size 10 from Exp(n=10,rate=0.2)

B=1000  # run 1000 bootstrap samples
means <- rep(0,B)  # create dummy matric for means
for(i in 1:B){
  rsample <- sample(base.smpl,n,replace=T) # sample n=10 from population with replacement
  means[i] <- mean(rsample)  # calculate mean and save in vector
}
x_bar_bar <- mean(means)  # calculate mean and standard error estimates
```

```r
se_est <- sd(means)
se <- 5/sqrt(n) # calculate true standard error
x_bar_bar
```

```
## [1] 5.60815
```

```r
se_est
```

```
## [1] 1.340186
```

```r
se
```

```
## [1] 1.581139
```

```r
set.seed(30)
n <- 100
base.smpl <- rexp(n,0.2) # create sample of size 10 from Exp(n=10,rate=0.2)

B=1000  # run 1000 bootstrap samples
means <- rep(0,B)  # create dummy matric for means
for(i in 1:B){
  rsample <- sample(base.smpl,n,replace=T) # sample n=10 from population with replacement
  means[i] <- mean(rsample)  # calculate mean and save in vector
}
x_bar_bar <- mean(means)  # calculate mean and standard error estimates
se_est <- sd(means)
se <- 5/sqrt(n) # calculate true standard error
x_bar_bar
```

```
## [1] 4.482176
```

```r
se_est
```

```
## [1] 0.3986888
```

```r
se
```

```
## [1] 0.5
```

```r
set.seed(30)
n <- 500
base.smpl <- rexp(n,0.2) # create sample of size 10 from Exp(n=10,rate=0.2)

B=1000  # run 1000 bootstrap samples
means <- rep(0,B)  # create dummy matric for means
for(i in 1:B){
  rsample <- sample(base.smpl,n,replace=T) # sample n=10 from population with replacement
  means[i] <- mean(rsample)  # calculate mean and save in vector
}
x_bar_bar <- mean(means)  # calculate mean and standard error estimates
se_est <- sd(means)
se <- 5/sqrt(n) # calculate true standard error
x_bar_bar
```

```
## [1] 4.935456
```

```r
se_est
```

```
## [1] 0.2260088
```

```
se
```

```
## [1] 0.2236068
```