Homework 8 MSE

Exercise 1

Let $X_1, X_2, ..., X_n$ be an i.i.d. sample from the following discrete probability mass function:

x	1	3	5	7
$f(x)$	$\frac{2(1-p)}{3}$	$\frac{2p}{3}$	$\frac{(1-p)}{3}$	$\frac{p}{3}$

Let n_1 represent the number of 1s in the sample, n_3 represent the number of 3s in the sample, etc.

a) (1 pt) Find an expression for the Maximum Likelihood Estimator of p, $\hat{p}$.

b) (1 pt) Find an expression for Bias of the MLE, $Bias(\hat{p})$.

c) (1 pt) Find an expression for the Method of Moments estimator of p, $\tilde{p}$.

d) (0.5 pt) Find an expression for Bias of the MoM Estimator, $Bias(\tilde{p})$.

e) (1.5 pt) Find an expression for MSE of the MoM Estimator, $MSE(\tilde{p})$.

a).
$$L(p) = \prod_{i=1}^{n_1} \frac{2(1-p)}{3} \cdot \prod_{i=1}^{n_3} \frac{2p}{3} \cdot \prod_{i=1}^{n_5} \frac{(1-p)}{3} \cdot \prod_{i=1}^{n_7} \frac{p}{3} = \left(\frac{2}{3}\right)^{n_1+n_3} \cdot \left(\frac{1}{3}\right)^{n_5+n_7} \cdot (1-p)^{n_1+n_5} \cdot p^{n_3+n_7}$$
$$\Downarrow$$
$$\ln[L(p)] = (n_1+n_3)\cdot \ln\frac{2}{3} + (n_5+n_7)\cdot \ln\frac{1}{3} + (n_1+n_5)\cdot \ln(1-p) + (n_3+n_7)\cdot \ln p$$
$$\Downarrow$$
$$\frac{\partial \ln[L(p)]}{\partial p} = 0 + 0 - \frac{n_1+n_5}{1-p} + \frac{n_3+n_7}{p} \stackrel{set}{=\!=} 0$$
$$\Downarrow$$
$$\hat{p} = \frac{n_3+n_7}{n_3+n_7+n_1+n_5} = \frac{n_3+n_7}{n}$$

2

b).
$$Bias[\hat{p}] = E[\hat{p}] - p = E\left[\frac{n_3+n_7}{n}\right] - p = \frac{n \cdot \frac{2p}{3} + n \cdot \frac{p}{3}}{n} - p = 0$$

c).
$$E[X] = \sum X \cdot P(X=x) = \frac{2(1-p)}{3} + \frac{6p}{3} + \frac{5(1-p)}{3} + \frac{7p}{3} = \bar{X} \implies \tilde{p} = \frac{3\bar{X}-7}{6}$$

$\mu_X = \mu_{\bar{X}}$ ☆

d).
$$Bias[\tilde{p}] = E[\tilde{p}] - p = E\left[\frac{3\bar{X}-7}{6}\right] - p = \frac{1}{2}E[\bar{X}] - \frac{7}{6} - p$$
$$= \frac{1}{2}E\left[\frac{6p+7}{3}\right] - \frac{7}{6} - p$$
$$= p + \frac{7}{6} - \frac{7}{6} - p = 0$$

e).
$$MSE[\tilde{p}] = Var[\tilde{p}] + Bias^2[\tilde{p}]$$
$$= Var\left[\frac{3\bar{X}-7}{6}\right] + 0$$
$$= \frac{1}{4}Var[\bar{X}] \qquad \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$$
$$= \frac{1}{4} \cdot \frac{Var[X]}{n}$$

$$Var[X] = E[X^2] - (E[X])^2 = 1 \cdot \frac{2(1-p)}{3} + 9 \cdot \frac{2p}{3} + 25 \cdot \frac{(1-p)}{3} + 49 \cdot \frac{p}{3} - \left[\frac{6p+7}{3}\right]^2$$
$$= -4p^2 + 4p + \frac{32}{9}$$

$$\therefore MSE[\tilde{p}] = \frac{1}{4n}\left[-4p^2 + 4p + \frac{32}{9}\right] = \frac{1}{n}\left[-p^2 + p + \frac{8}{9}\right]$$

Exercise 2

Hannibal wants to estimate the true average brain mass, μ, but he is locked in a cell with no internet and only has a scale. He collects a small random sample of size $n = 5$ from nearby prison guards. He find that these brains have the following weights:

$$1350g, \quad 1400g, \quad 1300g, \quad 1460g, \quad 1350g$$

Suppose that human brains follow a normal distribution.

a) (0.5 pt) Compute the sample standard deviation, s, of these brains by hand (with a calculator). You may only use $+$, $-$, $\times$, $\div$, and $\sqrt{}$ on a calculator. Show **all** your work.

b) (1 pt) Construct a 90% confidence interval for the true mean weight of brains, μ.

c) (0.5 pt) Construct a 95% confidence interval for the true mean weight of brains, μ.

a). $\bar{X} = \frac{1350 + 1400 + 1300 + 1460 + 1350}{5} = 1372$

$$S^2 = \frac{1}{n-1}\sum_{n=1}^{5}(x_i - \bar{x})^2 = \frac{1}{4}\left[(1350-1372)^2 + (1400-1372)^2 + (1300-1372)^2 + (1460-1372)^2 + (1350-1372)^2\right] = 3670$$

$$S = \sqrt{S^2} = \sqrt{3670} = \boxed{60.58}$$

b). $n = 5$ $\quad \frac{\alpha}{2} = 0.05 \quad \bar{X} = 1372 \quad S = \sqrt{3670}$

$t_{\frac{\alpha}{2}} = 2.132$

$$CI: \left(\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right) = (1314.2, 1429.8)$$

c). $n = 5$ $\quad \frac{\alpha}{2} = 0.025 \quad \bar{X} = 1372 \quad S = \sqrt{3670}$

$t_{\frac{\alpha}{2}} = 2.776$

$$CI: \left(\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right) = (1295.8, 1446.2)$$

Exercise 3 (R) Show all code and relevant output

Using R, generate a random sample of size n=*10* from the following Normal Distribution: $N(\mu = 100, \sigma^2 = 25)$.

a) (0.25 pt) Use R to find the sample mean $\bar{X}$, and sample variance s^2. Show your output.

b) (0.5 pt) **Assume that the true variance is unknown.** Using the sample variance, generate a 95% confidence interval for μ.

c) (1 pt) Generate 500 independent samples of size n=*10* from the same Normal distribution above. Use each of these samples to generate a 95% confidence interval for μ. (you will generate 500 different intervals) How many of your confidence intervals captured the true parameter value for μ? Show your code and final output (answer) for this step. (You do not need to output all the confidence intervals)

d) (0.75 pt) Now assume that the true variance is **known**, ($\sigma^2 = 25$). Repeat part (c), generate 500 new samples and 500 new confidence intervals for μ. How many of your confidence intervals captured the true parameter value for μ? Show your code and final output for this step.

e) (0.5pt) Compare the intervals from part (c) and part (d), which one is larger on average?

a).

```r
#Original parameter
n=10
mu=100
sigma=5

#Generating data set
sample=rnorm(n, mu, sigma)

#a).
sample_mean=mean(sample)
sample_variance=var(sample)
```

```
> sample_mean
[1] 101.6266
> sample_variance
[1] 24.83376
```

b).

```r
#b).
alpha=1-0.95
t=qt(1-alpha/2,n-1)
LB=mean(sample)-t*sd(sample)/sqrt(n)
UB=mean(sample)+t*sd(sample)/sqrt(n)
CI=c(LB,UB)
```

```
> CI
[1]  98.06172 105.19147
```

c).

```r
#c).
size=500
num1=0
alpha=1-0.95
UB_ave=0
LB_ave=0
for (i in 1:size){
  sample=rnorm(n, mu, sigma)
  t=qt(1-alpha/2,n-1)
  LB=mean(sample)-t*sd(sample)/sqrt(n)
  UB=mean(sample)+t*sd(sample)/sqrt(n)
  LB_ave=LB_ave+LB
  UB_ave=UB_ave+UB
  if ((LB<mu) && (mu<UB)){num1=num1+1}
}
UB_ave=UB_ave/size
LB_ave=LB_ave/size
CI1_ave=UB_ave-LB_ave#c(LB_ave,UB_ave)
```

```
> num1/size
[1] 0.962
```

d).

```r
#d).
size=500
num2=0
alpha=1-0.95
UB_ave=0
LB_ave=0
for (i in 1:size){
  sample=rnorm(n, mu, sigma)
  z=qnorm(1-alpha/2)
  LB=mean(sample)-z*sigma/sqrt(n)
  UB=mean(sample)+z*sigma/sqrt(n)
  LB_ave=LB_ave+LB
  UB_ave=UB_ave+UB
  if ((LB<mu) && (mu<UB)){num2=num2+1}
}
UB_ave=UB_ave/size
LB_ave=LB_ave/size
CI2_ave=UB_ave-LB_ave#c(LB_ave,UB_ave)
```

```
> num2/size
[1] 0.95
```

e). The program for computing the average length of CI is included in c). and d). And from the Output we see CI1>CI2, which means the interval for **unknown is larger than the known**.

```
> CI1_ave
[1] 6.90212
> CI2_ave
[1] 6.19795
```