MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.s077 Spring 2018

Problem Set 2 due Tuesday 2/27, in class

Problem 1. Bias, variance, and MSE practice.

(a) Using the linearity of expectation, we have,

$$\mathbb{E}[\widehat{M}] = \mathbb{E}\left[\frac{1 + X_1 + \cdots + X_n}{n+1}\right] = \frac{\mathbb{E}[1 + X_1 + \cdots + X_n]}{n+1} = \frac{1 + \mu n}{n+1}.$$

Hence, the bias is,

$$\mathbb{E}[\hat{M}] - \mu = \frac{1-\mu}{n+1} \to 0, \text{ as } n \to \infty.$$

(b) Now, we will make use of the following facts:

- For independent random variables, the sum of their variances is equal to the variance of their sum.
- For any random variable X, and any constant c, $\text{var}(cX) = c^2\text{var}(X)$.
- A constant random variable is independent of any other random variable.

Using these facts, we have

$$\begin{aligned}\text{var}(\hat{M}) &= \text{var}\left(\frac{1 + X_1 + \cdots + X_n}{n+1}\right) \\ &= \frac{1}{(n+1)^2}\text{var}(X_1 + \cdots + X_n) \\ &= \frac{n\sigma^2}{(n+1)^2}.\end{aligned}$$

(c) Using the fact that the MSE of an estimator is the sum of its bias squared and its variance, we have,

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

The bias term (first term above) is of order of $\frac{1}{n^2}$, while the variance is of order of $\frac{1}{n}$. Hence, the contribution of the bias is asymptotically smaller and negligible.

Problem 2. The most common estimation problem.

(a) First, since $X_i = \mu + \sigma W_i$,

$$\widehat{M} = \frac{\sum_{i=1}^n X_i}{n} = \frac{n\mu + \sigma \sum_{i=1}^n W_i}{n} = \mu + \sigma \frac{1}{n} \sum_{i=1}^n W_i.$$

In particular, $\widehat{M} - \mu = \frac{\sigma}{n} \sum_{i=1}^n W_i$. We now express $X_i - \widehat{M}$ as follows.

$$\begin{aligned} X_i - \widehat{M} &= \mu + \sigma W_i - \mu - \frac{\sigma}{n} \sum_{j=1}^n W_j \\ &= \sigma \left(\frac{(n-1)W_i - \sum_{j \neq i} W_j}{n} \right). \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}[(X_i - \widehat{M})^2] &= \mathrm{var}(X_i - \hat{M}) \\ &= \frac{\sigma^2}{n^2} \left((n-1)^2 + (n-1) \right) \\ &= \frac{\sigma^2 (n-1)}{n}, \end{aligned}$$

where, $(n-1)^2$ stands for the variance of $(n-1)W_i$, and, the remaining terms are just $\sum_j \mathrm{var} W_j$. Thus,

$$\mathbb{E}[V] = \frac{1}{n-1} \sum_{i=1}^n \frac{(n-1)\sigma^2}{n} = \sigma^2 = v.$$

(b) Begin by inspecting, $\hat{M} - \mu$. We have,

$$\hat{M} - \mu = \frac{1}{n} \sum_{i=1}^n (\mu + \sigma W_i) - \mu = \sigma \frac{(W_1 + \cdots + W_n)}{n}.$$

We next compute $\widehat{\sigma}$ as follows.

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(\mu + \sigma W_i - \frac{1}{n} \sum_{j=1}^n (\mu + \sigma W_j) \right)^2} = \sigma \sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(W_i - \frac{1}{n} \sum_{j=1}^n W_j \right)^2}.$$

Hence,

$$T = \frac{\sqrt{n}(\widehat{M} - \mu)}{\widehat{\sigma}} = \frac{W_1 + \cdots + W_n}{\sqrt{n} \cdot \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(W_i - \frac{1}{n} \sum_{j=1}^{n} W_j \right)^2}},$$

which does not depend on σ and μ.

(c) Note that, $0.025 + 0.025 = 0.05 = 5\%$. Hence,

$$\mathbb{P}(-2 < T < 3) = 0.95.$$

Plugging in $\dfrac{\sqrt{n}(\hat{M} - \mu)}{\hat{\sigma}}$ for T, we arrive at,

$$\mathbb{P}\left(\widehat{M} - \frac{3\widehat{\sigma}}{\sqrt{n}} \leq \mu \leq \widehat{M} + \frac{2\widehat{\sigma}}{\sqrt{n}} \right) = 0.95.$$

i.e.,

$$\left[\widehat{M} - \frac{3\widehat{\sigma}}{\sqrt{n}} \ , \ \widehat{M} + \frac{2\widehat{\sigma}}{\sqrt{n}} \right]$$

is a 95% confidence interval

Problem 3. Minimizing the absolute value of the error, and estimating the median.

(a) We first, write, $\mathbb{E}[|X - a|]$, as,

$$\begin{aligned} \mathbb{E}[|X - a|] &= \int_{-\infty}^{\infty} |x - a| f(x) \, dx \\ &= \int_{-\infty}^{a} (a - x) f(x) \, dx + \int_{a}^{\infty} (x - a) f(x) \, dx \end{aligned}$$

We now differentiate this quantity, using the Leibniz's rule given in the hint:

$$\frac{d}{da} \left(\int_{-\infty}^{a} (a - x) f(x) \, dx + \int_{a}^{\infty} (x - a) f(x) \, dx \right) = \int_{-\infty}^{a} f(x) \, dx - \int_{a}^{\infty} f(x) \, dx.$$

To find the desired minimum, we set the derivative to zero and obtain

$$\int_{-\infty}^{a} f(x) \, dx = \int_{a}^{\infty} f(x) \, dx.$$

Finally, using the fact that, PDF integrates to one, we get,

$$\int_{-\infty}^{a} f(x) \, dx = \frac{1}{2} = \int_{a}^{\infty} f(x) \, dx,$$

and therefore the optimal a must be the median.

(b) If $i = 1$, the sample median is equal to the single data point x_1. Therefore, $\mathbb{E}[\widehat{M}] = \mathbb{E}[X_1]$. For general distributions, the mean and the median are different. For example if $f_X(x) = 2x$ for $x \in [0, 1]$, the mean is $2/3$ and the median is $1/\sqrt{2}$.

(c),(d),(e),(f) The remaining parts of this problem is computational, and the corresponding solutions are given separately.

Problem 4. Estimating a functional of a distribution.

(a) $\mathbb{E}[K^2] = \text{var}(K) + (\mathbb{E}[K])^2 = np(1-p) + (np)^2$. Hence,

$$\mathbb{E}[\widehat{A}] = \mathbb{E}\left[\frac{K^2}{n^2}\right] = \frac{p(1-p)}{n} + p^2,$$

and

$$\text{bias}(\widehat{A}) = \mathbb{E}[\widehat{A}] - p^2 = \frac{p(1-p)}{n}.$$

(b) $\widehat{A}_b = \widehat{A} - \frac{\hat{p}(1-\hat{p})}{n}$.

$$\begin{aligned}\text{bias}(\widehat{A}_b) &= \mathbb{E}[\widehat{A}_b - p^2] \\ &= \mathbb{E}[\widehat{A} - p^2] - \mathbb{E}\left[\frac{\hat{P}(1-\hat{P})}{n}\right] \\ &= \frac{p(1-p)}{n} - \mathbb{E}\left[\frac{\hat{P}(1-\hat{P})}{n}\right] \\ &= \frac{\mathbb{E}[\widehat{P}^2] - p^2}{n}\end{aligned}$$

where the last equality holds because $\mathbb{E}[\widehat{P}] = p$. Now, $\mathbb{E}[\widehat{P}^2] = \frac{\mathbb{E}[K^2]}{n^2} = \frac{p(1-p)}{n} + p^2$, as shown in the solution to part (a). Thus,

$$\text{bias}(\hat{A}_b) = \frac{p(1-p)}{n^2},$$

which is smaller order of magnitude than $\text{bias}(\widehat{A})$.

(c) Using the suggested hint, and assuming that, $n = 2k$ is even, let us consider the estimator, $\widehat{A}_c$, given by,

$$\hat{f}(X_1, X_2, \ldots, X_n) = \frac{1}{k}\sum_{i=1}^{k} X_{2i-1}X_{2i}.$$

Note that, each $X_{2i-1}X_{2i}$ is Bernoulli with parameter p^2, and thereore variance $p^2(1-p^2)$. Thus,

$$\mathrm{var}(\widehat{A}_c) = \frac{1}{k^2} kp^2(1-p^2) = \frac{p^2(1-p^2)}{k} = \frac{2p^2(1-p^2)}{n}.$$

(d) Since $(X_1, \ldots, X_{n/2})$ is independent from $(X_{n/2+1}, \ldots, X_n)$, we see that K_1 and K_2 are independent. Thus,

$$\mathbb{E}\left[\frac{K_1}{n/2} \cdot \frac{K_2}{n/2}\right] = \mathbb{E}\left[\frac{K_1}{n/2}\right] \mathbb{E}\left[\frac{K_2}{n/2}\right].$$

Next

$$\mathbb{E}\left[\frac{K_1}{n/2}\right] = \mathbb{E}\left[\frac{K_2}{n/2}\right] = \frac{np/2}{n/2} = p,$$

and

$$\mathbb{E}\left[\frac{K_1}{n/2} \cdot \frac{K_2}{n/2}\right] = p^2.$$

Thus, this estimator for p^2 is unbiased.

(e) The idea is to express the mean and the bias of $\widehat{A}_s$ as functions of a, b, c, p. We can then choose a, b, c so that the bias is always zero.
Using $\mathbb{E}[K] = np$ and $\mathbb{E}[K^2] = np + n(n-1)p^2$, we have,

$$\begin{aligned}\mathbb{E}[\hat{A}_s] &= a + b\mathbb{E}[K] + c\mathbb{E}[K^2] \\ &= a + bnp + cnp + cn(n-1)p^2.\end{aligned}$$

For any given n, this is a quadratic polynomial in p. We require this polynomial to be equal to p^2, for all p. This leas to:

1. $a = 0$,
2. $bn + cn = 0$, or $b = -c$,
3. $cn(n-1) = 1$, so that $c = \frac{1}{n(n-1)}$.

Using the second condition, we get, $b = -\frac{1}{n(n-1)}$. Hence, the desired estimator is

$$\widehat{A}_s = -\frac{K}{n(n-1)} + \frac{K^2}{n(n-1)} = \frac{K(K-1)}{n(n-1)}.$$

Problem_3_sol

February 15, 2018

Problem 3 The data for this problem is provided as a csv file called data_problem_3.csv. You need to provide answers to parts (b), (c), (d) and (e). Refer to the Problem Set 2 pdf.

3(b) i. Use numpy to calculate the mean of the dataset ii.Use numpy to calculate the median of the dataset

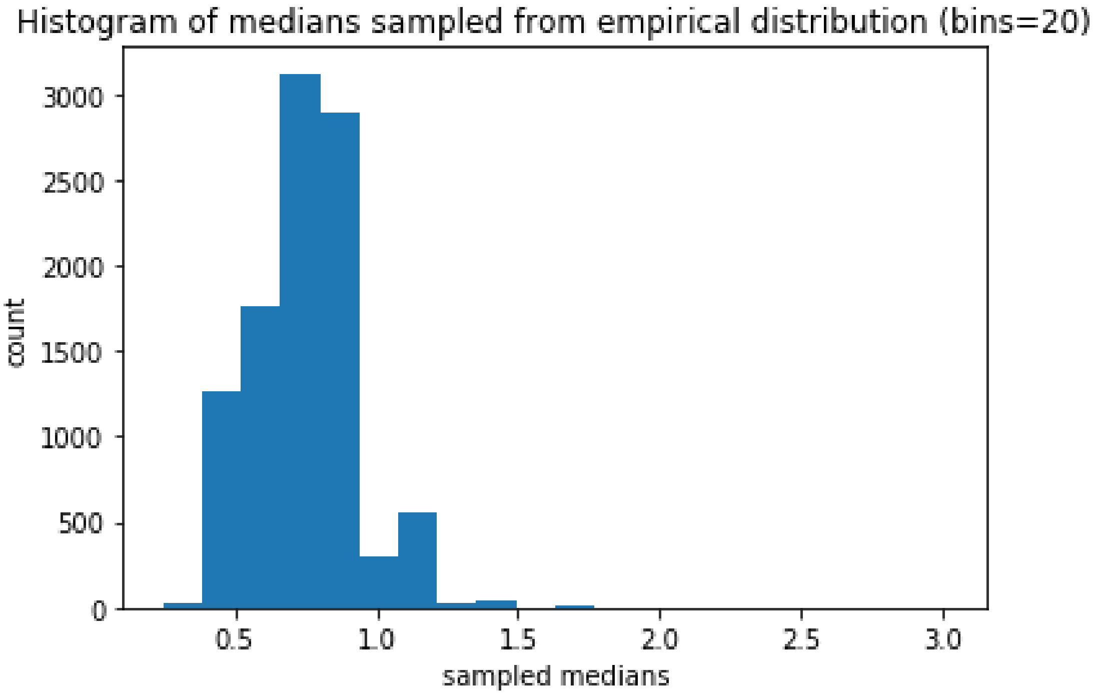
```
In [52]: # import libraries
         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt

         # read data as pandas dataframe
         data = pd.read_csv("data_problem_3.csv", header=None)
         data = data[0].values # now your data is just an array
```

```
In [53]: mu = np.mean(data)
         M = np.median(data)

         print("The empirical mean is {}.".format(mu))
         print("The empirical median is {}.".format(M))
```

```
The empirical mean is 4.2990193319925.
The empirical median is 0.7620781394499716.
```

3(c) i. Sample 101 data points (with replacement) and record the median. Repeat 10,000 times to get 10,000 estimates of the median. ii. Plot the resulting histogram of estimates using bins=20.

```
In [54]: bootstrap_M = np.array([np.median(np.random.choice(data, 101)) for k in range(10000)]

         plt.hist(bootstrap_M, bins=20);
         plt.title("Histogram of medians sampled from empirical distribution (bins=20)");
         plt.xlabel("sampled medians");
         plt.ylabel("count");
```



3(d) i. Calculate the estimated bias (recall that your reference value is the median from part (b). ii. Calculate the standard error (which is just the standard deviation of the values you plotted in the histogram).

```
In [55]: bias = np.mean(bootstrap_M) - M
         se = np.std(bootstrap_M)

         print("The estimated bias is {}.".format(bias))
         print("The estimated standard error is {}.".format(se))
```

```
The estimated bias is 0.0017794616066717506.
The estimated standard error is 0.17959154662974072.
```

3(e) Use the estimated values (the ones you plotted in the histogram) to produce estimates of the 95% confidence intervals. You may find the sort() function useful for this exercise.

```
In [56]: bootstrap_M.sort()
         l_tail = bootstrap_M[250]
         r_tail = bootstrap_M[9749]
         b = r_tail - M
         a = M - l_tail
         print("The 95% confidence interval for the median is ({},{}).".format(M - b, M + a))
```

```
The 95% confidence interval for the median is (0.39917850591665127,1.0848708339415882).
```