MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.s077 Spring 2018
Partial Notes for Lectures 3 and 4, Due Tuesday 2/27, in class
and **Problem Set 2.**

Readings: Slides from Lectures 3 and 4. We do not provide detailed notes for these lectures. For most topics, the slides, together with these notes, already contain all of the useful information.
Sources:
[BT] Bertsekas & Tsitsiklis, *Introduction to Probability*
[W] Wasserman, *All of Statistics*, available for online reading through the MIT library.

Problem 1. Bias, variance, and MSE practice.
Let $X_1, \ldots, X_n$ be i.i.d. random variables with common and unknown mean μ, and a known variance σ^2. Instead of using the sample mean to estimate μ, we use instead the estimator

$$\hat{M} = \frac{1 + X_1 + \cdot + X_n}{n+1}.$$

This estimator artifically pretends that we have an additional measurement, equal to 1. A possible reason for doing so might be that we believe μ to be close to 1, and wish to bias the estimator in that direction.

(a) Find $\mathbb{E}[\hat{M}]$, as a function of μ and n. Find the bias. What happens to the bias as $n \to \infty$?

(b) Find the variance of $\hat{M}$, as a function of σ and n.

(c) Explain why, as $n \to \infty$, the contribution of the bias to the MSE is negligible compared to the contribution of the variance.
Hint: Just keep track of the order of magnitude of different terms. For example, a term of order $1/n^4$ is considered negligible compared to a term of order $1/n^3$.

Problem 2. The most common estimation problem.
The most common estimation problem, which arises whenever you carry out repeatedly an experiment to measure a parameter in the presence of noise, involves the following model:

$$X_i = \mu + \sigma W_i, \qquad i = 1, \ldots, n,$$

where the W_i are zero mean, i.i.d., from a known distribution, and μ is the parameter to be estimated. Here, σ is an unknown noise intensity parameter. Without loss of generality, we assume that the variance of the W_i is equal to 1, so that the variance of the noise term σW_i is $v = \sigma^2$.

The W_i are often assumed to be standard normal, but other choices are possible; for example, the W_i might be distributed according to the PDF

$$f(w) = \frac{1}{2} e^{-|w|}.$$

We use the natural estimator

$$\hat{M} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

but we are also interested in confidence intervals (CIs). As discussed in lecture, if σ is known and the W_i are standard normal, we can use the fact that

$$Z = \frac{\sqrt{n}(\hat{M} - \mu)}{\sigma}$$

is also a standard normal. Then, the normal tables yield

$$\mathbb{P}\big(|Z| \leq 1.96\big) = 0.95,$$

which translates to

$$\mathbb{P}\Big(\hat{M} - 1.96\, \frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{M} + 1.96\, \frac{\sigma}{\sqrt{n}}\Big) = 0.95,$$

and provides a 95%-confidence interval

$$\Big[\hat{M} - 1.96\, \frac{\sigma}{\sqrt{n}},\ \hat{M} + 1.96\, \frac{\sigma}{\sqrt{n}}\Big].$$

If σ is not known, we can estimate the variance using the estimator

$$\hat{V} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{M})^2,$$

let $\hat{\sigma} = \sqrt{\hat{V}}$, and report the **approximate CI**

$$\Big[\hat{M} - 1.96\, \frac{\hat{\sigma}}{\sqrt{n}},\ \hat{M} + 1.96\, \frac{\hat{\sigma}}{\sqrt{n}}\Big].$$

The purpose of this exercise is to show that even when σ is unknown, one can actually construct **exact CIs**, and that this is possible without assuming that the W_i are normal.

(a) Show that $\hat{V}$ is unbiased, i.e., that $\mathbb{E}[\hat{V}] = v$, no matter what μ and v are, and no matter what the distribution of the W_i is.
Hint: Express $X_i - \widehat{M}$ in terms of σ and the W_i.
Note: This explains the denominator term $n-1$ in the definition of $\hat{V}$. If we were to use n instead, then $\hat{V}$ would have an expectation of $(n-1)v/n$, and would be biased

(b) We will rely on the statistic, similar to the above defined Z, given by

$$T = \frac{\sqrt{n}(\hat{M} - \mu)}{\hat{\sigma}}.$$

Express T as a function $T = g(W_1, \ldots, W_n)$ of the original random variables W_i, and show that the function $g(\cdot)$ does not involve μ or σ.

(c) In light of the result of part (b), as long as we know the distribution of the W_i, we can calculate the distribution of T: sometimes this is done analytically; if not, we can obtain it through simulation. Thus, suppose that the distribution of T has been obtained, and that we find that $\mathbb{P}(T < 2) = 0.025$ and $\mathbb{P}(T > 3) = 0.025$. Use this information to construct a 95%-CI for μ.

Comments: For the special case where the W_i are normal, the distribution of T is available in closed form, and is known as the *Student t-distribution with $n-1$ degrees of freedom*. Of course, the distribution depends on n. When n is very large, $\hat{\sigma}$ will be very close top σ, with high probability, and therefore T will be very close to Z. Since Z is normal, it follows that, for large n, T can be well approximated by a normal, and one can just use the normal tables. For example, when $n = 50$, and fif we are interested in 95% CIs, the relevant comparison is between the facts $\mathbb{P}(|Z| \geq 1.96) = 0.05$ and $\mathbb{P}(|T| \geq 2.01) = 0.05$. Using 1.96 instead of 2.01 does not make a big practical difference.

Problem 3. Minimizing the absolute value of the error, and estimating the median. Let X be a random variable, and let m be its median, i.e.,

$$\mathbb{P}(X \leq m) = \frac{1}{2}.$$

We assume that m is uniquely defined, which is always the case when X has a continuous distribution, and is also almost always the case when X has a discrete distribution.

The following is a useful fact about the median: it minimizes the risk function $\mathbb{E}\big[|X-a|\big]$ over all a. We provide here a proof for the case where X has a continuous distribution, described by a PDF $f(\cdot)$. (The proof for general random variables

involves similar ideas but is a bit more cumbersome.) For any a, we have

$$\mathbb{E}\big[|X-a|\big] = \int_{-\infty}^{a} (a-x) f(x)\, dx + \int_{a}^{\infty} (x-a) f(x)\, dx.$$

We now recall the Leibniz rule from calculus:

$$\frac{d}{d\alpha} \int_{-\infty}^{\alpha} g(\alpha, \beta)\, d\beta = g(\alpha, \alpha) + \int_{-\infty}^{\alpha} \frac{\partial}{\partial \alpha} g(\alpha, \beta)\, d\beta.$$

Applying the Leibniz rule we find that

$$\frac{d}{da}\mathbb{E}\big[|X-a|\big] = -\int_{-\infty}^{a} f(x)\, dx + \int_{a}^{\infty} f(x)\, dx.$$

At the optimum, the derivative must be zero, which implies that

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

On the other hand, the sum of the two integrals above is equal to 1. Therefore, each integral must be equal to $1/2$, so that the (optimal) a must be equal to the median.

Let us now consider the problem of estimating the median m, based on i.i.d. samples $X_1, \ldots, X_n$, with the same distribution as X. One possible approach is to use the **plugin** estimator of m: let $\hat{M}$ be the **sample median**, i.e., the median of the empirical distribution of X. Assuming that n is odd, we just sort the data points in order of increasing value, and pick the middle element. (E.g., if $n = 11$, we pick the 6th element.)

Alternatively, we can use the fact that the median minimizes the risk $\mathbb{E}\big[|X - a|\big]$ over all a. Then, the the **empirical risk minimization** approach leads us to consider an estimate $\hat{m}$ that minimizes the empirical risk $\widehat{\mathbb{E}}\big[|x_i - a|\big]$ over all a, or equivalently, that minimizes

$$\sum_{i=1}^{n} |x_i - a|.$$

According to the result stated earlier, the solution is the median of the empirical distribution, i.e., the same answer as that obtained through the plugin approach.

Let use $\hat{M}$ to denote the sample median. (We use an upper case symbol, to emphasize the fact that it is a function of the random data, hence a random variable.)

(a) Is $\hat{M}$ guaranteed to be an unbiased estimator of m? Prove that this is the case or provide a counterexample. *Hint:* Consider the case where $n = 1$.

In the remaining parts of this problem, we explore the bias and the mean squared error of the sample median. You are given a data set with $n = 101$ samples, in a data file (data_problem_3.csv) provided on the computation portal. Your task is to calculate the numerical values of the four estimates introduced in parts (b)-(e) of this problem. You should submit the .ipynb file with the code and numerical estimates along with your solution write-up. We provide a skeleton .ipynb file on the portal. You can access it as follows:

- Log on to the portal (http://mit-6s077.mit.edu)
- After you have successfully logged in, in the same browser window, copy and paste the following assignment url: (http://mit-6s077.mit.edu/hub/user-redirect/git-sync?repo=https://github.com/jehangiramjad/mit-6s077&subPath=Pset2/Problem_3.ipynb)

(b) Calculate the numerical value $\hat{m}$ of $\hat{M}$ on this data set.

(c) Following the bootstrap philosophy, resample from the empirical distribution $k = 10,000$ times. The jth time that you resample, for $j = 1, \ldots, 10,000$, you draw 101 data points, and calculate the empirical median $\hat{m}_j$ of those 101 data points. Plot a histogram of the 10,000 values $\hat{m}_j$ that you obtained.

Note that the sample median $\hat{m}$ is the true median of the empirical distribution, whereas the $\hat{m}_j$ are estimates of the median of the empirical distribution. The $\hat{m}_j$ are drawn by applying our estimator to the empirical distribution, and therefore their histogram tells us the sampling distribution, when the underlying distribution of the X_i is the empirical distribution $\widehat{\mathbb{P}}$. The bootstrap philosophy rests on the fact that $\widehat{\mathbb{P}}$ is representative of the true distribution $\mathbb{P}$. Thus, the sampling distribution (and therefre, the bias, the standard error, etc.) of $\hat{M}$ when applied to $\mathbb{P}$ should be well represented by the histogram of the $\hat{m}_j$.

(d) Based on the histogram of the $\hat{m}_j$, calculate an estimate of the bias and the standard error of the estimator $\hat{M}$.

(e) Based on the histogram of the $\hat{m}_j$, calculate a 95% confidence interval for m.

Problem 4. Estimating a functional of a distribution.
Let $X, X_1, \ldots, X_n$ be independent random variables, drawn according to a distribution $\mathbb{P}$. Let $a = h(\mathbb{P})$ be a functional; that is, h is a function (or a rule), which

to any $\mathbb{P}$ in a certain family of distributions, assigns a real number a. Examples of functionals are the mean of $\mathbb{P}$, the median of $\mathbb{P}$, the expectation of X^4, the expectation $\mathbb{E}[g(X)]$ (for a given function g), etc.

Suppose that X takes values in a given discrete set $\{z_1, \ldots, z_k\}$. In this case, $\mathbb{P}$ is completely determined by $(p_1, \ldots, p_k)$, where $p_i = \mathbb{P}(X = z_i)$. Accordingly, a functional is just an ordinary function that maps the vector $(p_1, \ldots, p_k)$ to a real number.

In the special case where X is discrete as above, and the functional is of the form $h(\mathbb{P}) = \mathbb{E}[g(X)]$, we obtain $h(\mathbb{P}) = \sum_i g(z_i)p_i$, which is a **linear** function of the p_i. Conversely, it is not hard to see that any linear function of the p_i can be expressed as an expectation of some random variable $g(X)$. If we wish to estimate a linear functional $a = h(\mathbb{P}) = \mathbb{E}[g(X)]$, on the basis of data $X_1, \ldots, X_n$, the plugin method leads to the estimator

$$\widehat{A} = h(\widehat{\mathbb{P}}) = \widehat{\mathbb{E}}[g(X)] = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

We know that this estimator is unbiased.

The plugin method also applies to nonlinear functionals: we can always let $\widehat{A} = h(\widehat{\mathbb{P}})$. For example, if $h(\mathbb{P})$ stands for the median of a distribution $\mathbb{P}$, a natural estimator is the sample median, i.e, the median $h(\widehat{\mathbb{P}})$ of the empirical distribution. Unfortunately, such estimates are sometimes biased, and the contribution of the bias to the overall MSE may be substantial. In such cases, it is worth trying to estimate and remove the bias. In the sequel, we run through a simple example of this type.

Let the $X_1, \ldots, X_n$ be independent Bernoulli random variables, with parameter p: $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1 - p$. Thus, $\mathbb{P}$ is completely determined by p. The empirical distribution is also Bernoulli, with parameter $\hat{p}$, where $\hat{p}$ is the realized value of the random variable $\widehat{P} = \sum_{i=1}^{n} X_i/n$. Let $K = \sum_{i=1}^{n} X_i$, so that $\widehat{P} = K/n$. Note that K is a binomial random variable, with parameters n and p; its mean and variance are np and $np(1-p)$, respectively.

Suppose that we are interested in estimating $a = p^2$. This is a nonlinear function of p, which makes it a nonlinear functional of the underlying distribution $\mathbb{P}$. Why would we want to do that? Think of a lottery (or a financial product) that pays 1 dollar if a certain biased coin (or the daily return of a stock) is heads (respectively, positive) for two consecutive trials (respectively, days). The fair value of this lottery (or financial product) is p^2. We would like to estimate the "fair value" of this lottery (or financial product) on the basis of the data $X_1, \ldots, X_n$.

(a) Note that $p^2 = (\mathbb{E}[X])^2$. Furthermore, $\widehat{P} = K/n$ is the mean $\widehat{\mathbb{E}}[X]$ of the

empirical distribution of X. Thus, the plugin estimator of $p^2 = (\mathbb{E}[X])^2$ is the square of the estimator of p:

$$\widehat{A} = (\widehat{\mathbb{E}}[X])^2 = \widehat{P}^2 = \frac{K^2}{n^2}.$$

Find the mean and the bias of $\widehat{A}$, as a function of p and n.

(b) To remove bias, we consider the bias-adjusted estimator

$$\widehat{A}_b = \hat{P}^2 - \frac{\widehat{P}(1-\widehat{P})}{n}.$$

Show that the bias of this estimator is of smaller order of magnitude compared to that of the estimator in part (a).

A nonlinear functional, such as p^2 in our case, cannot be expressed in the form $\mathbb{E}[g(X_1)]$ for some function g. However, it may be the case that it can be expressed in the form $\mathbb{E}\big[g(X_1, \ldots, X_n)\big]$, in which case it can lead to unbiased estimators. In what follows, assume that n is even.

(c) Note that $\mathbb{E}[X_1 X_2] = p^2$. Use this fact to construct an unbiased estimator of p^2 that relies on the products X_1X_2, X_3X_4, X_5X_6, etc. (There are $n/2$ such products.) Calculate its variance.

(d) In a variant of the previous method, we split the data into two halves, and let $K_1 = \sum_{i=1}^{n/2} X_i$, $K_2 = \sum_{i=(n/2)+1}^{n} X_i$. Each K_i can be used in a separate estimator of p. We combine them to form the estimator

$$\frac{K_1}{n/2} \cdot \frac{K_2}{n/2}.$$

Show that this estimator is unbiased.

(e) The preceding two estimators are somewhat unnatural, as the different X_i are not treated symmetrically. Symmetry considerations suggest that we should rely just on K, not on the detailed data X_i. This leads us to consider a symmetric estimator of the form

$$\widehat{A}_s = a + bK + cK^2.$$

Here a, b, c are constants that can depend on n. Find the values of these constants so that $\widehat{A}_s$ is unbiased, no matter what p is.

Discussion. We have found a number of plausible ways of estimating p^2. Which one is preferable? To answer this question, we need to consider the bias and the stndard error of each estimator. The bias is either zero [parts (c)-(e)], or can be obtained in closed form [parts (a) and (b)], so we focus on the standard error, and the different ways that it can be obtained. Note that the standard error will in general depend on p, so, in order to have the full picture, we should in principle find the standard error for each possible p. On the other hand, this is not necessary if the standard error is relatively constant within the range of "plausible" values of p.

1. **Exact calculation.** Because K has a binomial distribution, a somewhat tedious calculation leads to exact, closed form expressions for the variance of the different estimators. However, for the case of the estimators in parts (a) and (e), finding the variance of the estimator requires the expectation of K^4. This expectation can be obtained in closed form, but the formulas get pretty messy and tend to obscure the important insights.

2. **Approximate calculation.** In order to simplify calculations, we can use a normal approximation. According to the CLT, we know that we can write

$$\frac{K}{n} \approx p + \frac{\sigma Z}{\sqrt{n}},$$

where Z is a standard normal random variable and $\sigma = \sqrt{p(1-p)}$. Suppose, furthermore, that we already know that p is not too unfair, and lies in the range $[0.4, 0.6]$, so that p^2 lies in the range $[0.16, 0.36]$. Note that for any p in the given range, the variance $\sigma^2 = p(1-p)$ of X_i takes values between 0.24 and 0.25, and can be approximated, for simplicity by 0.25, i.e., $\sigma \approx 1/2$. This allows us to write

$$\frac{K}{n} \approx p + \frac{Z}{2\sqrt{n}}.$$

Using also the known fact, for the standard normal, that $\mathbb{E}[Z^4] = 3$, all required moments of K can be obtained. Note also that

$$\frac{K^4}{n^4} \approx p^4 + 4p^3\Big(\frac{Z}{2\sqrt{n}}\Big) + 6p^2\Big(\frac{Z}{2\sqrt{n}}\Big)^2 + 4p\Big(\frac{Z}{2\sqrt{n}}\Big)^3 + \Big(\frac{Z}{2\sqrt{n}}\Big)^4.$$

We have $\mathbb{E}[Z] = \mathbb{E}[Z^3] = 0$, and two of the terms above disappear. We then observe that the last term is of order $1/n^2$, which is much smaller than the third term, which is of order $1/n$. Thus, the last term can be ignored.

3. **Simulation.** We can consider a reaonably fine grid of the values of p, e.g., with spacing 0.01 or 0.2. For each such value of p, we can generate random data, and evaluate the estimator. We repeat this many times (for the given p), to get a histogram of the estimator. This histogram then provides us with an estimate of the mean, bias, and variance of the estimator, for each p.

4. **Parametric bootstrap.** This is essentially the same as the preceding simulation method except that we carry out the simulations for a single value of p. In particular, instead of simulating the estimator for every value of p on a grid, we use the actual data and the estimator K/n to obtain an estimate $\hat{p}$ of p. We then simulate (several times) the estimator with p set equal to $\hat{p}$, look at the resulting histogram, and thus obtain the mean, bias, and variance of the estimator. This method relies on the assumptions that the true p is within a "small" neighborhood of $\hat{p}$ and that within this neighborhood, the bias and the variance do not change much.

5. **Non parametric bootstrap.** This is similar to parametric bootstrap. However, instead of of carrying out a simulation based on newly drawn i.i.d. Bernoulli data, we draw Bernoulli data according to the empirical distribution. In general, parametric and nonparametric bootstrap can lead to different answers. For our example, however, it turns out that the two methods coincide. In parametric bootstrap, we draw Bernoulli r.v.s with parameter $\hat{p}$. In nonparametric bootstrap, each new sample is generated from the empirical distribution. The fraction of 1s in the empirical distribution is also $\hat{p}$. Thus, a sample from the empirical distribution is again Bernoulli with parameter $\hat{p}$.

Some additional reading: Heavy-tailed distibutions may be treacherous.
Pretty much everything we have done relies on an assumption that the random variables involved have a finite variance. Without this assumption, we would not be able to make any useful statements on the relation between the sample mean and the true mean.

On the other hand if the tails of a distribution fall off very slowly, the variance may turn out to be infinite. For example, suppose that a PDF $f(x)$ is proportional to $1/x^3$, when x is large. Using the fact that

$$\int_a^\infty x^2 \cdot x^{-3}\, dx = \int_a^\infty \frac{1}{x}\, dx = \log(\infty) - \log(a) = \infty,$$

it follows that the variance is infinite. The same is true if $f(x)$ falls off at the rate $x^{-\alpha}$ with $\alpha < 3$. We consider such distributions to be heavy-tailed.

In practice, distributions with infinite variance are not uncommon. For example, wealth distribution, or city size, or even financial shocks, have plausible models that are heavy-tailed. Of course, if we pick a person at random, we know that their wealth is bounded above by 500 trillion.[1] Hence the wealth distribution, in reality, has a finite variance. On the other hand, this variance is very large, and the far-right tail of the distribution makes most of the contribution to the variance.[2] For this reason, a heavy-tailed model, with infinite variance, provides better insights than those provided by the finite variance analysis.

Estimating the mean of a heavy-tailed distribution is not easy, even with very large sample sizes. The numerical example below captures the essence of the issues that arise.

Consider a population consisting of 1011 individuals. Suppose that:
(i) Each of 1000 individuals has a wealth of 1 (measured in some convenient unit).
(ii) Each of 10 individuals has a wealth of 100 .
(iii) One individual has a wealth of 1000.
The mean wealth is 3, the variance 1089, and the standard deviation is about 33.

Suppose that we draw a sample of $n = 100$ people. Then, the standard error is $33/\sqrt{n} = 3.3$. Now, with a sample of size 100, there is about 90% chance that the single very wealthy individual will not be picked. Out of the 100 moderately wealthy individuals, we will (most likely) pick 0, 1, or 2 of them.[3] The resulting estimate will be either 1, or $(99\times 1+1\times 100)/100 \approx 2$, or $(98\times 1+2\times 100)/100 \approx 3$.

[1] At present, total wealth in the world is estimated to be close to 300 trillion.

[2] One percent of the world's population owns about half of the world's wealth.

[3] To be precise, the probability p_i that we do not see the very wealthy person, and we see exactly i moderately wealthy persons is $p_0 = 0.33$, $p_1 = 0.33$, $p_2 = 0.16$, $p_3 = 0.05$.

So, even though the sample mean is unbiased (in expectation, on the average), it is most likely to be an underestimate. The issue is that a substantial contribution to the mean wealth (one third of the contribution, to be exact) comes from a single very wealthy individual. But it would take a very large number of samples until that individual is discovered.

To recapitulate, let us say, loosely speaking, that a distribution has heavy tails if the tails of the distribution make a large contribution to the mean. Estimating the mean of such a distribution is difficult, and in some sense impossible, unless almost all of the population is sampled.

On the other hand, one could argue that for such distributions, the mean is a rather uninteresting quantity, as it is very sensitive to the properties (the wealth, in this case) of very few individuals. Much more meaningful information is provided by the median and various quantiles of the distribution, and these are much easier to estimate. In our example, the median is equal to 1, and with a sample of $n = 100$, we are essentially certain that the sample median will also be equal to 1.