14.32 Recitation 2 Notes

Ian Sapollnik

September 15, 2023

1 Bessel's correction

We saw in lecture that an unbiased estimator of of the variance is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. You might be wondering why we divide by $n-1$ here instead of dividing by n like we do with a sample mean. This is such a common question that the idea of dividing by $n-1$ even has a formal name: Bessel's correction. To see why this is necessary, let's focus on some definitions first. For a given random variable, μ and σ^2 are the population mean and variance respectively. For our sample of i.i.d. data $\{x_1, x_2, \ldots, x_n\}$, we define $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ as the sample mean and s^2 the sample variance. In proving why Bessel's correction is necessary, you should become more familiar working with sample and population means and variances.

Let's start by expanding out $(n-1)\, s^2$, the summation part of the sample variance. We have

$$\begin{aligned}
(n-1)\, s^2 =& \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \\
=& \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2 \\
=& \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
=& \sum_{i=1}^{n} x_i^2 - n\bar{x}^2
\end{aligned}$$

Recall that we say that an estimator $\hat{\theta}$[1] is unbiased if $\mathbb{E}\left[\hat{\theta}\right] = \theta$, where θ is the true population value. So let's take expectations of the above expression and see what happens. In doing so, we will make use of the convenient property that for any random variable Y, $\mathbb{E}\left[Y^2\right] = \mathbb{E}\left[Y\right]^2 + \mathrm{Var}\left(Y\right)$.[2]

$$\begin{aligned}
(n-1)\, \mathbb{E}\left[s^2\right] =& \mathbb{E}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] \\
=& \sum_{i=1}^{n} \mathbb{E}\left[x_i^2\right] - n\mathbb{E}\left[\bar{x}^2\right] \\
=& \sum_{i=1}^{n}\Big(\mathbb{E}\left[x_i\right]^2 + \mathrm{Var}\left(x_i\right)\Big) - n\Big(\mathbb{E}\left[\bar{x}\right]^2 + \mathrm{Var}\left(\bar{x}\right)\Big)
\end{aligned}$$

Hopefully you should see that these expectations and variances are easy to work with, we already know what they are! They are just the population parameters: $\mathbb{E}\left[x_i\right] = \mu$ and $\mathrm{Var}\left(x_i\right) = \sigma^2$. It follows as well that $\mathbb{E}\left[\bar{x}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_i\right] = \frac{n}{n}\mu = \mu$ and $\mathrm{Var}\left(\bar{x}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(x_i\right) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$. With these equalities in hand, let's plug back into the expression above and see what we get.

[1] In econometrics it's common to denote estimators with a hat on top.

[2] To see why this is true, take the definition of variance $\mathrm{Var}\left(Y\right) = \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y\right]\right)^2\right]$,expand out the square and rearrange.

$$\begin{aligned}(n-1)\,\mathbb{E}\left[s^2\right] =&\sum_{i=1}^{n}\left(\mu^2+\sigma^2\right)-n\left(\mu^2+\frac{\sigma^2}{n}\right)\\ =&n\mu^2+n\sigma^2-n\mu^2-\sigma^2\\ =&\sigma^2\left(n-1\right)\end{aligned}$$

We can see that if we cancel out the $n-1$ terms on either side of the equation, we find that s^2 is indeed unbiased: $\mathbb{E}\left[s^2\right]=\sigma^2$. If instead we had divided by n, then we'd have $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]=\mathbb{E}\left[\frac{n-1}{n}s^2\right]=\frac{n-1}{n}\sigma^2$, a biased estimator.

So the math works out, but what is the intuition here? It turns out that an important yet potentially underlooked part of this problem is that we are simultaneously estimating the mean and the variance. If, for example, we were given data with known population mean μ but unknown population variance, the estimator $s_\mu^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$ is indeed unbiased; no Bessel's correction required! [3] The problem is that in general we know neither μ nor σ^2, and using our estimator $\bar{x}$ in place of μ within our estimator s^2 introduces bias. Why is that happening? One way of thinking about it (in words) is that for any data set the observed values will fall, on average, closer to the sample mean $\bar{x}$ than they do to the population mean μ. So the variance calculated using deviations from the sample mean will slightly understate the true variance of the population. Let's show that this is true using math. Like we said before, we would ideally like each term in the summation to be $x_i-\mu$, but we settle for $x_i-\bar{x}$ since we don't know μ. Let's define $d=(x_i-\mu)-(x_i-\bar{x})$. d is the difference between what we would like in our estimator and what we actually put into the estimator. Clearly, $d=\bar{x}-\mu$; we are off by the difference between our sample and population means. Keep in mind that because our data are independently and identically distributed, the variance of the sum is the sum of the variances (i.e. there is no covariance between any two x_i). Let's define one more, biased, estimator: $s_n^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$ is the variance estimator that uses n in the denominator instead of $n-1$. We can then say

$$\begin{aligned}\mathbb{E}\left[s_n^2\right]=&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]\\ =&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\left(x_i-\mu\right)-d\right)^2\right]\\ =&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\left(x_i-\mu\right)^2-2d\left(x_i-\mu\right)+d^2\right)\right]\\ =&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right]-2\mathbb{E}\left[d\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)}_{=d}\right]+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[d^2\right]\\ =&\sigma^2-2\mathbb{E}\left[d^2\right]+\mathbb{E}\left[d^2\right]\\ =&\sigma^2-\mathbb{E}\left[\left(\bar{x}-\mu\right)^2\right]\end{aligned}$$

So we've again confirmed that s_n^2 is biased, but we've gone a step further because we now know the size of the bias. $\mathbb{E}\left[s_n^2\right]$ is off by $\mathbb{E}\left[\left(\bar{x}-\mu\right)^2\right]$. The closer our sample mean is to the population mean, the less this bias will matter. But in small sample sizes, if we draw a sample mean that's far from the population mean, we'll be off by quite a bit if we used this biased estimator. So we're almost done, but we can do even better by figuring out what $\mathbb{E}\left[\left(\bar{x}-\mu\right)^2\right]$ is. You should hopefully see that we've already calculated this though,

[3] You should verify this for yourself using similar algebra as above!

it's just the variance of the sample mean. We showed above that $\text{Var}\left(\bar{x}\right) = \frac{\sigma^2}{n}$. So

$$\begin{aligned}\mathbb{E}\left[s_n^2\right] =& \sigma^2 - \frac{\sigma^2}{n} \\ =& \left(\frac{n-1}{n}\right)\sigma^2\end{aligned}$$

So $\frac{n}{n-1}s_n^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = s^2$ is indeed an unbiased estimator. Note that while s_n^2 is biased, it is still consistent. And the size of the bias decreases as your sample size grows, since $\frac{n-1}{n} \to 1$ as $n \to \infty$.