Bessel's Correction and Jensen's Inequality

Stephen Huan[1]

[1]Georgia Institute of Technology

November 24, 2021

Table of Contents

Setup

Suppose we have $X_1, X_2, \ldots, X_n$ sampled i.i.d from the random variable X with unknown $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.

Our goal is to find good estimators for the mean and variance *without* assumptions about the underlying distribution.

Table of Contents

Mean Estimator

To estimate the mean, the most natural thing is to take the sample mean. Let $S_n = (X_1, X_2, \ldots, X_n)$:

$$E[S_n] = \sum_{i=1}^{n} \frac{1}{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Note that our estimator $E[S_n]$, defined as a linear combination of random variables, is also a random variable. We can therefore take its expectation and variance.

Expectation of the Mean Estimator

$$
\begin{aligned}
E[E[S_n]] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] && \text{definition of } E[S_n] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[X_i] && \text{linearity of expectation} \\
&= \frac{1}{n}(n\mu) && \text{each } X_i \text{ i.i.d from } X \\
&= \mu
\end{aligned}
$$

This shows our estimator is an *unbiased* estimator of the true parameter, i.e. $E[E[S_n] - \mu] = \mu - \mu = 0$.
Being unbiased is nice, but how fast is the convergence?

Variance of the Mean Estimator

$$
\begin{aligned}
\mathrm{Var}[\mathrm{E}[S_n]] &= \mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} X_i] && \text{definition of } \mathrm{E}[S_n] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] && \text{variance of sum i.i.d. r.v.s} \\
&= \frac{1}{n^2}(n\sigma^2) && \text{each } X_i \text{ i.i.d from } X \\
&= \frac{1}{n}\sigma^2
\end{aligned}
$$

As expected, as the number of samples goes up, the mean estimator varies less — our estimate becomes more precise.

Table of Contents

Variance Estimator

With the success of the mean estimator, why not use the sample variance as an estimator for σ^2?

$$\mathrm{Var}[S_n] = \mathrm{E}[(S_n - \mathrm{E}[S_n])^2] = \mathrm{E}[S_n^2] - \mathrm{E}[S_n]^2$$

We'll just focus on computing the expectation of this estimator:

$$\mathrm{E}[\mathrm{Var}[S_n]] = \mathrm{E}[\mathrm{E}[S_n^2]] - \mathrm{E}[\mathrm{E}[S_n]^2] \quad \text{linearity of expectation}$$

Focusing on the second term,

$$\mathrm{Var}[\mathrm{E}[S_n]] = \mathrm{E}[\mathrm{E}[S_n]^2] - \mathrm{E}[\mathrm{E}[S_n]]^2 \quad \text{definition of variance}$$

$$\frac{1}{n}\sigma^2 = \mathrm{E}[\mathrm{E}[S_n]^2] - \mu^2 \quad \text{from previous}$$

$$\mathrm{E}[\mathrm{E}[S_n]^2] = \frac{1}{n}\sigma^2 + \mu^2$$

Expectation of the Variance Estimator

Working on the first term,

$$
\begin{aligned}
\mathrm{E}[\mathrm{E}[S_n^2]] &= \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n} X_i^2] && \text{definition of } \mathrm{E}[S_n^2] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[X_i^2] && \text{linearity of expectation} \\
&= \frac{1}{n} n(\sigma^2 + \mu^2) && \text{each } X_i \text{ i.i.d from } X
\end{aligned}
$$

where we used $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$, so $\mathrm{E}[X^2] = \mathrm{Var}[X] + \mathrm{E}[X]^2 = \sigma^2 + \mu^2$

$$= \sigma^2 + \mu^2$$

Bias of the Variance Estimator

Putting the two together,

$$\begin{aligned} \mathrm{E}[\mathrm{Var}[S_n]] &= \mathrm{E}[\mathrm{E}[S_n^2]] - \mathrm{E}[\mathrm{E}[S_n]^2] \\ &= (\sigma^2 + \mu^2) - (\frac{1}{n}\sigma^2 + \mu^2) \\ &= \boxed{\frac{n-1}{n}\sigma^2} \end{aligned}$$

This is an *underestimate* for the true variance! In order to correct it, the simplest thing is to multiply the estimator by $\frac{n}{n-1}$, which will unbias the estimator by linearity of expectation. This correction is called *Bessel's correction*.

Corrected Variance Estimator

That is, instead of computing the sample variance

$$\mathrm{Var}[S_n] = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}[S_n])^2$$

we multiply by $\frac{n}{n-1}$, yielding the unbiased estimator:

$$\frac{n}{n-1}\mathrm{Var}[S_n] = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mathrm{E}[S_n])^2$$

Understanding Bias

Why is the sample variance biased?

One explanation is "degrees of freedom", as the knowing the mean takes 1 degree of freedom off. This makes it $n - 1$ degrees of freedom instead of n. This type of reasoning is powerful, but feels heuristical to me.

Instead, I think a more compelling explanation is that the sample mean is necessarily a better minimizer of sample variance than the true mean. The sample mean happens to be the unique minimizer of the sum of squares error. Since we're using the sample mean instead of the true mean, we'll always *underestimate* the true variance. If we somehow knew the true mean, and used it to calculate sample variance, the estimation *would* be unbiased and we wouldn't need to correct.

Estimating Standard Deviation

But what if we wanted to estimate the standard deviation σ instead of σ^2? Well, one natural thing to do would be to take the square root of our unbiased variance estimator,

$$\hat{\sigma} = \sqrt{\hat{\sigma^2}}$$

However, note that in general $\mathsf{E}[f(X)] \neq f(\mathsf{E}[X])$

$$\mathsf{E}[\hat{\sigma}] \neq \sqrt{\mathsf{E}[\hat{\sigma^2}]} = \sigma$$

Our estimator is biased again! But in which direction?

Table of Contents

Convex Functions

Convexity

A function $f(x)$ is *convex* if its second derivative is always positive, i.e. $f''(x) \geq 0$. There are many alternative definitions.

Theorem

Jensen's inequality. If f is convex,

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)]$$

We will prove for discrete random variable X, but the theorem holds for continuous random variables as well.

Proof of Jensen's Inequality

Proof.

We use an alternative definition of convexity, that

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2)$$

for all $\lambda_1, \lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$.

We want to show that

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)]$$

Assuming that X is a discrete random variable,

$$\begin{aligned} \mathrm{E}[X] &= \sum_{x \in X} x\, p(x) \\ &= p(x_1)x_1 + p(x_2)x_2 + \cdots + p(x_n)x_n \\ &= \lambda_1 x_1 + \cdots + \lambda_n x_n \end{aligned}$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$

Proof of Jensen's Inequality

Proof.

We want to show:

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)]$$

$$f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n)$$

We can just induct out each $\lambda_i x_i$ term from the left side:

$$f(\lambda_1 x_1 + (1-\lambda_1)[\frac{\lambda_2}{1-\lambda_1}x_2 + \cdots + \frac{\lambda_n}{1-\lambda_1}x_n]) \leq$$

$$\lambda_1 f(x_1) + (1-\lambda_1) f(\frac{\lambda_2}{1-\lambda_1}x_2 + \cdots + \frac{\lambda_n}{1-\lambda_1}x_n)$$

by convexity. By induction, the theorem is true for all n. □

Special Case of Jensen's Inequality

For example, take the convex function $f(x) = x^2$.

By Jensen's inequality, $f(\mathrm{E}[X]) = \mathrm{E}[X]^2 \leq \mathrm{E}[f(X)] = \mathrm{E}[X^2]$. This tells us $\mathrm{E}[X^2] - \mathrm{E}[X]^2 \geq 0$, which we already knew since variance must be positive!

It also tells us why our sample variance is an underestimate for the true variance. Recall our estimator is $(\sigma^2 + \mu^2) - \mathrm{E}[\mathrm{E}[S_n]^2]$. If we used μ instead of $\mathrm{E}[S_n]$, $(\sigma^2 + \mu^2) - \mathrm{E}[\mu^2] = \sigma^2$ so we have an unbiased estimator as expected. However, we use $\mathrm{E}[S_n]$, and by Jensen's inequality, $\mathrm{E}[\mathrm{E}[S_n]^2] \geq \mathrm{E}[\mathrm{E}[S_n]]^2 = \mu^2$.

So we always subtract *more* than we should with the sample mean, underestimating the true variance.

Table of Contents

Jensen's Inequality for Concave Functions

Concavity

A function f is *concave* if its second derivative is always negative, i.e. $f''(x) \leq 0$. This is basically the "opposite" of convexity (not literally logically negative).

We can also adjust the inequality for concave functions trivially. Suppose $f(x)$ is concave. Then $g(x) = -f(x)$ is convex.

Applying Jensen's inequality on g,

$$g(\mathrm{E}[X]) \leq \mathrm{E}[g(x)]$$
$$-f(\mathrm{E}[X]) \leq \mathrm{E}[-f(x)]$$
$$f(\mathrm{E}[X]) \geq \mathrm{E}[f(x)]$$

Bias of the Standard Deviation Estimator

Finally, we can analyze the estimator of standard deviation.

If $\hat{\sigma^2}$ is our unbiased estimator for variance, then the estimator for standard deviation is the square root:

$$\hat{\sigma} = \sqrt{\hat{\sigma^2}}$$

Note that $f(x) = \sqrt{x}$ is a concave function, so

$$\mathsf{E}[\hat{\sigma}] = \mathsf{E}[\sqrt{\hat{\sigma^2}}] \leq \sqrt{\mathsf{E}[\hat{\sigma^2}]} = \sigma$$

Thus, our standard deviation estimator is always an *underestimate* of the true standard deviation. There is no generally unbiased estimator; one has to know the distribution of X to construct one. This is why variance is better!

References

Wikipedia articles on:

1. Bessel's Correction
2. Jensen's Inequality
3. Standard Deviation Estimation