On Bessel's correction, aka the unbaised variance

Ruimin Pan

Dec 2024

keywords: unbiased variance, Bessel's correction, expectation, Pythagorean theorem

Abstract

When we try to compute the sample mean and sample variance of a bunch of data, one word often pops up out of nowhere: bias. In particular, we see reminders like this all the time: in order to get an unbiased estimation of the variance, we need to divide the sum of all squared differences by $\mathbf{n-1}$ instead of $\mathbf{n}$. We'll explain in this article what 'biased variance' is and why dividing by $n-1$ will correct it. And of course, as mentioned in the title, we'll give details about 'Bessel's correction'...

1 Introduction

In everyday life, we often (actively or passively) get a group of numbers: the math final results for class of 2014, the height of year 7 girls in St. Nicolas High School (don't go search, I made it up!), the prices of new phones sold in a year, etc.

Naturally we are curious about extracting indicative information from these numbers: how did class of 2014 do in math? Do they measure up to class of 2013? Are the year 7 girls taller than boys as a whole? Are they taller than previous year 7 girls at St. Nicolas High School? Are the new phones more expensive this year? If you want

to dive a little deeper, what's the spread of these numbers? That is, is there a huge discrepency between good math students and bad math students, even if on average the math result in 2014 is pretty good? Are the year 7 girls all similar height or are there a lot of girl giants and little Thumbelinas? Did the phone companies focus on selling medium range phones or are they trying to rip us off with fancy diamond-covered bricks as well as cheap good-for-nothing selfie cameras?

Two numbers are among the most common when it comes to processing any of these data: **sample mean** and **sample variance**. The first measures the performance of your data as a whole (on average), the second one measures the spread of the data.

Sample mean is what we normally call 'average':

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

where X_i represents the ith data points, n is the total number of data points: number of students in class of 2014, number of year 7 girls, number of new phones sold this year. In summary, to calculate the sample mean, we just add everything up and divide the sum by the total number of data points.

When it comes to sample variance, you'll see two different schools of thinking:

- biased variance using n: $\sigma_b^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- unbiased variance using $n-1$: $\sigma_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

The biased variance above seems straightforward; the unbiased variance on the other hand, introduces this $n-1$ factor. Wait a minute, we have n data points, aren't we missing something by using $n-1$ instead? In Section 2, we'll mathematically prove that the one with $n-1$ above is indeed the 'unbiased' estimation of the population mean. Then in Section 3, I'll give a couple of different intuitions to help you understand it in a more natural way.

2 Mathematical proof

In statistics, the tradition is to assume concepts like **population mean** and **population variance**. They represent the inherent underlying patterns of the data. Meanwhile, 'sample mean' and 'sample variance' are our estimation from a small sample, of the 'population mean' and 'population variance', respectively. If we think of the quantity we are trying to measure as a random variable: eg. the height of a year 7 girl, the maths final result, the price of a new phone, the population mean and population variance are known if we know the probability distribution function (pdf) of the random variable (RV). The reality is, we don't necessarily know the pdf of the random variable all the time. In most situations, we don't even know the inherent 'population mean' or 'population variance'. The best we could do, is to collect a small number of samples and use these samples to estimate the population mean and population variance. One important thing to keep in mind: the sample mean ($\bar{\boldsymbol{X}}$) and sample variance (biased and unbiased: $\boldsymbol{\sigma_b^2}$ and $\boldsymbol{\sigma_u^2}$) calculated using samples (data points) of the random variable are themselves **random variables**.

In probability theories, we use capital letter to mark the random variable (RV) and lower case to mark the actual values (samples) of the random variable. For example, if the RV we are trying to measure is the height of a year 7 girl, let's denote the random variable with H. When we sample the random variable H, we might get the following 3 measurements: $h = 150cm$, $h = 148cm$, $h = 162cm$. Using the formula we introduced before, we can calculate the sample mean to be $\bar{H} = \frac{150+148+162}{3} = 153.33$, the biased sample variance to be

$$\sigma_b^2 = \frac{(150 - 153.33)^2 + (148 - 153.33)^2 + (162 - 153.33)^2}{3} = 38.2$$

and the unbiased sample variance to be

$$\sigma_u^2 = \frac{(150 - 153.33)^2 + (148 - 153.33)^2 + (162 - 153.33)^2}{2} = 57.3$$

2.1 Define 'unbiased'

When we say an estimation M is unbiased, we meant something simple: since M is itself a random variable (RV), it has its own **expectation**. If the expectation of M matches the quantity we are trying to estimate using M, then M is an unbiased estimation. In other words, if our estimation matches the true quantity **'on average'**, this is an unbiased estimation. A good example is the typical sample mean formula. Let me demonstrate how the comon sample mean (the estimation of the population mean) is an unbiased estimation using the height measurement example above:

$$\begin{aligned} E(\bar{H}) &= E(\frac{\sum_{i=1}^{3} H_i}{3}) \\ &= \frac{\sum_{i=1}^{3} E(H)}{3} \\ &= \frac{3 \times E(H)}{3} \\ &= E(H) \\ &= \text{population mean of year 7 girl's height } \mu_H \end{aligned} \tag{2}$$

In equation (2) above, we used the property that expectation is a linear operation to swap with the sum operation. We also used the fact that each measurement represents the same random variable so their expectations are the same. Since the expectation of the sample mean is the same as the population mean, this is indeed an unbiased estimation.

One subtle but important footnote: just because the estimation M is unbiased, doesn't necessarily mean it's a good estimation. In the case of the average year 7 girl's height, we can have many unbiased estimations but they are not all created equal! **In fact, any measurement we have of a year 7 girl's height is an unbiased estimation of the average height (population mean)!** This is because $E(H_i) = E(H) = \mu_H$, i.e., each of the measurement is of the same random variable therefore they share the same mean value, as we alluded to in the previous equation.

The only reason why we don't just measure one height and claim this is good enough since this is already an unbiased estimate of the average year 7 girl's height: taking the average across a large number of samples keeps the estimate unbiased while bringing down its variance (deviation from the mean). This way, not only is the expected estimate a good match for the true value, we are also certain this estimate doesn't deviate too far in general from the true value.

2.2 Unbiased variance

Now we are ready to show mathematically that the expectation of the 'unbiased variance' is indeed the true population variance. Using the formula for unbiased variance above, we have:

$$\begin{aligned} E(\sigma_u^2) &= E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2] \\ &= \frac{1}{n-1}\sum_{i=1}^{n} E[(X_i - \bar{X})^2] \\ &= \frac{1}{n-1}[\sum_i E(X_i^2) + \sum_i E(\bar{X}^2) - 2\sum_i E(X_i\bar{X})]) \end{aligned} \tag{3}$$

Again, we used the linearity of the expectation operation in equation (3).

Now let's consider the three terms in (3) one by one. First let's examine $E(X_i^2)$. If we denote the population mean using μ and the population variance using σ^2, we know from the definition of population variance:

$$\begin{aligned} \sigma^2 &= E[(X-\mu)^2] \\ &= E(X^2) + \mu^2 - 2\mu E(X) \\ &= E(X^2) - \mu^2 = E(X_i^2) - \mu^2 \end{aligned}$$

Then it's natural to deduce:

$$E(X_i^2) = \sigma^2 + \mu^2 \tag{4}$$

Here we used $E(X_i^2) = E(X^2), i = 1, \ldots, n$ because each measurement X_i is drawn from the same distribution as the random variable X. For the second term $E(\bar{X}^2)$ in equation 3, we have

$$\begin{aligned}
E(\bar{X}^2) &= E[(\frac{\sum_i X_i}{n})^2] \\
&= \frac{1}{n^2} E[(\sum_i X_i)^2] \\
&= \frac{1}{n^2} \sum_i \sum_k E(X_i X_k) \\
&= \frac{1}{n^2}(nE(X_i^2) + (n^2 - n)[E(X_i)]^2) \qquad (5) \\
&= \frac{1}{n^2}[n(\sigma^2 + \mu^2) + (n^2 - n)\mu^2] \qquad (6) \\
&= \frac{\sigma^2}{n} + \mu^2 \qquad (7)
\end{aligned}$$

In step (5), we used the independence property between individual samples. In step (6), we used the result above for $E(X_i^2)$ in equation (4).

The last term from (3) can be calculated using a similar process:

$$\begin{aligned}
E(X_i \bar{X}) &= E(X_i \frac{\sum_k X_k}{n}) \\
&= \frac{1}{n} E(\sum_k X_i X_k) \\
&= \frac{1}{n}(E(X_i^2) + (n - 1)[E(X_i)]^2) \qquad (8) \\
&= \frac{1}{n}(\sigma^2 + \mu^2 + (n - 1)\mu^2) \qquad (9) \\
&= \frac{\sigma^2}{n} + \mu^2 \qquad (10)
\end{aligned}$$

In step (8), we used the independence between individual samples. In step (9), we used the result in (4).

Plugging our results from (4), (7) and (10) back into (3), we have:

$$\begin{aligned} E(\sigma_u^2) &= \frac{1}{n-1}[n(\sigma^2+\mu^2) + n(\frac{\sigma^2}{n}+\mu^2) - 2n(\frac{\sigma^2}{n}+\mu^2)] \\ &= \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 \\ &= \sigma^2 \end{aligned} \tag{11}$$

There it is. We just mathematically proved that the 'unbiased' formula gives us a truly unbiased variance estimate.

Although we can't argument with the maths, it does bother a lot of people where this $n-1$ comes from? Why is it that we use n for sample mean and $n-1$ for sample variance? Out of all the n data points (samples), which sample should we ignore?

By the way, what we did above is officially called **Bessel's correction**, named after German mathematician Friedrich Bessel.

3 Intuitions

When you read introductory books to statistics, they'll often passingly mention how we have lost 'one degree of freedom' when we use $\bar{X}$ instead of population mean μ when calculating the sample variance. Therefore, when we centralize the variance calculation using $\bar{X}$, we really only have $n-1$ dimensions. It's true we can fully recover all sample points by using $X_1, \ldots, X_{n-1}$ and $\bar{X}$ thus render X_n redundant. This leads to the conclusion that $n-1$ is the real dimension of our centralized data. However, this explanation feels so vague that it kind of hints something but seems so farfetched and handwavy at the same time.

Good news is, there are good explanations and intuitions out there. Here I'll give 2 examples that actually gave *me* the 'aha' moment. Both examples use geometry to illustrate the relationship between samples and their average (sample mean). In one way or another, these intuitions try to construct a right angle triangle and use

Pythagorean theorem to derive Bessel's correction. The differences are only in the choice of triangles.

3.1 A tale of two spaces

This explanation is by far the most interesting and most spot-on, for me at least. Matter of fact, this one is the reason why I decided to write this article. The only problem is, we need to build a brand new paradigm to understand it. At times, it feels like we are 'cracking a nut with a sledgehammer', or 'killing a fly with a cannon', or 'breaking a butterfly on a wheel', or 'boiling the ocean' (my favorite), ...

My justification, though, is the book where this explanation came from: "Statistical Methods: The Geometric Approach" by Saville and Wood [1] from 1991. It looks at statistics from a very unusual angle and somehow appeals to the geometric part of my brain. Even without our goal of correcting sample variance bias, this book is worth a read.

In this book, Saville and Wood suggests that we can look at the sample space as a real n-dimensional 'space' where each sample represents 1 dimension. For example, the girl's height measurements we had earlier gives us a point in the 3D space $\vec{h} : (h_1 = 150, h_2 = 148, h_3 = 162)$. Here we used the raw data points as our coordinate system: the first measurement is our first dimension, the second measurement is our second dimension and so on. Fig. 1 below shows the vector $\vec{h}$ and its projection in the $h_1 - h_2$ plane.

Just like in a real 3D coordinate system, we can project this vector $\vec{h}$ onto any other **orthonormal** (orthogonal with modulus = 1) coordinate systems. The goal of the projection: divide the n-dimensional sample space into two parts, **model space** and **error space**. Model space should contain part of the sample space representing the mean (expection) of the random variable H; error space has the zero-mean (centralized) values so we can exploit some nice properties of them.

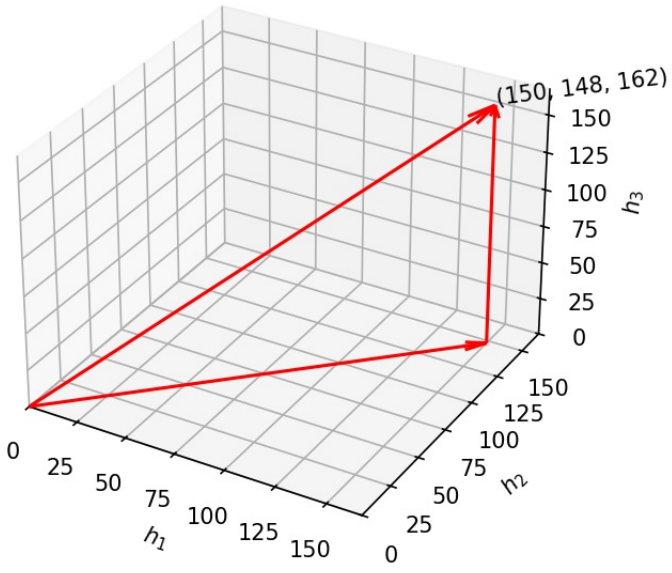

Figure 1: The sample vector $\vec{h}(h_1 = 150, h_2 = 148, h_3 = 162)$

Let's first pick an orthonormal basis for the 3D sample space:

$$\frac{1}{\sqrt{3}}\begin{bmatrix}1\\1\\1\end{bmatrix}, \quad \frac{1}{\sqrt{2}}\begin{bmatrix}1\\0\\-1\end{bmatrix}, \quad \frac{1}{\sqrt{6}}\begin{bmatrix}1\\-2\\1\end{bmatrix}$$

If we name the above vectors $\mathbf{u_1}$, $\mathbf{u_2}$ and $\mathbf{u_3}$, it's straightforward to confirm that these 3 vectors are orthogonal to each other and their modulus are all equal to 1. Therefore, they span the 3D sample space nicely. Now let's project our sample vector $\vec{h} : (h_1 = 150, h_2 = 148, h_3 = 162)$ onto these 3 vectors. As shown in Fig. 2, the blue vector is the original sample vector and the 3 red vectors illustrate the projection of the blue sample vector onto the 3 new coordinates. The first thing we'll notice in Fig. 2, the coordinate represented by $\mathbf{u_1}$ $([1, 1, 1]/\sqrt{3})$ seems to align very well with the raw data (blue vector): the two vectors are almost overlapping and the projection on $\mathbf{u_1}$ is a large value. In fact, the projections of the blue vector on the two other coordinates are so small they are barely visible in the lower left corner of the diagram.

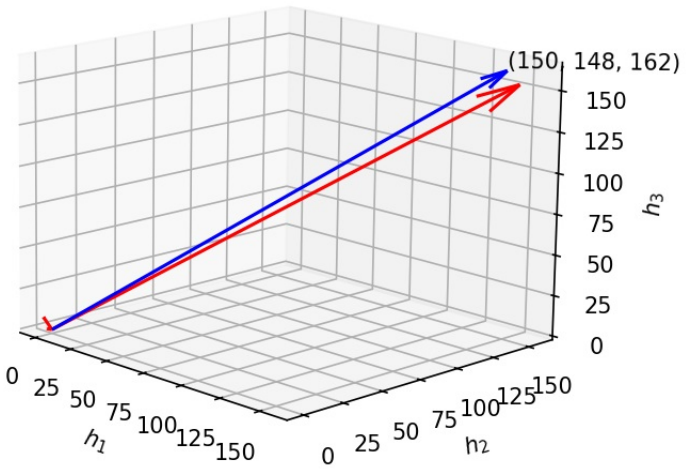

Figure 2: The sample vector projected onto the new coordinate system

Let's examine the actual projected values:

$$\begin{bmatrix} 150 \\ 148 \\ 162 \end{bmatrix} \cdot \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{460}{\sqrt{3}}$$

$$\begin{bmatrix} 150 \\ 148 \\ 162 \end{bmatrix} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \frac{-12}{\sqrt{2}}$$

$$\begin{bmatrix} 150 \\ 148 \\ 162 \end{bmatrix} \cdot \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = \frac{16}{\sqrt{6}}$$

Let's denote these projected values as $\mathbf{p_1}$, $\mathbf{p_2}$ and $\mathbf{p_3}$. Compared to $460/\sqrt{3}$, $-12/\sqrt{2}$ and $16/\sqrt{6}$ are very small, this explains why they are almost invisible in Fig. 2. Now let's prove a few important things so we can divide this 3D sample space into **model space** with the mean value and **error space** with zero-mean.

3.1.1 Model space and error space

First let's examine the 3 projected values p_1, p_2 and p_3. Note the projected value p_i is itself a **random variable** because it's a linear combination of 3 **indepedent, identically distributed (i.i.d.)** random variables h_k. The three measurements h_1, h_2 and h_3 are independent of each other: whatever we measure for h_2 is not affected by our measurement of h_1 and vice versa. They are identically distributed because they all follow the same underlying distribution of the year 7 girl's height.

If we check the expectation (mean) of p_i, it's easy to show that

$$\begin{aligned} E(p_1) &= E(h_1/\sqrt{3} + h_2/\sqrt{3} + h_3/\sqrt{3}) \\ &= E(h_1)/\sqrt{3} + E(h_2)/\sqrt{3} + E(h_3)/\sqrt{3} \\ &= 3E(H)/\sqrt{3} \\ &= \sqrt{3}\mu \end{aligned} \tag{12}$$

$$\begin{aligned} E(p_2) &= E(h_1/\sqrt{2} + h_2 \times 0 - h_3/\sqrt{2}) \\ &= E(h_1)/\sqrt{2} - E(h_3)/\sqrt{2} \\ &= E(H)/\sqrt{2} - E(H)/\sqrt{2} \\ &= 0 \end{aligned} \tag{13}$$

$$\begin{aligned} E(p_3) &= E(h_1/\sqrt{6} - 2h_2/\sqrt{6} + h_3/\sqrt{6}) \\ &= E(H)/\sqrt{6} - 2E(H)/\sqrt{6} + E(H)/\sqrt{6}) \\ &= 0 \end{aligned} \tag{14}$$

From the above expectation calculation, we can tell that p_1u_1 forms the model space by itself since the projection p_1 of the original data vector on u_1 is the only one with non-zero mean. The other two projections p_2u_2 and p_3u_3 span the error space because the projections on u_2 and u_3 both have zero expectations (means).

3.1.2 The Pythagorean triangle

Using the projected values above, we can express the decomposed vectors after the projection:

$$\frac{460}{\sqrt{3}} \cdot \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 153.33 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\frac{-12}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = -6 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

$$\frac{16}{\sqrt{6}} \cdot \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} = 2.67 \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

The above 3 vectors are the decomposed vectors after the projection. Let's call them $\mathbf{v_1}$, $\mathbf{v_2}$ and $\mathbf{v_3}$ and denote the original vector $(150, 148, 162)$ with $\vec{h}$. According to our definition: $\mathbf{v_1} = \mathbf{p_1 u_1}, \mathbf{v_2} = \mathbf{p_2 u_2}, \mathbf{v_3} = \mathbf{p_3 u_3}$ where p_i is the projected value of $\vec{h}$ on unit vector u_i. It is trivial to show that

$$\vec{h} = \mathbf{v_1} + \mathbf{v_2} + \mathbf{v_3}$$

Or more specifically,

$$\vec{h} - \mathbf{v_1} = \mathbf{v_2} + \mathbf{v_3}$$

This leads to, according to Pythagorean theorem,

$$\|\vec{h} - \mathbf{v_1}\|^2 = \|\mathbf{v_2} + \mathbf{v_3}\|^2 = \|\mathbf{v_2}\|^2 + \|\mathbf{v_3}\|^2 \tag{15}$$

where $\|\cdot\|^2$ is the L_2 norm of a vector. In equation (15), we used the fact that $v_2 \perp v_3$: v_2 and v_3 are orthogonal to each other. We'll use this result later to demonstrate the unbiased variance calculation.

3.1.3 The variance of the projected value

The projected value p_i is the inner product of the unit basis vector u_i and the raw sample vector $\vec{h}$. i.e., $p_i = \sum_{k=1}^{3} h_k u_{ik}$, where $i \in (1, 2, 3)$ represents the ith unit basis vector and $k \in (1, 2, 3)$ represents the kth element of $\vec{h}$ and u_i. For example,

$$p_1 = \sum_{k=1}^{3} h_k u_{1k} = 150/\sqrt{3} + 148/\sqrt{3} + 162/\sqrt{3} = 460/\sqrt{3}$$

and

$$p_2 = \sum_{k=1}^{3} h_k u_{2k} = 150/\sqrt{2} + 148 \times 0 - 162/\sqrt{2} = -12/\sqrt{2}$$

When a random variable is the linear combination of independent random variables, e.g.,

$$p = \sum_{k=1}^{3} h_k u_k$$

we have the following:

$$\begin{aligned} Var(p) &= \sum_{k=1}^{3} u_k^2 Var(h_k) && (16) \\ &= Var(H) \sum_{k=1}^{3} u_k^2 && (17) \\ &= Var(H) && (18) \\ &= \sigma^2 && (19) \end{aligned}$$

Step (16) uses the independence property of h_k's. The proof of it is available in all introductory books for statistics so we'll omit it here for simplicity. Step (17) uses the fact that h_k's are identically distributed therefore they have the same variance. In step

(18), we used the property of the unit basis vector u_i: $\|u_i\|^2 = 1$. The above result applies to all 3 projected values p_1, p_2 and p_3.

3.1.4 unbiased estimate of σ^2

With the results from equation(13), (14) and (19), we can show that for $j = 2, 3$ (the error space), we have

$$\begin{aligned} \sigma^2 &= Var(p_j) \\ &= E([p_j - E(p_j)]^2) \\ &= E(p_j^2) \end{aligned} \tag{20}$$

Because we used the fact that $E(p_j) = 0$ from (13) and (14), the above result is not true for $j = 1$. Equation (20) reveals something surprising yet familiar: for projections in the error space,

$$E(p_j^2) = \sigma^2$$

That is, either of the two ($j = 2, 3$) squared projection values is an unbiased estimate of the population variance σ^2. Following the example of the unbiased estimate of the population mean in Section 2.1, if we average the two unbiased estimates p_2^2 and p_3^2, not only is the average still an unbiased estimate of σ^2, it is also the best unbiased estimate we can get because it gives an estimate with very low deviation from the average value, just as we argued in the last paragraph of Section 2.1. The conclusion now is, we get a pretty good estimation of the population variance σ^2 when we use $\frac{p_2^2+p_3^2}{2}$.

If we still remember equation (15), derived using Pythagorean theorem, we have

$$\begin{aligned}
\frac{\sum_{i=1}^{3}(h_i - \bar{H})^2}{2} &= \frac{\|\vec{h} - \mathbf{v_1}\|^2}{2} \\
&= \frac{\|\mathbf{v_2}\|^2 + \|\mathbf{v_3}\|^2}{2} \\
&= \frac{p_2^2 + p_3^2}{2} \\
&= \text{the best unbiased population variance estimate} \qquad (21)
\end{aligned}$$

In equation (21), we used the fact that all unit basis vectors $u_i, i = 1, 2, 3$ have modulus 1. Note the 3 elements in the model space vector v_1 are simply the average of all measurements 153.33 (see first paragraph of Section 3.1.2). Using the year 7 girl's height measurements, equation (21) demonstrates how we can get an optimal sample variance that is unbiased.

In summary, from an n-dimensional sample space, we can always construct a new orthonormal coordinates made up of a 1D model space representing the mean and an $(n-1)$D error space representing the centralized (zero-mean) samples. If we add up all the squared projection value $p_i^2, i = 2, \ldots, n$ in the error space and divide the sum by the dimension of the error space $n - 1$, we'll get the best possible population variance estimate that is unbiased, given the n samples. In other words, **the unbiased sample variance (Bessel's correction) can be thought of simply as the average $\mathbf{L_2}$ norm of error space components of the original sample vector**.

The figure below shows an example with projection values $(5, 12, 13)$ in the new coordinate system. The blue vector is projected onto $[u_1, u_2, u_3]$ as the 3 red vectors. If u_1 in the figure makes up the model space representing the mean, the $u_2 - u_3$ plane represents the error space, the magenta vector is an indicator of the unbiased variance of our random variable that has been measured 3 times. In fact, half of the L_2 norm (see equation(21))of the magenta vector will be our best unbiased sample variance given the 3 measurements.

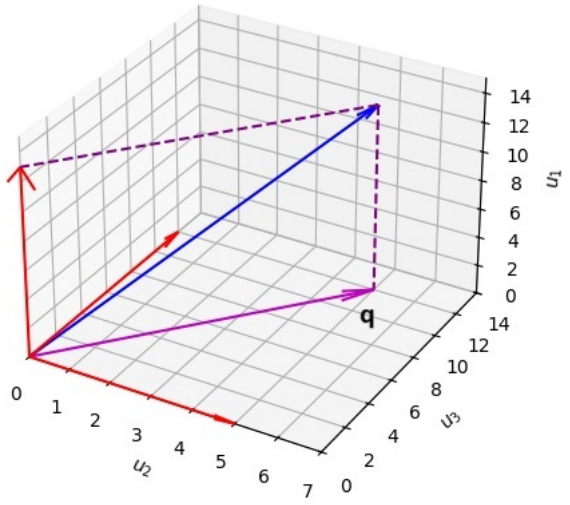

Figure 3: $\frac{1}{2}\|q\|^2$ is the unbiased sample variance

This way of looking at random variables and their samples is somewhat different from our usual mental model. However, the concept of model space and error space through projection is inate for most people: we tend to think of a random variable as something that just vibrates around a hidden mean value (1D model space) and how much it vibrates (error space) tells you how 'random' it is. And that is the measure of variance, which should be estimated only in the error space, hence the lost of 1 dimension (model space).

3.2 Two steps from hell

This section is named after one of my favorite music production companies, it's also meant as a tongue in cheek word play for what I'm about to describe: a geometric method to derive Bessel's correction with two stages, first we look at the difference between samples and their average, then we examine the difference between this average to the actual population mean.

Step 1: The formula for the *biased* sample variance given by

$$\sigma_b^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{22}$$

describes the difference between the samples and their average (sample mean) $\bar{X}$. If we think of the sample values as a vector $\vec{X}$, then equation(22) can be rewritten as:

$$\|\vec{a}\|^2 = n\sigma_b^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \|\vec{X} - \mathbf{1_n}\bar{X}\|^2 \tag{23}$$

where $\mathbf{1_n}$ is the n-dimensional uniform vector $[1, ..., 1]$. Equation (23) involves 2 vectors: $\vec{X}$ and $\mathbf{1_n}\bar{X}$. If we reuse the year 7 girl's height example earlier, $n = 3$,

$$\vec{X} = \begin{bmatrix} 150 \\ 148 \\ 162 \end{bmatrix}, \quad \mathbf{1_n}\bar{X} = \begin{bmatrix} 153.33 \\ 153.33 \\ 153.33 \end{bmatrix}$$

and we denote their difference as $\vec{a}$.

Now we need to recognize something critical for our next argument: $\vec{a}$ is perpendicular to $\mathbf{1_n}\bar{X}$. This is because $\mathbf{1_n}\bar{X}$ is the projection of $\vec{X}$ onto the vector $\frac{\mathbf{1_n}}{\sqrt{n}}$. We can easily confirm this with the above example:

$$\vec{a} \cdot \mathbf{1_n}\bar{X} = 0 \Longrightarrow \vec{a} \perp \mathbf{1_n}\bar{X}$$

Now we are ready to examine the second stage.

Step 2: if we construct another vector $\mathbf{1_n}\mu$ where μ is the population mean, just like we did with $\mathbf{1_n}\bar{X}$, it's easy to see that these two vectors are in the same direction. The only difference between them is the amplitude. We define their difference as $\vec{b} = \mathbf{1_n}\bar{X} - \mathbf{1_n}\mu$ and we have:

$$\|\vec{b}\|^2 = \|\mathbf{1_n}\bar{X} - \mathbf{1_n}\mu\|^2 = n(\bar{X} - \mu)^2 \tag{24}$$

Fig. 4 below illustrates the original sample vector $\vec{X}$, the sample mean vector $\mathbf{1_n}\bar{X}$ and the population mean vector $\mathbf{1_n}\mu$. As proved in step 1, the triangle by the 3 vectors:

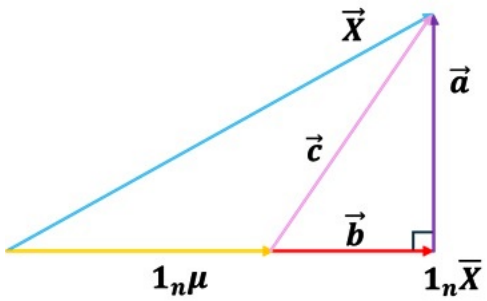

Figure 4: The right angle triangles

$\vec{X}$, $\mathbf{1_n}\bar{X}$ and $\vec{a}$ is a right angle triangle. Note in this figure we assumed μ is smaller than $\bar{X}$. In reality, μ can be smaller or larger, but it does not affect the generality of our derivation.

Now let's pay attention to the magenta vector in Fig. 4: $\vec{c} = \vec{X} - \mathbf{1_n}\mu$. By definition, we have

$$\|\vec{c}\|^2 = \|\vec{X} - \mathbf{1_n}\mu\|^2 = \sum_{i=1}^{n}(X_i - \mu)^2 \tag{25}$$

We can confirm that $\frac{\|\vec{c}\|^2}{n}$ is an unbiased estimator of the population variance σ^2:

$$\begin{aligned} E(\frac{\|\vec{c}\|^2}{n}) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i - \mu)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(E(X_i^2) - \mu^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}(E(X^2) - E(X)^2) \\ &= E(X^2) - E(X)^2 = \sigma^2 \text{ the population variance} \end{aligned} \tag{26}$$

In equation (26), we used the i.i.d. property of the samples. Pythagorean theorem tells us that $\|\vec{c}\|^2 = \|\vec{a}\|^2 + \|\vec{b}\|^2$ (Fig. 4), which means (see equation (23) and (24)):

$$\|\vec{c}\|^2 = n\sigma_b^2 + n(\bar{X} - \mu)^2 \Longrightarrow \frac{\|\vec{c}\|^2}{n} = \sigma_b^2 + (\bar{X} - \mu)^2$$

Take expectation on both sides and referring to equation (26), we get

$$\begin{aligned} \sigma^2 &= E(\sigma_b^2) + E[(\bar{X} - \mu)^2] \\ &= E(\sigma_b^2) + Var(\bar{X}) \\ &= E(\sigma_b^2) + \frac{\sigma^2}{n} \end{aligned} \tag{27}$$

Then we have

$$E(\sigma_b^2) = \frac{n-1}{n}\sigma^2, \tag{28}$$

which is Bessel's correction. In equation (27), we the fact that

$$\begin{aligned} Var(\bar{X}) &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X) \quad \text{i.i.d. property of } X_i \\ &= \frac{1}{n^2} \cdot nVar(X) \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{29}$$

This explanation is based on a blog by Rui Sun [2]. I essentially just expanded some of the formulae and added a drawing (Fig. 4). The main ideas is, to find out how your samples deviate from the actual population mean, first find out how the samples vary from the sample mean, then add on top of that the deviation of the sample mean from the population mean. It might not be 'two steps from hell', but it definitely takes some intuitive understanding of population mean, population variance, sample mean and sample variance.

Phew... That was a lot of work! If there's one take home message you should get from Section 3, I'd say: when you look at things from a geometric perspective, think Pythagoras...

4 Summary

Hopefully the mathematical proof in Section 2 and the two intuitions in Section 3 not only convinced you that unbiased variance is the way to go, you also gained some insight into the behaviour of a random variable and its estimates. A few things we should probably summarize here as a final note:

- Any sample of a random variable provides an unbiased estimate of its population mean.
- We use average instead of just a single sample to estimate the population mean because we want an unbiased estimate of the population mean while minimizing the deviation of our estimate.
- We can think of the n-dimensional sample space as two parts: model space made of 1 sample mean vector and the error space made of $n-1$ zero mean vectors. Only vectors in the error space should be used to estimate the variance, hence the $n-1$ factor in the unbiased sample variance.
- Bessel's correction gives an unbiased sample variance, however it does not give unbiased standard deviation. For information on unbiased standard deviation, check out this Wiki page.

References

[1] D. Saville, G. R. Wood, *Statistical Methods: The Geometric Approach*, Springer, New York, 1991.

[2] R. Sun, "Why's there an (n-1) in unbiased variance estimation?" *https://ruishu.io/2023/04/28/why-n-minus-1/*, accessed May 2024.