Sample variance

introduction

This write-up tries to explain why there is a factor of $n-1$ in the formula for the sample variance. It's not intuitive, especially since its twin, the population variance, has a factor of n.

I'm a stats amateur, so although I've tried hard to eliminate errors, one or more may remain. My primary source was

`https://en.wikipedia.org/wiki/Bias_of_an_estimator`

which I have tried to expand so as to leave no unexplained steps.

We suppose there is a population from which we obtain samples of a random variable, X, which are iid (independent and identically distributed), with a mean of μ and variance σ^2.

The definitions of the population mean and variance are

$$\mu = \frac{1}{N}\sum_{k=1}^{N} X_k \qquad \sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(X_k - \mu)^2$$

Usually, we'd just be given the value of μ and σ^2, as in "you draw from a normal distribution with mean of m and variance of v".

We take samples of size n from this population, and define the mean of the values in the sample as

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

Naively, let us define the sample variance similarly

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

We will show that this second definition must be adjusted. It is biased, by which we mean that if samples are drawn repeatedly and S^2 is calculated, the mean of those values will not be equal to the population variance σ^2.

The denominator must be $n-1$ rather than n. The reason turns out to be closely related to the standard error of the mean.

It is traditional to use E, the expected value, or expectation. Clearly, if we take a bunch of samples, we will get a value for the sample mean $\bar{X}$ which is close to μ but may vary.

The expected value of $\bar{X}$ is just μ, since if we take $m = N/n$ samples

$$E\ [\ \bar{X}\ ]\ = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{N}\sum_{i=1}^{N} X_i = \mu$$

The problem is that

$$E\ [\ S^2\ ]\ \neq \sigma^2$$

R code

Here is an R function to calculate SoSqD, the *sum of squared differences*:

```
SoSqD <- function(v) {
  m = mean(v)
  v = (v - m)^2
  return(sum(v)) }
```

Try it out:

```
>
> SoSqD(1:3)
[1] 2
>
```

Here is a second R function to exercise it. The function takes an operation. Then it draws n numbers from the normal distribution with mean $= 0$ and sd $= 1$, and feeds them to the operation we have specified.

```
f <- function(op,n) {
  N = 1000
  L = c()
  for (i in 1:N) {
    v = rnorm(n)
    L = append(L,op(v)) }
  print(mean(L)) }
```

It does this N times and prints the mean of the values produced.

This is, apparently, idiomatic R. We make an empty vector, then append to it and reassign the result to the vector itself.

I've used a random seed so you should see the same results.

```
> set.seed(131)
```

When I test this with the mean function the results are very close to zero, as we'd expect:

```
> for (i in 1:3) { f(mean,10) }
[1] 0.004191045
[1] -0.006282398
[1] 0.008173401
>
```

That's pretty close to the real mean. But when tested with the SoSqD function:

```
> for (i in 1:3) { f(SoSqD,10) }
[1] 8.912271
[1] 9.502785
[1] 9.015266
> for (i in 1:3) { f(SoSqD,5) }
[1] 3.983878
[1] 4.025661
[1] 3.919965
> for (i in 1:3) { f(SoSqD,4) }
[1] 3.03644
[1] 3.002947
[1] 2.940534
>
```

If we were to divide the output from SoSqD by n we would have what is defined to be the population variance. But that would underestimate the sample variance (our samples were taken from a population with variance equal to 1).

That would be a systematic difference from what we want to estimate. It looks like we should divided by $n - 1$ rather than n so we get the "right" answer. Let's see why.

derivation

This section shows an analysis of the bias. All the sums here are the same, so we will suppress the index on $\sum$.

The wikipedia derivation starts with the expectation applied as an operator to both sides of the definition for the sample variance from above. The first term is $E[S^2]$ and so on.

I'm just going to run through the algebraic manipulations without the E. We start with the sample variance as we defined it provisionally:

$$S^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$$

Our trick is to subtract and then add back μ:

$$= \frac{1}{n} (\sum (X_i - \mu) - (\bar{X} - \mu))^2$$

Multiplying out

$$= \frac{1}{n} (\sum (X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2)$$

Distributing the sum

$$= \frac{1}{n} \sum (X_i - \mu)^2 - 2\frac{1}{n} \sum (X_i - \mu)(\bar{X} - \mu) + \frac{1}{n} \sum (\bar{X} - \mu)^2$$

The first key manipulation is that the difference $\bar{X} - \mu$ can come out of the sums

$$= \frac{1}{n} \sum (X_i - \mu)^2 - 2\frac{1}{n} (\bar{X} - \mu) \sum (X_i - \mu) + \frac{1}{n} (\bar{X} - \mu)^2 \sum 1$$

And the third term would then be just $(\bar{X} - \mu)^2$.

Clearly μ is just a number. I originally justified this by saying that $\bar{X}$ is just a number as well, but I puzzled over it, since for a sequence of samples $\bar{X}$ is obviously a random variable

aside

Let's just take a quick look at $\sum X_i - \mu$.

$$\begin{aligned}\sum(X_i - \mu) &= \sum X_i - \sum \mu \\ &= n\bar{X} - n\mu \\ &= n(\bar{X} - \mu)\end{aligned}$$

So what about

$$\frac{1}{n}\ \sum(X_i - \mu)(\bar{X} - \mu)$$

We can certainly multiply out and distribute *that* sum

$$= \frac{1}{n}\ \sum X_i\bar{X} - \frac{1}{n}\ \sum X_i\mu - \frac{1}{n}\ \sum \mu\bar{X} + \frac{1}{n}\ \sum \mu^2$$

The second and fourth terms are no problem

$$-\mu\ (\frac{1}{n}\ \sum X_i - \mu\frac{1}{n}\sum 1) = -\mu\ (\bar{X} - \mu)$$

The problem is really what to do with

$$\frac{1}{n}\ \sum X_i\bar{X}$$

or even

$$\frac{1}{n}\ \sum \mu\bar{X}$$

And what I came up with was to bring back the index

$$\sum_{i=0}^{n} \bar{X}$$

and realize that there is no i in $\bar{X}$.

For any individual sample, for any particular calculation of S^2, $\bar{X}$ *is* a fixed constant so

$$\frac{1}{n} \sum_{i=0}^{n} \bar{X} = \bar{X}$$

$$\frac{1}{n} \sum \mu \bar{X} = \mu \bar{X}$$

$$\frac{1}{n} \sum X_i \bar{X} = (\bar{X})^2$$

So the first and third terms are

$$= \frac{1}{n} \sum X_i \bar{X} - \frac{1}{n} \sum \mu \bar{X}$$

$$= \bar{X}^2 - \mu \bar{X} = \bar{X}(\bar{X} - \mu)$$

Combining all four terms we get $(\bar{X} - \mu)^2$. Then, pick up the leading factor of -2 and cancel with the third term to get a single copy of $(\bar{X} - \mu)^2$.

back to the main argument

We are left with

$$S^2 = \frac{1}{n} \sum (X_i - \mu)^2 - (\bar{X} - \mu)^2$$

Let's put back the E. Each time we calculate S^2 we're going to get something a little different. The average or expected value depends on how $\bar{X}$ varies.

$$E[S^2] = E\ [\ \frac{1}{n} \sum (X_i - \mu)^2 - (\bar{X} - \mu)^2\ ]$$

The E operator is distributive ($E[X + Y] = E[X] + E[Y]$):

$$E[S^2] = E\ [\ \frac{1}{n} \sum (X_i - \mu)^2\ ]\ - E\ [\ (\bar{X} - \mu)^2\ ]$$

For the first term on the right-hand side, we could rewrite what's in the brackets as

$$\frac{1}{n}\sum(X_i - \mu)^2 = E\ [\ (X_i - \mu)^2\ ]$$

but this is just the *population* variance σ^2.

The article makes this substitution without comment

$$= E\ [\ \sigma^2\ ]\ = \sigma^2$$

We can understand this as meaning that σ^2 has a fixed value, so its expected value is that value, in the same way as $E[\ \mu\ ] = \mu$.

So now

$$E[\ S^2\ ] = \sigma^2 - E\ [\ (\bar{X} - \mu)^2\ ]$$

says that the expected value of S^2 as we have defined it is unfortunately not equal to σ^2 (the population value), which is what we'd really like to estimate by analyzing these samples.

There is a subtractive term which is non-zero because on the average, the sample mean is not precisely equal to the population mean. We should anticipate that as n increases, then S^2 will tend to σ^2.

If we look at the correction term, the claim is going to be that

$$E\ [\ (\bar{X} - \mu)^2\ ]\ = \frac{1}{n}\ \sum(\bar{X} - \mu)^2$$

is equal to

$$= \frac{1}{n}\ \sigma^2$$

This can be understood as follows. In taking multiple samples of size n, we compute $\bar{X}$ for each. In other words, *now* $\bar{X}$ is a random variable.

The correction term is exactly the variance of $\bar{X}$.

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\frac{\sum X_i}{n})$$

where the right-hand side uses the definition of $\bar{X}$.

One fact about variance is that for a constant c, $\mathrm{Var}(cX) = c^2\ \mathrm{Var}(X)$. (See the addendum).

So we can continue with

$$= \frac{1}{n^2}\ \mathrm{Var}(\sum X_i)$$

Another property of variance is that (for uncorrelated variables) it adds:

$$\mathrm{Var}(A + B) = \mathrm{Var}(A) + \mathrm{Var}(B)$$

So

$$= \frac{1}{n^2}\ [\sum \mathrm{Var}(X_i)\ ]$$

$$= \frac{1}{n^2}\ [\ \sigma^2 + \sigma^2 \cdots + \sigma^2\ ]$$

$$= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$$

This is a fundamental result. The variance of the mean is equal to the population variance of the random variable, divided by the number of items in the sample.

The standard deviation of the mean (called the standard error of the mean), is the square root of the variance of the mean, and it goes like $1/\sqrt{n}$.

Putting it all together

$$E[S^2] = \sigma^2 - E\ [\ (\bar{X} - \mu)^2\ ]$$

$$= \sigma^2 - \text{Var}(\bar{X})$$

$$= \sigma^2 - \frac{1}{n}\sigma^2$$

$$= \frac{n-1}{n}\sigma^2$$

Hence, to correct the left-hand side, we should multiply by $n/(n-1)$, but we decide instead to just divide by $n-1$ in our original formula.

addendum

Variances add, provided the variables are independent.

Proof.

The mean of the sum $X + Y$ is equal to the sum of the individual means $\bar{X} + \bar{Y}$. This is easily shown by distributing $1/n$ over the two sub-sums separately. Then

$$\text{Var}(X+Y) = \frac{1}{n}\sum((X_i + Y_i) - (\bar{X} + \bar{Y}))^2$$

$$= \frac{1}{n}\sum(X_i - \bar{X} + Y_i - \bar{Y})^2$$

One way is to just multiply out. The product has a total of 16 terms. Each term individually squared is present once, the six cross-terms are each present twice.

A better way is to group the terms

$$= \frac{1}{n}\sum((X_i - \bar{X}) + (Y_i - \bar{Y}))^2$$

$$= \frac{1}{n}\sum(X_i - \bar{X})^2 + (Y_i - \bar{Y})^2 + 2(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\ \text{Cov}(XY)$$

If the variance is taken of a constant times a random variable, then the constant comes out of the variance as the square. The reason is that the average *also* is multiplied by c.

First of all, the mean of cX for constant c and random variable X is $c\bar{X}$. *Proof.* Just factor c out of each term in the sum. Then

$$\mathrm{Var}(cX) = \frac{1}{n}\sum(cX - c\bar{X})^2$$

$$\frac{c^2}{n}\sum(X - \bar{X})^2$$

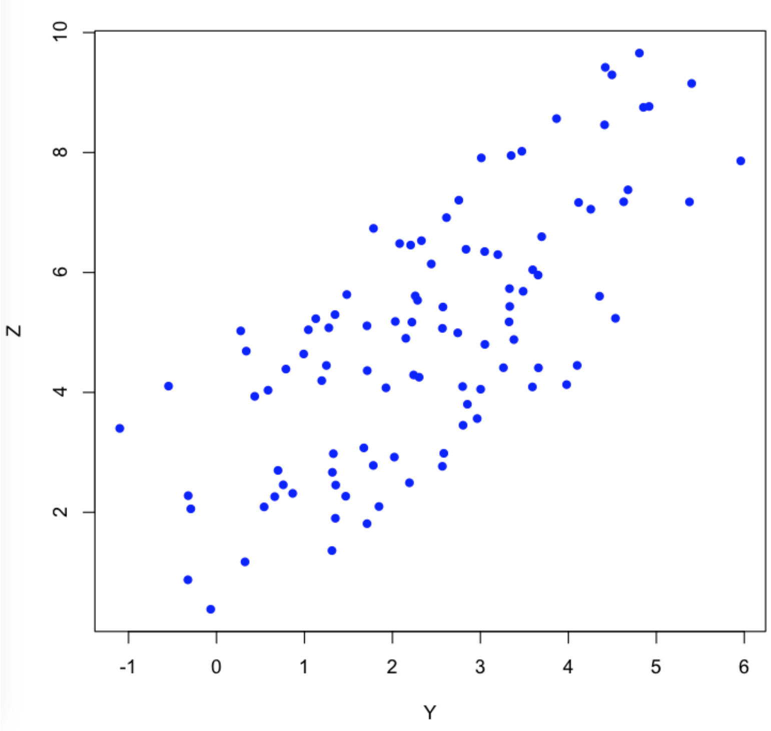
$$c^2\ \mathrm{Var}(X)$$

Here is some R code to test the first result. X is drawn from a random uniform distribution, adjusted to have mean ≈ 2.5, stdev ≈ 1.5.

Y is the same size as X, drawn from the normal with roughly the same mean and stdev.

Z is constructed by adding $X + Y$.

```
> set.seed(131)
> X = 1:100
> X = X/20
> Y = rnorm(100,2.5,1.45)
> Z = X + Y
> plot(Y,Z,pch=16,col=’blue’)
```



Z is correlated with both X and Y. We just consider Y, but the results would be similar for X.

```
> var(Y)
[1] 2.229434
> var(Z)
[1] 4.415937
> cov(Y,Z)
[1] 2.270602
>
```

We confirm that the variance of the sum is the sum of the variances, *plus* twice the covariance.

```
> var(Y+Z)
[1] 11.18658
> var(Y) + var(Z) + 2*cov(Y,Z)
```

```
[1] 11.18658
>
```

Even X and Y, which covary only a little bit by random chance, have this relationship:

```
> cov(X,Y)
[1] 0.04116795
> var(X+Y)
[1] 4.415937
> var(X) + var(Y)
[1] 4.333601
> var(X) + var(Y) + 2 * cov(X,Y)
[1] 4.415937
>
```