Statistics - Math 410, 11/7/2011

1 Estimating Variance

One of the main themes of this course is to estimate the mean μ of some variable X of a population. We typically do this by collecting a *sample* of n individuals out of the population, $\{x_1, \ldots, x_n\}$ and using the sample mean

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

as our estmate for μ. But in order to understand how good this estimate is, we also need to be able to make some estimation of the variance of X (occasionally we know the variance of X but not the mean μ, but most of the time, if we don't know μ, we also don't know the variance).

2 Definition of Variance and the "obvious" guess

Recall that the definition of the variance of X: Let the size of the population be N (typically much larger than our sample size). Label the value of X on the individuals in the population $y_1, \ldots, y_N$ (I'm using y so as not to confuse with the individuals in the sample as described above). Let the mean of X (calculated from the entire population) be μ.

Then

$$\sigma_X^2 = \mathrm{Var}(X) = \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{n}. \tag{1}$$

Of course making this calculation involves examining the *entire* population (as does computing μ), which is often impossible or impractical.

The "obvious" guess for how to estimate $\mathrm{Var}(X)$ is to use formula (1) replacing the entire population with the sample data $\{x_1, \ldots, x_n\}$, replacing N with n and replacing μ with $\overline{x}$. In other words, we might guess that

$$s_n^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n} \tag{2}$$

is the best way to estimate σ_X^2 from our data.

In fact this guess is wrong. It is what is called *biased* and will consistently *under*estimate the variance of X.

The *right* way to estimate the variance from a sample is

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}. \tag{3}$$

This is of course still an estimate, but it is centered around the actual population variance, unlike the estimate (2).

3 What's wrong with the biased estimate (2)?

The easiest way to see that there is something wrong with (2) is to look at examples for small values of n.

1. $n = 1$. Here we take a sample of size 1. Then $\overline{x} = x_1$ and (2) is *always* 0. So in this case, it consistently underestimates the variance (well, unless the variable X happens to be constant).

   It is easy to see that the problem is caused by the fact that we had to use the sample mean instead of the population mean.

2. Here is an example with $n = 2$. Let the random variable X be the result of a fair toin coss, that is a Bernoulli R.V. with $\pi = .5$. If you like, the population here is all coin tosses, and X is 1 if the toss results in a head, and 0 if tails.

   We know the expected value of X is $\mu = E(X) = .5$. We also know (from an easy calculation) that $\mathrm{Var}(X) = .25$.

   So let's calculate the possibilities for s_2^2 by looking at all possible samples of two coin tosses. Our possible samples are $\{0,0\}, \{0,1\}, \{1,0\}, \{1,1\}$, each of which occurs with probability 1/4.

{0,0}: In this case, $x_1 = x_2 = 0$. so $\overline{x}$ is also 0, and

$$s_2^2 = \frac{(0-0)^2}{2} + \frac{(0-0)^2}{2} = 0.$$

{0,1}: In this case $\overline{x} = .5$, so

$$s_2^2 = \frac{(0-.5)^2}{2} + \frac{(1-.5)^2}{2} = .25$$

{1,0}: As in the previous case, $s_2^2 = .25$.

{1,1}: As in the first case, $s_2^2 = 0$.

So if we let S_2^2 be the random variable defined on all possible samples of size 2 by formula (2), we notice from the 4 cases considered above that S_2^2 has two equally likely values: 0 and .25. So

$$E(S_2^2) = .125.$$

Note that this is half of $\mathrm{Var}(X)$, so is a rather bad estimate.

(One can repeat this with a Bernoulli variable with some other value of π. Then $\mathrm{Var}(X) = \pi(1-\pi)$. The four calculations of s_2^2 are the same, but they occur with probabilities $(1-\pi)^2, \pi(1-\pi), \pi(1-\pi)$ and π^2 respectively. So

$$E(S_2^2) = .5(\pi)(1-\pi)$$

Again, this is half of the actual value of $\mathrm{Var}(X)$, so a rather poor estimate.)

3. $n = 3$. If you found the previous example convincing, you may want to skip this one. Again take X Bernoulli with $\pi = .5$. As before $E(X) = .5$, $\mathrm{Var}(X) = .25$. Now take three individuals in your sample. The possible samples are

$$\{0,0,0\}, \{0,0,1\}, \{0,1,0\}, \{0,1,1\}, \{1,0,0\}, \{1,0,1\}, \{1,1,0\}, \{1,1,1\}.$$

The values one gets for s_2^2 are

$$0, \frac{2}{9}, \frac{2}{9}, \frac{2}{9}, \frac{2}{9}, \frac{2}{9}, \frac{2}{9}, \frac{2}{9}, 0.$$

Each of these outcomes are equally likely, so

$$E(S_3^2) = \frac{1}{8} \cdot 2 \cdot 0 + \frac{1}{8} \cdot 6 \cdot \frac{2}{9} = \frac{1}{6}.$$

Again, $E(S_3^2) < \mathrm{Var}(X)$.

The problem that comes up in all of these examples is caused by the need to use the sample mean rather than the population mean. The "distance" (in the least squares sense) from the sample data to the sample mean will be less than the distance to the population mean.

There is a pattern to be observed in the above examples. In each case (there are only 3),

$$E(S_n^2) = \frac{n-1}{n}\mathrm{Var}(X). \tag{4}$$

This is suggestive.

(Here

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

where X_i is the variable (defined on samples of size n) which is the value of X on the ith individual in the sample, and $\overline{X}$ is the mean of the X_i.)

4 The unbiased estimator for Var(X)

Our guess from equation (4) is that

$$\frac{n}{n-1} E(S_n^2) = \mathrm{Var}(X).$$

In other words, if we define

$$S^2 = \frac{n}{n-1} S_n^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1} \tag{5}$$

then we expect

$$E(S^2) = \mathrm{Var}(X).$$

This is not especially difficult to prove.

Theorem 1. *Let S^2 be defined as in (5). Then $E(S^2) = Var(X)$. In other words, S^2 is an unbiased estimator for $Var(X)$.*

Proof. We'll need to use:

$$E(aX + bY) = aE(X) + bE(Y)$$

$$\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$$

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$$

(assuming of course that X_1 and X_2 are independent). We also use that $\mu = E(X) = E(\overline{X})$.

Then

$$E(S^2) = E\left(\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2\right) = \frac{1}{n-1} E\left(\sum_{i=1}^{n} (X_i^2 - 2X_i\overline{X} + \overline{X}^2)\right)$$

We use the fact that $\sum X_i = n\overline{X}$, and the last expression above is then equal to

$$\frac{1}{n-1}E(\sum_{i=1}^{n} X_i^2) - 2nE(\overline{X}^2) + nE(\overline{X}^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)\right]$$
$$= \frac{1}{n-1}\left[nE(X^2) - nE(\overline{X}^2)\right] = \frac{n}{n-1}(E(X^2) - E(\overline{X}^2))$$

Now we remember that $\mathrm{Var}(X) = E(X^2) - E(X)^2$ so using that in the expression immediately above gives

$$= \frac{n}{n-1}\left[\mathrm{Var}(X) + E(X)^2 - \mathrm{Var}(\overline{X}) - E(\overline{X})^2\right]$$
$$= \frac{n}{n-1}\left[\mathrm{Var}(X) + \mu^2 - \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) - \mu^2\right]$$
$$= \frac{n}{n-1}\left[\mathrm{Var}(X) - \frac{n}{n^2}\mathrm{Var}(X)\right] = \frac{n}{n-1}\left[\frac{n-1}{n}\mathrm{Var}(X)\right] = \mathrm{Var}(X).$$

□

5 Summary

Here is what to keep in mind in applications.

1. If you are lucky enough to already know the population variance for some reason, then you don't need to approximate it.

2. If you know the population mean, μ, but not the variance, then

   $$\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

   is your best unbiased approximation to the population variance.

3. If you don't know the population mean or the population variance, then the sample variance

   $$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}$$

   is your best unbiased approximation to the population variance.

4. If you forget, and use

$$\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n} = \frac{n-1}{n} s^2$$

instead of s^2, it is not likely to make much difference if n is large, since $\frac{n-1}{n}$ is then close to 1. But you will be using an estimate that tends to underestimate the population variance.