

## 1 Section 7.5: Variance and Standard Deviation

If we have a random variable  $X$ , then the expected value gives us the “average” result. It is the what we expect to happen on average if the trial is repeated many times. But this only gives us a piece of the puzzle.

Ex: Let  $X, Y, Z$  be random variables for the following experiments.  $X$  is distributed binomially with  $n = 10$  and  $p = 0.3$ .  $Y$  is the distribution associated to picking a number between 0 and 6 randomly and recording the outcome. And  $Z$  is a random variable which always has outcome 3. Find  $E(X), E(Y), E(Z)$ .

Notice that the expected value for all three distributions was the same but they have some very different properties. Discuss some of the differences between them among yourselves.

Now compute  $P(2 \leq X \leq 4)$  and do the same for  $Y$  and  $Z$ . Take a few minutes to do that.

$$P(2 \leq X \leq 4) = P(X=2) + P(X=3) + P(X=4) = \binom{10}{2} \left(\frac{3}{10}\right)^2 \left(\frac{7}{10}\right)^8 + \binom{10}{3} \left(\frac{3}{10}\right)^3 \left(\frac{7}{10}\right)^7 + \binom{10}{4} \left(\frac{3}{10}\right)^4 \left(\frac{7}{10}\right)^6 \approx 0.7$$

On the other hand

$$P(2 \leq Y \leq 4) = \frac{3}{7} \text{ and } P(2 \leq Z \leq 4) = 1$$

So how would you describe these situations.

We often want to have an idea of how “spread out” a random variable is. By spread out, we want to know how far away from the mean it can get and how frequently. This gives us the following definition.

The **Variance** of a random variable  $X$  with mean  $\mu$  is

$$\text{Var}(X) = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \cdots + (x_N - \mu)^2 p_N$$

Notice that we need to square so that all the terms are positive.

Example: If  $X$  is a random variable representing the roll of a fair 6-sided die, find  $\text{Var}(X)$ .

First we have to find the mean or expected value. We can compute that

$$E(X) = 1 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$$

Now we can compute

$$\text{Var}(X) = (1 - 3.5)^2 \frac{1}{6} + \cdots + (6 - 3.5)^2 \frac{1}{6} = \frac{35}{12}$$

This formula is a little clunky and awkward, so we can use the alternative formula which is a lot more convenient

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

Take a minute and use this formula to compute the variance in the previous example. Verify that you got the same answer.

Another common measure of spread is called the **standard deviation** which you have probably heard of. The formula for standard deviation is much simpler.

$$\sigma_X = \sqrt{\text{Var}(X)}$$

This is called sigma sub  $X$  and it will be used to denote standard deviation. The variance and standard deviation give the exact same information, just written differently. For this reason we also refer to the variance as  $\sigma_X^2$ .

Ex: If  $X$  is the result of playing the VA match 3 lottery, find  $\sigma_X$ .

We can remember that there are two outcomes. The first is losing a dollar which happens with probability 999/1000 and we could also win \$499 which happens the rest of the time. We can calculate or remember that  $E(X) = -1/2$ . We can similarly compute that

$$E(X^2) = (499)^2 \frac{1}{1000} + (-1)^2 \frac{999}{1000} = 250$$

Thus  $\sigma_X^2 = 250 - 1/4 = 999/4$  and so  $\sigma_X = \sqrt{999}/2$ .

If  $X$  is a binomial distribution with parameters  $n$  and  $p$  and  $q = 1 - p$  then

$$\text{Var}(X) = npq = np(1 - p) \text{ and } \sigma_X = \sqrt{npq}$$

We also want to be able to compute these same ideas for sets of data. First we want to look at the population. Suppose  $\{x_1, x_2, \dots, x_N\}$  is a population with population mean  $\mu$ . Then the **population variance** and **population standard deviation** are given by

$$\sigma^2 = \frac{(x_1 - \mu)^2 + \dots + (x_N - \mu)^2}{N}$$

and  $\sigma = \sqrt{\sigma^2}$ . If the data is grouped together by frequency with  $x_1, \dots, x_r$  as the data with frequencies  $f_1, \dots, f_r$  then we can also get

$$\sigma^2 = (x_1 - \mu)^2 \frac{f_1}{N} + \dots + (x_r - \mu)^2 \frac{f_r}{N}$$

so again we see similarities between probability and relative frequency.

Now, suppose  $\{x_1, \dots, x_n\}$  is a sample with mean  $\bar{x}$  then the sample variance and standard deviation are denoted by  $s^2$  and  $s$  and have formulas

$$s^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$

and if the data is grouped by frequency we get

$$s^2 = \frac{1}{n - 1} [(x_1 - \bar{x})^2 f_1 + \dots + (x_r - \bar{x})^2 f_r]$$

and the standard deviation is the square root of the variance.

Why do we divide by  $n - 1$ . The answer is complicated but it essentially comes down to this,  $s$  is a random variable for some population and this formula makes  $E(s) = \sigma$  which is what we want.

One last thing. This should be giving us a measure of how spread our data is. Thus the smaller the variance the less likely we have values far from the mean. There is a theorem which makes this more precise.

Chebychev's Inequality:

Suppose that we have a probability distribution with mean  $\mu$  and standard deviation  $\sigma$ . Then the probability that a randomly chosen outcome lies between  $\mu - c$  and  $\mu + c$  is at least  $1 - (\sigma^2/c^2)$ .

Check that this makes some sense. When  $c$  is big then the probability should be big and these probabilities scale with  $\sigma$ . This bound can be unhelpful though. What is the probability that a data point lies between  $\mu - \sigma$  and  $\mu + \sigma$ ? It must be greater than 0 which we already know.

Example: Suppose that you know a test has a mean of 77 and a standard deviation of 8. If you know nothing else, what can you say about your probability of passing.

To pass you must get at least a 60. So we want to look at  $77 - 17$  to  $77 + 17$ . Chebychev's inequality tells us the probability that your score is in this range is at least  $1 - (8^2/17^2) = 225/289$ . But but you could also score higher than that so you can say that you have at least that good of a chance and probably better.