

## Lecture 13: October 19, 2009

CS 330 Discrete Structures  
Fall Semester, 2009

Did you hear about the statistician who drowned in a river with average depth of three inches?

### 1 How indicative of “real life” is the average?

The mean or expected value is often not enough when we are trying to draw some meaningful conclusions about the nature of a distribution. Suppose you were told that the average score on the CS 330 midterm was 50, you still do not know if everyone in the class scored a 50, or if half the class scored 0 and the other half 100, and so on.

In our analysis of the algorithm for determining the largest value in an array, we determined the average number of executions of the assignment statement  $m \leftarrow k$  to be  $H_n - 1$ . This, on its own, doesn't give us a lot of information about the behavior of the algorithm. We need some more information. So we now ask the question, how close to the mean are the actual values that I have?

#### 1.1 The difference

One way to measure the “averageness” is to calculate the expected value of the deviation, that is, the average value of  $x - a$ , where  $x$  is the variable with average value  $a$ . Because the average of a difference is the difference of the averages (why?), the average value of  $x - a$  is zero. That tells us that the variable  $x$  is just as often above  $a$  as it is below  $a$ , which we knew because  $a$  is the average of  $x$ .

#### 1.2 Absolute difference

We might want to phrase our last statement as: “What is the expected value of  $|x - a|$ , where  $a$  is the mean?” However, the function  $|x - a|$  does not suit our needs because it is not differentiable and hence cannot be subjected to several useful mathematical operations.

#### 1.3 Variance and Standard Deviation

We modify our goal and say: “What is the expected value of  $(x - a)^2$ ?” This value is referred to as the *variance* of the given distribution.

For any distribution, we know that the expected value of the random variable  $x$  is  $E(x) = \sum_{-\infty}^{\infty} x \cdot Pr(x)$ , where  $Pr(x)$  is the probability that the random variable takes the value  $x$ .

The variance is the expected value of  $(x - a)^2$ , and is given by  $E[(x - a)^2] = E(x^2) - a^2$ . The variance gives us a measure of the spread of values around the mean, except that it squares the deviation from the mean.

The *standard deviation*, often denoted by  $\sigma$ , is another common measure, and  $\sigma = \sqrt{Variance(x)}$ .

**Note** In general,  $E[f(x)] = \sum_{-\infty}^{\infty} f(x)Pr(x)$ .

## 1.4 Moments of a distribution

The *moments* of a distribution are certain other measures that provide some more insight into the shape of the distribution. The  $k$ th moment of a distribution is defined to be  $\sum_{-\infty}^{\infty} (x - a)^k Pr(x)$ , where  $a$  is the mean. The *variance* is simply the 2nd moment of a distribution.

## 2 Analyzing algorithms, finding the variance

Now that we know what variance is, we can try to determine the variance of  $i$ , the number of times the assignment statement executes in the algorithm to determine the largest element of an array.

The **probability generating function** is:

$$\mathcal{P}_n(x) = p_n(0)x^0 + p_n(1)x^1 + p_n(1)x^2 + \cdots = \sum_{i=0}^{\infty} p_n(i)x^i = \frac{x+n-1}{n}\mathcal{P}_{n-1}(x).$$

We also determined that

$$\mathcal{P}'_n(1) = e_n = H_n - 1,$$

where  $e_n$  is the expected number of executions of the assignment statement when the array size is  $n$ .

We know that the variance is given by  $E(i^2) - [E(i)]^2$ ,  $i$  being the number of executions of the assignment statement. And now we need to determine  $E(i^2)$ . Let us consider

$$\mathcal{P}''_n(x) = \sum_{i=2}^{\infty} (i)(i-1)x^{i-2}p(i) = \sum_{i=2}^{\infty} i^2 \cdot x^{i-2}p(i) - \sum_{i=2}^{\infty} i \cdot x^{i-2}p(i).$$

Comparing this with  $E(i^2)$  gives us:

$$E(i^2) = (\mathcal{P}''_n(1) + 0^2 \cdot p(0) + 1^2 \cdot p(1)) + \sum_{i=2}^{\infty} i \cdot p(i) = \mathcal{P}''_n(1) + \mathcal{P}'_n(1).$$

So, the variance is

$$\mathcal{P}''_n(1) + \mathcal{P}'_n(1) - (\mathcal{P}'_n(1))^2.$$

$$\mathcal{P}''_n(x) = \frac{2}{n}\mathcal{P}'_{n-1}(x) + \frac{x+n-1}{n}\mathcal{P}''_{n-1}(x)$$

$$\begin{aligned} \mathcal{P}''_n(1) &= \frac{2}{n}\mathcal{P}'_{n-1}(1) + \mathcal{P}''_{n-1}(1) \\ &= \frac{2}{n}(H_{n-1} - 1) + \mathcal{P}''_{n-1}(1) \\ &= \frac{2}{n}(H_{n-1} - 1) + \frac{2}{n-1}(H_{n-2} - 1) + \mathcal{P}''_{n-2}(1) \\ &= 2\left(\sum_1^n \frac{1}{i}H_{i-1} - \sum_1^n \frac{1}{i}\right) \end{aligned}$$

$$\begin{aligned}
&= 2\left(\sum_1^n \frac{1}{i}\left(H_i - \frac{1}{i}\right) - \sum_1^n \frac{1}{i}\right) \\
&= 2\left(\sum_1^n \frac{1}{i}H_i - \sum_1^n \frac{1}{i^2} - H_n\right)
\end{aligned}$$

By doing some simple regrouping of terms in the summation, we get

$$\begin{aligned}
S_n &= \sum_1^n \frac{1}{i}H_i \\
&= H_n^2 - S_n + \sum_1^n \frac{1}{i^2} \\
&= \frac{1}{2} \times \left(H_n^2 + \sum_1^n \frac{1}{i^2}\right)
\end{aligned}$$

(Imagine a square  $n \times n$  array in which position  $i, j$  has the value  $\frac{1}{ij}$  in it;  $H_n^2$  is the sum of all  $n^2$  array positions. But the array is symmetric around the main diagonal, so the sum of the elements on or above the diagonal equals the sum of the elements on or below the diagonal;  $S_n$  is that sum. In other words,  $H_n^2$  would be  $2S_n$ , but since  $S_n$  includes the diagonal elements  $1/i^2$ , those elements are included twice in  $2S_n$ . Hence  $H_n^2 = 2S_n - \sum 1/i^2$ .)

Using this result, we obtain

$$\mathcal{P}_n''(1) = H_n^2 - \sum_1^n \frac{1}{i^2} - 2H_n$$

Further, we use  $\sum_1^\infty \frac{1}{i^2} = \frac{\pi^2}{6}$  to find

$$\sigma^2 = H_n - O(1)$$

This is the variance for the number of times the assignment statement executes when we want to find the greatest value in an array. It is somewhat expected that the deviation from the mean will be greater as  $n$  increases, and that is what this result also tells us.