Illinois Institute of Technology
Department of Computer Science

Lecture 13: March 4, 2019

CS 330 Discrete Structures
Spring Semester, 2019

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

The average human being has one breast and one testicle.

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 now ask, how close to the mean are the actual values that we see in practice?

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 $\mathbf{E}(\mathbf{x}) = \sum_{\mathbf{x}} \mathbf{x}\,\mathrm{Pr}(\mathbf{x})$, 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

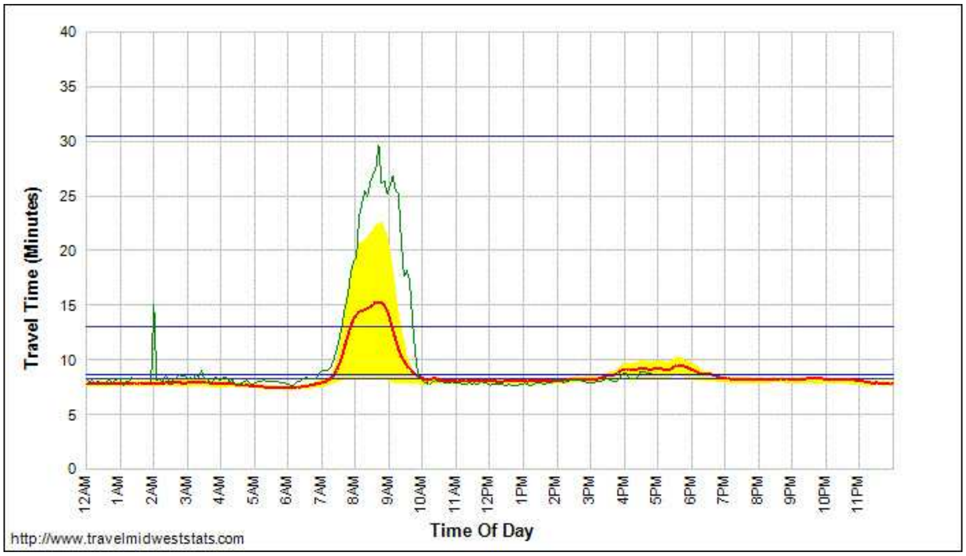
$$\mathbf{E[(x-a)^2] = E(x^2 - 2ax + a^2) = E(x^2) - 2aE(x) + a^2 = E(x^2) - a^2}.$$

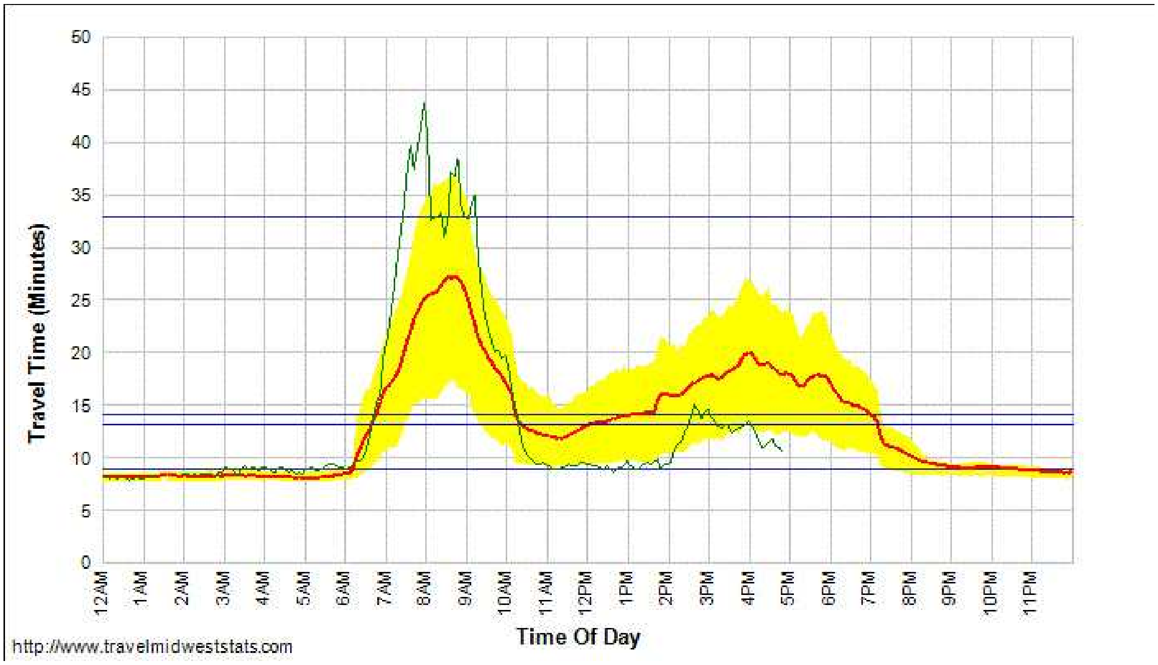
The variance gives us a measure of the spread of values around the mean, except that it squares the deviation from the mean.

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

In general, $E[f(x)] = \sum_x f(x)Pr(x)$.

Here is a nice example: The following two graphs show the commuting times from the north side of Chicago to the Loop. The first is along Lake Shore Drive from Bryn Mawr to Randolph, a distance of 7.62 miles and the second is along the Kennedy Expressway from Montrose to the Circle, a distance of 8.23 miles. In both cases the green line is the current travel time (on Monday, September 26, 2011 at 4:45pm), the red line is the average travel time for all samples ever collected, the yellow fill indicates one standard deviation from the mean (so about 68% of all travel times are in the yellow areas), the blue line is average overall travel time for all days/times, and the dark blue line indicate speed thresholds (55 mph, 35 mph, and 15 mph).





The distance from the red line to the (upper, say) edge of the yellow area is the *standard deviation* of travel time. At 8:30am for Lake Shore Drive it is about 7 minutes, while for the Kennedy it is about 10 minutes. The smaller value means that the average for the Drive is more indicative of what a driver actually experiences than for it is for the Kennedy. On the other hand, the standard deviations at, say, 3am are so small that the average travel times is a good indicator of what a driver would actually experience.

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 kth moment of a distribution is defined to be $\sum_x (x-a)^k Pr(x)$, 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:

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

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

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

We know from last time that the value of $\mathcal{P}_n'(1) = H_n - 1$, so all we have to evaluate is $\mathcal{P}_n''(1)$. Recall also from last time that

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

so

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

Now,

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

where $S_n = \sum_1^n \frac{1}{i}H_i$. Imagine an $n \times n$ array in which position (i, j) contains the value $\frac{1}{ij}$. Now multiply out all the terms in the product

$$H_n^2 = \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right).$$

That product 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 $2S_n$

Thus,

$$S_n = \frac{1}{2} \times \left(H_n^2 + \sum_1^n \frac{1}{i^2} \right).$$

From above,

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

so,

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

We know that $\sum_1^\infty \frac{1}{i^2} = O(1)$ from Lecture 2 (in fact, very advanced techniques can prove that $\sum_1^\infty 1/i^2 = \pi^2/6$) so that we have

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

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.

Exercise Find the variance for the "coupon collector's problem" discussed at the end of the previous lecture.