Probability and Mathematical Statistics in Data Science

Lecture 16: Section 6.1: Variance and Standard Deviation
Section 6.2: Simplifying the Calculation

Variation in Measurements

First Weight-Loss Program (Wide Variability)

-15, -10, -5, 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55

Mean Weight Loss: 20 lbs

Second Weight-Loss Program (Narrow Variability)

13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27

Mean Weight Loss: 20 lbs

How should we measure variability

- The expected value (μ_X) of a random variable X is a measure of *center*.
- Expectation is a weighted average indicating the center of the distribution of mass.
- How can we describe how the values of a random variable *vary* about this center of mass? How far does a typical value land from the center?

Variance of a random variable

- The difference between the values X takes and the mean is called the deviation from the mean or the average: ($D = X - \mu_X$)
- The *variance* of a random variable is defined by:

$$Var(X) = E(D^2) = E[(X - E(X))^2]$$

- Note that variance is an expectation of a function of X
- We could use the absolute deviation from the mean:
- $| D | = | X - \mu_X |$ but it isn't as nice a function as the square of the deviation from the mean.
- The only problem with using the variance is the units are off because we squared the deviation. In order to get the proper units back, we have to now take the square root.

Standard deviation of a random variable

- The *standard deviation* of a random variable is the *square root of the variance* of the random variable.

$$SD(X) = \sqrt{Var(X)} = \sqrt{E(D^2)} = \sqrt{E[(X - E(X))^2]}$$

- The variance is more convenient for computations because it doesn't have square roots. However, since the units are squared, it is difficult to interpret. Better to think about SD(X)

- SD(X) is a *give-or-take* number telling us how far the values of X are from μ_X on average, that is, it gives us a measure of the variability of the random variable.

The Variance of a Random Variable (X)

Let X have pmf $p(x)$ and expected value μ. Then the **variance** of X, denoted by $V(X)$ or σ_X^2, or just σ^2, is

$$V(X) = \sum_D (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

The **standard deviation** (SD) of X is

$$\sigma_X = \sqrt{\sigma_X^2}$$

Hypertension Example

$$\sigma^2 = \sum_{i=1}^{N} [X_i - E(X)]^2 P(X_i)$$

$Pr(X = r)$	.008	.076	.265	.411	.240
r	0	1	2	3	4

$$E(X) = 0(.008) + 1(.076) + 2(.265) + 3(.411) + 4(.240) = 2.80$$

Thus on average about 2.8 hypertensives would be expected to be brought under control for every 4 who are treated.

Find the standard deviation for the expected number of patients to be brought under control for every 4 who are treated.

Example: Videos

- A library has an upper limit of 6 on the number of videos that can be checked out to an individual at one time. Consider only those who check out videos, and let X denote the number of videos checked out to a randomly selected individual. The pmf of X is as follows:

x	1	2	3	4	5	6
$p(x)$	.30	.25	.15	.05	.10	.15

The expected value of X is easily seen to be $\mu = 2.85$. The variance of X is then

$$V(X) = \sigma^2 = \sum_{x=1}^{6} (x - 2.85)^2 \cdot p(x)$$

$$= (1 - 2.85)^2(.30) + (2 - 2.85)^2(.25) + \cdots + (6 - 2.85)^2(.15) = 3.2275$$

The standard deviation of X is $\sigma = \sqrt{3.2275} = 1.800$.

A shortcut formula for Variance

- The calculation of variance can be simplified:

$$\begin{aligned} Var(X) &= E((X-\mu_X)^2) \\ &= E(X^2 - 2X\mu_X + \mu_X^2) \\ &= E(X^2) - 2\mu_X E(X) + \mu_X^2 \\ &= E(X^2) - 2\mu_X^2 + \mu_X^2 \\ &= E(X^2) - \mu_X^2 \end{aligned}$$

$$V(X) = \sigma^2 = \left[\sum_D x^2 \cdot p(x)\right] - \mu^2 = E(X^2) - [E(X)]^2$$

Variance of Linear Function of Random Variables

- How would variance change if we add a constant to the values of the random variable?

$$\text{Var}(X) + c = \text{Var}(X)$$

- How would variance change if we multiply values of the random variable by a constant?

$$V(aX + b) = a^2 \cdot V(X)$$

Properties of Standard Deviation

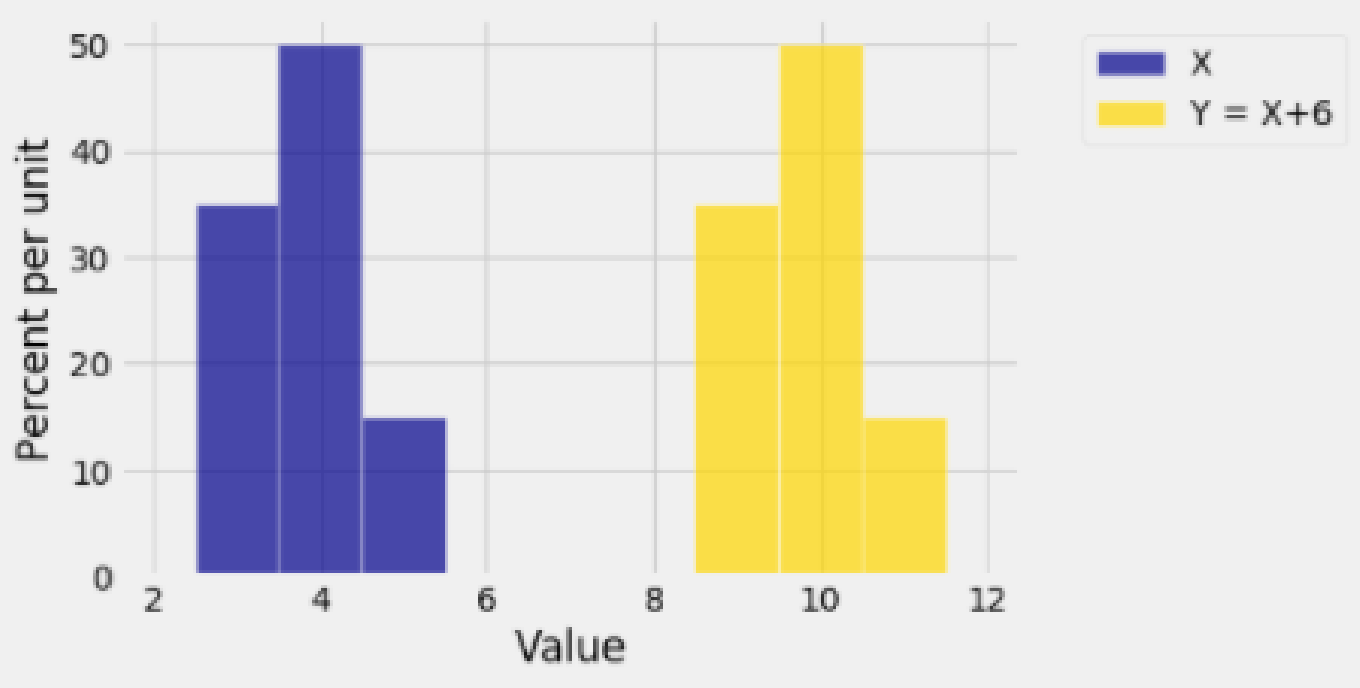
1. SD (constant) = 0

2. $SD(X + c) = SD(X)$, where c is a constant

3. $SD(cX) = |c|SD(X)$, thus $SD(-X) = SD(X)$

4. If X and Y are *independent*, then $Var(X \pm Y) = Var(X) + Var(Y)$

Example: Variability Unchanged by Adding Constant to Random Variable



Question

The pmf of the amount of memory X (GB) in a purchased flash drive was given in Example 3.13 as

x	1	2	4	8	16
$p(x)$	.05	.10	.35	.40	.10

Compute the following:

a. $E(X)$

b. $V(X)$ directly from the definition

c. The standard deviation of X

d. $V(X)$ using the shortcut formula