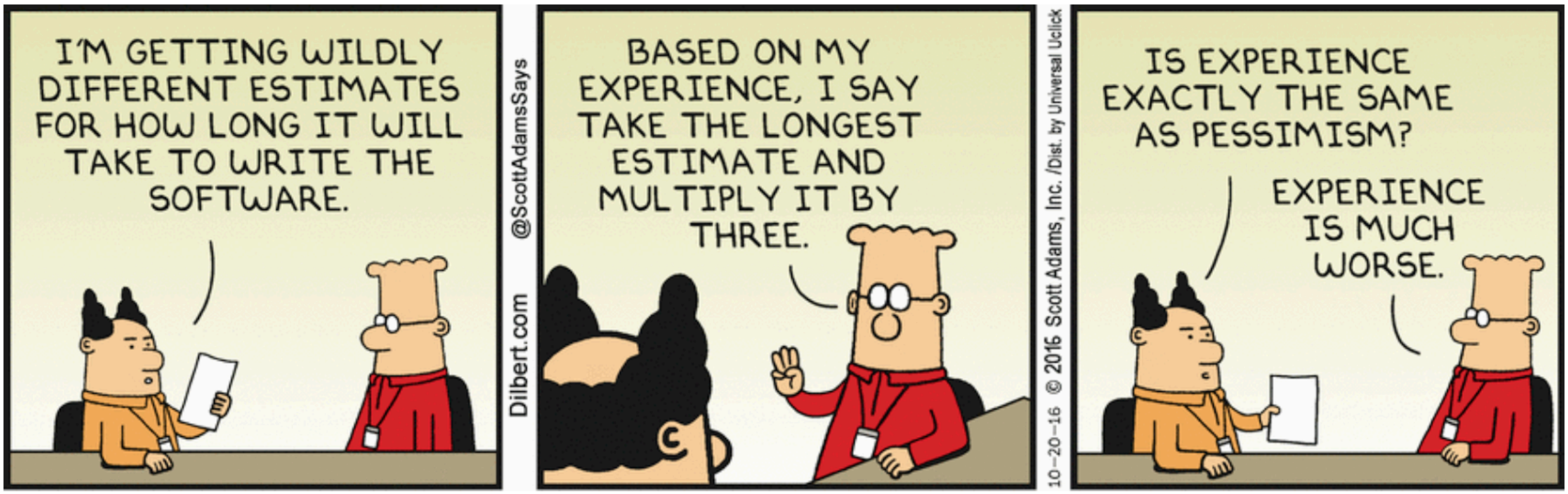
Stat 88: Probability & Mathematical Statistics in Data Science



Lecture 20: 3/8/2021

Sections 6.1, 6.2 (variance and standard deviation)

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 taken by the random variable ***vary*** about this center of mass? How far does a typical value land from the center?

- The difference between the values X takes and the mean is called the deviation from the mean or the average: ($D = X - \mu_X$)

- We could take each value of X, see how far it is from μ_X, and compute the (weighted) average of this distance.
- Why weighted? Values that are more likely should be counted more.

Measuring variability

- Suppose that X is a rv that takes values -1 and 1 with equal probability.

- We know what a measure of the variability should be, let's see if it works.

- Write out $E(D) = E((X - E(X))$. What problem do you see?
- How can we fix it?

Example

- Consider the random variable with distribution shown below:

x	1	2	3
$P(X = x)$	0.2	0.5	0.3
$(x - \mu_X)^2$			

- Find $E(X) = \mu_X$
- Write down the values of $(x - \mu_X)^2$ and find $E(X - \mu_X)^2$

Variance of a random variable

- 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.

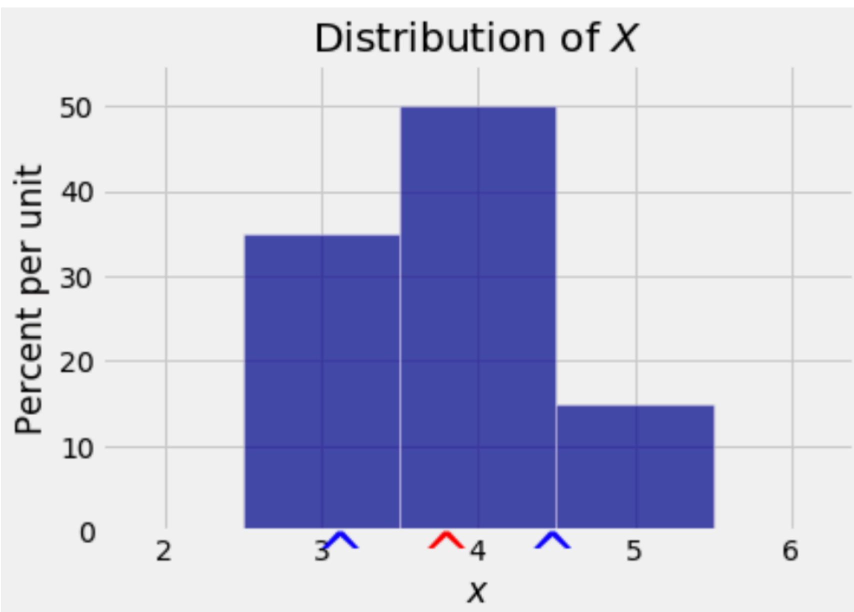
$$SD(X) = \sqrt{Var(X)} = \sqrt{E(D^2)} = \sqrt{E[(X - E(X))^2]}$$

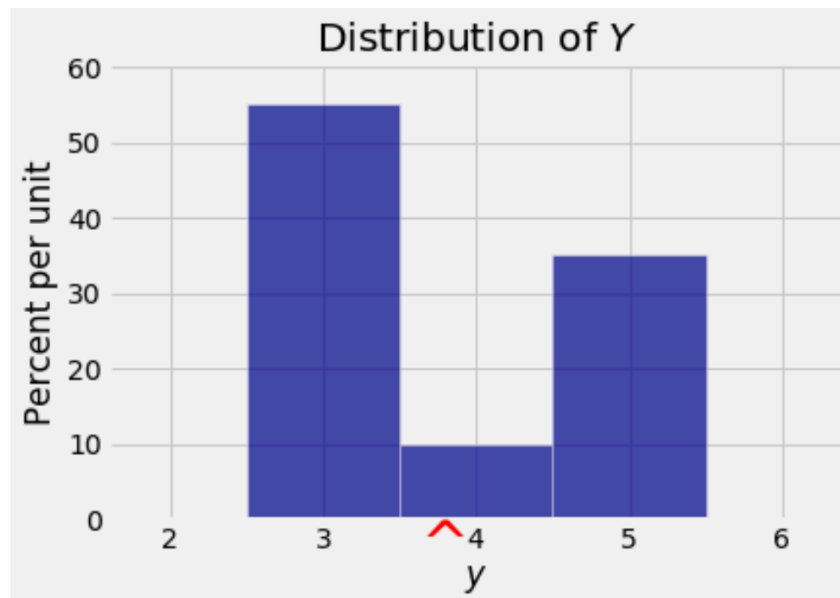
- The variance is more convenient for computations because it doesn't have square roots. However, since the units are squared, it is difficult to interpret. Better to think about SD
- You can think of SD as the RMS deviation from the mean (***root-mean-square***)
- *Deviation* is the amount above or below the expected value. How big is it likely to be?
- The likely or typical size of the deviation is given by the *standard deviation.*
- SD 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.

A tale (tail?) of two random variables.

x	3	4	5
$P(X = x)$	0.35	0.5	0.15
$(x - \mu_X)^2$			

y	3	4	5
$P(Y = y)$	0.55	0.1	0.35
$(y - \mu_Y)^2$			





A tale (tail?) of two random variables.

x	3	4	5
$P(X = x)$	0.35	0.5	0.15
$(x - \mu_X)^2$			

y	3	4	5
$P(Y = y)$	0.55	0.1	0.35
$(y - \mu_Y)^2$			

- For both X and Y, compute their means and variances, then the SDs.

Comparing X and Y

- For both X and Y, compute their means and variances, then the SDs.

Shortcuts and alternative formulas

- $E(X) = \sum x \cdot P(X = x) = \mu$

- $E(X^2) = \sum x^2 \cdot P(X = x)$

- $SD(X) = \sqrt{\sum (x - \mu)^2 \cdot P(X = x)}$

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

Shortcuts and alternative formulas

- If X is a random variable, and $E(X)$ is its mean

- Alternative formula for $Var(X)$:

$$E(X^2) - [E(X)]^2$$

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

Properties of SE

1. $SD(\text{constant}) = 0$
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 $Var(X \pm Y) = Var(X) + Var(Y)$

Properties of SE

1. SE(constant) = 0
2. $SE(X + c) = SE(X)$, where c is a constant
3. $SE(cX) = |c|SE(X)$, thus $SE(-X) = SE(X)$
4. If X and Y are *independent*, then

$$SE(X+Y) = \sqrt{SE(X)^2 + SE(Y)^2}$$