Math 316

Section 3.4 Density Curves

Recall histograms: For a numerical outcome such as weight or height (rather than a categorical outcome such eye color or gender), we can divide up the outcomes into ranges of values. The greater the number of ranges (the smaller the width of each range), the "smoother" the histogram becomes. The smooth curve is the density curve (or probability density curve or probability density function). It's a way to describe relatively how likely a particular outcome (actually a range of values of an outcome) is to occur. Frequencies: **Table 2.2.6/7, Figure 2.2.7**. More useful of course are relative frequencies: **Figure 2.2.6**. Finally: **Example 3.4.1**. **Online: Probability distribution of coin flips**.

The area under the curve between two points is:

- The fraction of all observation in the range, if we have actual observations.
- The probability that a particular value would be in that range, if we are trying to predict what would happen.

The total area under the curve is 1, since 100% of the data are somewhere under the curve. In **Examples 3.4.2 – 3.4.4** we see that 42% of the women have a blood glucose level between 100 and 150 mg/dl; described another way, the probability of a particular woman having blood glucose level between 100 and 150 is 0.42.

We talk about ranges of values rather than a specific value. Why? In other words, why would I ask a question like "What is the probability that the height of someone in our class is between 70 and 72 inches?" rather than "What is the probability that the height of someone in our class is exactly 71 inches?" See **The Continuum Paradox** at the bottom of page 100. The probability that a height is exactly 71 inches is 0. This is similar to how "the probability that a randomly chosen number is *exactly* 9381.0981837…" is 0. (There are an infinite number of possibilities, so the likelihood that a height is exactly 71 or the number is exactly 9381.0981837… is one 1 in infinity, which is the same as 0.)

Section 3.5 Random Variables

A random variable Y is a value that cannot be predicted ahead of time ("random") and that varies ("variable") from subject to subject. For example, if I were to ask each student in class his/her height, that value/observation would be a random variable. There can be discrete (specific) random variables (**Examples 3.5.1 – 3.5.3**) or continuous (any value in a certain range) random variables (**Example 3.4.4**, **Example 3.5.4**). Reminder: we can have *theoretical probabilities* (**Example 3.5.1**) or *frequency-based* probabilities—see top of page 85—in which the probability is based on past results (**Examples 3.5.2 – 3.5.4**).

We are interested in the <u>mean</u> μ_Y of a random variable Y is the <u>expected value</u> $E(Y)$. Where Y is the random variable and $y_1, y_2, \ldots$ are the possible values Y can take (recall upper case means "variable" and lower case means "values that variable can have"), that is, where

Possible outcome or value of Y	Probability of that outcome
y_1	$\Pr\{Y = y_1\}$
y_2	$\Pr\{Y = y_2\}$
$\vdots$	$\vdots$

then

$$E(Y) = \mu_Y = y_1 \cdot \Pr\{Y = y_1\} + y_2 \cdot \Pr\{Y = y_2\} + \cdots = \sum y_i \cdot \Pr\{Y = y_i\}$$

and

$$\begin{aligned}\sigma_Y^2 &= (y_1 - \mu_Y)^2 \cdot \Pr\{Y = y_1\} + (y_2 - \mu_Y)^2 \cdot \Pr\{Y = y_2\} + \cdots \\ &= \sum (y_i - \mu_Y)^2 \cdot \Pr\{Y = y_i\}\end{aligned}$$

Remember that Σ is a Greek "S" for "Sum." These two formulas for μ_Y and σ_Y^2 are given on pages 103 and 104.

Suppose we play a game. Flip a 2 coin twice. The amount you win is as in the table below. Let's simulate what might happen in **Class Example 1: Flip 2 coins**.

Here's what theoretically "should" happen:

Outcome	Relative frequency
T T ($0)	$1/4$
H T or T H ($1)	$2/4$
H H ($5)	$1/4$

On average we expect win (let $Y =$ winnings)

$$\mu_Y = E(Y) = (\$0)\left(\frac{1}{4}\right) + (\$1)\left(\frac{2}{4}\right) + (\$5)\left(\frac{1}{4}\right) = \$1.75$$

and the variance in winnings should be

$$\sigma_Y^2 = (0 - 1.75)^2\left(\frac{1}{4}\right) + (1 - 1.75)^2\left(\frac{2}{4}\right) + (5 - 1.75)^2\left(\frac{1}{4}\right) = \frac{251}{98}$$

so $\sigma_Y = \sqrt{\frac{251}{98}} \approx 1.98$. Notice that $.75 \le 1.98 \le 3.25$.

Observations, as seen in the above example:

- The expected value is not necessarily a value than can even occur. It is simply the average value you expect if you were to repeat the experiment several times.
- The expected value is a weighted average of the possible outcomes $y_1, y_2, \ldots$ and thus will be between the smallest and largest possible outcomes.

Another (kind of "mathy") idea worth knowing a bit about.

In general, there are four results related to changing random variables:

1. $\mu_{X \pm Y} = \mu_X \pm \mu_Y$
2. $\mu_{a+bY} = a + b\mu_Y$
3. $\sigma^2_{a+bY} = b^2\sigma^2_Y \Rightarrow \sigma_{a+bY} = b\sigma_Y$
4. $\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$

In Rules 2 & 3, Y represents the original data, and $a + bY$ represents the data that has been multiplied by b and had a added to it. We saw these two rules on the last page (the table at the top) of Handout 2.1 – 2.9. We also saw these rules evident in the **Online: Histograms…** from Chapter 2.

An example related to Rules 1 and 4:

Suppose X is a value from $\{1,2,2\}$. Then $\mu_X = 1\frac{2}{3}$ and $\sigma^2_X = \frac{2}{9}$ (so $\sigma_X = \sqrt{\frac{2}{9}} \approx .47$).

Suppose Y is a value from $\{3,3,5,5\}$. Then $\mu_Y = 4$ and $\sigma^2_Y = 1$ (so $\sigma_y = 1$).

Consider $X + Y$ which comes from the 12 possible sums of X and Y values $\{\, 4, 4, 6, 6, 5, 5, 7, 7, 5, 5, 7, 7 \,\}$.
Then $\mu_{X+Y} = 5\frac{2}{3}$ and the variance is $\sigma^2_{X+Y} = 1\frac{2}{9}$.

We see that $\mu_{X+Y} = \mu_X + \mu_Y$ and $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$. (You can actually prove that this happens in general.)

In general (so not for the specific problem above), what do

$$\mu_{X+Y} = \mu_X + \mu_Y \quad \text{and} \quad \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

look like?