Data Analytics Workshop, JHU Pre-College Program
Day 8: Expected Value (optional notes)

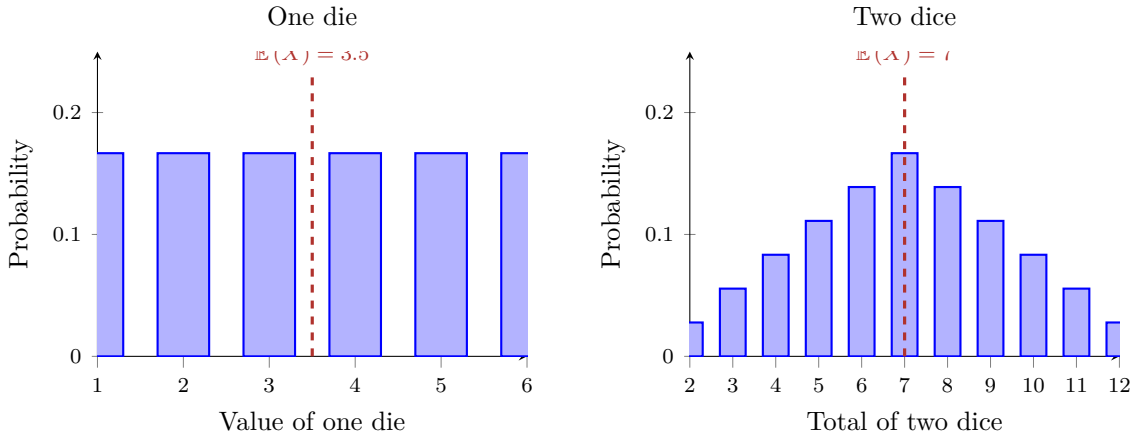
TL;DR The mean is an average of data we already collected. The **expected value** is the average we *expect* in the long run, computed from the distribution itself, before any data is collected.

(Expected value)

*If a random variable X can take the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, its **expected value** is*

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n.$$

- It is a weighted average of the possible values, weighted by how likely each one is. One die has $\mathbb{E}(X) = 3.5$, even though you can never roll a 3.5.
- Expected values add: $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. So three dice have expected total $3 \cdot 3.5 = 10.5$, exactly where the average settled on Day 2.
- The **law of large numbers**: as we collect more and more data, the mean of the data gets closer and closer to the expected value. The SEM from Day 4 says how close.
- Expected value tells us whether a game or bet is fair: a fair game has expected winnings of 0.



The expected value is the balance point of the distribution

1 Expected Value

On Day 4 we computed the mean of data we had already collected. But often we want to know what to expect *before* collecting anything. If I hand you a die, what should you expect to roll, on average? We do not need to roll it a thousand times to answer this, since we already know the distribution: each face comes up with probability $\frac{1}{6}$.

Definition 1.1 (Expected value)

If a random variable X can take the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, its ***expected value*** *is*

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i \, p_i.$$

That is, multiply each value by how likely it is, and add them up. Notably, this is a weighted average, just like the grade formula on Day 1, except now the weights are probabilities.

Example 1.1

For one die,

$$\mathbb{E}(X) = 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} = \frac{21}{6} = 3.5.$$

You will never actually roll a 3.5! The expected value does not have to be a possible outcome. It is the long run average, and the balance point of the distribution, like the mean was the balance point of the data.

Example 1.2

For a Bernoulli random variable that equals 1 with probability p (success) and 0 with probability $1 - p$ (failure),

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

So the expected number of hits in one at bat, for our batter from Day 3 who hits 30% of the time, is 0.3.

2 Expected Values Add

The most useful fact about expected value is this.

Proposition 2.1 (Linearity of expectation)

For any random variables X and Y and any constant c,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) \qquad \text{and} \qquad \mathbb{E}(cX) = c\,\mathbb{E}(X).$$

This looks just like the rules for sums on Day 1, and that is no coincidence: an expected value *is* a sum. What is surprising is that this works even when X and Y are not independent. Let's see how much work it saves.

Example 2.1

The expected total of two dice is $3.5 + 3.5 = 7$*, which matches the peak of the distribution on the first page. The expected total of three dice is* $3 \cdot 3.5 = 10.5$*. This is exactly the number the running average settled on in the Day 2 simulation! Computing it directly would mean listing all* $6^3 = 216$ *outcomes.*

Example 2.2

Our batter gets 10 *at bats per game and hits each with probability* 0.3*. The number of hits is a sum of* 10 *Bernoulli random variables, so the expected number of hits per game is* $10 \cdot 0.3 = 3$*. That is exactly where the binomial distribution peaked on Day 3. In general, a binomial distribution with* n *trials and success probability* p *has expected value* np*.*

3 The Law of Large Numbers

How does expected value connect back to data? The mean of our data is a statistic, and it is our best guess at the expected value.

Theorem 3.1 (Law of large numbers, informally)

If we collect more and more independent samples of a random variable X*, the mean of the samples gets closer and closer to* $\mathbb{E}(X)$*.*

This is what you saw on Day 2: the average of three dice wandered around at first and then settled near 10.5. The SEM from Day 4 tells us how close to expect: with N samples, the mean is usually within about $2 \cdot \frac{\sigma}{\sqrt{N}}$ of the expected value. So more data does not change the expected value; it changes how well our data can find it.

4 Is It a Fair Game?

Expected value is exactly how casinos (and insurance companies) make money. A game is **fair** if your expected winnings are 0.

Example 4.1

You pay \$1 *to roll a die. If you roll a* 6*, you get* \$5 *back; otherwise you get nothing. Your profit is* \$4 *with probability* $\frac{1}{6}$ *and* $-$\$1 *with probability* $\frac{5}{6}$*, so*

$$\mathbb{E}(\textit{profit}) = 4 \cdot \tfrac{1}{6} + (-1) \cdot \tfrac{5}{6} = -\tfrac{1}{6} \approx -\$0.17.$$

On average you lose about 17 *cents per game. You might win the first few, but by the law of large numbers, play long enough and you will lose about* 17 *cents a game. For the game to be fair, a* 6 *would need to pay back* \$6 *(check this!).*

Remark 1

Expected value can be surprising, and that is the point of Discussion 5. Straws weighing anywhere from 0 to 20 pounds are piled on a camel until the total goes over 100 pounds. A single straw weighs 10 pounds on average, so you might expect the final straw to weigh 10 pounds on average too. Does it? Run the camel straw simulator a few hundred times and see what the data says before you decide.

5 Practice

Problem 5.1 *1) A spinner lands on 1 with probability $\frac{1}{2}$, on 2 with probability $\frac{1}{3}$, and on 6 with probability $\frac{1}{6}$. What is its expected value?*

2) What is the expected total of 20 dice?

3) A raffle sells 500 tickets for \$2 each, and one ticket wins \$600. What is the expected profit of buying one ticket? Is it fair?

Solution:

1) $1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{3} + 6 \cdot \frac{1}{6} = \frac{1}{2} + \frac{2}{3} + 1 = \frac{13}{6} \approx 2.17$.
2) By linearity, $20 \cdot 3.5 = 70$.
3) You win \$598 with probability $\frac{1}{500}$ and lose \$2 with probability $\frac{499}{500}$, so the expected profit is $\frac{598-998}{500} = -\$0.80$. Not fair: the raffle collects \$1000 and pays out \$600, so on average each ticket loses $\frac{400}{500} = \$0.80$.