Mean and Standard Deviation of the Binomial Distribution

Reading

- Section 3.4.1, 3.4.2, 3.4.3

Practice Problems

3.6.4 (Page 163) 3.27, 3.28, 3.29, 3.30, 3.36

Notes

Mean and Standard Deviation for the Binomial

We have seen how to find the mean and standard deviation of combinations of variables, when those variables are independent of each other. We will now use that knowledge to find formulas for the mean and standard deviation of the binomial.

> Consider a binomial setting with parameters n and p, and denote by X the number of successes.
>
> - Denote by S_1 the "number of successes in the first trial".
> - Denote by S_2 the "number of successes in the second trial".
>
> and so on. Then:
>
> - All the S_i are independent of each other.
> - Each S_i follows the distribution given by the table:
>
> | S | 0 | 1 |
> |---|---|---|
> | Prob | 1-p | p |
>
> - From this table we find:
>
> $$\mu_{S_i} = p$$
> $$\sigma^2_{S_i} = p(1-p)$$

To see why these are true:

- The probability table for the S_i is pretty easy to see, since the i-th trial is a trial with the two options success/failure, and we count the number of successes; so either 0 or 1.

- Using the table we compute:

$$\mu_S = (1-p)\cdot 0 + p\cdot 1 = p$$
$$\sigma_S^2 = (1-p)(0-p)^2 + p(1-p)^2 = p(1-p)\left[p + (1-p)\right] = p(1-p)$$

We now use the S_i to say something about X:

- X relates to the S_i via:

$$X = S_1 + S_2 + \cdots + S_n$$

- Using this we find formulas for the mean and standard deviation of X:

$$\mu_X = np$$
$$\sigma_X = \sqrt{np(1-p)}$$

- If we define by $\hat{p}$ the “percent of successes”, namely

$$\hat{p} = \frac{X}{n}$$

  then we have the formulas:

$$\mu_{\hat{p}} = p$$
$$\sigma_{\hat{p}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

The first part is straightforward. Because the S_i are independent of each other, the formulas we learned earlier allow us to compute the mean and standard deviation of X using the ones for S_i:

$$\mu_X = \mu_{S_1} + \mu_{S_2} + \cdots + \mu_{S_n} = p + p + \cdots + p = np$$
$$\sigma_X^2 = \sigma_{S_1}^2 + \sigma_{S_2}^2 + \cdots + \sigma_{S_n}^2 = p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p)$$

The formulas for $\hat{p}$ follow from the fact that it is just a linear transformation from X, just dividing everything by n.

Approximation via Normal Distribution

These formulas become useful when n is large, because in that case we can approximate the binomial distribution with normal:

> When n "sufficiently large", then the binomial follows an approximately normal distribution. So we have:
>
> $$X \sim N\left(np, \sqrt{np(1-p)}\right)$$
> $$\hat{p} \sim N\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$
>
> The **rule of thumb** for when n is sufficiently large is that we should have:
>
> $$np \geq 10$$
> $$n(1-p) \geq 10$$
>
> Both of these conditions should be true. We only need to check the smallest of p, $1-p$, since if that one results in 10 or more then the other will do as well.

Using these we can quickly do some computations, without having to use the explicit formula for $P(X = k)$, which becomes very hard to use for large n.

Here's an example:

> In a multiple-choice test there are 100 questions. We pick answers at random and there are 5 possible answers to each question, so we have a 20% chance to answer each question correctly. What are the chances, that we will get at least 25 answers correct?

This is a binomial setting, since there is a fixed number of questions $n = 5$, and for each one we either get it right with probability $p = 0.2$ or we get it wrong, and since we pick answers at random the trials are independent. X measures the number of correct answers.

We start by computing the mean and standard deviation of X. We have:

$$\mu_X = np = 100 \cdot 0.2 = 20$$
$$\sigma_X = \sqrt{np(1-p)} = \sqrt{20 \cdot 0.8} = 4$$

We then check the rule of thumb: We need both np and $n(1-p)$ to be at least 10. But np is clearly the smallest of the two, and it is already ≥ 10, so we are OK and can use the normal approximation.

Therefore we can approximate X by $N(20, 4)$. The question therefore becomes, in $N(20, 4)$ how much data is above $x = 25$.

This is now a problem about a normal distribution, and we know well how to solve those problems. We would compute:

$$z = \frac{x - \mu}{\sigma} = \frac{25 - 20}{4} = 1.25$$

Then look that up in our table to get $p = 0.894$. Since this measures how many are below that value, we need to look at the rest, so $1 - 0.894 = 0.106$, or 10.6%. So there is roughly a 10% chance that we would score more than 25 points at random like that.

Continuity Correction One important topic to discuss is that of **continuity correction**. This is relevant when n is relatively small, like in this example.

The problem is this: We are approximating the binomial distribution with a normal distribution. But the binomial distribution corresponds to integers only, while the normal distribution allows for all numbers. So for instance according to the normal distribution there should be a number of students who scored between 24 and 25. But that is not possible. This is a discrepancy we need to somehow correct.

The fix is to divide the space between 24 and 25 in half, and count the upper half as part of 25, and the lower half as part of 24. What this means is that in these problems you want to often start halfway to the previous or next value, depending on the question.

In other words, in this instance we should be using $x = 24.5$ rather than $x = 25$. You can think of it as saying that we should include as part of 25 values that would have "rounded up to 25".

With that in mind, the computation would have been:

$$z = \frac{24.5 - 20}{4} = 1.125$$
$$p = 0.8697$$
$$1 - p = 0.1303$$

So with this computation, the answer would be closer to 13%, rather than 10%. This would be a better estimate in this case.

For comparison, the perfect answer, the one that would be computed if we did the exact formula for $P(X = k)$ for all numbers from 25 to 100, would have given us a percent of 13.135%.