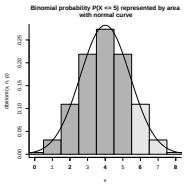


The normal and binomial

The binomial model — using n independent trials, each with success probability p to count something — can be used in many different guises. For fun today (and which days of Math 113 aren't fun?) we'll assume we are doing some medical research.

1 Binomial and Normal probabilities

A new cold medicine is being introduced. The manufacturer claims that in 100 cases, on average 60 patients will get better in 3 days. However, we want to test whether the manufacturer isn't exaggerating their claim. We do a test. We take 50 people with a cold, and administer the treatment. After 3 days we find 20 have gotten better. How likely is this given a manufacturer's claim?

We will answer the following questions:

What is the probability of exactly 20 getting better?
What is the probability of 20 or fewer getting better?

For both, a binomial model is used. Let X be the number that get better. If the manufacturer is right, the X should be binomial with

Question 1: Well, X is binomial with $n =$? and $p =$? (What does independent mean here.)

Since you all got that last one correct ($n = 50$, $p = 0.6$) we turn our questions into ones using math notation:

$$P(X = 20), \quad P(X \leq 20)$$

These are answered with `dbinom` and `pbinom`:

```
> n = 50
> p = 0.6
> dbinom(20, n, p)

[1] 0.001987

> pbinom(20, n, p)

[1] 0.003360
```

Both probabilities are pretty small. Perhaps the manufacturer exaggerates.

Another manufacturer markets a different cold medicine. They say that 50% of people will be better in 72 hours, and 68% of the people will improve in a range from 60 to 84 hours. We take the medicine and don't get better for 96 hours. How unlikely is that?

To answer this, we need a probability model. We assume that the distribution of the time a person will improve is normal with a mean of 72 and a standard deviation of 12. This is consistent with the numbers we know. Let X be a random variable with this distribution. Then we can answer our question by finding

$$P(X \geq 96)$$

The function `pnorm` will help. It answers $P(X \leq b)$ so we need to turn it around:

```
> mu = 72
> sigma = 12
> 1 - pnorm(96, mu, sigma)

[1] 0.02275
```

(For the normal, we don't have to worry about the difference between $P(X > b)$ and $P(X \geq b)$.)

Again, not too likely if the manufacturer is telling the truth.

For the following exercises, first write the answer in math form, then calculate a number. "Math form" means, define a random variable, then express the answer in a form like $P(X < a), P(X \leq a), P(X \geq a)$ etc.

Question 2: Suppose it is known that on average 23% of a medicine's user will not follows the usage instructions. In a study of 150 patients, what is the probability that 35 or more will not follow the instructions? What about 23 or fewer?

Question 3: It is claimed that *on average* 1% of users of some medication will feel severer discomfort. If 250 people are given the medication, what is the probability that 1 or fewer will feel this severe discomfort?

Question 4: A manufacturer assumes that the average squirt of some nasal decongestant delivers 0.5 ml of medication on average. They assume the distribution of squirt-size is normal with this mean and a standard deviation of 0.3 ml. For a randomly chosen user, calculate the probability they get fewer than 0.1 ml of medication. Then find the probability they get more than 1.0 ml.

Question 5: A fever medication claims to lower a 101 degree fever by an average amount of 2 degrees in 1 hour, with a standard deviation of 0.2 degrees. If the distribution of decline is normally distributed, find the probability that a patient's fever is lowered less than 1.5 degrees.

2 The normal approximation to the binomial

The computer makes calculating the normal and binomial distributions easy using the functions `pnorm()` and `pbinom()`. Historically, this was not the case and it was quite useful that

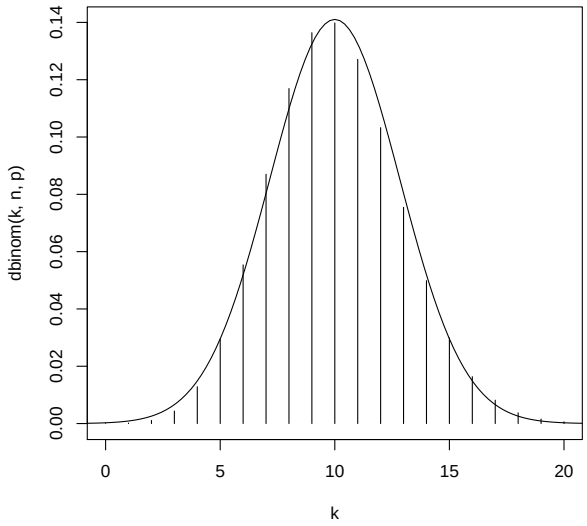
the normal could give probabilities that approximate the binomial. (It still is, as we shall see when we do confidence intervals for proportions.)

We can make a graph that shows what is going on. First, let's recall the formulas for the mean and standard deviation of the binomial distribution with parameters n and p:

$$\mu = np, \quad \sigma = \sqrt{np(1-p)}.$$

The relationship can be seen by plotting the binomial with parameters n and p and the normal with parameters μ and σ. Let's use $n = 50$ and $p = 0.20$:

```
> n = 50
> p = 0.2
> mu = n * p
> sigma = sqrt(n * p * (1 - p))
> k = 0:20
> plot(k, dbinom(k, n, p), type = "h")
> curve(dnorm(x, mu, sigma), add = T)
```



The tips of the lines have a height giving the binomial probabilities. They match up quite well with the normal distribution. This gives rise to the relationship between the probabilities:

> The probability the binomial is less than k is approximately the probability that normal is less than k.

That is in formulas

> `pbinom(k, n, p)` is approximately `pnorm(k,mu,sigma)`.

To observe, we need to put in some numbers. Let's try $k = 10$, $n = 50$ and $p = 0.20$:

```
> k = 10
> n = 50
> p = 0.2
> mu = n * p
> sigma = sqrt(n * p * (1 - p))
> pbinom(k, n, p)

[1] 0.5836

> pnorm(k, mu, sigma)

[1] 0.5
```

Are they close?

Question 6: Repeat the above using $k = 5, 8, 12,$ and 15. Are they always close?

Question 7: Repeat again, this time using $k = 200$ and $k = 210$ with $n = 500$ and $p = 0.4$.

Question 8: The normal approximation is really only valid when $np \geq 5$ and $n(1 - p) \geq 5$. As otherwise, the binomial plot has its peak too close to 0 or n to be "bell-shaped." Verify this by comparing the value of the binomial and normal when $n = 100$, $p = 0.01$ and $k = 1$ and 3.

3 The continuity correction

The book uses a "continuity correction" when discussing the normal approximation. The plot in Figure 1 shows where it comes from. Look at the $P(X \leq 5)$. This is found by adding the area in the boxes for $X = 0, 1, 2, 3, 4$ and 5. The area of these boxes is approximated by the area under the normal curve to the left of 5 1/2 – where the box for $X = 5$ ends. So in formulas, a more accurate approximation is:

$$P(X \leq b) = P(Y \leq b + \frac{1}{2}), \quad \text{and } P(X \geq a) = P(Y \geq a - \frac{1}{2})$$

To see with $k = 10$, $n = 50$ and $p = 0.20$ we can see

```
> k = 10
> n = 50
> p = 0.2
> mu = n * p
> sigma = sqrt(n * p * (1 - p))
> pbinom(k, n, p)
```

```
[1] 0.5836

> pnorm(k + 1/2, mu, sigma)

[1] 0.5702
```

To go probabilities like $P(X \geq 10)$, we have

```
> k = 10
> 1 - pbinom(9, n, p)

[1] 0.5563

> pnorm(k - 1/2, mu, sigma)

[1] 0.4298
```

(Why is it 9 for `pbinom()`?)

Question 9: Repeat the last two exercises, only use this "continuity correction". Is the approximation more accurate?

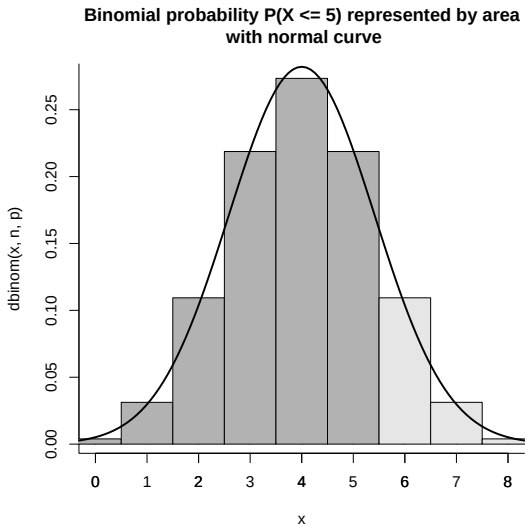


Figure 1: Probability $P(X \leq 5)$ for a binomial random variable with $n = 8$ and $p = 1/2$. This probability is given exactly by the area of the six darker-shaded rectangles, as each has a base of 1 and a height of $P(X = k)$. This area is approximated by the area to the left of 5 1/2 under the normal curve.

4 Sample proportions

Suppose a manufacturer claims that 60% of people will get better in 3 days if they take a medication. So what? Is this actually better than if the patients *did not take* the medication? How might one tell? Suppose two studies are done, one with 100 people who take the medication and one with 75 people who do not take the medication. Let X the number who get better and took the medication, Y the number who get better and did not take the medication.

We assume both X and Y are binomial random variables, with $p = 0.60$, but different values of n.

Suppose $X = 55$ and $Y = 41$. Clearly Y is less than X, but Y only had $n = 75$ and X had $n = 100$. What is a better way to compare X and Y? Proportions. That is we compare $X/100 = 0.55$ to $Y/75 = 0.547$. Basically they are the same.

Notationally, let $\hat{p} = X/n$ be a *sample proportion*. If X is binomial with n and p, then $n\hat{p}$ is binomial. Another way to think about it is in terms of $\hat{p}$ which is *like a binomial* only it has:

the mean of $\hat{p}$ is p;
the standard deviation of $\hat{p}$ is $\sqrt{p(1-p)/n}$.

There is a slight difference as we have divided by n. Usually we assume the normal approximation applies — without the continuity correction — and say that $\hat{p}$ has a normal distribution with these value of μ and σ for the parameters.

So the probability that $\hat{p}$ for X (with $n = 100$) has a value of .55 or less would be answered with

```
> p = 0.6
> n = 100
> mu = p
> sigma = sqrt(p * (1 - p)/n)
> pnorm(0.55, mu, sigma)
```

```
[1] 0.1537
```

Question 10: Repeat the above, only use Y to give $\hat{p}$. (That is $n = 75$). Are the answers different? Why?

Question 11: It is claimed that 5% of patients may experience nausea with a certain medication. If 1000 people take the medication, find the probability that more than 60 experience nausea.

Question 12: It is estimated that a vaccine can be fatal in 1 out 1,000,000 (0.0001%) people. As part of a prevention program, it is suggested that 6,000,000 newborns be given this vaccine. Find the probability that more than 0.0002% will have a fatal reaction.

Question 13: An allergy medication is assumed to be effective for only 40% of the population. Suppose, a study of 200 patients is conducted using this medication. What is the probability that only 35% or fewer found it effective?