Chapter 12: The Binomial distribution

Corinne Riddell

October 2, 2020

The binomial setting and binomial distributions

- An elementary school administers eye exams to 800 students. How many students have perfect vision?
- A new treatment for pancreatic cancer is tried on 250 patients. How many survive for five years?

What are the common threads to each of these questions?

- Something happens n number of times, or across n individuals
- The outcome for each "trial" is binary
- Can think of this like flipping a coin n times, and for each trial recording whether the coin came up heads

The binomial setting and binomial distributions

- An elementary school administers eye exams to 800 students. How many students have perfect vision?
  - n = 800
  - outcome: Does student i have perfect vision?
- A new treatment for pancreatic cancer is tried on 250 patients. How many survive for five years?
  - n = 250
  - outcome: Does patient i survive for 5 years?

Format of data for the binomial setting

1. There are a fixed number of n observations.
2. The n observations are independent. This means that the result of one observation does not affect the outcome for any other observation.
3. Each observations is either a "success" (1) or a "failure" (0). These are general terms – sometimes "success" means death rather than survival.
4. The probability of success, call it p is the same for each observation. This follows directly from item #2.

Example

An elementary school adminsters eye exams to 800 students. How many students have perfect vision?

Let X represent the number of students with perfect vision. Then

$$X \sim Binom(n = 800, p = ?)$$

where n is the number of students in the school, and p is the probability of having perfect vision. Here we don't know p because it wasn't provided in the question.

- Is X a discrete or a continuous random variable?
- What is the range of values that X can take, hypothetically?

Example 2

Forty-five percent of the population is blood type O. Consider the next five blood donations from unrelated individuals. The number who have type O is the count X of successes in 5 independent observations. Here,

$$X \sim Binom(n = 5, p = 0.45)$$

- Is X a discrete or a continuous random variable?
- What is the range of values that X can take, hypothetically?

Example 3

- A researcher has access to 40 men and 40 women and selects 10 of them at random to participate in an experiment. The number of women selected can be represented by X.
- Is X binomially distributed?
- Read the question carefully. What is the probability of selecting a woman when there are 40 individuals. If a woman is chosen, what is the probability of selecting a women the second time?

Example 4

A pharmaceutical company inspects a simple random sample of 10 empty plastic containers from a shipments of 10,000. They are examined for traces of benzene. Suppose that 10% of the containers in the shipment contain benzene. Let X represent the number of containers contaminated with benzene. Is X binomially distributed?

- Issue: Each time you sample one bottle, it affects the chance that the next bottle will be contaminated (by reducing the number of bottles in the population). However given that the population is size 10,000 and the sample size is 20, the effect of one sample's success status on the next bottle's success status is negligible.
- Here the distribution of X is *approximately* Binomial:

  $$X \dot{\sim} Binom(10, 0.10)$$

  where $\dot{\sim}$ is read as "approximately distributed as".

Binomial approximation when N is much larger than n

- Choose a simple random sample of size n from a population with proportion p of success.
- If the population size (N) is much larger than the sample size n, then the count X of successes in the sample has an approximately binomial distribution with parameters n and p.

Definition: sampling distribution

A **sampling distribution** is the distribution (shown using a histogram) of a **sample statistic** after taking many samples.

We often are most interested in the sampling distribution for a **sample proportion** (denoted by $\hat{p}$) or of a **sample mean** (denoted by $\bar{x}$).

Let's investigate the sampling distribution of X from the previous example.

Sampling distribution of a count in R

First, set up a large population of size 10,000 where 10% of the containers are contaminated by benzene. We call `benzene` a "success" since it is coded as 1. We can see that 10% of the containers are contaminated and 1000 bottles are "successes"

```r
# Students, don't worry about these three lines of code to set up the data frame.
container.id <- 1:10000
benzene <- c(rep(0, 9000), rep(1, 1000))
pop_data <- data.frame(container.id, benzene)

# Calculate the population number of bottles contaminated by benzene and the
# population mean proportion
pop_stats <- pop_data %>% summarize(pop_num_successes = sum(benzene),
                                    pop_mean = mean(benzene))

pop_stats
```

```
##   pop_num_successes pop_mean
## 1              1000      0.1
```

Sampling distribution of a count in R

Take a sample of size 10 from the population. Note that $n = 10$ is much smaller than $N = 10,000$.

- How many contaminated bottles are we expecting in the sample?
- Given that we sample 10, what is the full range of possible values we could see for X, the number of successes and p the proportion of successes?
- Which values from this full range are most likely?

```r
# first sample
set.seed(445)
sample_data <- pop_data %>% sample_n(10)
sample_data %>% summarize(sample_num_successes = sum(benzene),
                          sample_mean = mean(benzene))
```

```
##   sample_num_successes sample_mean
## 1                    2         0.2
```

Sampling distribution of a count in R

We only took one sample, and got 2 successes for a sample mean of 20%. Is that usual or unusual?

To see what is most likely, we need to imagine repeatedly taking samples of size 10 from the population and calculating the number of successes and the proportion of successes for each repeated sample.

The distribution of the number of successes across many samples is called the **sampling distribution** for X.

The distribution of the proportion of successes across many samples is called the **sampling distribution** for p.

For the next few slides, we focus on the sampling distribution for X.

Sampling distribution of a count in R

This code takes 1000 samples each of size 10. It then calculates the mean sample proportion and number of successes for each sample and stores all the results in a data frame. You don't need to know how the code works.

```r
# students, you don't need to understand the new parts of this code (i.e., how
# to write a function, or use `replicate`, `lapply`, or `bind_rows`)

calc_sample_stats <- function(df) {
  df %>%
```

```
    summarize(sample_proportion = mean(benzene),
              sample_num_successes = sum(benzene))
}

many.sample.stats <- replicate(1000, sample_n(pop_data, 10), simplify = F) %>%
  lapply(., calc_sample_stats) %>%
  bind_rows() %>%
  mutate(sample.id = 1:n())
```

Sampling distribution of a count in R

Here are the first five rows of the data frame we made on the previous slide. Each row represents an independent sample from the population.

```
head(many.sample.stats)
```

```
##   sample_proportion sample_num_successes sample.id
## 1               0.1                   1         1
## 2               0.0                   0         2
## 3               0.1                   1         3
## 4               0.2                   2         4
## 5               0.2                   2         5
## 6               0.2                   2         6
```

Sampling distribution of a count in R

We want to know: Of the 1000 samples, what percent observed 0 contaminated bottles? What percent observed 1 contaminated bottle? And so on. We can used `dplyr` functions to calculate this and plot the results in a histogram.

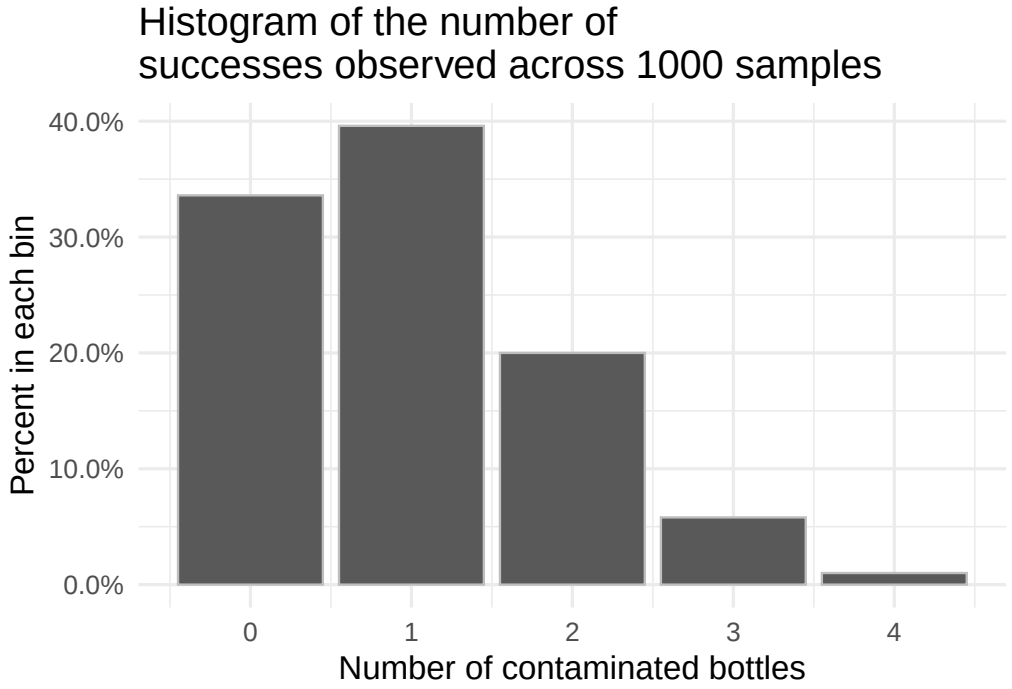
```
aggregated.stats <- many.sample.stats %>%
  group_by(sample_num_successes) %>%
  summarize(percent = n()/1000) #n() counts the number of rows within each group
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
aggregated.stats
```

```
## # A tibble: 5 x 2
##   sample_num_successes percent
##                  <dbl>   <dbl>
## 1                    0   0.336
## 2                    1   0.396
## 3                    2   0.2
## 4                    3   0.058
## 5                    4   0.01
```

Sampling distribution of a count in R



Sampling distribution of a count in R

- This histogram *approximates* the shape of the binomial distribution with n = 10 and p = 0.1.
- Observing one success is the most likely outcome. Why is that?
- Why don't we see the full range of hypothetical outcomes?

Worked probabilities, x = 0

We sampled n=10 bottles where the probability of success on any one pick is 10%.

- What is the chance of observing zero contaminated bottles?
- This means the first bottle is not contaminated and the second bottle is not contaminated ... and the tenth bottle is not contaminated

$P(X_1 = 0 \text{ and } X_2 = 0 \text{ and...and } X_{10} = 0)$

$= P(X_1 = 0) \times P(X_2 = 0) \times ... \times P(X_{10} = 0)$, applying the multiplication rule for independent events

$= (0.90)^{10}$

$= 0.3486784 = 34.9\%$

Worked probabilities, x = 1

- What is the chance of observing exactly one contaminated bottle?
- Suppose that the first bottle was contaminated, then all the rest had to be not contaminated. What is the probability of observing this specific sequence of events?

$P(X_1 = 1 \text{ and } X2 = 0 \text{ and } X3 = 0 \text{ and...and } X_{10} = 0)$

$= P(X_1 = 1) \times P(X_2 = 0) \times P(X_3 = 0)... \times P(X_{10} = 0)$

$= (0.1)^1(0.90)^9$

$= 0.03874205 = 3.87\%$

- Thus, there is 3.87% chance that the first bottle chosen is contaminated and bottles 2-9 are not contaminated.
- This is only one specific way of observing exactly one contaminated bottle.
- What is another way? How many ways are there to observed exactly one contaminated bottle when there are ten bottles?

Worked probabilities, x = 1 (continued)

There are ten ways to observe exactly one contaminated bottle:

- 1, 0, 0, 0, 0, 0, 0, 0, 0, 0
- 0, 1, 0, 0, 0, 0, 0, 0, 0, 0
- 0, 0, 1, 0, 0, 0, 0, 0, 0, 0
- ...
- 0, 0, 0, 0, 0, 0, 0, 0, 0, 1

Each of these ten ways has the same probability of occurring.

Thus,

P(observed exactly 1 contaminated bottle)

$= P$(1st bottle is contaminated, and rest are not OR 2nd bottle is contaminated, and rest are not OR... OR 10th bottle is c

$= (0.1)^1(0.9)^9 + (0.1)^1(0.9)^9 + ... + (0.1)^1(0.9)^9$, using the addition rule for disjoint events

$= 10 \times (0.1)^1(0.9)^9$

$= 0.3874205 = 38.7\%$

Use R's dbinom() to check your probability calculation

Finally, we can check our calculations using the `dbinom()` function in R. This function calculates the probability of observing `x` successes when $X \sim Binom(n, p)$

```
dbinom(x = 1, size = 10, prob = 0.1)
```

```
## [1] 0.3874205
```

This is exactly the answer we obtained.

Worked probabilities, x = 2

What is chance of observing exactly two contaminated bottles?

Following the same line of thinking, suppose that the first two bottles were contaminated. The chance of this happening is:

$(0.1)^2(0.9)^8 = 0.004303672$

But how many ways are there to observe exactly two contaminated bottles? You could write out all the possibilties like last time, but there are a lot more!

We can get our calculators or R to perform this calculation for us. On our calculator, we need the button $\binom{n}{k}$, pronounced "n choose k". This button asks how many ways there are to have k successes when there are n individuals. In R we need the function `choose(n, k)`

```
choose(10, 2)
```

```
## [1] 45
```

There are 45 ways to observe exactly two contaminated bottles when you have ten bottles observed.

Make sure you can also perform this calculation on your calculator!

Worked probabilities, x = 2

To get the probability of observing exactly 2 contaminated bottles, we multiply 45 by the probability of observing the first two bottles as being contaminated:

$45 \times (0.1)^2(0.9)^8 = 0.1937102 = 19.4\%$

Check using R:

```
#fill in during class
```

All of the combinations with 10 bottles

Each of these is written as $\binom{10}{k}$, where k is 0, 1, 2, ..., 10. This is known as the **binomial coefficient**.

Let's compute `choose(n, k)`, for n=10, and k=0, 1, 2, ..., 10:

```
choose(10, 0) # how many ways to choose 0 successes from 10 bottles?
```

```
## [1] 1
```

```
choose(10, 1) # how many ways to choose 1 success from 10 bottles?
```

```
## [1] 10
```

```
choose(10, 2) # how many ways to choose 2 successes from 10 bottles?
```

```
## [1] 45
```

```
choose(10, 3)
```

```
## [1] 120
```

```
choose(10, 4)
```

```
## [1] 210
```

```
choose(10, 5)
```

```
## [1] 252
```

```
choose(10, 6)
```

```
## [1] 210
```

```
choose(10, 7)
```

```
## [1] 120
```

```
choose(10, 8)
```

```
## [1] 45
```

```
choose(10, 9)
```

```
## [1] 10
```

```
choose(10, 10)
```

```
## [1] 1
```

Notice the symmetric structure of `choose(n, k)`. Why is it symmetric?

Bringing it all together: Binomial probability distribution function

If X has a binomial distribution with n observations and probability p of success on each observation, then the possible values of X are 0, 1, 2, ..., n.

The **probability distribution function** for X is given by the formula:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- You can use this formula to compute (by hand) the probability that $X = x$ for every x between 0 and n.
- Read $\binom{n}{x}$ as "n choose x". It counts the number of ways in which x successes can be arranged among n observations.
- $p^x(1-p)^{n-x}$ is the chance of observing x successes and $n-x$ failures for one specific ordering of these successes and failure. Thus, we multiply by $\binom{n}{x}$ to count each way we can have x successes and $n-x$ failures.

Binomial probability in R using `dbinom()` and `pbinom()`

- Recall for Normal distributions we used `pnorm()` to calculate the probability **below** a given number.
- For discrete distributions we can calculate the probability of observing a specific value. For example, we can ask: What is the probability that exactly 3 of the ten bottles were contaminated when the risk of contamination was 10%?
- We can also ask, what is the probability that 3 or less of the ten bottles were contaminated when the risk of contamination was 10%?
- `dbinom()` is used to compute *exactly* 3
- `pbinom()` is used to compute 3 *or less*

Notice how these function return different probabilities:

```
dbinom(x = 3, size = 10, prob = 0.1)
```

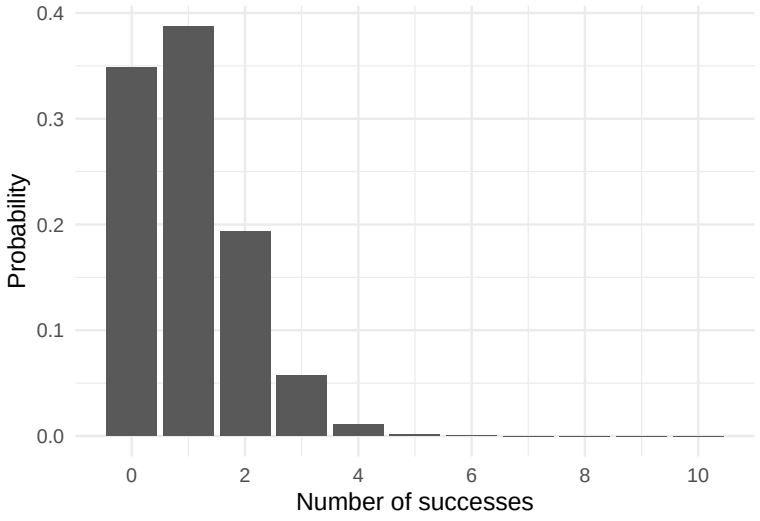
```
## [1] 0.05739563
```

```
pbinom(q = 3, size = 10, prob = 0.1)
```

```
## [1] 0.9872048
```

Histogram of binomial probabilities

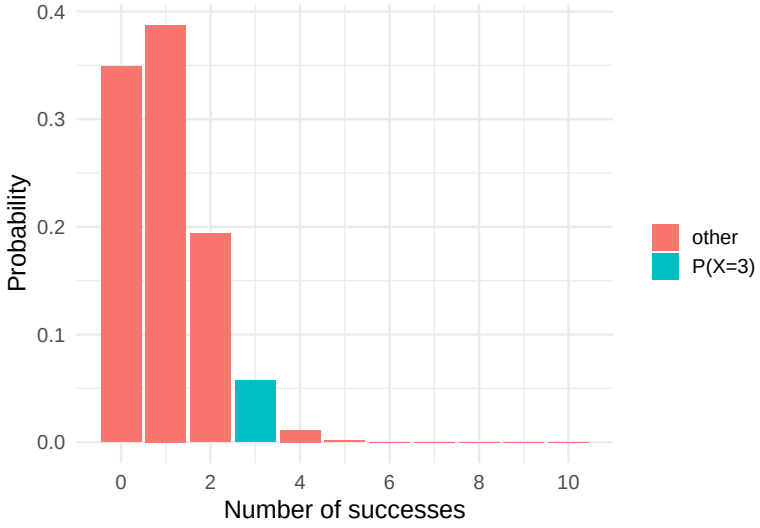
This histogram shows the probability of observing each value of X. That is, it shows the $P(X = x)$, for x in 0,1,2, ... 10, when $X \sim Binom(n = 10, p = 0.1)$



Exact discrete probability, graphed

```
dbinom(x = 3, size = 10, prob = 0.1)
```

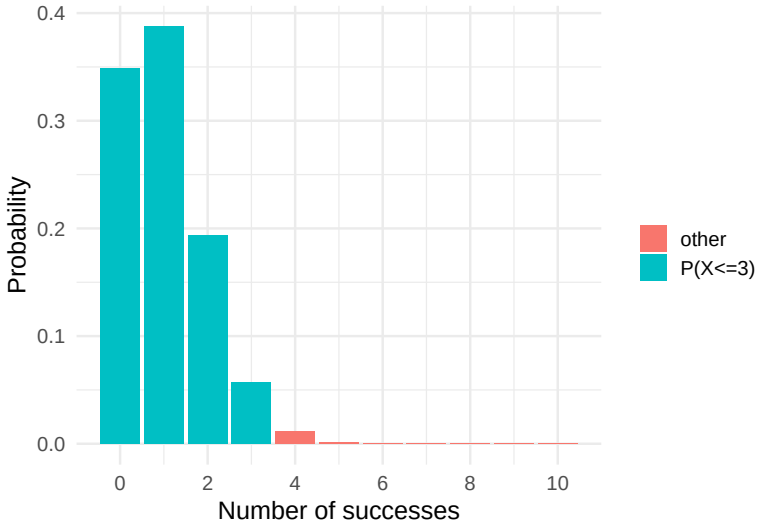
```
## [1] 0.05739563
```



Cumulative discrete probability, graphed

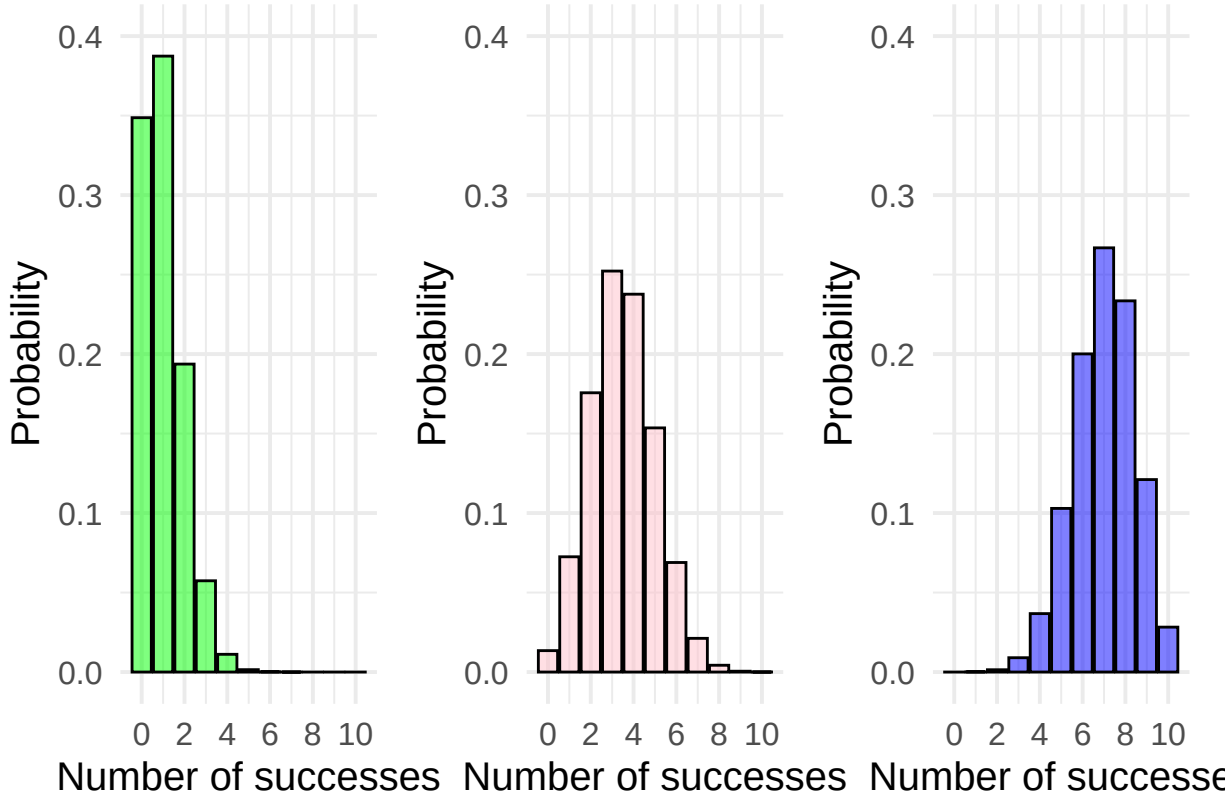
```
pbinom(q = 3, size = 10, prob = 0.1)
```

```
## [1] 0.9872048
```



Histogram of binomial probabilities with different values for p

Here we have n = 10, and p = 0.10 (green), 0.35 (pink), and 0.7 (blue)



Binomial mean and standard deviation

If a count X has the binomial distribution with n number of observations and p as the probability of success, then the population mean and population standard deviation are:

$$\mu = np$$

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

- For this class, you don't need to know why this is true, but if you are interested you could watch this video.

Example of mean and SD calculations

Recall our example of the number of bottles contaminated in benzene, where $X \sim Binom(n = 10, p = 0.1)$.

$\mu = np = 10 \times 0.1 = 1$

$\sigma = \sqrt{np(1-p)} = \sqrt{10 \times 0.1(1-0.1)} = 0.9487$

Thus, we **expect** to find one container contaminated with benzene per sample, on average. The standard deviation can be thought of, very roughly, as the expected deviation from this mean if you were to take many random samples.

An approximation to the binomial distribution when n is large

Imagine a setting where $X \sim Binom(n = 2000, p = 0.62)$. Then:

$$P(X = x) = \binom{2000}{x} 0.62^x (1-0.62)^{2000-x}$$

And:

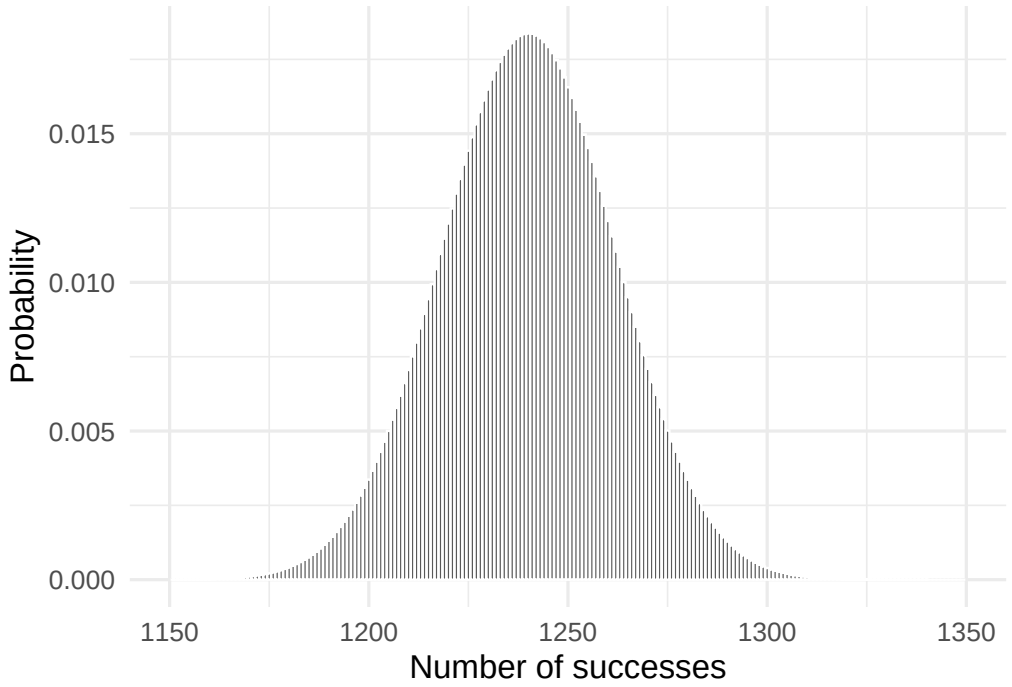
$$P(X <= x) = \sum_{i=0}^{x} \binom{2000}{i} 0.62^i (1-0.62)^{2000-i}$$

If you needed to calculate this by hand for k = 100 it would take a very long time.

An approximation to the binomial distribution when n is large

Consider the probability distribution function for $P(X = k) = \binom{2000}{k} 0.62^k (1-0.62)^{2000-k}$

What shape does this remind you of?



An approximation to the binomial distribution when n is large

The previous graph is unimodal and symmetric. Let's calculate μ and σ:

$\mu = np = 2000 \times 0.62 = 1240$

$\sigma = \sqrt{np(1-p)} = \sqrt{2000 \times 0.62 \times (1-0.62)} = 21.70714$

How much data is within 1 SD of the mean?

1240 +/- 1 SD gives the range {1218.293, 1261.707}

Thus, we can use R to add up all the probabilities between X = 1218 and X = 1262 to give an approximate guess to the area one standard deviation from the mean:

This code cycles through the probabilities to add them up

```
#students, no need to know how to write this code.
cumulative.prob <- 0

for(i in 1218:1262){
  cumulative.prob <- cumulative.prob + point.probs.2k[i]
}

cumulative.prob
```

```
## [1] 0.6994555
```

How much data is within 2 SD of the mean?

1240 +/- 2 SD gives the range {1196.586, 1283.414}

Thus, we can use R to add up all the probabilities between X = 1197 and X = 1283 to give an approximate guess to the area 1 SD from the mean:

This code cycles through the probabilities to add them up

```r
#students, no need to know how to write this code.
cumulative.prob.2 <- 0

for(i in 1197:1283){
  cumulative.prob.2 <- cumulative.prob.2 + point.probs.2k[i]
}

cumulative.prob.2
```

```
## [1] 0.9547453
```

- You could also perform the check for 3 SD

The Normal approximation to Binomial distributions

From the previous calculations, you might see that the shape looks Normal and that the distribution nearly meets the 68%-95%-99.7% rule. Thus, it is approximately Normal.

This means that you can use the Normal distribution to perform calculations when data is binomially distributed with large sample size n.

Example calculation of the Normal approximation to the Binomial

Suppose we want to calculate $P(X >= 1250)$ using the Normal approximation.

```r
# write the Normal code
1- pnorm(q = 1250, mean = 1240 , sd = 21.70714)
```

```
## [1] 0.3225149
```

Check how well the approximation worked:

```r
# write the binomial code and see how well the approximation is
1 - pbinom(q = 1249, size = 2000, prob = 0.62)
```

```
## [1] 0.3313682
```

Important note: To calculate $P(x >= 1250)$ we take $1 - P(X <= 1249)$ using the binomial code. This is **different** from the Normal code where we can use $1 - P(X < 1250)$ because for the Normal code $P(X = 1250) = 0$ by definition.

Normal approximation for binomial distributions

Suppose that a count X has the binomial distribution with n observations and success probability p. When n is large, the distribution of X is approximately Normal. That is,

$$X \dot{\sim} N(\mu = np, \sigma = \sqrt{np(1-p)})$$

As a rule of thumb, we will use the Normal approximation when n is so large that $np >= 10$ and $n(1-p) >= 10$.

This approximation is most accurate for p close to 0.5, and least accurate for p close to 0 or 1.

Normal approximation with continuity correction

This approximation can be improved a tiny bit!

As you know, counts can only take integer values (whole numbers), but continuous data can take any real value. The proper continuous equivalent to a count is the interval around the count with size 1. For example,

the continuous equivalent to a 1250 count is the interval between 1249.5 and 1250.5. Thus, we should compute P(X >= 1249.5) rather than P(X > 1250) for an even more accurate answer.

Here is the more precise estimate for the example:

```
1- pnorm(q = 1249.5, mean = 1240 , sd = 21.70714)
```

```
## [1] 0.3308222
```

Normal approximation with continuity correction

This correction makes a bigger difference when the sample size n is small.

For example, we can compare how much better the approximated probabilities are where a) $n = 100$ and b) $n = 1000$

Scenario a) smaller n

```
pbinom(q = 8, size = 100, p = 0.1)
```

```
## [1] 0.3208739
```

```
pnorm(q = 8, mean = 10, sd = sqrt(100*0.1*(0.9))) # no continuity correction
```

```
## [1] 0.2524925
```

```
pnorm(q = 8.5, mean = 10, sd = sqrt(100*0.1*(0.9))) # continuity correction applied
```

```
## [1] 0.3085375
```

Applying the continuity correct increased the approximated probability from 0.252 to 0.309, which is much closer to the true value of 0.321.

Scenario b) larger n

```
pbinom(q = 100, size = 1000, p = 0.1)
```

```
## [1] 0.5265991
```

```
pnorm(q = 100, mean = 100, sd = sqrt(1000*0.1*(0.9))) # no continuity correction
```

```
## [1] 0.5
```

```
pnorm(q = 100.5, mean = 100, sd = sqrt(1000*0.1*(0.9))) # continuity correction applied
```

```
## [1] 0.5210164
```

Applying the continuity correct increased the approximated probability from 0.5 to 0.521, which is much closer to the true value of 0.527.

Review

- Ch. 11 was all about the Normal distribution. We learned about the properties of the Normal curve, and how to use R to calculate cumulative probabilities and generate random Normal values. We learned that the Normal distribution can be described by its mean and standard deviation.
- So far, Ch. 12 is all about the Binomial distribution. We learned that Binomially-distributed variables must meet certain assumptions and that their distributions can be described by n and p. We also learned how to calculate the probability of observing $X = x$ exactly (`dbinom()`) or the cumulative probability less than some x (`pbinom()`) and when we can apply the Normal approximation