L16: The Binomial distribution continued

Objectives for today

- wrap up the worked examples
- an aside about pascal's triangle
- discuss exact vs. cumulative probabilities for the binomial
- introduce some R code for binomial distributions
- recap Normal and Binomial

continuation of worked examples

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:

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## [1,]    1   10   45  120  210  252  210  120   45    10     1
```

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

An aside: Pascal's triangle

$$\binom{0}{0}$$

$$\binom{1}{0} \binom{1}{1}$$

$$\binom{2}{0} \binom{2}{1} \binom{2}{2}$$

$$\binom{3}{0} \binom{3}{1} \binom{3}{2} \binom{3}{3}$$

An aside: Pascal's triangle

1

1 1

1 2 1

1 3 3 1

TED ed Video about Pascal's triangle
https://www.youtube.com/watch?v=XMriWTvPXHI

Binomial probability

Recal from lecture 15:

If X has the binomial distribution with n observations and probability p of success on each observation, the possible values of X are 0, 1, 2, . . . , n. If k is any one of these values,

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

Binomial probability in R

Binomial probability in R using dbinom()

- 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%?
- `dbinom()` is used to compute *exactly* 3

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

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

Binomial probability in R using pbinom()

- Recall for Normal distributions we used pnorm() to calculate the probability below a given number.
- For our Binomial, 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%?
- pbinom() is used to compute 3 *or less*

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

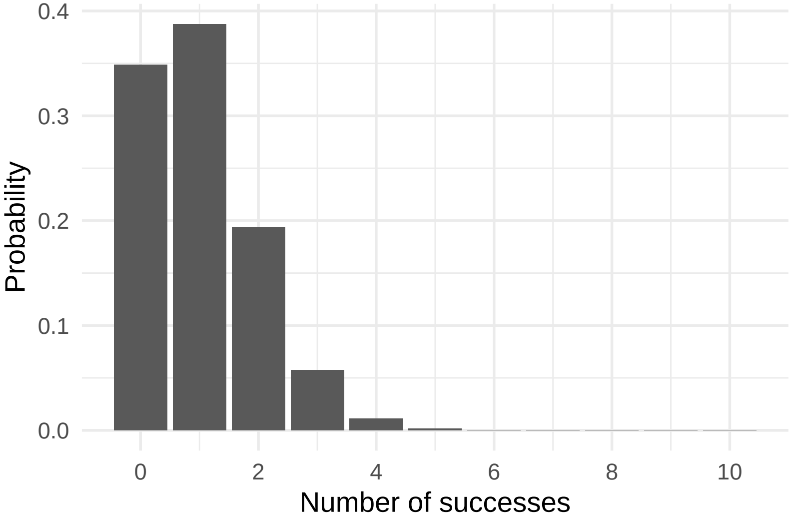
```
## [1] 0.05739563
```

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

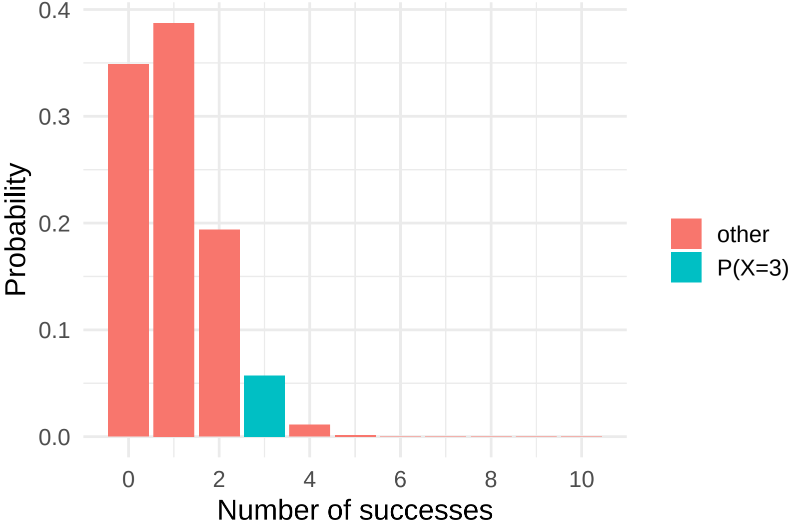
```
## [1] 0.9872048
```

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 $X \sim Binom(n = 10, p = 0.1)$

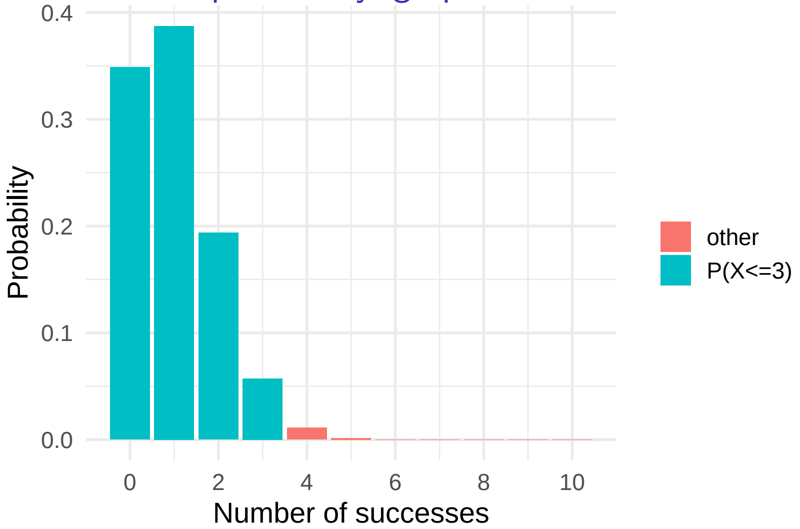


Exact discrete probability, graphed




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

Cumulative discrete probability, graphed




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

Mean and Variance of a Binomial

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:

$$\mu = np$$

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

Example of mean and SD calculations

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

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

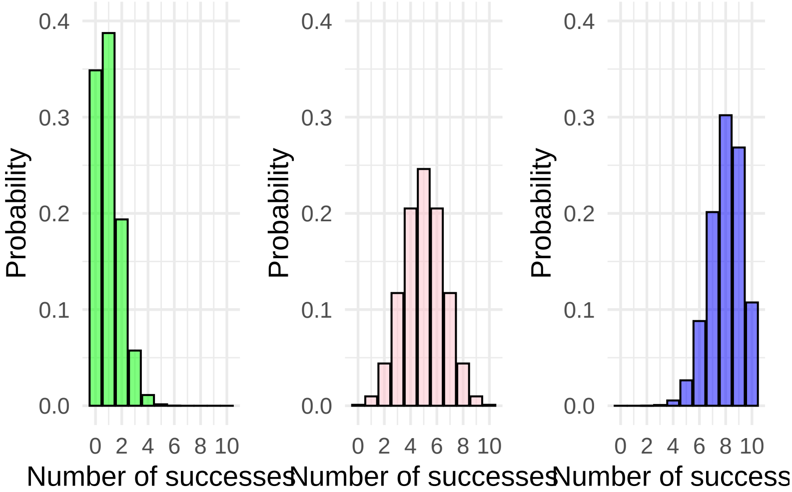
$\sigma = \sqrt{np(1-p)} = \sqrt{10 \times 0.1(1-0.1)} = 0.9487$

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.

Normal approximation of a binomial

Histogram of binomial probabilities with different values for p

Here we have $n = 10$, and $p = 0.10$ (green), 0.5 (pink), and 0.8 (blue)



Histogram of binomial probabilities with different values for p

How does the shape change when the probability is closer to .5?

What do you think happens when n gets larger?

An approximation to the binomial distribution when n is large

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

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

And:

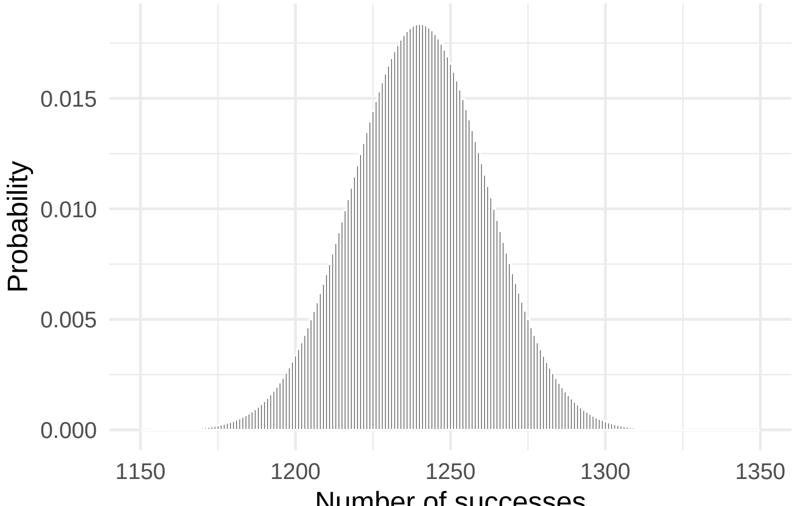
$$P(X \leq k) = \sum_{i=0}^{k} \binom{2000}{i} 0.62^k (1 - 0.62)^{2000-k}$$

If you were asked to calculate this by hand for, say, k = 100, it would take a very long time.

An approximation to the binomial distribution when n is large

Consider the probability distribution 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 1 SD 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

```
#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 n.

Example calculation of the Normal approximation to the Binomial

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

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

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

Check how well the approximation worked:

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

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

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 general rule, we will use the Normal approximation when n is so large that $np \geq 10$ and $n(1-p) \geq 10$.

It is most accurate for p close to 0.5, and least accurate for p closer 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, 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.

This correction makes a bigger difference when n is small.

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

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

Recap

Properties of the Normal distribution

- the mean μ can be any value, positive or negative
- the standard deviation σ must be a positive number
- the mean is equal to the median (both $= \mu$)
- the standard deviation captures the spread of the distribution
- the area under the Normal distribution is equal to 1 (i.e., it is a density function)
- a Normal distribution is completely determined by its μ and σ

The 68-95-99.7 rule for all Normal distributions

- Approximately 68% of the data fall within one standard deviation of the mean
- Approximately 95% of the data fall within two standard deviations of the mean
- Approximately 99.7% of the data fall within three standard deviations of the mean

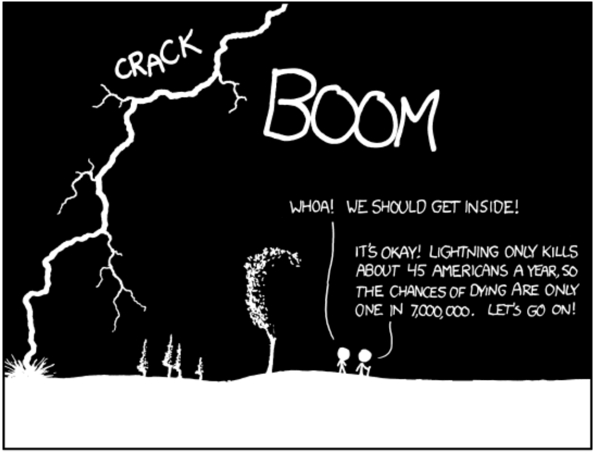
Properties of the Binomial distribution

- The random variable must assume one of two possible and mutually exclusive outcomes
- Each trial of the BRV results in either a success or failure
- Each trial must be independent of every other trial
- Derived from the experiment: counting the number of occurrences of an event in n independent trials
- Random Variable: $X =$ number of times the event happens in the fixed number of trials (n)
- Parameters
  - $n =$ number of trials
  - $p =$ probability of success (event happening)

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()`)
- Next lecture we will introduce the Poisson distribution.

Comic Relief



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.