Week Two: Inference for Proportions

This Week (and next)

- Binomial & Multinomial Distributions
- Maximum Likelihood Estimation
- Intervals and Testing

Next Week: Bayesian Inference for Proportions

Tuesday:

- Watch Week 3 videos and submit HW 2 (video notes)
- Week 3 activity

Thursday:

- Lab 2

Binomial Distribution

Recall that 7 out of 16 of us selected Sweet Peaks as our favorite ice cream shop in Bozeman (or Montana).

This data can be modeled with a Binomial distribution, where

$$P(Y=y)=\frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}, \quad \text{where } y=0,1,2,...$$

Q: Why do we care? *Our goal is to estimate the proportion of MSU students that would select Sweet Peaks as their favorite ice cream shop. Doing this - especially in a way that accounts for uncertainty - requires a statistical probability distribution (likelihood) and model parameters (π).*

Given a binomial distribution with specified n and π, we can estimate to probability of observing a number of successes. This is my 10th year at MSU and their have been 8 Cat-Griz football games.

Exercise: Assume that there is no difference in team ability ($\pi = \frac{1}{2}$).

- 1. Estimate the probability that MSU wins the 2025 game.

$\frac{1}{2}$

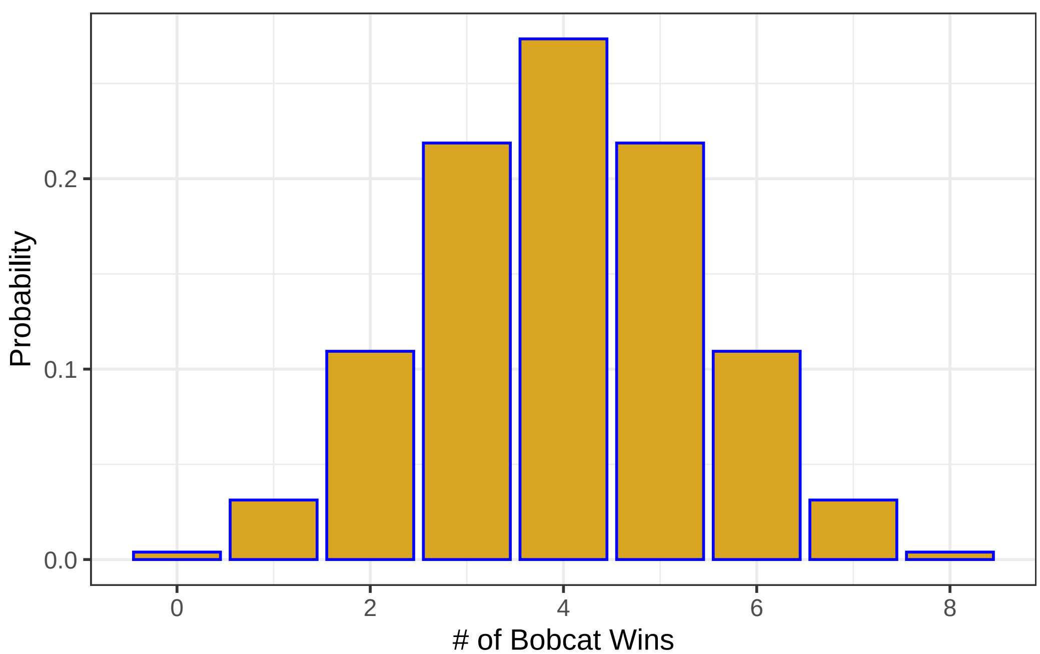
- 2. Estimate the probability that MSU won 6 of 8 games, hint: `dbinom`.

`dbinom(6,8,.5)` = 0.109375

- 3. Create a figure to show the probability that MSU won 0, 1, ..., 8 games.

```
library(tidyverse)

tibble(wins = 0:8,
       prob = dbinom(0:8, 8, .5)) |>
  ggplot(aes(y = prob, x = wins)) +
  geom_col(color ='blue', fill = 'goldenrod') +
  ylab('Probability') +
  xlab("# of Bobcat Wins") +
  theme_bw()
```

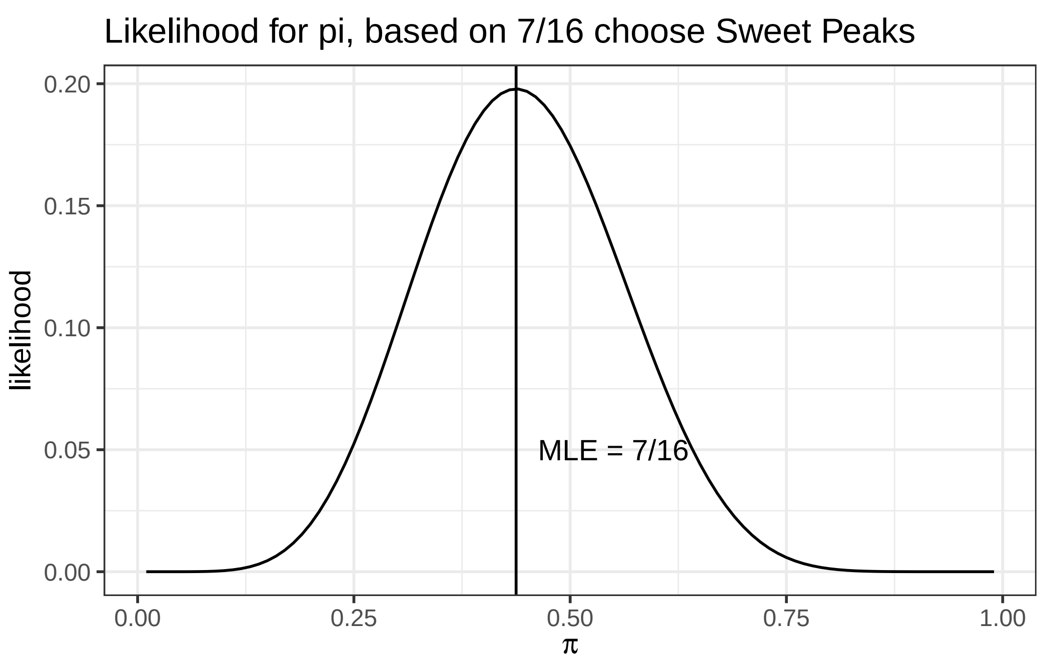


Maximum Likelihood Estimation

Generally our goal is to estimate π from a set of binary responses, as opposed to estimating the number of successes given π.

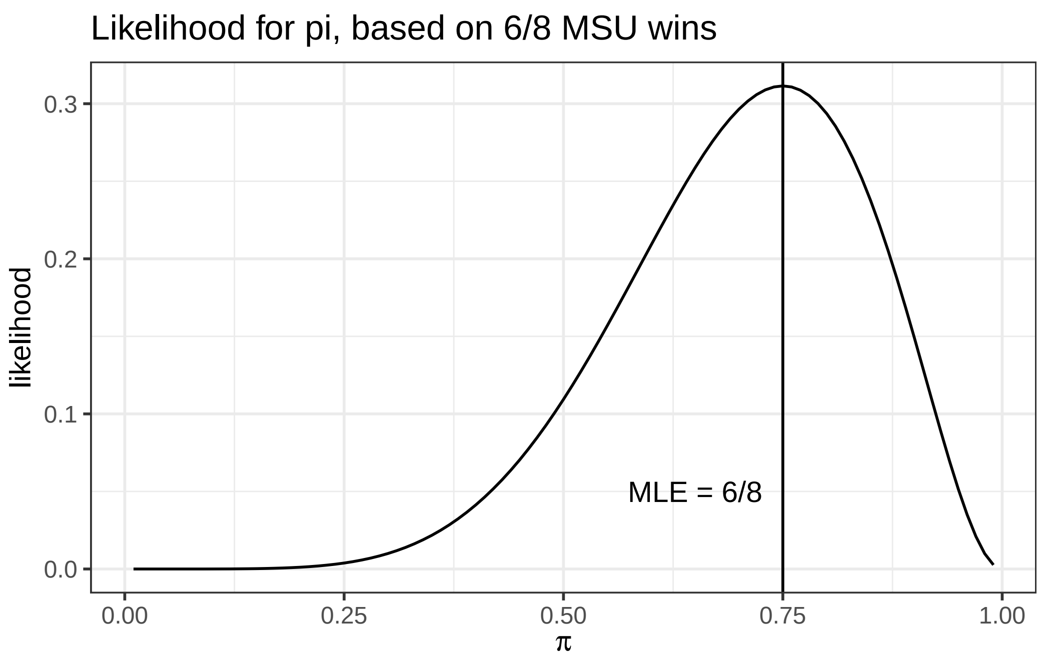
On the first day of class, we intuitively estimated π (or p) to be $\frac{7}{16}$. It turns out that this is also the maximum likelihood estimator for π.

```
pi_vals <- seq(.01, .99, by = .01)
like_vals <- dbinom(7, 16, pi_vals)
tibble(pi = pi_vals,
       likelihood = like_vals) |>
  ggplot(aes(y = likelihood, x = pi)) +
  geom_line() +
  theme_bw() +
  xlab(expression(pi)) +
  ggtitle("Likelihood for pi, based on 7/16 choose Sweet Peaks") +
  geom_vline(xintercept = 7/16) +
  annotate('text', x = .55, y = .05, label = 'MLE = 7/16')
```



Construct a likelihood profile for the Cat-Griz setting.

```
pi_vals <- seq(.01, .99, by = .01)
like_vals <- dbinom(6, 8, pi_vals)
tibble(pi = pi_vals,
       likelihood = like_vals) |>
  ggplot(aes(y = likelihood, x = pi)) +
  geom_line() +
  theme_bw() +
  xlab(expression(pi)) +
  ggtitle("Likelihood for pi, based on 6/8 MSU wins") +
  geom_vline(xintercept = 6/8) +
  annotate('text', x = .65, y = .05, label = 'MLE = 6/8')
```



Testing

Testing and estimation are two related, but different things.

Figure 1: Testing: Yes or No

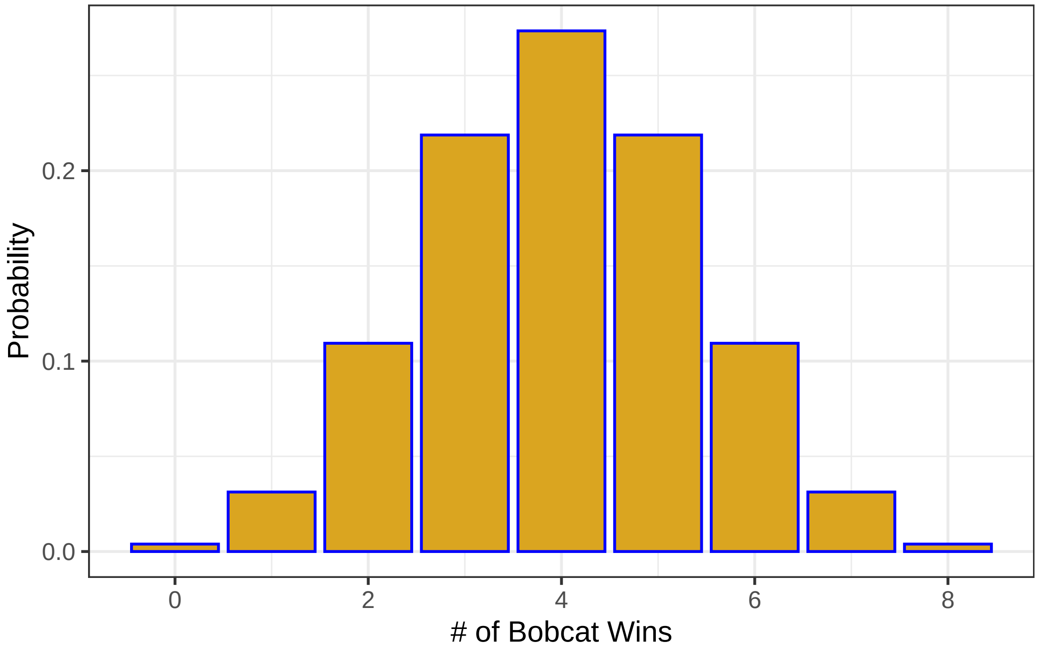
Figure 2: Estimation: What values are reasonable

- Traditionally much of applied statistics was focused on testing and associated p-values.
- In this class, we will talk about both but testing and estimation but emphasis will be on estimation.

Testing

Given the record of the Bobcats (6 wins in 8 games) over the last 10 years, we may question whether the ability level of the teams is actually the same.

This could be formulated as a testing problem in which we ask how unusual would it be for the bobcasts to win 6 out of 8 games if the teams had the same ability level. Note the similarity to the preivously created figure.



The `binom.test()` function can be used for this purpose, interpret the results.

```r
binom.test(6, 8, p = .5, alternative = "greater")
```

```
	Exact binomial test

data:  6 and 8
number of successes = 6, number of trials = 8, p-value = 0.1445
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.4003106 1.0000000
sample estimates:
probability of success 
                  0.75 
```

Estimation: Uncertainty intervals

A common way to construct confidence intervals uses asymptotic theory (large samples and CLT) such that

$\hat{\pi} \pm z_{\alpha/2}(SE)$.

With a binomial distribution, the standard error (SE) can be calculated as $\sqrt{\pi(1-\pi)/n}$. We don't know π so the estimator can be used.

Use this framework to construct a confidence interval for π in our ice cream example.

```
pi_hat <- 7/16
n <- 16

multiplier <- qnorm(.975) * sqrt(pi_hat * (1 - pi_hat)/ n)

lower <- pi_hat - multiplier
upper <- pi_hat + multiplier
```

This results in a 95% confidence interval from 0.19 to 0.68.

There are some known issues with this procedure that we will explore in the future, but as a thought exercise

- What do the intervals look like when Y = 0?

not good, a point mass at 0

- What happens when n is small?

also not good, intervals values can be smaller than 0 or greater than 1

Multicategory Outcomes

Thus far, we have simplified the ice cream example to a binary situation (Sweet Peaks or not).

However, the data was collected as a multicategory setting with many possible answers.

An extension to the binomial distribution, known as the multinomial distribution is appropriate in this case.

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_k = y_k) = \frac{n!}{y_1!y_2!...y_k!}\pi_1^{y_1}\pi_2^{y_2}...\pi_k^{y_k}, \quad \text{where } y_i = 0, 1, 2, ...$$