Standard Normal
Are data normally normally distributed?

Sample mean: 66.78
Sample standard deviation: 3.37

- $(66.78 - 1 \times 3.37, 66.78 + 1 \times 3.37)$
- $(66.78 - 2 \times 3.37, 66.78 + 2 \times 3.37)$
- $(66.78 - 3 \times 3.37, 66.78 + 3 \times 3.37)$

How many of the measurements lie within these intervals?

- 30 lie in $(63.4, 70.2) = 65.2 \%$
- 45 lie in $(60.0, 73.5) = 97.8\%$
- 46 lie in $(56.7, 76.9) = 100\%$
The Standard Normal Distribution

- Mean is 0.
- Standard deviation is 1.
- The data are symmetrically dispersed around the bell.
- About 68% of the data lie within 1 standard deviation of the mean. About 95% of the data lie within 2 standard deviation of the mean.
z-transformation

- Standard units are the number of standard deviations away from the mean
  - Some values will be negative, corresponding to values below the mean
  - Some values will be positive, corresponding to values above the mean
- The z-transformation transforms a normally distributed random variable with mean $\mu$ and standard deviation $\sigma$ into a standard normal random variable, meaning it has mean 0 and standard deviation 1.
- First compute how far away it is from the mean, and then compute the ratio of that deviation to the standard deviation
z-transformation

We use the formula for Z transformation: \[ Z = \frac{X - \mu}{\sigma} \]

In sum, the Z formula transforms the original normal distribution in two ways:
1. The numerator of the formula, \(X - \mu\), shifted the distribution so it is centered on 0.
2. By dividing by the standard deviation of the original distribution, it compressed the width of the distribution so it has a standard deviation equal to 1.
Women’s heights in the US

• In the US, women’s heights are normally distributed with a mean of 65 inches and a standard deviation of 3.5 inches.

• I am 63.25 inches tall. How many standard deviations below the mean am I?
• How many standard deviations above the mean is 6 feet (72 inches)?

• What percentage of women in the US are shorter than me?
• What percentage of women in the US are taller than 6 feet?
Women’s heights in the US (cont’d)

- To find the percentages, all we have to do is integrate
  - The percentage of women in the US who are shorter than me is simply the area under the normal curve $N(63.25, 3.5)$ from $-\infty$ to 63.25.
  - Likewise, the percentage of women in the US who are taller than 72 inches is simply the area under the normal curve $N(63.25, 3.5)$ from 72 to $\infty$.

- Big problem: Integrating a normal distribution is not easy!
- Strategy: to answer queries such as the ones about women’s heights, pose the equivalent query in terms of the standard normal instead.
Standard Normal Table

In the olden days (back when I was a student), we used the standard normal table to answer queries.

- \( P[-1 \leq Z \leq 1] = P[Z \leq 1] - P[Z \leq -1] = P[Z \leq 1] - (1 - P[Z \leq 1]) \)

- \( P[-1 \leq Z \leq 1] = .8413 - (1 - .8413) = 68.26\% \)
- \( P[-2 \leq Z \leq 2] = .9773 - (1 - .9773) = 95.46\% \)
- \( P[-3 \leq Z \leq 3] = .9986 - (1 - .9986) = 99.72\% \)
Back to Women’s heights in the US

Apply the z-transformation:

- $Z = (63.25 - 65) / 3.5 = -\frac{1}{2}$
- $Z = (72 - 65) / 3.5 = 2$

Therefore, I am $\frac{1}{2}$ a standard deviation below the mean.
Women who are 6 feet tall are two standard deviations above the mean.
Standard Normal Table

- The probability a woman in the US is shorter than 63.25 inches is equal to the probability a z-score is less than or equal to $-\frac{1}{2}$.
  - In the textbook, $P[Z \leq -\frac{1}{2}] = P[-Z \geq \frac{1}{2}] = 1 - P[Z \leq \frac{1}{2}] = 1 - 0.6915 = 0.3085$.
  - In R: `pnorm(-1/2) = 0.31`

- Hence, about 31% of women in the US are shorter than me.
**Standard Normal Table**

- The probability a woman in the US is taller than 72 inches is equal to the probability a z-score is greater than 2.
  - In the textbook, $P[Z \leq 2] = 0.9773$. So $P[Z > 2] = 1 - 0.9773 = 0.0227$
  - In R: `pnorm(2, lower.tail = FALSE) = 0.023`

- Hence, about 2.3% of women in the US are taller than 72 inches.
Another Example: 2016 Olympic Runners

- In the 2016 summer Olympics:
  - Usain Bolt won the gold medal in the men’s 100M race with a time of 9.81 seconds.
  - Elaine Thompson won the gold medal in the women’s 100M race with a time of 10.71 seconds.

- Statistics for the races:
  - The men’s race had an average time of 9.94 seconds, with a standard deviation of 0.08 seconds.
  - The women’s race had an average time of 10.98 seconds, with a standard deviation of 0.34 seconds.
  - Assume the results were normally distributed.

- Which racer did better in comparison to the rest of their field?
2016 Olympic Runners (cont’d)

- We cannot compare these data easily unless we first standardize them.
- Indeed, by applying the z-transformation, all the units cancel, so we are comparing “apples to apples” instead of “apples to oranges.”
- First, we can see how Usain Bolt did relative to his competition.
  - $Z = (9.81 \text{ sec} - 9.94 \text{ sec}) / 0.08 \text{ sec} = -1.63$
- Second, we can see how Elaine Thompson did relative to her competition.
  - $Z = (10.71 \text{ sec} - 10.98 \text{ sec}) / 0.34 \text{ sec} = -0.79$
- More extreme z-scores imply rarer events. So, the runner with the more extreme z-score performed better comparatively.
  - As measured by z-scores, Usain Bolt performed better relative to his field than Elaine Thompson did relative to hers.
Visualizing Runner Times

9.81 Bolt
9.94 Mean

10.71 Thompson
10.98 Mean
Interval Estimates
Point Estimate

Given a statistical model of a population, a point estimate is a single value used to estimate a model parameter.

Examples:

- A sample mean is often used to estimate the mean (i.e., the “true” mean) of a normal distribution.
- Likewise, a sample proportion is often used to estimate the probability of success of a binomial distribution.

But even the very best, data-driven estimate is usually wrong!
Confidence Interval

- Goal is to find not just a single point estimate, but an interval estimate
- An interval estimate quantifies the uncertainty in an estimator
- Delimited by an upper and lower bound
- $\propto$ is called the confidence level, and $\propto$ is usually small
  - For example: $\propto = 0.05$ implies a 95% confidence interval
  - Likewise: $\propto = 0.1$ implies a 90% confidence interval

A confidence interval is a **NOT** pair $\theta_{lo} \leq \theta_{hi}$ such that $\Pr[\theta_{lo} \leq \theta \leq \theta_{hi}] \geq 1 - \propto$

A confidence interval is a pair $\theta_{lo} \leq \theta_{hi}$ such that $\Pr[\theta_{lo} \leq \text{estimator} \leq \theta_{hi}] \geq 1 - \propto$
A Motivational Story

- Let’s say you are running for mayor.
  - You hire a polling agency to determine if you are likely to win or not.
  - After sampling 100 likely voters, the polling agency reports that 55 of those 100 support you.
- But more than 100 people live in your town! Are you going to win?
  - The polling agency also reports a 95% confidence interval around the number 55: e.g., (53, 57).
    - Polling agencies tend to report margins of error (MoE = 2), but these notions are closely related.
  - Here’s what this means:
    - If the polling agency were to take this poll for you repeatedly, then (on average) the true proportion of voters who support you would fall within this interval 95% of the time.
  - Here’s what this doesn’t mean:
    - There is a 95% chance the proportion of the population who support you lies in this interval.
- But wait...how did the polling agency come up with this confidence interval/MoE?
Standard Normal Table

In the olden days (back when I was a student), we used the standard normal table to answer queries.

Find $z_{lo}$ and $z_{hi}$ s.t. $P[z_{lo} \leq Z \leq z_{hi}] = 1 - \alpha$

- If $1 - \alpha = 90\%$, then $|z_{\alpha/2}| = z_{1-\alpha/2} = 1.645$
- If $1 - \alpha = 95\%$, then $|z_{\alpha/2}| = z_{1-\alpha/2} = 1.96$
- If $1 - \alpha = 98\%$, then $|z_{\alpha/2}| = z_{1-\alpha/2} = 2.33$
- If $1 - \alpha = 99\%$, then $|z_{\alpha/2}| = z_{1-\alpha/2} = 2.58$

$z_{lo}$ and $z_{hi}$ are called critical values.
Building a 95% Confidence Interval

- Let $X$ be the number of people who support you in a poll, and let $p_{\text{hat}} = X / n$.
- By the central limit theorem (YAY!), for large $n$, $p_{\text{hat}}$ is approximately normal, with mean $\mu$ and standard deviation $\sigma$.
- Perfect. Let’s apply a z-transformation: $\Pr[z_{ \text{lo}} \leq Z \leq z_{ \text{hi}}] = .95$.
  - $\Pr[z_{ \text{lo}} \leq (p_{\text{hat}} - \mu) / \sigma \leq z_{ \text{hi}}] = .95$
  - $\Pr[\mu + z_{ \text{lo}} \sigma \leq p_{\text{hat}} \leq \mu + z_{ \text{hi}} \sigma] = .95$
  - N.B. $z_{\text{lo}} = -1.96$ and $z_{\text{hi}} = 1.96$, since $\alpha = 0.05$. 
Building a 95% Confidence Interval (cont’d)

- Let $X$ be the number of people who support you in a poll.
- By the central limit theorem (YAY!), for large $n$, the estimator $X/n$ is approximately normal, with mean $\mu$ and standard deviation $\sigma$.
- Perfect. Let’s apply a z-transformation: $\Pr[z_{lo} \leq Z \leq z_{hi}] = .95$.
  - $\Pr[z_{lo} \leq (X/n - \mu)/\sigma \leq z_{hi}] = .95$
  - $\Pr[\mu + z_{lo} \sigma \leq X/n \leq \mu + z_{hi} \sigma] = .95$
  - N.B. $z_{lo} = -1.96$ and $z_{hi} = 1.96$, since $\alpha = 0.05$.
- Sounds good, but what are the mean $\mu$ and standard deviation $\sigma$ of $X/n$?
  - $\mu = \mathbb{E}[X/n] = 1/n \mathbb{E}[X]$  
  - $\sigma^2 = \text{Var}[X/n] = 1/n^2 \text{Var}[X]$
- This begs the question: what are the mean and variance of $X$?
The mean and variance of $X/n$

- Recall that $X$ is the number of people who support you in a poll.
- $X$ is the result of a random experiment (the poll), so it is a random variable.
- Great! How is it distributed? Easy: $X$ is binomially distributed:
  - Mean = $np$
  - Variance = $np(1-p)$
- The expected value of $\hat{p}$ is $p$.
  - $\mu = \mathbb{E}[X/n] = np/n = p$
- The variance of $\hat{p}$ is $(p)(1 - p)/n$.
  - $\sigma^2 = \text{Var}[X/n] = 1/n^2 \text{Var}[X] = 1/n^2 (np)(1 - p)$
Z-transformation

- Recall:
  - $\Pr[z_{lo} \leq (X/n - \mu)/\sigma \leq z_{hi}] = .95$
  - $\Pr[\mu + z_{lo} \sigma \leq X/n \leq \mu + z_{hi} \sigma] = .95$

- The expected value of $X/n$ is $p$ and the variance of $X/n$ is $(p)(1 - p)/n$. So:
  - $\Pr[p + z_{lo} \ p \ (1 - p)/n \leq X/n \leq p + z_{hi} \ p \ (1 - p)/n] = .95$

- New problem: We don’t know $p$!
Z-transformation

- Recall:
  - $\Pr[z_{lo} \leq (X/n - \mu)/\sigma \leq z_{hi}] = .95$
  - $\Pr[\mu + z_{lo} \sigma \leq X/n \leq \mu + z_{hi} \sigma] = .95$

- The expected value of $X/n$ is $p$ and the variance of $X/n$ is $(p)(1 - p)/n$. So:
  - $\Pr[p + z_{lo} p (1 - p)/n \leq X/n \leq p + z_{hi} p (1 - p)/n] = .95$

- New problem: We don’t know $p$!

- We estimate $p$ by $p_{\text{hat}}$ and $\sigma$ by $\sigma_{\text{hat}}$:
  - $\Pr[p_{\text{hat}} + z_{lo} \sigma_{\text{hat}} \leq \mu \leq p_{\text{hat}} + z_{hi} \sigma_{\text{hat}}] = .95$
  - $\sigma_{\text{hat}} = p_{\text{hat}} (1 - p_{\text{hat}})/n$

- For our sample, $p_{\text{hat}} = X/n = 0.55$, and
  - $\sigma_{\text{hat}}^2 = (.55)(.45)/100 = 0.002475$
  - $\sigma_{\text{hat}} = \sqrt{0.002475} = 0.05$
An Interval Around Proportions

Hence, we build our 95% confidence interval as follows:

- \( \Pr[p_{\hat{\text{lo}}} + \sigma_{\hat{\text{lo}}} \leq \frac{X}{n} \leq p_{\hat{\text{hi}}} + \sigma_{\hat{\text{hi}}} ] = .95 \)
  - Lower Bound: 0.55 + (-1.96)(0.05) = .45
  - Upper Bound: 0.55 + (1.96)(0.05) = .65

- The value (1.96)(0.05) \( \approx 0.1 \) is called the margin of error.

Reminder, restated in terms of margin of error:

- If we conduct this poll repeatedly, the true proportion of your supporters would be expected to fall within in the margin of error 95 times out of 100.
Examples
Favorite Ice Cream Flavor among U.S. Adults

- A Harris Poll surveyed 2242 adults in the United States about their favorite flavor of ice cream.
  - 27% of people said chocolate was their favorite flavor.
  - 23% of people said vanilla was their favorite flavor.
- Is the preference for chocolate over vanilla significant? That is, do people really prefer chocolate?
- Plan of attack:
  - Compare the confidence intervals for chocolate and vanilla.
  - If the confidence intervals overlap, we cannot conclude that there is a difference between the proportion of people who prefer chocolate and the proportion who prefer vanilla.
Favorite Ice Cream Flavor among U.S. Adults

• We need to calculate the standard error to find a confidence interval:
  ○ (.27)(2242) = 605 people preferred chocolate
  ○ (.23)(2242) = 516 people preferred vanilla
  ○ Var[p1] = (0.27)(1 - 0.27) / 605 = 0.000326
  ○ Var[p2] = (0.23)(1 - 0.23) / 516 = 0.000343
  ○ The standard error is the square root of the variance:
    ■ SE[p1] = sqrt(0.000326) = 0.018
    ■ SE[p2] = sqrt(0.000343) = 0.019

• So, the confidence intervals, at the 95% level, are:
  ○ For chocolate: [0.27 - (1.96)(0.018), 0.27 + (1.96)(0.018)] = [0.234, 0.305]
  ○ For vanilla: [0.23 - (1.96)(0.019), 0.23 + (1.96)(0.019)] = [0.193, 0.267]

• The confidence intervals overlap!
  ○ This means we cannot conclude, at the 95% confidence level, that people prefer chocolate to vanilla.
  ○ But maybe people prefer chocolate to vanilla at the 90% confidence level. (Do they? Find out for yourself.)
Some words of warning!

- Our methods make use of the central limit theorem.
  - In the example, we do not assume anything about how ice cream preferences are distributed.
  - However, by the CLT, we do assume the sampling distribution is approximately normal.
  - The survey was large enough for the CLT to apply. However, the CLT might not have applied if the sample size were smaller. (Rule of thumb: CLT applies whenever $n \geq 30$.)

- **Very common mistake:** a 95% confidence interval between 1 and 2 is often interpreted as a 95% chance the parameter lies between 1 and 2.
  - This is a misconception! A 95% interval means that if we repeated the experiment 100 times, 95 of the resulting 100 intervals would contain the true population parameter.
  - These two interpretations are not the same. Be careful to avoid this potential pitfall!
Examples, Revisited
Favorite Ice Cream Flavor among U.S. Adults

- A Harris Poll surveyed 2242 adults in the United States about their favorite flavor of ice cream.
  - 27% of people said chocolate was their favorite flavor.
  - 23% of people said vanilla was their favorite flavor.
- Is the preference for chocolate over vanilla significant? That is, do people really prefer chocolate?
- Plan of attack:
  - Look at the difference in the proportions. It is 4%.
  - Calculate the confidence interval around this difference.
  - If it contains 0, then we cannot conclude that there is a difference.
Favorite Ice Cream Flavor among U.S. Adults

- We need to calculate the standard error to find a confidence interval:
  - \( (0.27)(2242) = 605 \) people preferred chocolate
  - \( (0.23)(2242) = 516 \) people preferred vanilla
  - \( \text{Var}[p_1 - p_2] = (0.27)(1 - 0.27)/605 + (0.23)(1 - 0.23)/516 = 0.00067 \)
  - The standard error is the square root of the variance: \( \sqrt{0.00067} = 0.026 \)

- So, the confidence interval, at the 95% level, is:
  - \([0.04 - (1.96)(0.026), 0.04 + (1.96)(0.026)]\)
  - \([-0.011, 0.09]\)
  - \([-1.1\%, 9\%]\)

- 0% is included in this interval!
  - This means we cannot conclude, at the 95% confidence level, that people prefer chocolate to vanilla.
  - But maybe people prefer chocolate to vanilla at the 90% confidence level. (Check this!)
The 2016 Presidential Election

- YouGov surveyed 1300 people in the United States and asked “Who will you vote for in the election for President in November?”
  - The poll was conducted on October 7th and 8th
  - 44% of people planned to vote for Clinton
  - 38% of people planned to vote for Trump

- Given these results, how confident are you that Clinton will win the election?

- Plan of attack:
  - Look at the difference in the proportions. It is 6%.
  - Calculate the confidence interval around this difference.
  - If it contains 0, then we cannot conclude that there is a difference.
The 2016 Presidential Election

● We need to calculate the standard error to find a confidence interval:
  ○ (.44)(1300) = 572 people preferred Clinton
  ○ (.38)(1300) = 494 people preferred Trump
  ○ $\text{Var}[p1 - p2] = (0.44)(1 - 0.44) / 572 + (0.38)(1 - 0.38) / 494 = 0.0009$
  ○ The standard error is the square root of the variance: $\sqrt{0.0009} = 0.0301$

● So, the confidence interval, at the 95% level, is:
  ○ 0.06 - (1.96)(0.0301) to 0.06 + (1.96)(0.0301)
  ○ 0.001 to 0.119
  ○ 0.1% to 11.9%

● 0% is not included in this interval!
  ○ This means we can conclude, at the 95% confidence level, that Clinton is preferred to Trump.
  ○ But people probably don’t prefer Clinton to Trump at the 99% confidence level. (Check this!)
Student $t$-Distribution
When does the CLT kick in?

Sampling Distribution of the Mean with $n = 10$

Sampling Distribution of the Mean with $n = 30$

Sampling Distribution of the Mean with $n = 100$

Sampling Distribution of the Mean with $n = 1000$
What is a large enough sample size?

- The distribution looks roughly normal with a sample size of 30.
- So, as a rule of thumb, people often say that 30 is a large enough sample size for the central limit theorem to apply.
- However, the histogram becomes more and more bell-shaped as the sample size increases.
- So all things being equal, larger sample sizes are always better than smaller ones!
What if the sample size is not large enough?

- We can extend our methodologies by introducing a new distribution, called the *(Student)* t-distribution, which approximates the normal.
  - The student in question was one William S. Gosset.
  - He discovered this (family of) distribution(s) in 1908, while employed as a statistician by the Guinness brewing company, who forbade him from publishing under his own name.
  - He wrote under the pen name “Student” instead.

- The t-distribution allows us to perform statistical inference even when the sample size is not large enough to apply the central limit theorem.
The *t*-Distribution

- The *t*-distribution has 1 parameter: the **degrees of freedom**.
- As \( n \) goes to infinity, the *t*-distribution converges to the normal.
- Thus, using the *t*-distribution when the sample size is small is consistent with using the normal distribution when the sample size is large.
Building Confidence Intervals

- The only difference between building confidence intervals using the \( t \)-distribution and building them using the normal is: the critical values for the normal distribution are familiar numbers to most statisticians. Recall:
  - Find \( z_{lo} \) and \( z_{hi} \) s.t. \( P[z_{lo} \leq Z \leq z_{hi}] = 1 - \alpha \%
    - If \( 1 - \alpha = 90\% \), then \( |z_{a/2}| = z_{1-\alpha/2} = 1.645 \)
    - If \( 1 - \alpha = 95\% \), then \( |z_{a/2}| = z_{1-\alpha/2} = 1.96 \)
    - If \( 1 - \alpha = 98\% \), then \( |z_{a/2}| = z_{1-\alpha/2} = 2.33 \)
    - If \( 1 - \alpha = 99\% \), then \( |z_{a/2}| = z_{1-\alpha/2} = 2.58 \)
  - \( z_{lo} = z_{a/2} \) & \( z_{hi} = z_{1-\alpha/2} \) are called critical values.

- For the \( t \)-distribution, we have to look these values up.
  - The magnitude of these values for \( \alpha = .05 \), for example, is higher than 1.96.
  - There is more uncertainty because of the smaller sample size.
An Example

- Let’s say we polled 30 people to see who they wanted to win the 2016 presidency, and 60% of people said Clinton.
- We would build a 95% confidence interval as follows:
  - The estimate is 0.6.
  - The SE = \sqrt{(.6)(.4)/30} = 0.09.
  - We look up the critical t-values:
    - $qt(0.025, 29) = -2.05$ and $qt(0.975, 29) = 2.05$
- The 95% CI is $(0.6 - (2.05)(0.09), 0.6 + (2.05)(0.09)) = (0.42, 0.78)$
Deriving Standard Error
Standard Error

- The **sample mean** is the average of all the sample values.
- The **sample variance** is the average of the squared deviations from the mean.
- The **sample standard deviation** is the square root of the sample variance.
- The **standard error** is the standard deviation of the sampling distribution.

- In this class, we are learning to build confidence intervals for sample means and sample proportions.
- This means we need to be able to calculate the standard error of sample means and sample proportions.
The Standard Error of the Sample Mean

- Let $X_M$ represents the sample mean:
  - $X_M = (X_1 + \ldots + X_N) / n$

- First, we need the variance of the sample mean:
  - $X_i$ are normally distributed with variance $\sigma$
  - $\text{Var}[X_M] = \text{Var}[(X_1 + \ldots + X_N) / n] = (1 / n^2) (n) \text{Var}[X_1] = \sigma^2 / n$

- We now have a formula for the standard error of the sample mean:
  - $\text{SE}[X_M] = \sigma / \sqrt{n}$
The Standard Error of the Sample Proportion

- Let $\hat{P}$ represent the sample proportion:
  - $\hat{P} = X / n$, where $X$ is the binomial random variable distributed according to $(n, p)$

- First, we need the variance of the sample proportion:
  - $X \sim B(n, p)$
  - $\text{Var}[\hat{P}] = \text{Var}[X / n] = (1 / n^2)(np)(1 - p) = p (1 - p) / n$

- We now have a formula for the standard error of the sample proportion:
  - $\text{SE}[\hat{P}] = \sqrt{p (1 - p) / n}$
Difference of Two Sample Means

- Let’s find the standard error of the difference between two sample means.
  - Let \( X_M - Y_M \) represent the difference between two sample means.
- \( X_M - Y_M = \frac{X_1 + \ldots + X_N}{n} - \frac{Y_1 + \ldots + Y_M}{m} \)
  - \( X_i \) and \( Y_i \) are normally distributed with variance \( \sigma_x \) and \( \sigma_y \)
  - \( \text{Var}[X_M - Y_M] = \text{Var}[X_M] + \text{Var}[Y_M] \)
    - \( = \text{Var}\left[\frac{X_1 + \ldots + X_N}{n}\right] + \text{Var}\left[\frac{Y_1 + \ldots + Y_M}{m}\right] \)
    - \( = \left(\frac{1}{n^2}\right) n \text{Var}[X_1] + \left(\frac{1}{m^2}\right) m \text{Var}[Y_1] \)
    - \( = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} \)
- We now have a formula for the standard error of the difference between two sample means:
  - \( \text{SE}[X_M] = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} \)
Difference of Two Sample Proportions

• Let’s find the standard error of the difference between two sample proportions.
  ○ Let $P_1$ represent one proportion and $P_2$ represent a second proportion.

• $P_1 = \frac{X}{n}$ and $P_2 = \frac{Y}{m}$
  ○ $X \sim B(n, p_1)$ and $Y \sim B(m, p_2)$
  ○ $\text{Var}[P_1 - P_2] = \text{Var}[X / n] + \text{Var}[Y / m]$
  ○ $= \left(\frac{1}{n^2}\right)(np_1)(1-p_1) + \left(\frac{1}{m^2}\right)(mp_2)(1-p_2)$
  ○ $= \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$

• We now have a formula for the standard error of the difference between two sample proportions:
  ○ $\text{SE}[P_1 - P_2] = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$
Extras
The 2016 Presidential Election

- YouGov surveyed 1300 people in the United States and asked “Who will you vote for in the election for President in November?”
  - The poll was conducted on October 7th and 8th
  - 44% of people planned to vote for Clinton
  - 38% of people planned to vote for Trump

- Given these results, how confident are you that Clinton will win the election?

- Plan of attack:
  - Compare the confidence intervals for Clinton and Trump.
  - If the confidence intervals overlap, we cannot conclude that there is a difference between the proportion of people who prefer Clinton and the proportion who prefer Trump.
The 2016 Presidential Election

- We need to calculate the standard error to find a confidence interval:
  - (.44)(1300) = 572 people preferred Clinton
  - (.38)(1300) = 494 people preferred Trump
  - Var[p1] = (0.44)(1 - 0.44) / 572 = 0.000431
  - Var[p2] = (0.38)(1 - 0.38) / 494 = 0.000477
  - The standard error is the square root of the variance:
    - SE[p1] = sqrt(0.000431) = 0.021
    - SE[p2] = sqrt(0.000477) = 0.022

- So, the confidence intervals, at the 95% level, are:
  - For Clinton: [0.44 - (1.96)(0.021), 0.44 + (1.96)(0.021)] = [0.40, 0.48]
  - For Trump: [0.38 - (1.96)(0.022), 0.38 + (1.96)(0.022)] = [0.34, 0.42]

- The confidence intervals overlap! (Don’t say we didn’t warn you!)
  - It turns out that if confidence intervals don’t overlap, then we can conclude that the difference is statistically significant; but if they do overlap, then we cannot draw any statistical conclusions.