Class information

- Drew D. Creal
- Email: dcreal@chicagobooth.edu
- Office: 404 Harper Center
- Office hours: email me for an appointment
- Office phone: 773.834.5249
Course schedule

- Week # 1: Plotting and summarizing univariate data
- Week # 2: Plotting and summarizing bivariate data
- Week # 3: Probability 1
- Week # 4: Probability 2
- Week # 5: Probability 3
- Week # 6: In-class exam and Probability 4
- Week # 7: Statistical inference 1
- Week # 8: Statistical inference 2
- Week # 9: Simple linear regression
- Week # 10: Multiple linear regression
Outline of today’s topics

I. Statistical Inference: Two Examples
II. Sampling Distributions
III. Sampling Distribution for $p$
IV. Confidence Interval for $p$
V. Summary: Estimating $p$ for i.i.d. Bernoulli data
VI. Sampling Distribution for $\mu$
VII. Confidence Interval for $\mu$
VIII. Confidence Intervals other than 95%
IX. The Plug-in Predictive Interval for the i.i.d. Normal Model
Statistical Inference:

two motivating examples
Statistical Inference: Two Examples

We have discussed possible models for data. Examples are:

1. \textit{i.i.d.} Bernoulli model
2. \textit{i.i.d.} Normal model
3. random walk model
4. regression model

Each of these models has unknown \textit{parameters}.

How do we know the values for the parameters?
Example 1: Voting and the \textit{i.i.d.} Bernoulli model

Suppose we have a large population of voters in Chicago and suppose each citizen will vote either Democrat or Republican.

We would like to know the fraction that will vote Democrat. But the \textbf{population} is too large: \textit{We can’t ask them all!}

We \textbf{model} voters as \textit{i.i.d.} Bernoulli. Let $X_i = 1$ if the $i$-th person votes Democrat and zero otherwise.

$$X_i \sim \text{Bernoulli}(p) \ i.i.d.$$
Example 2: Stock returns and the \textit{i.i.d.} Normal model

We are considering investing in a certain stock.

If we hold this asset for a long time, what is the average monthly return we should expect?

We model these returns as \textit{i.i.d.} normal. Let $X_i$ be the return in month $i$.

$$X_i \sim N(\mu, \sigma^2) \text{ i.i.d.}$$

We don’t know $\mu$!!! (or $\sigma^2$)

Can we learn about $\mu$ by taking a \textit{sample} of returns?
In both cases, we want to use *observed data* to learn something about the **parameters** of a model.

How do we do this?

**Example 1:** We conduct a poll. We take a random sample of voters and we use the proportion of Democrats in the sample as a guess or “estimate” of the true proportion in the whole population.

**Example 2:** We look at a sample of past returns and use the sample average as a guess for the expected return.
Example 1: Voting and the i.i.d. Bernoulli model

\( p \): fraction of Democrats in the whole population. This parameter is unknown.

\( \hat{p} \): before we have taken the poll, we plan to use the proportion of Democrats in our sample of size \( n \) as our guess for the true value of \( p \).

\[
\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

We hope that \( \hat{p} \approx p \).
Example 1: Voting and the \textit{i.i.d.} Bernoulli model

We treat our sample of voters from the poll \( \{X_1, X_2, \ldots, X_n\} \) as the outcomes of \textit{i.i.d.} Bernoulli r.v.’s from the population.

\textbf{KEY IDEA:} Recognize that there are many possible “samples” of voters that we could poll.

The value we get for \( \hat{p} \) may be different for each sample.

Prior to conducting the poll, \( \hat{p} \) is a r.v.!! Why?

After conducting the poll, we observe an outcome for the random variable \( \hat{p} \).
Example 1: Voting and the $i.i.d.$ Bernoulli model

In real life, we (typically) only observe one sample, e.g. we only conduct one poll.

Given this sample, we get a specific value (a number) for $\hat{p}$.

In general, the value we get for $\hat{p}$ from our sample will not be equal to the true, unknown value of $p$.

Question: can we use our model to make a statement about how close $\hat{p}$ will be to $p$?
Example 2: Stock returns and the \textit{i.i.d.} Normal model

\(\mu\) : the “true” expected return on an asset. This parameter is \textit{unknown}.

\(\bar{x}\) : before we get the data, we plan to use the average return in a sample of historical monthly returns as a guess for \(\mu\).

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

We hope that \(\bar{x} \approx \mu\).
Example 2: Stock returns and the \( i.i.d. \) Normal model

We treat our return data \( \{X_1, X_2, \ldots, X_n\} \) as though it is an \( i.i.d. \) draw from a \( N(\mu, \sigma^2) \) distribution.

**KEY IDEA:** Before we collect our data, \( \bar{x} \) is a r.v.

This is because it will be different for each sample of returns we might observe!!

Once we collect the data, we will observe an outcome for the random variable \( \bar{x} \).
Example 2: Stock returns and the \textit{i.i.d.} Normal model

In real life, we only observe one sample of returns for any individual asset.

Given this sample, we get a single value (i.e. a specific number) for $\bar{x}$.

Question: can we use our model to make a statement about how close $\bar{x}$ will be to $\mu$?
Statistical terminology

An *estimator* is a function of the data that is used to infer the value of an unknown parameter in a statistical model. An estimator is a random variable because it is a function of the data, which are random variables.

You can think of an estimator as a “rule” that determines how we will use our observed data to calculate our guess of the unknown parameter.

- Example #1: \( \hat{p}(X) = \frac{1}{n} \sum_{i=1}^{n} X_i \).
- Example #2: \( \bar{x}(X) = \frac{1}{n} \sum_{i=1}^{n} X_i \).
An *estimate* is the observed outcome of the estimator for a specific data set. In other words, an estimate is a realization of the random variable (the estimator).

Once you observe a specific data set $X_i = x_i$, you will be able to compute the estimate which is just a number.

- Example #1: $\hat{p} = 0.52$.
- Example #2: $\bar{x} = 0.05$. 
Statistical Inference

**Statistical inference** combines data analysis with probability. Here’s the logic:

We have a model in mind for our data but do not know what the parameters are.

**STEP 1:** Assume the data are *i.i.d.* draws from a model with unknown parameters.

**STEP 2:** Come up with an **estimator**, i.e. a summary statistic that tells us about some aspect of the model.

**STEP 3:** Before we see the data, our estimator is a r.v. due to the fact that there are many samples we might observe.

**STEP 4:** Assuming our model is true, we use probability to say how close our estimator will be to the actual (unknown) value of the parameter.
Warning: don’t get confused over notation and jargon.

Remember, to do statistical inference, we think of our sample statistics as “estimators” where our estimators are random variables because we have yet to see the data.

Once we see the data, we will write stuff like: \( \hat{p} = 0.2 \)

This statement says that the outcome of the r.v. \( \hat{p} \) was 0.2.

A potential point of confusion is that we use \( \hat{p} \) to denote the estimator and the estimate.

It is an abuse of notation.
Sampling Distributions
The Sampling Distribution of an Estimator

We will proceed as follows...

(1.) First, we build a model for some data.
(2.) Secondly, we choose an estimator of an unknown parameter.
(3.) Finally, after we have observed our sample, we get a numeric value that we call our estimate.

But, our estimate will (likely) be wrong!!

Remember from example 1, the proportion of Democrats in our sample will be different from that in the whole population.
The *sampling error* is the difference between the estimate and the true value of the parameter.

- You can think of the “sampling error” as the randomness introduced into our “guess” due to taking one sample instead of another.

- In the voting example, the value of $\hat{p}$ will generally be different than the true value of $p$.

  $$\varepsilon = \hat{p} - p$$

- In the stock return example, the value of $\bar{x}$ will generally be different than $\mu$.

  $$\varepsilon = \bar{x} - \mu$$
The Sampling Distribution of an Estimator

Key question: Can we use our model to make a statement about how close our estimator will be to the true value of the parameter?

We can ask questions like:

- On average, will our estimator tend to be too high or too low?
- Will the variance of the sampling error be large or small?

To answer these questions, we use the *sampling distribution* of our estimator.
The **sampling distribution** of an estimator is a probability distribution that describes all the possible outcomes of the estimator we might see if we could “repeat” our sample over and over again.

**Key idea:** imagine repeatedly drawing hypothetical samples of data from a large population defined by our model. For each sample, compute the value of the estimator. A histogram of the values is (approximately) the sampling distribution.
The Sampling Distribution for $p$
The Sampling Distribution for $p$

Example 1: Voting and the *i.i.d.* Bernoulli model

Let $X_i = 1$ if the $i$-th sampled voter is a Democrat, 0 if Republican. Let $Y$ denote the number of Democrats in the sample.

\[
\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{Y}{n}, \quad Y \sim \text{Binomial}(n, p)
\]

$\hat{p}$ is just a $\text{Binomial}(n, p)$ random variable divided by $n$.

This is the **sampling distribution** of the estimator (in this example).

The sampling distribution of the estimator depends on:

1. The sample size $n$.
2. The unknown parameter (in this case $p$).
Example 1: Suppose voters are $X_i \sim i.i.d. \text{Bernoulli}(0.6)$. We take a sample of size $n = 200$ out of a large population.

One possible value of $\hat{p}$ from one possible sample.

True value of $p = 0.6$

$\hat{p} = 0.6012$

Another possible value of $\hat{p}$ from a different sample

True value of $p = 0.6$

$\hat{p} = 0.5979$

$\hat{p} = 0.6012$

A third possible value of $\hat{p}$ from a third sample

$\hat{p} = 0.5979$

$\hat{p} = 0.5989$

$\hat{p} = 0.6012$

The green lines are 15 different values of $\hat{p}$ from 15 different samples we might observe.
The Sampling Distribution for $p$

If we create a histogram of a large number of values of $\hat{p}$ we might observe from different potential samples, we will find that (approximately)

$$\hat{p} \sim N \left( p, \frac{p(1 - p)}{n} \right)$$

(NOTE: In a minute, we will see how to approximate the sampling distribution using the Central Limit Theorem (CLT).)
The Sampling Distribution for $p$

Example 1: Voting and the i.i.d. Bernoulli model

In this example, we can compute the mean and variance of our estimator to summarize the distribution:

$$E[\hat{p}] = E \left[ \frac{Y}{n} \right]$$

$$= \frac{1}{n} E \left[ \sum_{i=1}^{n} X_i \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} E[X_i]$$

$$= \frac{np}{n} = p$$

Before we take our sample, we expect the proportion of Democrats in our sample will be $p$ (the true population value).
The Sampling Distribution for $p$

Example 1: Voting and the $i.i.d.$ Bernoulli model

Now, let's look at the variance of the sampling distribution for this example.

$$V[\hat{p}] = V\left[ \frac{Y}{n} \right] = \frac{1}{n^2} V[Y]$$

$$= \frac{np(1 - p)}{n^2}$$

$$= \frac{p(1 - p)}{n}$$

(NOTE: the $n$ in the denominator will cause the variance to decrease if we sample a larger number of people.)
Example 1: Voting and the \textit{i.i.d.} Bernoulli model

We can combine this result and also apply the Central Limit Theorem. For large \(n\), the sampling distribution is (approximately):

\[
\hat{p} \sim N \left( p, \frac{p(1-p)}{n} \right)
\]

This gives us a simple way of thinking about what kind of estimate we might see.
The Sampling Distribution for $p$

**Example 1: Voting and the i.i.d. Bernoulli model**

Let’s repeat the logic we used:

We **model** voters in the U.S. as Bernoulli($p$). We want to learn about the unknown parameter $p$ based on a sample of voters.

**STEP 1:** Assume the data are i.i.d. Bernoulli($p$).

**STEP 2:** We plan to use the sample proportion of Democrats, $\hat{p}$, as an estimator of the unknown parameter $p$.

**STEP 3:** Before we see the data, $\hat{p}$ is (approximately) normally distributed with

$$E[\hat{p}] = p \text{ and } V[\hat{p}] = \frac{p(1-p)}{n}$$
The Sampling Distribution for $p$

In our Bernoulli example, if the population we are sampling from is described by:

$$X_i \sim \text{Bernoulli}(p) \ i.i.d.$$  

The sampling distribution of our estimator $\hat{p}$ is (approximately)

$$\hat{p} \sim \mathcal{N} \left( p, \frac{p(1-p)}{n} \right)$$

(NOTE #1: This is the sampling distribution for this example not all examples.)

(NOTE #2: In this example, the stand. dev. of the sampling dist. is $\sqrt{\frac{p(1-p)}{n}}$.)
The Sampling Distribution of an Estimator

Note:

This is where people often get confused by two things!

1. A lot of people think the true value of the parameter $p$ is “random.” NO! It’s just a number, like 0.5 or 0.265, we just do not know what it is.

2. Similarly, people think “how can $\hat{p}$ be random?! I can compute it from the data!!”

But the sampling distribution we will use is based on the insight:

Before we see the data, $\hat{p}$ is a random variable because it depends on which sample we observe.
ASIDE: Unbiasedness

In this example, notice how the mean of the sampling distribution is equal to the true value of the parameter.

\[ E[\hat{p}] = p \]

In other words, the sampling error \( \varepsilon \) is zero on average

\[ E[\varepsilon] = E[\hat{p}] - p = 0 \]

Intuitively, our estimate can turn out to be too big or too small, but it is not “systematically” too high or too low.

An estimator that has this property is said to be \textit{unbiased}.

(NOTE: The expectation (or average) is being taken over hypothetical random samples we might observe from the model.)
Confidence Interval for $p$
Confidence Interval for $p$

Our estimate $\hat{p}$ gives a specific value or “guess” as to what the unknown $p$ is.

An alternative question is: how uncertain are we about $p$?

The *confidence interval* is the standard solution.

It follows directly from the idea of a sampling distribution.
**95% Confidence Interval for \( p \)**

**STEP 1:** First, we use our sampling distribution for the estimator \( \hat{p} \) to make a probability statement:

\[
Pr \left( p - 2 \sqrt{\frac{p(1-p)}{n}} < \hat{p} < p + 2 \sqrt{\frac{p(1-p)}{n}} \right) \approx 0.95
\]

**STEP 2:** By applying the CLT and using the normal dist, we can construct an interval

\[
\left( p - 2 \sqrt{\frac{p(1-p)}{n}}, p + 2 \sqrt{\frac{p(1-p)}{n}} \right)
\]

**STEP 3:** In practice we do not know \( p \), we consequently replace it by our guess \( \hat{p} \) to get:

\[
\left( \hat{p} - 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)
\]

Interpretation: this interval will include the true value of \( p \) in 95% of the samples we might observe.
95% Confidence Interval for $p$

Before we see the data, we know that the interval

$$
\left( \hat{p} - 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)
$$

will include the true value of $p$ in 95% of the samples we might observe.

We are therefore 95% “confident” in our interval.

The 95% confidence interval for $p$ is:

$$
\left( \hat{p} - 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)
$$
95% Confidence Interval for $p$

Our logic:

1. Suppose the true value is $p = 0.6$

2. Consider a sample of size $n = 200$ from Bernoulli$(0.6)$.

3. For this sample, our estimate is $\hat{p} = 0.59$ and $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.035$.

4. Given $\hat{p} = 0.59$, our confidence interval is $(0.52, 0.66)$

5. This interval includes the true value of $p$. 

The 95% confidence interval $(0.52, 0.66)$ based on the first sample.
95% Confidence Interval for $p$

Our logic:

1. Consider a second sample of size $n = 200$ from Bernoulli(0.6).
2. For this sample, our estimate is $\hat{p} = 0.62$ and $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.034$.
3. Given $\hat{p} = 0.62$, our confidence interval is $(0.55, 0.69)$
4. This interval includes the true value of $p$. 

The 95% confidence interval (0.55, 0.69) based on the second sample.
95% Confidence Interval for $p$

Our logic:

1. Consider a third sample of size $n = 200$ from Bernoulli($0.6$).

2. For this sample, our estimate is $\hat{p} = 0.51$ and $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.035$.

3. Given $\hat{p} = 0.51$, our confidence interval is $(0.45, 0.59)$.

4. This interval does **NOT** include the true value of $p$. 

The 95% confidence interval (0.44, 0.58) based on the third sample.

True value of $p = 0.6$
95% Confidence Interval for $p$

Our confidence interval (CI) is based on the idea that in 95% of the samples we might observe the true value of $p$ will lie in the interval

\[
(\hat{p} - 2\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}})
\]

This follows from the Central Limit Theorem.

To make this statement we use two approximations.

1. Use of central limit theorem.
2. Use of estimate $\hat{p}$ in standard deviation formula.
Standard Error

The *standard error* is the special name given to the standard deviation of the sampling distribution of an estimator.

- The name “standard” error is unfortunate because it can be awfully confusing at first.
- In the Bernoulli example, the standard error is: 
  \[ S.E. = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \]
- IMPORTANT: The formula for the standard error changes depending on the sampling distribution.
“Republican Chris Christie continues to hold a three-point advantage...The latest Rasmussen Reports telephone survey in the state, conducted Thursday night, shows Christie with 46% (\(\hat{\rho} = 0.46\)) of the vote and Corzine with 43%.”

“This statewide telephone survey of 1,000 Likely Voters in New Jersey was conducted by Rasmussen Reports October 29, 2009. The margin of sampling error for the survey is +/- 3 percentage points with a 95% level of confidence.”

\[
2\sqrt{\frac{0.46(1 - 0.46)}{1000}} = 0.0315
\]
Here, \( p \) = the probability that a randomly chosen voter will vote for Christie.

The 95% confidence interval for the true value of \( p \) is

\[ 0.46 \pm 0.0315 \]

This is equal to the “estimate \( \pm 2\times \) standard error.”

(Remember: we are using the CLT which says that we are approximating the true sampling distribution using a normal distribution.)
Is that a big interval?

Well, notice that the closest competitor at the time (Corzine) is at 43% which is within the 95% confidence interval.

Although Christie did win the election (48.5% of the vote), many news agencies said the race was still up for grabs right up until election day.
Summary: Estimating $p$ for IID Bernoulli data

For $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$ i.i.d.

Let $\hat{p} = \frac{X_1 + X_2 + \ldots + X_n}{n}$ be our estimator of $p$.

Our estimator is **unbiased**: $E[\hat{p}] = p$

It has variance: $V[\hat{p}] = \frac{p(1-p)}{n}$

The sampling distribution is (approximately):

$$\hat{p} \sim N \left( p, \frac{p(1-p)}{n} \right)$$
Summary: Estimating $p$ for IID Bernoulli data

For $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$ i.i.d.

The standard error of a sample proportion is

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$ i.i.d.

The (approximate) 95% confidence interval for $p$ is given by:

$$\left( \hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$
Example: the IID defect model

- Recall our defects example from Lecture # 4.
- 35 out of the 200 parts are defective or 17%.

We modeled this as:

\[ X_i \sim \text{Bernoulli}(p) \]

where \( X_i = 1 \) if the \( i \)-th part is defective.

In Lecture # 4, we “guessed” a value for \( p \) of 0.20.
Example: the IID defect model

Now, we estimate $p$ to be:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{35}{200} = 0.175$$

Given our model for the process, how wrong could we be about $p$?

$$2 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 2 \sqrt{\frac{0.175 \times 0.825}{200}} = 0.054$$

The 95% confidence interval is $0.175 \pm 0.054 = (0.12, 0.23)$.

(NOTE: The true value of $p$ when I simulated the “data” was 0.2, which is inside the interval.)
Interpreting the size of a confidence interval

Question: how much do I know about the parameter?

Answer:

If the confidence interval is small: I know a lot.

If the confidence interval is large: I know little.
Interpreting the size of a confidence interval

Example:
Suppose \( \hat{p} = 0.2 \) and \( n = 100 \).

\[
\text{S.E.}(\hat{p}) = 0.04 \quad \text{such that} \quad 95\% \ \text{CI is: } (.12, .28)
\]

Instead, suppose \( \hat{p} = 0.2 \) and \( n = 10,000 \).

\[
\text{S.E.}(\hat{p}) = 0.004 \quad \text{such that} \quad 95\% \ \text{CI is: } (.192, .208)
\]

\textbf{NOTE:} We increased \( n \) by a factor of 100 and decreased the standard error by 1/10.
Population versus sample
Most statistics textbooks emphasize the population/sample viewpoint.

In the voting example, we are trying to learn about a “population” that is too big to study directly.

In our voting example we took a random sample from a large population.

In that example, the difference between the population proportion \( p \), which we wish to know, and the sample proportion \( \hat{p} \) is pretty clear.
Population and Sample

In the defects example, we took a sample of parts and computed the fraction that were defective.

But……how do we interpret what $p$ is in this example?

In the defects example, the concept of “population” is a bit contrived because there is no “population” of parts that we are picking a sample from.

If you want, you can imagine our process as generating an infinite stream of parts such that $p$ is the long run fraction of defectives in the “infinite population” of parts.
Population and Sample

In the example on asset returns of a stock, we took a sample of past historical returns and computed the average return over that time period.

What is the population of returns in this case?

Most statistics textbooks emphasize the population/sample viewpoint. Sometimes it is a good way to understand things intuitively, but sometimes it is not.
The Sampling Distribution and Confidence Interval for $\mu$
Confidence Interval for $\mu$

Suppose you work for General Mills. The company claims to put 350 grams of cereal in each box.

You are responsible for quality control and need to measure the amount of cereal put in each box in grams. You take a sample of $n = 500$ boxes.
Confidence Interval for $\mu$

This is a different situation than the voter example we discussed above.

Our data is numeric: weight in grams.

Rather than a categorical variable with two possible outcomes, we are modeling a numeric variable.

It doesn’t make sense to use the Bernoulli distribution.
Confidence Interval for $\mu$

We could model the process as:

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2) \ i.i.d.$$  

**Given our model, a parameter of interest is $\mu$.**

A reasonable estimator of $\mu$ is the sample mean, $\overline{x}$.

When we finally observe our data, we get a value for $\overline{x}$; that is our estimate of the unknown parameter $\mu$.

What are the properties of our estimator?
The Sampling Distribution for $\mu$

Example 2: Cereal and the i.i.d. Normal model

Let’s compute the mean and variance of the sampling distribution.

Our estimator is:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} (X_1 + X_2 + \ldots + X_n)$$

In HWK #2, we used our linear formulas to compute the mean

$$E[\bar{x}] = \frac{1}{n} E \left[ \sum_{i=1}^{n} X_i \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} E[X_i]$$

$$= \frac{n \mu}{n} = \mu$$
Example 2: Cereal and the \textit{i.i.d.} Normal model

In HWK #2, we also used our linear formulas to compute the variance

\[
V[\bar{x}] = V \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right]
\]

\[
= \frac{1}{n^2} V \left[ \sum_{i=1}^{n} X_i \right]
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} V [X_i]
\]

\[
= \frac{n \sigma^2}{n^2} = \frac{\sigma^2}{n}
\]

(NOTE: the \( n \) in the denominator will cause the variance to decrease if we sample a larger number of people.)
Summary: Estimating $\mu$ for IID Normal data

For $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ i.i.d.

Given a sample of $n$ observations, let $\bar{x} = \frac{X_1+X_2+\ldots+X_n}{n}$ be our estimator of $\mu$.

Our estimator is **unbiased**: $E[\bar{x}] = \mu$

It has variance: $V[\bar{x}] = \frac{\sigma^2}{n}$

The sampling distribution is: $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
Confidence Interval for $\mu$

In our cereal example, the sample mean is $\bar{x} = 345$. But, how different is this from the true $\mu$?
Confidence Interval for \( \mu \)

How different is our estimate \( \bar{x} \) from the true \( \mu \)?

**STEP 1:** Using the sampling distribution, we can make a probability statement about \( \bar{x} \):

\[
Pr \left( \mu - 2 \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 2 \frac{\sigma}{\sqrt{n}} \right) = 0.95
\]

**STEP 2:** Consequently, we can construct an interval

\[
\left( \mu - 2 \frac{\sigma}{\sqrt{n}}, \mu + 2 \frac{\sigma}{\sqrt{n}} \right)
\]

**STEP 3:** Since we don't know \( \mu \) or \( \sigma \), we replace them by their sample values to get

\[
\left( \bar{x} - 2 \frac{s_x}{\sqrt{n}}, \bar{x} + 2 \frac{s_x}{\sqrt{n}} \right)
\]

Interpretation: this interval will include \( \mu \) in 95% of the samples we might observe.
Standard Error

The standard error of the sample mean is

$$\text{S.E.}(\overline{X}) = \frac{s_x}{\sqrt{n}}$$

- The “standard error” is still the standard deviation of the sampling distribution.
- However, the sampling distribution for $\overline{X}$ is different than the sampling distribution for $\hat{p}$, making the standard errors different.
- Here, the standard error uses the sample standard deviation $s_x$. 
For $X_1, X_2, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$ i.i.d.

The (approximate) 95% confidence interval for $\mu$ is given by:

$$\bar{x} \pm 2 \times \text{S.E.}(\bar{x}) = \left( \bar{x} - 2 \frac{s_x}{\sqrt{n}}, \bar{x} + 2 \frac{s_x}{\sqrt{n}} \right)$$
Confidence Interval for $\mu$

Example 2: In our cereal example, the sample mean and stand. dev. are $\bar{x} = 345$ and $s_x = 15$.

The standard error is:

$$S.E.(\bar{x}) = \frac{s_x}{\sqrt{n}} = \frac{15}{\sqrt{500}} = 0.67$$

The confidence interval is:

$$\bar{x} \pm 2 \times S.E.(\bar{x}) = 345 \pm 1.34 = (343.66, 346.34)$$

If your company was claiming to put 350 grams of cereal in each box, the interval would be accurate enough to suggest the distribution is centered far from the target.
Confidence Interval for $\mu$

Here, we plot the 95% confidence interval (343.66, 346.34) to see how uncertain we are about $\mu$. 
Confidence Interval for $\mu$

**Question:** Do we really need to assume our data is *normal*?

**NO!** The Central Limit theorem tells us that when $n$ is large, 

$$\bar{x} \sim N \left( \mu, \frac{\sigma^2}{n} \right)$$

This justifies the confidence interval $\bar{x} \pm 2 \frac{s_x}{\sqrt{n}}$ even when the data is not normally distributed if:

1. the sample size $n$ is large
2. the data are *i.i.d.*
Confidence Interval for $\mu$

Example: Let us find a 95% confidence interval for the Canadian return data with $\bar{x} = 0.00907$ and $s_x = 0.03833$. 

![Confidence Interval Diagram](image-url)
Summary
Statistics terms: repeat for emphasis...

A parameter refers to an unknown quantity we are interested in. Two examples:

- The proportion of Democratic voters $p$ in the U.S. population.
- The expected return on a stock, $\mu$.

Before we have observed a specific data set, we can consider all the possible sequences of data that we might possibly see from our model.

An estimator is the value of the estimate for different possible sequences of data that might happen.
**Statistics terms:** repeat for emphasis...

Remember that an estimator is a random variable because we could observe many possible samples. For each sample, the estimator is likely to be different.

An estimate is something we compute from observed data and use as a “guess” for the unknown parameter. Two examples:

- Take a sample of voters and use the proportion of Democrats in the sample $\hat{p}$ to estimate the unknown $p$.
- Look at the sample mean $\bar{x}$ of a set of historical returns to estimate the unknown $\mu$. 
Statistics terms: continued...

The sampling distribution of an estimator is a probability distribution that describes the different values we might see for our estimator.

Two examples:

- If $X_i \sim \text{Bernoulli}(p)$ i.i.d.: $\hat{p} \sim N \left(p, \frac{p(1-p)}{n}\right)$

- If $X_i \sim N(\mu, \sigma^2)$ i.i.d.: $\bar{x} \sim N \left(\mu, \frac{\sigma^2}{n}\right)$

Key: The sampling distribution depends on the unknown parameter.
The standard error of our estimate is the “standard deviation” of the sampling distribution.

Two examples:

1. \[ S.E. (\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]
2. \[ S.E. (\bar{x}) = \frac{s_x}{\sqrt{n}} \]
How to Compute Confidence Intervals other than 95\%
Confidence Intervals other than 95%

The confidence intervals we have considered both look like:

\[ P(\text{estimate} - 2 \times \text{S.E.} < \text{par} < \text{estimate} + 2 \times \text{S.E.}) = 0.95 \]

Here, “par” is the unknown parameter of interest, e.g. \( p \) or \( \mu \).

Why do our 95% CI’s have this “estimate ± 2 × S.E.” form?

The normal distribution! The probability that a normal r.v. is within TWO standard deviations of its mean is 0.95.
Confidence Intervals other than 95%

- 2 (or 1.96) is called the “95% critical value” of the normal distribution.
- There is a 95% probability that a normal random variable is within 2 stand. devs. of its mean.
Confidence Intervals other than 95%

Suppose we want a 90% confidence interval instead. In other words, we want an interval that (before the sample is drawn) is 90% likely to contain the unknown parameter.
Confidence Intervals other than 95%

There is a 90% probability that a normal random variable is within 1.64 standard deviations of its mean!
How did we get -1.64 and 1.64?

- Let’s look at the standard normal CDF.
- There is a 5% probability that a standard normal random variable is less than -1.64.

In Excel, “=NORMSINV(0.05) = -1.64”
How did we get -1.64 and 1.64?

- Let’s look at the standard normal CDF.
- There is a 95% probability that a standard normal random variable is less than 1.64.

In Excel, “=NORMSINV(0.95) = 1.64”
Confidence Intervals other than 95%

We then say that 1.64 is the “90% critical value” of the normal distribution.

A 90% confidence interval would then look like:

\[ \text{estimate} \pm 1.64 \times \text{S.E.} \]

Notice that this interval is narrower than our 95% C.I..

We are making a trade here. We have narrowed our interval in exchange for admitting there is a greater chance of it being wrong.
Confidence Intervals other than 95%

Let $\alpha$ be a number between 0 and 1.

For an estimator whose sampling distribution is (approximately) normal, the $100 \times \alpha\%$ confidence interval for the true parameter value is:

$$\text{estimate} \pm z\text{-value} \times \text{S.E.}$$

$$z\text{-value} = \text{NORMSINV}\left(\frac{1-\alpha}{2}\right) \quad \text{(NOTE: In Excel)}$$

NOTE: Sometimes the critical value is called a “z-value” because it comes from the standard normal distribution. You can also look up these critical values in most stats textbooks.
Confidence Intervals other than 95%

Don’t worry too much about this. We will deal mostly with 95% confidence intervals.

By the way, this is also why I don’t like the term “z value.” People use this term to refer to BOTH the actual number of S.E.’s an estimator is from the mean, AND to refer to the “cutoff point” (under the bell curve, how many standard deviations away from the mean correspond to a certain probability). I will always call the latter a “critical value.”
ASIDE: Quantile function

- Notice that when using the CDF, we give it a value $x$ and it returns the probability.
- The CDF is just a function $F(x)$.
- We can invert this function $F^{-1}(x)$.
- In other words, we give $F^{-1}(x)$ a probability and it returns $x$.
- This is what the function NORMSINV/NORMINV does.
- It is called a *quantile function*. 
The Plug-in Predictive Interval for the IID Normal Model
The Plug-in Predictive Interval for the IID Normal Model

Let’s go back to our cereal box example.

Suppose a machine is about to fill the next box.

How much cereal will go in the box?
The Plug-in Predictive Interval for the IID Normal Model

According to our model,

\[ X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2) \quad i.i.d. \]

the amount of cereal in the next box will be an independent draw from the same normal distribution the observed data came from.

(NOTE: For the confidence interval, we did not really need to assume the data were normally distributed (because of the CLT). Here we do need this assumption!)
The Plug-in Predictive Interval for the IID Normal Model

We have our usual picture:

If the data is $N(\mu, \sigma^2)$, there is a 95% chance the next observation will be in the interval:

$$\mu \pm 2\sigma$$
The Plug-in Predictive Interval for the IID Normal Model

Our estimate for $\mu$ is 345 (the sample mean) and our estimate for $\sigma$ is 15 (the sample standard deviation).

If we plug in our estimates for the true values we obtain

$$345 \pm 30$$

which looks reasonable.
The Plug-in Predictive Interval for the IID Normal Model

For *i.i.d.* normal data, the 95% plug in predictive interval is

\[ \bar{x} \pm 2 \times s_x \]

**NOTE:**

This is nothing new - it is the “empirical rule” we’ve been using all along! If we believe things are normal, then 95% of values we see should be within 2 s.d.’s of the mean.

Because of our model, we can apply the interval to future data and predict!
The Plug-in Predictive Interval for the IID Normal Model

- The dashed green lines are the plug in predictive interval.
- They answer the question: “what will the next observation be?”
- The solid red lines are the confidence interval for $\mu$.
- They answer the question: “what do we think $\mu$ is?”
The Plug-in Predictive Interval for the IID Normal Model

Comparing the predictive intervals and the confidence intervals.
IMPORTANT: Inference versus Prediction

- In the confidence interval, we are using all the data to learn about an unknown parameter \( \mu \).

- In the predictive interval, we are trying to make a prediction for a single random outcome.

- The formulas for these two intervals look similar. But, they are doing very different things.

- The confidence interval shrinks as \( n \) grows because with more data we gather more information about the unknown parameter \( \mu \).

- The prediction interval does not shrink. Having more data does not make us any more certain about what happens for any single outcome.