Outline

1. Sampling
2. Central Limit Theorem
3. Proportions
Outline

1. Sampling
2. Central Limit Theorem
3. Proportions
Populations and samples

When we use statistics, we are trying to find out information about a whole *population* by gathering information about just some of them (a *sample*).
When we use statistics, we are trying to find out information about a whole *population* by gathering information about just some of them (a *sample*).
Our process

population
Our process

population

select

sample
Our process

1. Select
2. Sample
3. Compute statistics
Our process

- Select a sample from the population
- Compute statistics
- Approximate parameters
Population Mean vs. Sample Mean

For example, we use the sample mean to approximate the population mean.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Considers</th>
<th>is a...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>All members of the population</td>
<td>parameter</td>
</tr>
<tr>
<td>$\bar{x}$</td>
<td>Just our random sample</td>
<td>statistic</td>
</tr>
</tbody>
</table>
## Parameters and Statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>Standard Deviation</th>
<th>these are . . .</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>$\sigma$</td>
<td>parameters</td>
</tr>
<tr>
<td>Sample</td>
<td>$\bar{X}$</td>
<td>$s^2$</td>
<td>$s$</td>
<td>statistics</td>
</tr>
</tbody>
</table>
### Parameters and Statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>Standard Deviation</th>
<th>these are...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>$\sigma$</td>
<td>parameters</td>
</tr>
<tr>
<td>Sample</td>
<td>$\overline{X}$</td>
<td>$s^2$</td>
<td>$s$</td>
<td>statistics</td>
</tr>
</tbody>
</table>

These are what we care about.
### Parameters and Statistics

These are what we care about.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>Standard Deviation</th>
<th>these are...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>$\sigma$</td>
<td>parameters</td>
</tr>
<tr>
<td><strong>Sample</strong></td>
<td>$\bar{x}$</td>
<td>$s^2$</td>
<td>$s$</td>
<td>statistics</td>
</tr>
</tbody>
</table>

These are just our estimates of the top row.
To keep our variables straight, we typically use the following conventions:

- We use Greek letters for parameters describing the whole population
  - E.g., $\mu$, $\sigma$, $\sigma^2$
- We use lower-case Roman letters for statistics computed from the sample
  - E.g., $\bar{x}$, $s$, $s^2$
- We use upper-case Roman letters for random variables
  - E.g., $X$, $Z$
Notation convention

To keep our variables straight, we typically use the following conventions:

- We use Greek letters for parameters describing the whole population
  - E.g., $\mu$, $\sigma$, $\sigma^2$

- We use lower-case Roman letters for statistics computed from the sample
  - E.g., $\bar{x}$, $s$, $s^2$

- We use upper-case Roman letters for random variables
  - E.g., $X$, $Z$
Notation convention

To keep our variables straight, we typically use the following conventions:

- We use Greek letters for parameters describing the whole population
  - E.g., $\mu$, $\sigma$, $\sigma^2$
- We use lower-case Roman letters for statistics computed from the sample
  - E.g., $\bar{x}$, $s$, $s^2$
- We use upper-case Roman letters for random variables
  - E.g., $X$, $Z$
Statistics as random variables

- Every time you take a random sample and compute the sample mean $\overline{x}$, you’ll probably get a different answer.
- Thus the sample mean *is itself a random variable*, written as $\overline{X}$.
- Likewise the sample standard deviation is a random variable, written as $S$.

**Distinction**

- When we write $\overline{x}$ or $s$, we are thinking of an actual number computed from a specific sample.
- When we write $\overline{X}$ or $S$, we are thinking of the *process* of computing a sample mean or sample s.d., which gives us different results at different times.
Statistics as random variables

- Every time you take a random sample and compute the sample mean $\bar{x}$, you’ll probably get a different answer.
- Thus the sample mean is itself a random variable, written as $\bar{X}$.
- Likewise the sample standard deviation is a random variable, written as $S$.

**Distinction**

- When we write $\bar{x}$ or $s$, we are thinking of an actual number computed from a specific sample.
- When we write $\bar{X}$ or $S$, we are thinking of the process of computing a sample mean or sample s.d., which gives us different results at different times.
Outline

1. Sampling
2. Central Limit Theorem
3. Proportions
Since $\bar{X}$ is a random variable, we can ask what *its* probability distribution is.

What are its mean and standard deviation?

Let’s begin with a population distribution, take some random samples, and compute $\bar{X}$ for each sample.

By keeping track of lots of sample means, we’ll see what the distribution of $\bar{X}$ looks like.
Since $\overline{X}$ is a random variable, we can ask what its probability distribution is.

What are its mean and standard deviation?

Let’s begin with a population distribution, take some random samples, and compute $\overline{x}$ for each sample.

By keeping track of lots of sample means, we’ll see what the distribution of $\overline{X}$ looks like.
Central Limit Theorem

![Diagram of a normal distribution curve]

![Diagram of a rectangle]
Sampling Central Limit Theorem

Number of samples: 1
Number of samples: 2
Sampling Central Limit Theorem Proportions

Number of samples: 3
Number of samples: 4
Sampling Central Limit Theorem

Number of samples: 5
Sampling Central Limit Theorem Proportions

Number of samples: 6
Sampling

Central Limit Theorem

Number of samples: 7
Sampling Central Limit Theorem Proportions

Number of samples: 8
Number of samples: 9
Sampling Central Limit Theorem Proportions

Number of samples: 10
Number of samples: 11
Number of samples: 12
Number of samples: 13
Number of samples: 14
Number of samples: 15
Number of samples: 16
Sampling

Central Limit Theorem

Number of samples: 17
Number of samples: 18
Number of samples: 19
Number of samples: 20
Number of samples: 21
Sampling Central Limit Theorem

Number of samples: 22
Number of samples: 23
Number of samples: 24
Sampling Central Limit Theorem Proportions

Number of samples: 25
Number of samples: 26
Sampling Central Limit Theorem

Number of samples: 27
Sampling Central Limit Theorem

Number of samples: 28
Number of samples: 29
Sampling Central Limit Theorem

Number of samples: 30
Number of samples: 31
Number of samples: 32
Number of samples: 33
Number of samples: 34
Number of samples: 35
Number of samples: 36
Number of samples: 37
Number of samples: 38
Number of samples: 39
Sampling Central Limit Theorem

Number of samples: 40
Sampling Central Limit Theorem

Number of samples: 41
Sampling Central Limit Theorem

Number of samples: 42
Sampling

Central Limit Theorem

Proportions

Number of samples: 43
Number of samples: 44
Sampling Central Limit Theorem

Number of samples: 45
Number of samples: 46
Sampling Central Limit Theorem

Number of samples: 47
Number of samples: 48
Sampling Central Limit Theorem Proportions

Number of samples: 49
Number of samples: 50
Number of samples: 1000
Number of samples: 1000
Sampling Central Limit Theorem

Number of samples: 1000
Sampling Central Limit Theorem

Number of samples: 1
Number of samples: 2
Sampling Central Limit Theorem Proportions

Number of samples: 3
Number of samples: 4
Sampling 

Central Limit Theorem 

Number of samples: 5
Number of samples: 6
Number of samples: 7
Number of samples: 8
Sampling Central Limit Theorem

Number of samples: 9
Number of samples: 10
Sampling

Central Limit Theorem

Proportions

Number of samples: 12
Number of samples: 13
Number of samples: 14
Sampling

Central Limit Theorem

Number of samples: 15
Number of samples: 16
Sampling Central Limit Theorem Proportions

Number of samples: 17
Sampling Central Limit Theorem

Number of samples: 18
Number of samples: 19
Sampling Central Limit Theorem

Number of samples: 20
Number of samples: 21
Number of samples: 22
Number of samples: 23
Sampling Central Limit Theorem Proportions

Number of samples: 24
Number of samples: 25
Number of samples: 26
Number of samples: 27
Number of samples: 28
Number of samples: 29
Sampling

Central Limit Theorem

Proportions

Number of samples: 30
Number of samples: 31
Number of samples: 32
Sampling Central Limit Theorem Proportions

Number of samples: 33
Number of samples: 34
Number of samples: 35
Sampling Central Limit Theorem

Number of samples: 36
Central Limit Theorem

Number of samples: 37
Number of samples: 38
Number of samples: 39
Number of samples: 40
Number of samples: 41
Number of samples: 42
Number of samples: 43
Number of samples: 44
Sampling Central Limit Theorem

Number of samples: 45
Number of samples: 46
Number of samples: 47
Number of samples: 49
Sampling

Central Limit Theorem

Number of samples: 50
Sampling Central Limit Theorem

Number of samples: 1000
Sampling Central Limit Theorem

Number of samples: 1000
Sampling Central Limit Theorem

Number of samples: 1000
Sampling Central Limit Theorem Proportions
Sampling Central Limit Theorem

Number of samples: 1
Number of samples: 2
Sampling Central Limit Theorem Proportions

Number of samples: 3
Number of samples: 4
Sampling

Central Limit Theorem

Number of samples: 5
Number of samples: 6
Number of samples: 7
Sampling

Central Limit Theorem

Number of samples: 8
Number of samples: 9
Sampling

Central Limit Theorem

Number of samples: 10
Sampling Central Limit Theorem

Number of samples: 11
Sampling Central Limit Theorem

Number of samples: 12
Number of samples: 13
Number of samples: 14
Sampling Central Limit Theorem

Number of samples: 15
Sampling

Central Limit Theorem

Number of samples: 16
Number of samples: 17
Sampling Central Limit Theorem

Number of samples: 18
Number of samples: 19
Number of samples: 20
Number of samples: 21
Sampling

Central Limit Theorem

Number of samples: 22
Number of samples: 23
Sampling

Central Limit Theorem

Number of samples: 24
Central Limit Theorem

Number of samples: 25
Number of samples: 26
Number of samples: 27
Number of samples: 28
Sampling Central Limit Theorem

Number of samples: 29
Number of samples: 30
Number of samples: 31
Sampling Central Limit Theorem Proportions

Number of samples: 32
Number of samples: 33
Number of samples: 34
Number of samples: 35
Sampling Central Limit Theorem Proportions

Number of samples: 36
Sampling

Central Limit Theorem

Number of samples: 37
Number of samples: 38
Sampling

Central Limit Theorem

Number of samples: 39
Sampling Central Limit Theorem

Number of samples: 40
Number of samples: 41
Number of samples: 42
Number of samples: 43
Sampling Central Limit Theorem

Number of samples: 44
Number of samples: 45
Number of samples: 46
Sampling Central Limit Theorem

Number of samples: 48
Sampling Central Limit Theorem Proportions

Number of samples: 49
Sampling Central Limit Theorem

Number of samples: 50
Number of samples: 1000
Number of samples: 1000
Sampling Central Limit Theorem

Number of samples: 1000
We started with a distribution, took samples, and recorded our sample means.

The sample means $\bar{x}$ seemed to be normally distributed.

- The average of the sample means was the actual population mean.
- The standard deviation of the sample means seemed smaller than the population standard deviation.

All of this was true, no matter what the original distribution looked like.

These observations lead to a very, very important theorem!
We started with a distribution, took samples, and recorded our sample means.

The sample means $\bar{x}$ seemed to be normally distributed.

- The average of the sample means was the actual population mean.
- The standard deviation of the sample means seemed smaller than the population standard deviation.

All of this was true, no matter what the original distribution looked like.

These observations lead to a very, very important theorem!
Observations

- We started with a distribution, took samples, and recorded our sample means.
- The sample means $\bar{x}$ seemed to be normally distributed.
  - The average of the sample means was the actual population mean.
  - The standard deviation of the sample means seemed smaller than the population standard deviation.
- *All of this was true, no matter what the original distribution looked like.*
- These observations lead to a very, very important theorem!
Observations

- We started with a distribution, took samples, and recorded our sample means.
- The sample means $\bar{x}$ seemed to be normally distributed.
  - The average of the sample means was the actual population mean.
  - The standard deviation of the sample means seemed smaller than the population standard deviation.

  *All of this was true, no matter what the original distribution looked like.*

- These observations lead to a very, very important theorem!
Observations

- We started with a distribution, took samples, and recorded our sample means.
- The sample means $\bar{x}$ seemed to be normally distributed.
  - The average of the sample means was the actual population mean.
  - The standard deviation of the sample means seemed smaller than the population standard deviation.
- All of this was true, no matter what the original distribution looked like.
- These observations lead to a very, very important theorem!
Observations

- We started with a distribution, took samples, and recorded our sample means.
- The sample means $\bar{x}$ seemed to be normally distributed.
  - The average of the sample means was the actual population mean.
  - The standard deviation of the sample means seemed smaller than the population standard deviation.
- All of this was true, no matter what the original distribution looked like.
- These observations lead to a very, very important theorem!
Central Limit Theorem

Consider a population with mean $\mu$ and standard deviation $\sigma$. Let the random variable $\overline{X}$ be the sample mean of a randomly chosen sample of size $n$.

Then $\overline{X}$ is approximately a normal distribution

- with mean $\mu$
- and with standard deviation $\frac{\sigma}{\sqrt{n}}$.

In general, the Central Limit Theorem will apply if:

- the population distribution is normal, OR
- the population distribution is symmetric and $n \geq 10$, OR
- $n \geq 30$. 

Central Limit Theorem

Consider a population with mean $\mu$ and standard deviation $\sigma$. Let the random variable $\bar{X}$ be the sample mean of a randomly chosen sample of size $n$.

Then $\bar{X}$ is approximately a normal distribution

- with mean $\mu$
- and with standard deviation $\frac{\sigma}{\sqrt{n}}$

In general, the Central Limit Theorem will apply if:

- the population distribution is normal, OR
- the population distribution is symmetric and $n \geq 10$, OR
- $n \geq 30$. 
Central Limit Theorem

Consider a population with mean \( \mu \) and standard deviation \( \sigma \). Let the random variable \( \bar{X} \) be the sample mean of a randomly chosen sample of size \( n \).

Then \( \bar{X} \) is approximately a normal distribution
- with mean \( \mu \)
- and with standard deviation \( \frac{\sigma}{\sqrt{n}} \).

In general, the Central Limit Theorem will apply if:
- the population distribution is normal, OR
- the population distribution is symmetric and \( n \geq 10 \), OR
- \( n \geq 30 \).
Let’s be clear: we do not use the Central Limit Theorem to get a better approximation of $\mu$ by using lots of samples.

- We only ever do one sample; we only get one red dot.
- If you want a better approximation, and can afford it, use a larger sample size (i.e., increase $n$).

We only get one red dot (one sample mean $\bar{x}$); the Central Limit Theorem will tell us how good it is!
What the Central Limit Theorem *is* for

In the next two weeks, we will use the Central Limit Theorem to:

- calculate “error bars” to estimate how accurate our sample mean is.
- test our hypotheses about the population.
- calculate how firmly we can believe our statistical conclusions.

We won’t do any of that today, though!
Example

Question
A hardware store receives a shipment of bolts that are supposed to be 12 cm long. The mean is indeed 12 cm, and the standard deviation is 0.2 cm. For quality control, the hardware store chooses 100 bolts at random to measure. They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.
Example

Question
A hardware store receives a shipment of bolts that are supposed to be 12 cm long. The mean is indeed 12 cm, and the standard deviation is 0.2 cm. For quality control, the hardware store chooses 100 bolts at random to measure. They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.

Solution
The population has $\mu = 12$ and $\sigma = 0.2$, but we are finding the sample mean $\bar{x}$ of the sample of $n = 100$ bolts.
Example

Question

... They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.

Solution, cont.

So $\mu = 12$, $\sigma = 0.2$, and $n = 100$. According to the Central Limit Theorem, $\bar{X}$ is normally distributed with mean $\mu = 12$ and standard deviation $\sigma/\sqrt{n} = 0.2/\sqrt{100} = 0.02$. Thus

$$Pr(11.97 \leq \bar{X} \leq 12.04)$$
Example

Question

...They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.

Solution, cont.

So $\mu = 12$, $\sigma = 0.2$, and $n = 100$. According to the Central Limit Theorem, $\bar{X}$ is normally distributed with mean $\mu = 12$ and standard deviation $\sigma / \sqrt{n} = 0.2 / \sqrt{100} = 0.02$. Thus

$$\Pr(11.97 \leq \bar{X} \leq 12.04)$$
Example

Question

... They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.

Solution, cont.

So \( \mu = 12 \), \( \sigma = 0.2 \), and \( n = 100 \). According to the Central Limit Theorem, \( \overline{X} \) is normally distributed with mean \( \mu = 12 \) and standard deviation \( \sigma / \sqrt{n} = 0.2 / \sqrt{100} = 0.2 \). Thus

\[
\Pr(11.97 \leq \overline{X} \leq 12.04)
\]
Example

Question

... They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.

Solution, cont.

So $\mu = 12$, $\sigma = 0.2$, and $n = 100$. According to the Central Limit Theorem, $\overline{X}$ is normally distributed with mean $\mu = 12$ and standard deviation $\sigma / \sqrt{n} = 0.2 / \sqrt{100} = 0.02$. Thus

$$
\Pr(11.97 \leq \overline{X} \leq 12.04) = \Pr \left( \frac{11.97 - 12}{0.02} \leq Z \leq \frac{12.04 - 12}{0.02} \right)
$$
Example

Question

... They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.

Solution, cont.

So $\mu = 12$, $\sigma = 0.2$, and $n = 100$. According to the Central Limit Theorem, $\bar{X}$ is normally distributed with mean $\mu = 12$ and standard deviation $\sigma/\sqrt{n} = 0.2/\sqrt{100} = 0.02$. Thus

\[
\Pr(11.97 \leq \bar{X} \leq 12.04) = \Pr \left( \frac{11.97 - 12}{0.02} \leq Z \leq \frac{12.04 - 12}{0.02} \right) = \Pr(-1.5 \leq Z \leq 2.0)
\]
Example

**Question**

...They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.

**Solution, cont.**

So \( \mu = 12, \sigma = 0.2, \) and \( n = 100. \) According to the Central Limit Theorem, \( \bar{X} \) is normally distributed with mean \( \mu = 12 \) and standard deviation \( \sigma / \sqrt{n} = 0.2 / \sqrt{100} = 0.02. \) Thus

\[
\Pr(11.97 \leq \bar{X} \leq 12.04) = \Pr \left( \frac{11.97 - 12}{0.02} \leq Z \leq \frac{12.04 - 12}{0.02} \right) \\
= \Pr(-1.5 \leq Z \leq 2.0) \\
= \Pr(Z \leq 2.0) - \Pr(Z \leq -1.5)
\]
Example

Question

... They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.

Solution, cont.

So $\mu = 12$, $\sigma = 0.2$, and $n = 100$. According to the Central Limit Theorem, $\bar{X}$ is normally distributed with mean $\mu = 12$ and standard deviation $\sigma/\sqrt{n} = 0.2/\sqrt{100} = 0.02$. Thus

$$\Pr(11.97 \leq \bar{X} \leq 12.04) = \Pr \left( \frac{11.97 - 12}{0.02} \leq Z \leq \frac{12.04 - 12}{0.02} \right)$$

$$= \Pr (-1.5 \leq Z \leq 2.0)$$

$$= \Pr(Z \leq 2.0) - \Pr(Z \leq -1.5)$$

$$= 0.977250 - 0.06681$$

$$= 0.910443.$$
Example

Question

... They will call the shipment defective and return it to the manufacturer if the average length of the 100 bolts is less than 11.97 cm or greater than 12.04 cm. Find the probability that the shipment will be found satisfactory.

Solution, cont.

So \( \mu = 12, \sigma = 0.2, \) and \( n = 100. \) According to the Central Limit Theorem, \( \bar{X} \) is normally distributed with mean \( \mu = 12 \) and standard deviation \( \sigma/\sqrt{n} = 0.2/\sqrt{100} = 0.02. \) Thus

\[
\Pr(11.97 \leq \bar{X} \leq 12.04) = \Pr \left( \frac{11.97 - 12}{0.02} \leq Z \leq \frac{12.04 - 12}{0.02} \right)
\]

\[
= \Pr (-1.5 \leq Z \leq 2.0)
\]

\[
= \Pr(Z \leq 2.0) - \Pr(Z \leq -1.5)
\]

\[
= 0.977250 - 0.06681 = 0.910443.
\]
Outline

1. Sampling
2. Central Limit Theorem
3. Proportions
There are two types of statistical questions:

- Numerical questions, to which the answer is a number
  - “How much does this pumpkin weigh?”
  - “How old are you?”

- Category questions, to which the answer is yes or no
  - “Is this mouse infected with the disease?”
  - “Do you have blue eyes?”

Everything we’ve done so far has been fine for numerical questions.

How can we handle these category questions?
Two Types of Questions

There are two types of statistical questions:

- Numerical questions, to which the answer is a number
  - “How much does this pumpkin weigh?”
  - “How old are you?”

- Category questions, to which the answer is yes or no
  - “Is this mouse infected with the disease?”
  - “Do you have blue eyes?”

Everything we’ve done so far has been fine for numerical questions.

How can we handle these category questions?
Two Types of Questions

There are two types of statistical questions:

- **Numerical questions**, to which the answer is a number
  - “How much does this pumpkin weigh?”
  - “How old are you?”

- **Category questions**, to which the answer is yes or no
  - “Is this mouse infected with the disease?”
  - “Do you have blue eyes?”

Everything we’ve done so far has been fine for numerical questions.

How can we handle these category questions?
Two Types of Questions

There are two types of statistical questions:

- Numerical questions, to which the answer is a number
  - “How much does this pumpkin weigh?”
  - “How old are you?”
- Category questions, to which the answer is yes or no
  - “Is this mouse infected with the disease?”
  - “Do you have blue eyes?”

Everything we’ve done so far has been fine for numerical questions.

How can we handle these category questions?
The main issue: Proportion

The main question about a category is, “what percentage of the population belongs to the category?”

Definition

A *proportion* is the percentage of the group that belongs to the category you’re interested in.
- The *population proportion* is denoted by $p$.
- The *sample proportion* is denoted by $\hat{p}$.

Because it is a number between 0 and 1, it can also be viewed as the probability that a randomly selected individual belongs to the category.

We use the sample proportion to estimate the population proportion.
The main question about a category is, “what percentage of the population belongs to the category?”

**Definition**

A *proportion* is the percentage of the group that belongs to the category you’re interested in.

- The *population proportion* is denoted by $\rho$.
- The *sample proportion* is denoted by $\hat{\rho}$.

Because it is a number between 0 and 1, it can also be viewed as the probability that a randomly selected individual belongs to the category.

We use the sample proportion to estimate the population proportion.
The main issue: Proportion

The main question about a category is, “what percentage of the population belongs to the category?”

**Definition**

A *proportion* is the percentage of the group that belongs to the category you’re interested in.

- The *population proportion* is denoted by $p$.
- The *sample proportion* is denoted by $\hat{p}$.

Because it is a number between 0 and 1, it can also be viewed as the probability that a randomly selected individual belongs to the category.

We use the sample proportion to estimate the population proportion.
What proportion of the population is green?
What proportion of the population is green?
What proportion of the population is green?
What proportion of the population is green?
What proportion of the population is green?

15 green
20 total
What proportion of the population is green?

\[ \hat{p} = \frac{15}{20} = 0.75. \]
What proportion of the population is green?

15 green
20 total

\[ \hat{p} = \frac{15}{20} = 0.75. \]

So we guess \( p \approx 0.75. \)
What proportion of the population is green?

\[ \hat{p} = \frac{15}{20} = 0.75. \]

So we guess \( p \approx 0.75. \)
## Parameters and Statistics

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Variance</th>
<th>Standard Deviation</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>$\mu$</td>
<td>$\sigma^2$</td>
<td>$\sigma$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>Statistic</td>
<td>$\bar{x}$</td>
<td>$s^2$</td>
<td>$s$</td>
<td>$\hat{\rho}$</td>
</tr>
</tbody>
</table>
We need a CLT for proportions!

- Over the next weeks we’ll use the CLT to calculate how good $\bar{x}$ is.
- The CLT we saw only worked for the sample mean of numerical questions.
- We want a CLT that will work for $\hat{p}$, the sample proportion of category questions.

Since the CLT described $\bar{X}$, we need an analogue that describes $\hat{P}$, which is the number of individuals in a random sample belonging to the category, divided by the sample size $n$. 
Calculating $\hat{P}$

Suppose the actual population proportion is $p$, and we take a sample of $n$ individuals. Call it a “success” if an individual belongs to our category.

- For each individual, the probability of success is $p$.
- For each individual, the probability of failure is $q = 1 - p$.
- This is true for each of the $n$ individuals.
- We are counting up the number of successes.
Calculating $\hat{P}$

Suppose the actual population proportion is $p$, and we take a sample of $n$ individuals. Call it a “success” if an individual belongs to our category.

- For each individual, the probability of success is $p$.
- For each individual, the probability of failure is $q = 1 - p$.
- This is true for each of the $n$ individuals.
- We are counting up the number of successes.

This is a binomial distribution!!

- Let $B = B(n, p)$ denote the binomial random variable with $n$ trials and probability $p$.
- Then $B$ is the number of successes in a random sample.
- Thus the sample proportion $\hat{P}$ equals $\frac{\# \text{ in category}}{\# \text{ in sample}} = \frac{1}{n} B$. 
Calculating $\hat{P}$

Suppose the actual population proportion is $p$, and we take a sample of $n$ individuals. Call it a “success” if an individual belongs to our category.

- For each individual, the probability of success is $p$.
- For each individual, the probability of failure is $q = 1 - p$.
- This is true for each of the $n$ individuals.
- We are counting up the number of successes.

*This is a binomial distribution!!*

- Let $B = B(n, p)$ denote the binomial random variable with $n$ trials and probability $p$.

  - Then $B$ is the number of successes in a random sample.
  - Thus the sample proportion $\hat{P}$ equals $\frac{\# \text{ in category}}{\# \text{ in sample}} = \frac{1}{n} B$. 
Calculating $\hat{P}$

Suppose the actual population proportion is $p$, and we take a sample of $n$ individuals. Call it a “success” if an individual belongs to our category.

- For each individual, the probability of success is $p$.
- For each individual, the probability of failure is $q = 1 - p$.
- This is true for each of the $n$ individuals.
- We are counting up the number of successes.

*This is a binomial distribution!!*

- Let $B = B(n, p)$ denote the binomial random variable with $n$ trials and probability $p$.
- Then $B$ is the number of successes in a random sample.

Thus the sample proportion $\hat{P}$ equals $\frac{\# \text{ in category}}{\# \text{ in sample}} = \frac{1}{n} B$. 
Calculating $\hat{P}$

Suppose the actual population proportion is $p$, and we take a sample of $n$ individuals. Call it a “success” if an individual belongs to our category.

- For each individual, the probability of success is $p$.
- For each individual, the probability of failure is $q = 1 - p$.
- This is true for each of the $n$ individuals.
- We are counting up the number of successes.

*This is a binomial distribution!!*

- Let $B = B(n, p)$ denote the binomial random variable with $n$ trials and probability $p$.
- Then $B$ is the number of successes in a random sample.
- Thus the sample proportion $\hat{P}$ equals $\frac{\text{# in category}}{\text{# in sample}} = \frac{1}{n} B$. 
On the last slide, we saw that

$$\hat{P} = \frac{1}{n} B(n, p).$$

On the other hand, we know from §7.7 that $B(n, p)$ is approximately the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. In symbols,

$$B(n, p) \approx N(np, \sqrt{npq}).$$

Putting these together, we see that

$$\hat{P} = \frac{1}{n} B(n, p)$$
On the last slide, we saw that

$$\hat{P} = \frac{1}{n} B(n, p).$$

On the other hand, we know from §7.7 that $B(n, p)$ is approximately the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. In symbols,

$$B(n, p) \approx N(np, \sqrt{npq}).$$

Putting these together, we see that

$$\hat{P} = \frac{1}{n} B(n, p)$$
On the last slide, we saw that
\[ \hat{P} = \frac{1}{n} B(n, p). \]

On the other hand, we know from §7.7 that \( B(n, p) \) is approximately the normal distribution with mean \( \mu = np \) and standard deviation \( \sigma = \sqrt{npq} \). In symbols,
\[ B(n, p) \approx N(np, \sqrt{npq}). \]

Putting these together, we see that
\[ \hat{P} = \frac{1}{n} B(n, p) \]
On the last slide, we saw that

\[ \hat{P} = \frac{1}{n} B(n, p). \]

On the other hand, we know from §7.7 that \( B(n, p) \) is approximately the normal distribution with mean \( \mu = np \) and standard deviation \( \sigma = \sqrt{npq} \). In symbols,

\[ B(n, p) \approx N(np, \sqrt{npq}). \]

Putting these together, we see that

\[ \hat{P} = \frac{1}{n} B(n, p) \]

\[ \approx \frac{1}{n} N(np, \sqrt{npq}) \]
On the last slide, we saw that

$$\hat{P} = \frac{1}{n} B(n, p).$$

On the other hand, we know from §7.7 that $B(n, p)$ is approximately the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. In symbols,

$$B(n, p) \approx N(np, \sqrt{npq}).$$

Putting these together, we see that

$$\hat{P} = \frac{1}{n} B(n, p) \approx \frac{1}{n} N(np, \sqrt{npq}) = N\left(\frac{np}{n}, \frac{\sqrt{npq}}{n}\right)$$
On the last slide, we saw that

\[ \hat{P} = \frac{1}{n} B(n, p). \]

On the other hand, we know from §7.7 that \( B(n, p) \) is approximately the normal distribution with mean \( \mu = np \) and standard deviation \( \sigma = \sqrt{npq} \). In symbols,

\[ B(n, p) \approx N(np, \sqrt{npq}). \]

Putting these together, we see that

\[ \hat{P} = \frac{1}{n} B(n, p) \]

\[ \approx \frac{1}{n} N(np, \sqrt{npq}) \]

\[ = N \left( \frac{np}{n}, \frac{\sqrt{npq}}{n} \right) \]

\[ = N \left( p, \sqrt{\frac{pq}{n}} \right). \]
Central Limit Theorem for Proportions

Consider a population with proportion \( p \), and let \( q = 1 - p \). Let the random variable \( \hat{P} \) be the sample proportion of a randomly chosen sample of size \( n \).

Then \( \hat{P} \) is approximately a normal distribution

- with mean \( p \)
- and with standard deviation \( \sqrt{\frac{pq}{n}} \).
### Example

#### Question
Suppose that 75% of Concordia students like corn on the cob. If you randomly survey 30 students, what is the probability that your survey says at least 60% like corn on the cob?
Example

Question
Suppose that 75% of Concordia students like corn on the cob. If you randomly survey 30 students, what is the probability that your survey says at least 60% like corn on the cob?

Solution
The population proportion is $p = 0.75$, so $q = 0.25$. The sample size is $n = 30$. Thus, by the Central Limit Theorem for Proportions, the sample proportion $\hat{P}$ is approximately normal with mean $p = 0.75$ and standard deviation

$$\sqrt{\frac{pq}{n}} = \sqrt{\frac{0.75 \cdot 0.25}{30}} = 0.0790569.$$
Question
Suppose that 75% of Concordia students like corn on the cob. If you randomly survey 30 students, what is the probability that your survey says at least 60% like corn on the cob?

Solution, cont.
So the sample proportion $\hat{P}$ is approximately normal with mean 0.75 and standard deviation 0.0790569. Then

\[
Pr(\hat{P} \geq 0.6) \approx Pr \left( Z \geq \frac{0.6 - 0.75}{0.0790569} \right) = Pr (Z \geq -3.16228) = \boxed{0.97111}.
\]
Hints for the Homework

- Identify whether the problem is about means or proportions.
- Identify what is the population and what is the sample.