Samples v. populations

- Recall our terminology:
  - Our data forms a **sample**
  - Our target of inference is the **population** from which the sample is drawn
Samples v. populations

- Recall our terminology:
  - Our data forms a sample
  - Our target of inference is the population from which the sample is drawn
- The population is (generally) *unobservable*
Samples v. populations

- Recall our terminology:
  - Our data forms a sample
  - Our target of inference is the population from which the sample is drawn
- The population is (generally) unobservable
- Inference means we seek to generalize about the population from observing the characteristics of a sample

Probably the most commonly known use of sampling strategies involves the use of surveys of limited groups to infer characteristics of a much larger group. Also standard in auditing strategies, the typical, central concern with sampling is to achieve representativeness.
Samples v. populations

- Recall our terminology:
  - Our data forms a sample
  - Our target of inference is the population from which the sample is drawn
- The population is (generally) unobservable
- Inference means we seek to generalize about the population from observing the characteristics of a sample
- Probably the most commonly known use of sampling strategies involves the use of surveys of limited groups to infer characteristics of a much larger group
Samples v. populations

- Recall our terminology:
  - Our data forms a sample
  - Our target of inference is the population from which the sample is drawn
- The population is (generally) unobservable
- Inference means we seek to generalize about the population from observing the characteristics of a sample
- Probably the most commonly known use of sampling strategies involves the use of surveys of limited groups to infer characteristics of a much larger group
- Also standard in auditing strategies
Samples v. populations

- Recall our terminology:
  - Our data forms a sample
  - Our target of inference is the population from which the sample is drawn
- The population is (generally) unobservable
- Inference means we seek to generalize about the population from observing the characteristics of a sample
- Probably the most commonly known use of sampling strategies involves the use of surveys of limited groups to infer characteristics of a much larger group
- Also standard in auditing strategies
- The typical, central concern with sampling is to achieve representativeness
Random sampling

- Basic idea: Every member of the population has an equal chance of being drawn into the sample
Random sampling

- Basic idea: Every member of the population has an equal chance of being drawn into the sample
- Presumes that every member of the population can be identified before the sample is drawn
Random sampling

- Basic idea: Every member of the population has an equal chance of being drawn into the sample
- Presumes that every member of the population can be identified before the sample is drawn
- Randomization is achieved in various ways:
  1. Generating numbers from chance physical events, such as drawing numbered balls, rolling dice, flipping a coin, etc.
  2. Selecting numbers from a random number table
  3. Using a computer to generate random numbers (preferred!)
Random sampling

- Basic idea: Every member of the population has an equal chance of being drawn into the sample
- Presumes that every member of the population can be identified before the sample is drawn
- Randomization is achieved in various ways:
  1. Generating numbers from chance physical events, such as drawing numbered balls, rolling dice, flipping a coin, etc.
  2. Selecting numbers from a random number table
  3. Using a computer to generate random numbers (preferred!)
- Many variations on this method of *simple random sampling* exist, since it is hard to achieve the “equal chance” standard in practice
Sampling error

- We may expect a sample statistic to equal a population parameter, but in each individual sample, we are likely to observe differences
Sampling error

- We may expect a sample statistic to equal a population parameter, but in each individual sample, we are likely to observe differences.
- These differences will vary each time a sample is drawn, even if in the long run, the average difference will be zero.
Sampling error

- We may expect a sample statistic to equal a population parameter, but in each individual sample, we are likely to observe differences.
- These differences will vary each time a sample is drawn, even if in the long run, the average difference will be zero.
- These differences are referred to as **sampling error**: the difference caused by chance differences between the sample’s characteristics and those in the population.
Sampling error

- We may expect a sample statistic to equal a population parameter, but in each individual sample, we are likely to observe differences.
- These differences will vary each time a sample is drawn, even if in the long run, the average difference will be zero.
- These differences are referred to as sampling error: the difference caused by chance differences between the sample’s characteristics and those in the population.
- This notion (and even the terminology) should be familiar from polling results, which are almost always reported with a sampling error.
Example: Random draws from total candidate spending

First, take 100 draws of 10 observations each, and plot the 100 resulting means.
Example: Random draws from total candidate spending

Next, compare with the same but using 1,000 draws of 10 observations each
Example: Random draws from total candidate spending

Finally, compare with the same but using 30 draws of 100 observations each
Example: Random draws from total candidate spending

draws.spending <- NULL
for (i in 1:30) {
    draws.spending[i] <- mean(sample(spending, 100))
}
hist(draws.spending, freq=FALSE, xlim=c(5000,25000), breaks=10,
     xlab="Spending", main="30 Draws of 100 observations each")
lines(density(draws.spending), col="red")
abline(v=mean(spending), lty="dashed", col="blue")

Note:

- the initialization of draws.spending with NULL
- the for() loop
- abline(v=...) produced a vertical line
Sampling distribution of means

- Assume that we have the luxury (as in the example) of repeating a sample as many times as we wish
Sampling distribution of means

- Assume that we have the luxury (as in the example) of repeating a sample as many times as we wish.
- The set of means of each sample – our “sample means” – would provide a *frequency distribution*. 
Sampling distribution of means

- Assume that we have the luxury (as in the example) of repeating a sample as many times as we wish.
- The set of means of each sample – our “sample means” – would provide a frequency distribution.
- But using probability theory, we also know what the probability distribution of the sampling means will be: in particular, it will be normally distributed.
Characteristics of a sampling distribution of means

1. The sampling distribution of means will be approximately normally distributed. This is true regardless of the distribution of the data from which the samples are drawn.
Characteristics of a sampling distribution of means

1. The sampling distribution of means will be approximately normally distributed. This is true regardless of the distribution of the data from which the samples are drawn.

2. The mean of a sampling distribution of means will equal the population mean.
Characteristics of a sampling distribution of means

1. The sampling distribution of means will be approximately normally distributed. This is true regardless of the distribution of the data from which the samples are drawn.

2. The mean of a sampling distribution of means will equal the population mean.

3. The standard deviation of a sampling distribution is smaller than the standard deviation of the population. In other words, the sample mean is less variable than the scores from which it is a sample. This feature is the key to our ability to make reliable inferences from samples to populations.
Population, sample, and sampling distributions

(a) Population distribution

(b) Sample distribution
(one sample with $N = 200$)

(c) Observed sampling distribution
(for 100 samples)

(d) Theoretical sampling distribution
(for infinite number of samples)
Central Limit Theorem

- The mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed.
- Works even when the population distribution being sampled is not normally distributed. Example: proportion of heads in a fair coin toss, over many coin tosses (underlying distribution is Bernoulli).
Using the normal curve to assess sample mean probabilities

- Recall that if we define probability as the likelihood of occurrence, then the normal curve can be regarded as a probability distribution.

- Theory tells us that the distribution of sampling means will be normally distributed.

- The question: If we assume \( \mu \) equals some specific value, then how likely was it to have drawn a given sample mean \( \bar{X} \)?
Using the normal curve to assess sample mean probabilities

- Recall that if we define probability as the likelihood of occurrence, then the normal curve can be regarded as a probability distribution.

- Using the $\mu$ and $\sigma$ from a normal curve, we can then assess the probability of finding specific scores along this distribution (as we did last week).
Using the normal curve to assess sample mean probabilities

- Recall that if we define probability as the likelihood of occurrence, then the normal curve can be regarded as a probability distribution.
- Using the $\mu$ and $\sigma$ from a normal curve, we can then assess the probability of finding specific scores along this distribution (as we did last week).
- The same applies to the distribution of sampling means, since theory tells us that this will be normally distributed.
Using the normal curve to assess sample mean probabilities

- Recall that if we define probability as the likelihood of occurrence, then the normal curve can be regarded as a probability distribution.
- Using the $\mu$ and $\sigma$ from a normal curve, we can then assess the probability of finding specific scores along this distribution (as we did last week).
- The same applies to the distribution of sampling means, since theory tells us that this will be normally distributed.

The question: If we assume $\mu$ equals some specific value, then how likely was it to have drawn a given sample mean $\bar{X}$?
Probability and the sampling distribution of means

- Remember that $\sigma$ is the standard deviation of population scores
Probability and the sampling distribution of means

- Remember that \( \sigma \) is the standard deviation of population scores
- The standard deviation of the sampling distribution (which will be smaller) is denoted as \( \sigma_{\bar{X}} \)
Probability and the sampling distribution of means

- Remember that $\sigma$ is the standard deviation of population scores
- The standard deviation of the sampling distribution (which will be smaller) is denoted as $\sigma_{\bar{x}}$
- Because the sampling distribution of means is normally distributed, we can use $z$ scores to obtain the probability of obtaining any given sample mean
Sampling means example

- Imagine that UCD claims its graduates earn $25,000 annually.
- We decide to test this claim by sampling 100 graduates and measuring their incomes.
- We obtain a sample mean of $\bar{X} = 23,500$. How likely was it to obtain this sample mean (or less) if the true (population) earnings mean is $25,000?
Sampling means example

Imagine that UCD claims its graduates earn $25,000 annually.

We decide to test this claim by sampling 100 graduates and measuring their incomes.

We obtain a sample mean of $23,500. How likely was it to obtain this sample mean (or less) if the true (population) earnings mean is $25,000?

Question: what is the area of the shaded region? (since this tells us the probability of obtaining a sample mean of $23,500 or less)
Sampling means example cont.

1. Obtain the $z$ score for this value, using

$$z_i = \frac{X_i - \mu}{\sigma_{\bar{x}}}$$

- $\bar{x}$ is the sample mean ($23,500$)
- $\mu$ is the mean of means (the university’s claim of $25,000$)
- $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of means
Sampling means example cont.

1. Obtain the z score for this value, using

\[ z_i = \frac{X_i - \mu}{\sigma_{\bar{x}}} \]

- \( \bar{x} \) is the sample mean ($23,500)
- \( \mu \) is the mean of means (the university’s claim of $25,000)
- \( \sigma_{\bar{x}} \) is the standard deviation of the sampling distribution of means

2. Suppose we know that the standard deviation of the sampling procedure is \( \sigma_{\bar{x}} = $700 \). Then we translate to a z score as:

\[ z = \frac{23,500 - 25,000}{700} = -2.14 \]
3. Then we can consider the probability up to this value:

```r
> pnorm(-2.14)
[1] 0.01617738
> round(pnorm(-2.14), 2)
[1] 0.02
```
3. Then we can consider the probability up to this value:

   > pnorm(-2.14)
   [1] 0.01617738
   > round(pnorm(-2.14), 2)
   [1] 0.02

4. What do we conclude then about the original reported value of $25,000 from the university claim?

   Answer: we reject the university’s claim since it was very unlikely to have obtained this sample mean if the true population mean were actually $25,000.
3. Then we can consider the probability up to this value:
   
   > pnorm(-2.14)
   [1] 0.01617738
   > round(pnorm(-2.14), 2)
   [1] 0.02

4. What do we conclude then about the original reported value of $25,000 from the university claim?

   Answer: we reject the university’s claim since it was very unlikely to have obtained this sample mean if the true population mean were actually $25,000
Standard error of the mean

- In practice, we usually know very little about the sampling distribution of the mean, since we usually collect only a single sample

\[
\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}
\]

Example: IQ test is standardized to \( \mu = 100 \) and \( \sigma = 15 \). If we have \( N = 10 \), then

\[
\sigma_{\bar{x}} = \frac{15}{\sqrt{10}} = 15 / 3.1623 = 4.74
\]
Standard error of the mean

- In practice, we usually know very little about the sampling distribution of the mean, since we usually collect only a single sample.
- The result is that we generally will not know $\sigma_{\bar{x}}$.

Example: IQ test is standardized to $\mu = 100$ and $\sigma = 15$. If we have $N = 10$, then $\sigma_{\bar{x}} = \frac{15}{\sqrt{10}} = \frac{15}{3.162} = 4.74$. 
In practice, we usually know very little about the sampling distribution of the mean, since we usually collect only a single sample.

The result is that we generally will not know $\sigma_{\bar{x}}$.

But we can derive the standard deviation of a theoretical sampling distribution (if the means of infinite samples were obtained).

Example: IQ test is standardized to $\mu = 100$ and $\sigma = 15$. If we have $N = 10$, then $\sigma_{\bar{x}} = \frac{15}{\sqrt{10}} = 15 / 3.16 = 4.74$. 
In practice, we usually know very little about the sampling distribution of the mean, since we usually collect only a single sample. The result is that we generally will not know $\sigma_{\bar{x}}$ but we can derive the standard deviation of a theoretical sampling distribution (if the means of infinite samples were obtained). This quantity is called the standard error of the mean:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$
Standard error of the mean

- In practice, we usually know very little about the sampling distribution of the mean, since we usually collect only a single sample.
- The result is that we generally will not know $\sigma_\bar{x}$.
- but we can derive the standard deviation of a theoretical sampling distribution (if the means of infinite samples were obtained).
- This quantity is called the standard error of the mean:

$$
\sigma_\bar{x} = \frac{\sigma}{\sqrt{N}}
$$

- Example: IQ test is standardized to $\mu = 100$ and $\sigma = 15$. If we have $N = 10$, then $\sigma_\bar{x} = 15/\sqrt{10} = 15/3.1623 = 4.74$. 
Confidence intervals

- Using the standard error of the mean, we can determine a range of values within which our population mean is most likely to fall. This is the concept of a confidence interval.
Confidence intervals

- Using the standard error of the mean, we can determine a range of values within which our population mean is most likely to fall. This is the concept of a confidence interval.
- Put another way, we can estimate the probability that our population mean actually falls within a range of values.
Confidence intervals

- Using the standard error of the mean, we can determine a range of values within which our population mean is most likely to fall. This is the concept of a confidence interval.
- Put another way, we can estimate the probability that our population mean actually falls within a range of values.
- If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter.
Confidence intervals

- Using the standard error of the mean, we can determine a range of values within which our population mean is most likely to fall. This is the concept of a confidence interval.
- Put another way, we can estimate the probability that our population mean actually falls within a range of values.
- If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter.
- Confidence intervals are usually calculated so that this percentage is 95%, but can be others (e.g. 99%).
Confidence intervals

- Using the standard error of the mean, we can determine a range of values within which our population mean is most likely to fall. This is the concept of a confidence interval.
- Put another way, we can estimate the probability that our population mean actually falls within a range of values.
- If independent samples are taken repeatedly from the same population, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter.
- Confidence intervals are usually calculated so that this percentage is 95%, but can be others (e.g. 99%).
- Confidence intervals are more informative than the simple results of hypothesis tests (where we decide ”reject the null” or ”don’t reject the null”), since they provide a range of plausible values for the unknown parameter.
Example

- Suppose we want to find the mean left-right position of a member of the 785-seat European Parliament . . .
Example

▶ Suppose we want to find the mean left-right position of a member of the 785-seat European Parliament . . . but because of resource constraints we can only interview 25. So we select 25 at random, and find that the mean (1-100 point scale) is 46

▶ Question: Is the average MEP really left of center?
Example

- Suppose we want to find the mean left-right position of a member of the 785-seat European Parliament . . . but because of resource constraints we can only interview 25. So we select 25 at random, and find that the mean (1-100 point scale) is 46.
- Question: Is the average MEP really left of center?
- How to approach the problem:
Example

▶ Suppose we want to find the mean left-right position of a member of the 785-seat European Parliament . . . but because of resource constraints we can only interview 25. So we select 25 at random, and find that the mean (1-100 point scale) is 46

▶ Question: Is the average MEP really left of center?

▶ How to approach the problem:
  ▶ Let’s use $\bar{x} = 46$ as our estimate of $\mu$
Example

- Suppose we want to find the mean left-right position of a member of the 785-seat European Parliament . . . but because of resource constraints we can only interview 25. So we select 25 at random, and find that the mean (1-100 point scale) is 46.

- Question: Is the average MEP really left of center?

- How to approach the problem:
  - Let’s use $\bar{x} = 46$ as our estimate of $\mu$
  - Likewise, we measure $s_x = 12$, so estimate $\sigma_{\bar{x}} = 12/\sqrt{25} = 2.4$
Example

- Suppose we want to find the mean left-right position of a member of the 785-seat European Parliament . . . but because of resource constraints we can only interview 25. So we select 25 at random, and find that the mean (1-100 point scale) is 46

- Question: Is the average MEP really left of center?

- How to approach the problem:
  - Let’s use \( \bar{x} = 46 \) as our estimate of \( \mu \)
  - Likewise, we measure \( s_x = 12 \), so estimate \( \sigma_{\bar{x}} = 12/\sqrt{25} = 2.4 \)
  - We know that 95% of all values are within \( \pm \) (approximately) two standard deviations of the mean, in a normal distribution
Example

Suppose we want to find the mean left-right position of a member of the 785-seat European Parliament . . . but because of resource constraints we can only interview 25. So we select 25 at random, and find that the mean (1-100 point scale) is 46

Question: Is the average MEP really left of center?

How to approach the problem:

- Let’s use $\bar{x} = 46$ as our estimate of $\mu$
- Likewise, we measure $s_x = 12$, so estimate $\sigma_{\bar{x}} = \frac{12}{\sqrt{25}} = 2.4$
- We know that 95% of all values are within $\pm$ (approximately) two standard deviations of the mean, in a normal distribution
- In other words, the 95% C.I. $= \bar{X} \pm 1.96\sigma_{\bar{X}}$
Example

- Suppose we want to find the mean left-right position of a member of the 785-seat European Parliament ... but because of resource constraints we can only interview 25. So we select 25 at random, and find that the mean (1-100 point scale) is 46.

- Question: Is the average MEP really left of center?

- How to approach the problem:
  - Let’s use $\bar{x} = 46$ as our estimate of $\mu$.
  - Likewise, we measure $s_x = 12$, so estimate $\sigma_{\bar{x}} = 12/\sqrt{25} = 2.4$.
  - We know that 95% of all values are within $\pm$ (approximately) two standard deviations of the mean, in a normal distribution.
  - In other words, the 95% C.I. $= \bar{X} \pm 1.96\sigma_{\bar{X}}$.
  - For this problem, $Cl_{95} = 46 \pm 1.96 \times (2.4) = [41.2, 50.8]$.
Example cont.

- We can therefore conclude that with 95% confidence that the mean for the entire Parliament is 46, give or take 4.8
Example cont.

- We can therefore conclude that with 95% confidence that the mean for the entire Parliament is 46, give or take 4.8
- In this case, we could not conclude with 95% confidence that the average MEP is left of center, since the confidence interval includes the middle (50) value
Example cont.

- We can therefore conclude that with 95% confidence that the mean for the entire Parliament is 46, give or take 4.8.
- In this case, we could not conclude with 95% confidence that the average MEP is left of center, since the confidence interval includes the middle (50) value.
- The specified probability level (95%) is known as the level of confidence.
Example cont.

- We can therefore conclude that with 95% confidence that the mean for the entire Parliament is 46, give or take 4.8.
- In this case, we could not conclude with 95% confidence that the average MEP is left of center, since the confidence interval includes the middle (50) value.
- The specified probability level (95%) is known as the level of confidence.
- Our guesses will be wrong 100% – 95% = 5% of the time.
We can therefore conclude that with 95% confidence that the mean for the entire Parliament is 46, give or take 4.8.

In this case, we could not conclude with 95% confidence that the average MEP is left of center, since the confidence interval includes the middle (50) value.

The specified probability level (95%) is known as the level of confidence.

Our guesses will be wrong 100% − 95% = 5% of the time.

The precision of our estimate will be determined by the margin of error, which we obtain by multiplying our the standard error by the z score representing a desired level of confidence (e.g. $1.96 \times 2.4 = 4.7$ in our earlier example).
Example cont.

► We can therefore conclude that with 95% confidence that the mean for the entire Parliament is 46, give or take 4.8
► In this case, we could not conclude with 95% confidence that the average MEP is left of center, since the confidence interval includes the middle (50) value
► The specified probability level (95%) is known as the level of confidence
► Our guesses will be wrong $100\% - 95\% = 5\%$ of the time
► The precision of our estimate will be determined by the margin of error, which we obtain by multiplying our the standard error by the $z$ score representing a desired level of confidence (e.g. $1.96 \times 2.4 = 4.7$ in our earlier example
  ► for 68%, $z = \pm 1.00$
  ► for 95%, $z = \pm 1.96$
  ► for 99%, $z = \pm 2.58$
In most applications we never know $\sigma$ to use in the formula

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$
The *t* distribution

- In most applications we never know $\sigma$ to use in the formula
  \[ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \]
- Problem: If we simply substitute sample standard deviation $s$
  for $\sigma$ then we will underestimate $\sigma$
The $t$ distribution

- In most applications we never know $\sigma$ to use in the formula
  \[ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \]
- Problem: If we simply substitute sample standard deviation $s$ for $\sigma$ then we will underestimate $\sigma$
- As a result, we compensate by using $N - 1$ instead of $N$ as the denominator:

\[ \hat{\sigma} = \sqrt{\frac{\sum_i(X_i - \bar{X})^2}{N - 1}} \]
The $t$ distribution

- In most applications we never know $\sigma$ to use in the formula
  \[ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \]
- Problem: If we simply substitute sample standard deviation $s$ for $\sigma$ then we will underestimate $\sigma$
- As a result, we compensate by using $N-1$ instead of $N$ as the denominator:
  \[ \hat{\sigma} = \sqrt{\frac{\sum_i(X_i - \bar{X})^2}{N-1}} \]
- To get an unbiased estimate of the standard error of the mean from a sample, we use this formula instead:
  \[ s_{\bar{X}} = \frac{s}{\sqrt{N-1}} \]
The $t$ distribution cont.

- One result of this is that the sampling distribution of means is no longer perfectly normal – in other words $\frac{\bar{X} - \mu}{s_{\bar{X}}}$ does not quite follow the $z$ distribution.
- Instead, we call this the $t$ distribution, which is like a normal distribution but slightly wider (or fatter at the tails).
- The ratio we use instead of the $z$ ratio is called the $t$ ratio:

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

- Each $t$ distribution's exact shape will also depend on the degrees of freedom $df = N - 1$. The greater the $df$, the closer $t$ gets to the normal distribution.
Probabilities and the $t$ distribution

- Tables for area under the $t$ distribution are represented by both degrees of freedom and different levels of $\alpha$, which represent the level of confidence:

\[ \alpha = 1 - \text{level of confidence} \]

- Confidence intervals for the $t$ distribution are constructed using $t$ scores for given degrees of freedom and $\alpha$ levels:

\[ \text{confidence interval} = \bar{X} \pm ts_{\bar{X}} \]
Confidence intervals for proportions

- Often we may seek to estimate a population proportion on the basis of a random sample: e.g. proportion to vote Yes on the Lisbon Treaty referendum
Confidence intervals for proportions

Often we may seek to estimate a population *proportion* on the basis of a random sample: e.g. proportion to vote Yes on the Lisbon Treaty referendum.

We use a version of the standard error known as the standard error of the proportion:

\[
s_p = \sqrt{\frac{p(1 - p)}{N}}
\]

where

- \( s_p \) is the standard error of the proportion
- \( p \) is the sample proportion
- \( N \) is the total number of cases in the sample
Confidence intervals for proportions

- Often we may seek to estimate a population proportion on the basis of a random sample: e.g. proportion to vote Yes on the Lisbon Treaty referendum
- We use a version of the standard error known as the standard error of the proportion:

\[
s_p = \sqrt{\frac{p(1-p)}{N}}
\]

where
- \(s_p\) is the standard error of the proportion
- \(p\) is the sample proportion
- \(N\) is the total number of cases in the sample
- Because proportions have only a single parameter \(\pi\) of which \(p\) is an estimate, we can use the normal distribution and not \(t\)
Confidence intervals for proportions

- Often we may seek to estimate a population proportion on the basis of a random sample: e.g. proportion to vote Yes on the Lisbon Treaty referendum

- We use a version of the standard error known as the standard error of the proportion:

\[
s_p = \sqrt{\frac{p(1 - p)}{N}}
\]

where

- \(s_p\) is the standard error of the proportion
- \(p\) is the sample proportion
- \(N\) is the total number of cases in the sample

- Because proportions have only a single parameter \(\pi\) of which \(p\) is an estimate, we can use the normal distribution and not \(t\)

- So: 95\% CI for proportions = \(p \pm 1.96s_p\)
Simulation and bootstrapping

Used for:

▶ Gaining intuition about distributions and sampling

Both simulation and bootstrapping are numerical approximations of the quantities we are interested in. (Run the same code twice, and you get different answers)

We have already seen simulation in the illustrations of the Central Limit Theorem, in applications to estimating the mean of spending from sample means.
Simulation and bootstrapping

Used for:

- Gaining intuition about distributions and sampling
- Providing distributional information not distributions are not directly known, or cannot be assumed
Simulation and bootstrapping

Used for:

▶ Gaining intuition about distributions and sampling
▶ Providing distributional information not distributions are not directly known, or cannot be assumed
▶ Acquiring uncertainty estimates
Simulation and bootstrapping

Used for:

▶ Gaining intuition about distributions and sampling
▶ Providing distributional information not distributions are not directly known, or cannot be assumed
▶ Acquiring uncertainty estimates

Both simulation and bootstrapping are numerical approximations of the quantities we are interested in. (Run the same code twice, and you get different answers)

We have already seen simulation in the illustrations of the Central Limit Theorem, in applications to estimating the mean of spending from sample means.
Bootstrapping refers to repeated resampling of data points with replacement. Used to estimate the error variance (i.e. the standard error) of an estimate when the sampling distribution is unknown (or cannot be safely assumed). Robust in the absence of parametric assumptions. Useful for some quantities for which there is no known sampling distribution, such as computing the standard error of a median.
Bootstrapping illustrated

> ## illustrate bootstrap sampling
> # using sample to generate a permutation of the sequence 1:10
> sample(10)
> [1] 6 1 2 4 5 7 9 3 10 8
> # bootstrap sample from the same sequence
> sample(10, replace=T)
> [1] 3 3 10 7 5 3 9 8 7 6
> # bootstrap sample from the same sequence with probabilities that
> # favor the numbers 1-5
> # favor the numbers 1-5
> prob1 <- c(rep(.15, 5), rep(.05, 5))
> prob1
> [1] 0.15 0.15 0.15 0.15 0.15 0.05 0.05 0.05 0.05 0.05
> sample(10, replace=T, prob=prob1)
> [1] 10 4 7 6 5 2 9 5 1 5
Bootstrapping the standard error of the median

Using loops:

```r
bs <- NULL
for (i in 1:100) {
  bs[i] <- median(sample(spending, replace=TRUE))
}
quantile(bs, c(.025, .5, .975))
median(spending)
```
Bootstrapping the standard error of the median

Using `lapply` and `sapply`:

```r
resamples <- lapply(1:100, function(i) sample(spending, replace=TRUE))
bs <- sapply(resamples, median)
quantile(bs, c(.025, .5, .975))
```
Using a user-defined function:

```r
b.median <- function(data, n) {
  resamples <- lapply(1:n, function(i) sample(data, replace=T))
  sapply(resamples, median)
  std.err <- sqrt(var(r.median))
  list(std.err=std.err, resamples=resamples, medians=r.median)
}
summary(b.median(spending, 10))
summary(b.median(spending, 100))
summary(b.median(spending, 400))
median(spending)
```
Bootstrapping the standard error of the median

Using R’s `boot` library:

```r
library(boot)
samplemedian <- function(x, d) return(median(x[d]))
quantile(boot(spending, samplemedian, R=10)$t, c(.025, .5, .975))
quantile(boot(spending, samplemedian, R=100)$t, c(.025, .5, .975))
quantile(boot(spending, samplemedian, R=400)$t, c(.025, .5, .975))
```

**Note:** There is a good reference on using `boot()` from [http://www.mayin.org/ajayshah/KB/R/documents/boot.html](http://www.mayin.org/ajayshah/KB/R/documents/boot.html)