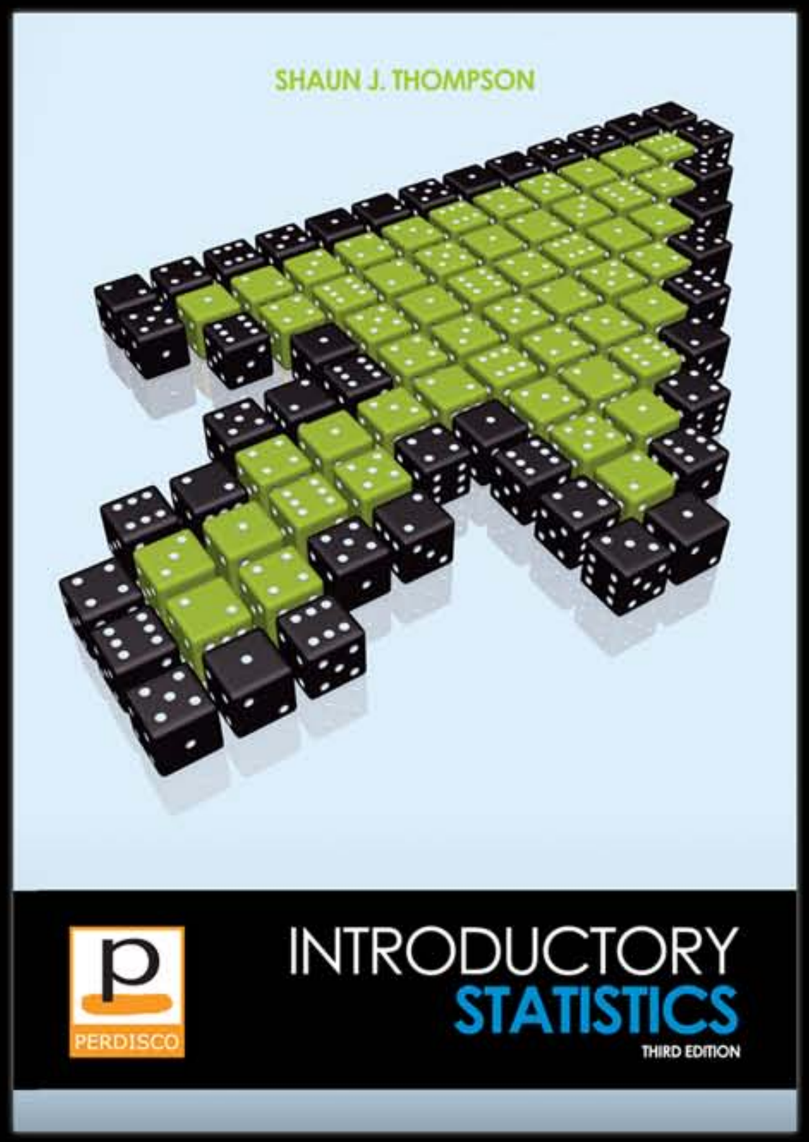


These lecture slides are designed to accompany:

Introductory Statistics, Third Edition

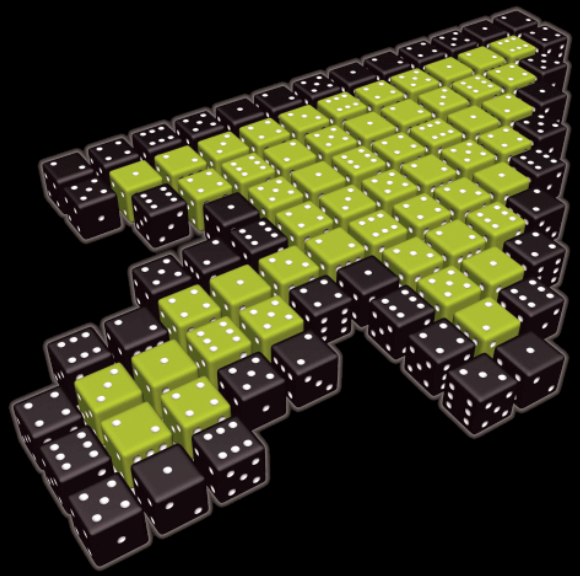
Other features include:

- Chapter summary videos
- MP3 audio podcasts
- VirtualTutor e-learning
- AntiCheat and AutoGrade homework
- Detailed instructor resources

To find out more, visit: perdisco.com/stats

Chapter 7

Estimation



Estimation as inference

- Statistical **estimation** is one of the major fields of statistical inference
- Remember that inference is where we use what we know about a sample to try and conclude something about the entire population
- In estimation, this means that we:
  1. Collect a sample
  2. Calculate a value for the sample statistic
  3. Construct a range of values that may contain the population parameter

Example

- A survey of 100 people finds that they donate an average of $57 a year to various charities
- This is a sample mean, and is used to estimate the true population mean, μ
- Such an estimate might look something like:

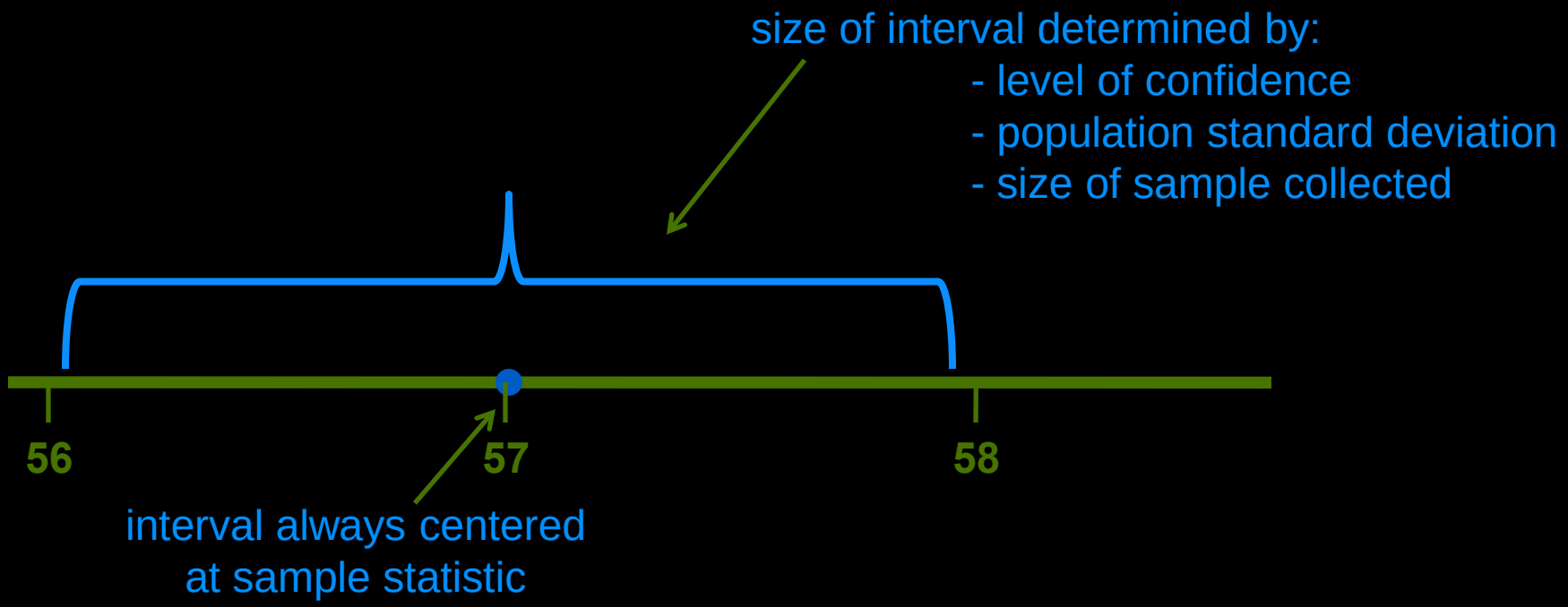
"We are 95% confident that the average amount that people donate each year is somewhere between $56.02 and $57.98"

Imprecision and uncertainty

- The range of values from $56.02 and $57.98 is known as a **confidence interval** for the mean
- Any such interval will always involve imprecision and uncertainty
- Imprecision:
  - An entire range of values is provided, instead of a single value
- Uncertainty:
  - You can't be certain that the range definitely contains the population mean

The nature of a confidence interval

- A confidence interval for a population mean (for example) will look something like this:



The size of a confidence interval

- The size of the interval reflects the imprecision in the interval – the bigger the interval, the less precise
- Three things affect the size of a confidence interval
- Level of confidence – the more confident you want to be, the wider your interval will have to be
- Standard deviation of population – the more values vary, the more sample statistics will vary, so bigger standard deviations mean less accurate estimates
- Sample size – larger samples behave ‘nicer’ and so tend to produce more accurate sample statistics

Constructing an interval – basic approach

- Earlier, we saw an example of an interval estimate:

"We are 95% confident that the average amount that people donate each year is somewhere between $56.02 and $57.98"

- But how did they arrive at this estimate?
  - You start by choosing how confident you want to be
  - Then collect a sample to calculate a sample statistic
  - Then use the sampling distribution to estimate parameter

Constructing an interval – basic approach

- In the donation survey, there was 95% confidence
- For the moment, let's assume we know $\sigma = \$5$
- We want to estimate the mean μ of all donations X
- A survey of $n = 100$ gave a sample mean of \$57

- Important: this sample mean came from the sampling distribution of the mean, $\bar{X}$
- What does this sampling distribution look like?

Constructing an interval – basic approach

- Well it has mean μ, standard deviation $\sigma/\sqrt{n} = 0.5$
- So given that μ is unknown, we don't exactly know $\overline{X}$
- But we know the sample mean of \$57 is probably not too far from the middle!
- In particular, because $\overline{X}$ is normal, we can say that 95% of the values in $\overline{X}$ will be less than 1.96 standard deviations from the mean
- That is, 95% of all sample means will be within $1.96 \times 0.5 = 0.98$ of μ

Constructing an interval – basic approach

- So let's assume that $\bar{x}$ = $57 is one of these!
- That is, we are asserting that the sample mean of $57 is less than 0.98 from the true population mean
- This is where the estimate comes from:

"We are 95% confident that the average amount that people donate each year is somewhere between $56.02 and $57.98"

- And that's why we're 95% 'confident' in the estimate
  - For 95% of samples we could get, we would 'capture' μ

Other levels of confidence

- What about other levels of confidence?
- We made it easy for ourselves by choosing 95% because a common property (which we used) is:

For any normal distribution, 95% of all values lie within 1.96 standard deviations of the mean

- This was crucial to calculating the estimate
- So what about 90%? Within how many standard deviations does 90% of a normal distribution lie?
- And what about all the other levels of confidence?

Using the standard normal distribution

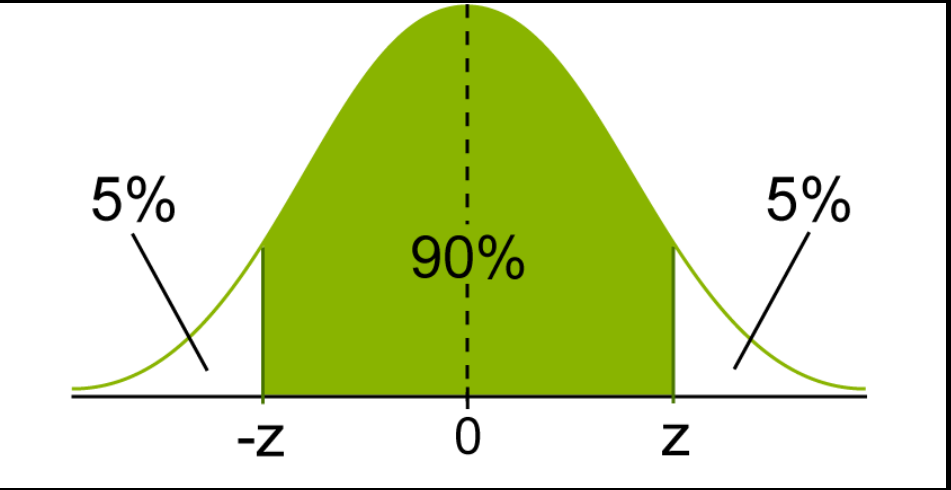
- $\overline{X}$ approximately follows the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$
- We'll relate this to Z via the transformation formula:

$$\overline{X} = \mu + Z\frac{\sigma}{\sqrt{n}}$$

- In other words, a value in $\overline{X}$ is within $\pm z$ standard deviations of the mean when a value in Z is within $\pm z$ of 0

Using the standard normal distribution

- So for what value z does 90% of Z fall between ± z?



- We can find this using the standard normal table!
- In fact, 90% of Z falls between 1.645 and -1.645
- So 90% of all sample means will fall within 1.645 standard deviations of the population mean!

Using the standard normal distribution

- And now we can use this to construct the estimate
- 90% of all sample means will fall within 1.645 standard deviations of μ
- For the donation survey, this means that 90% of all sample means fall within 1.645 x 0.5 = 0.823 of μ
- Let's assume that sample mean $57 is one of these!

"We are 90% confident that the average amount that people donate each year is somewhere between $56.18 and $57.82"

Different confidence intervals

- We've now constructed two different confidence intervals: a 95% interval and a 90% interval
- We have less confidence in the 90% interval, but it is more precise

- Using Z and the transformation formula, we can construct confidence intervals at any level of confidence that we want

Level of significance

- For a general level of confidence level, C%, we often convert this to a **level of significance**
- If we express C% as follows:

$$C\% = 100\% \times (1 - \alpha)$$

then α is known as the level of significance

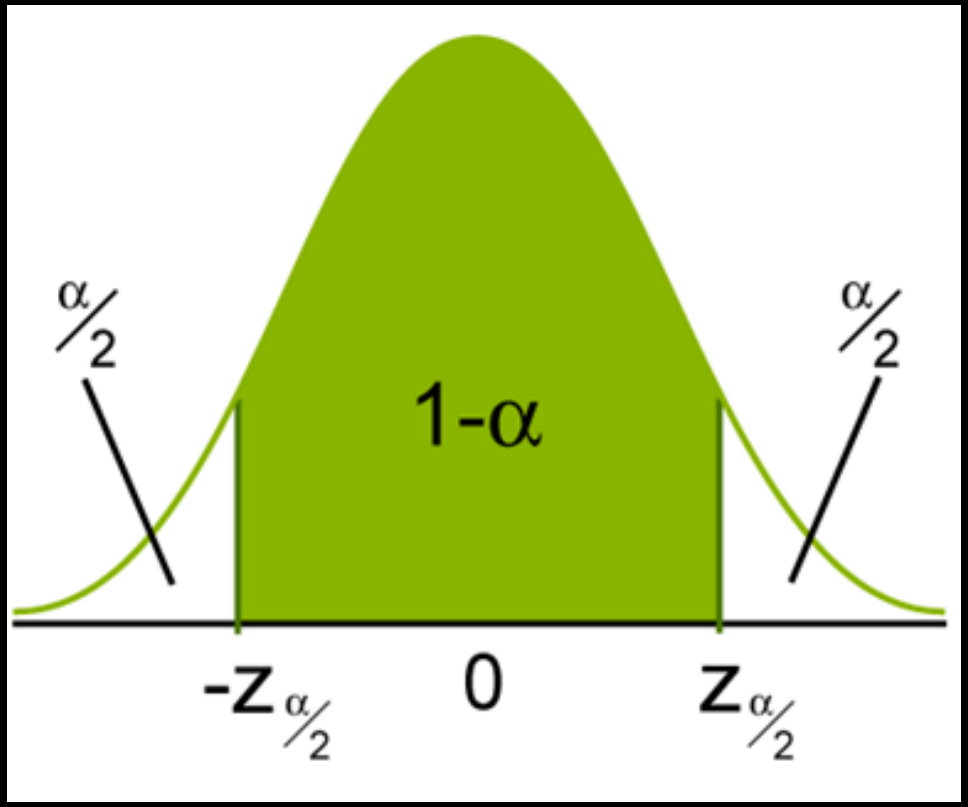
- Example: For a 95% confidence level, the level of significance is $\alpha = 0.05$

Critical values

- For a level of significance α, there are two **critical values**, $z_{\alpha/2}$ and $-z_{\alpha/2}$
- These values are defined such that as a proportion:
  - $\alpha/2$ of the standard normal distribution Z falls above $z_{\alpha/2}$
  - $\alpha/2$ of Z falls below $-z_{\alpha/2}$
  - so the chance that Z falls between $z_{\alpha/2}$ and $-z_{\alpha/2}$ is $1 - \alpha$
- Examples
  - For $\alpha = 0.05$, critical values are 1.96 and -1.96
  - For a 90% confidence level, $\alpha = 0.1$ and the critical values are 1.645 and -1.645

Critical values

- The critical values are the z-score boundary points that you use in your confidence interval



Critical values and the confidence interval

- We can now apply this to the estimate for μ
- For a general level of confidence C%
  - we write this as $C\% = 100 \times (1 - \alpha)\%$
  - the two critical values contain, as a percentage, C% of Z
- In other words, in the sampling distribution of the mean, C% of all sample means will lie within $z_{\alpha/2}$ standard deviations ($\sigma/\sqrt{n}$) of the mean
- So for C% of all sample means $\bar{x}$, it is true to say that μ will fall within $z_{\alpha/2}(\sigma/\sqrt{n})$ of $\bar{x}$

The confidence interval for the mean

- The 100% x (1 - α) confidence interval for μ is:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

lower bound of interval

upper bound of interval

Margin of error

- The interval's size is fixed by the value $z_{\alpha/2}(\sigma/\sqrt{n})$
- This is known as the margin of error for the estimate

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- So a simplified version of the confidence interval is:

$$\bar{x} - E \leq \mu \leq \bar{x} + E$$

Factors in the margin of error

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Three things affect the margin of error:
  - level of confidence – the larger the confidence, the larger the critical values, and the larger the margin of error
  - population standard deviation – a larger σ will mean a larger margin of error
  - sample size – a larger n means a smaller margin of error
- This is why, as discussed before, these three things affect the size of a confidence interval

Steps to constructing the interval for μ

- **Step 1**: Acknowledge the value you are trying to estimate, μ, and assume you know σ
- **Step 2**: Decide a level of confidence, C%, which determines a level of significance α
- **Step 3**: Collect a sample and calculate $\overline{x}$
- **Step 4**: Calculate the critical values $z_{\alpha/2}$ and $-z_{\alpha/2}$
- **Step 5**: Calculate the margin of error $z_{\alpha/2}(\sigma/\sqrt{n})$
- **Step 6**: Calculate the upper and lower bounds of the confidence interval

Determining the sample size

- When constructing a confidence interval, the larger the sample size, the more precise the estimate
- The margin of error formula can be rearranged to make the sample size the subject of the equation:

$$n = z_{\alpha/2}^{2} \times \frac{\sigma^{2}}{E^{2}}$$

- Using this, you can choose a desired margin of error E and level of significance α. For a given assumed value for σ, you can determine a suitable sample size n.

Example

- In the donations study, we assumed σ was \$5
- Suppose we want a 95% confidence interval
  - this means $\alpha = 0.05$ and gives a critical value $z_{\alpha/2} = 1.96$
- We want interval width to be no more than $E = 0.20$

$$n = z_{\alpha/2}^2 \times \frac{\sigma^2}{E^2} = 1.96^2 \times \frac{25}{0.04} = 2{,}401$$

- So a sample size of at least 2,401 is needed for this level of confidence and margin of error

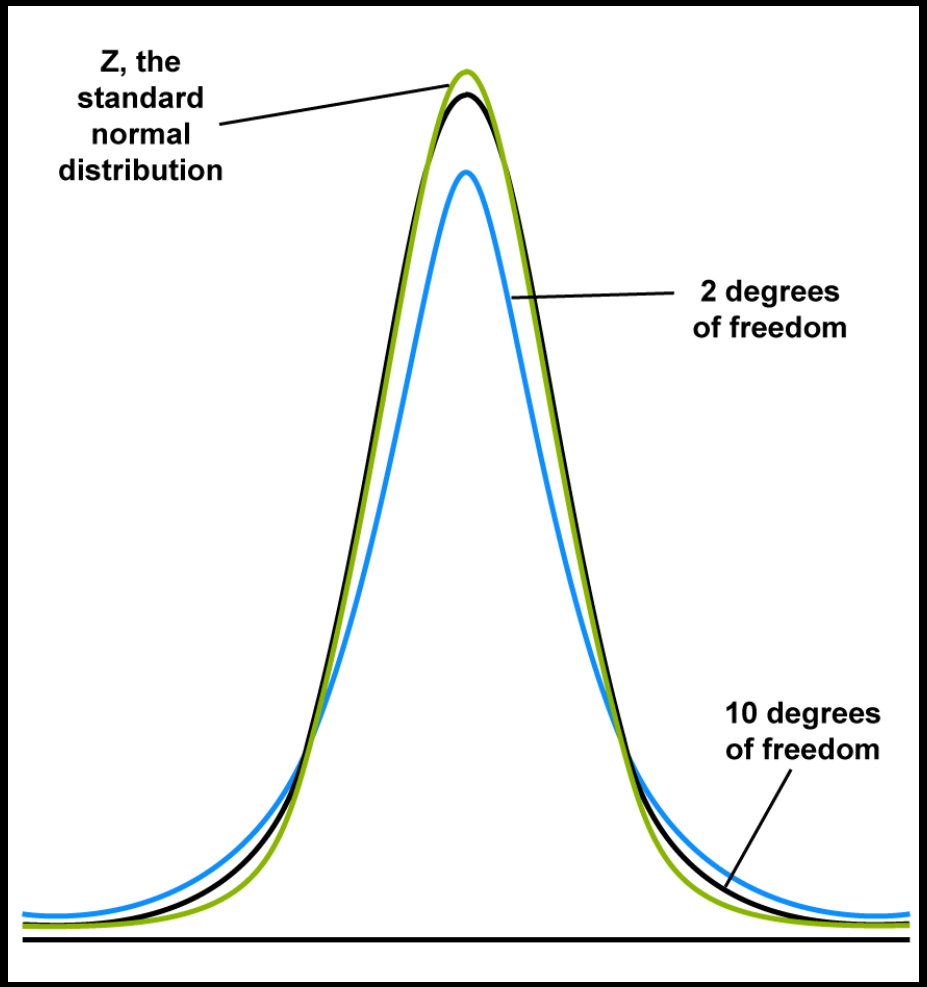
Confidence interval when σ is unknown

- So far we've been assuming that we know σ
- For example, we assumed that the standard deviation in donations was $\sigma = 5$
- It is not very realistic to assume we know this value
- What to do?
- We use the sample standard deviation, s, as an estimate for the population standard deviation, σ
- However, in this case the critical values will no longer come from Z!

t-distributions

- Instead, they come from **t-distributions**
- A family of distributions developed for this purpose
- Each t-distribution is characterized by a parameter, n, known as the **degrees of freedom**
- So which t-distribution do you use?
- If your sample has n items, to estimate μ you use the t-distribution with $(n - 1)$ degrees of freedom
- As n increases, the distributions look more and more like the standard normal distribution, Z

Some t-distributions



Critical values

- Like with Z, critical values of the t-distributions are found by referring to a table of critical values

- Example
  - Suppose a 95% confidence interval is being constructed, and a sample of $n = 25$ items is collected
  - A level of significance of $\alpha = 0.05$ is chosen and so there are $n - 1 = 24$ degrees of freedom
  - So critical values are $t_{\alpha/2} = t_{0.025}$ and $-t_{\alpha/2} = -t_{0.025}$

Table of critical values

- So how do we find $t_{0.025}$?
- We refer to the table of critical values for t
- Use the row with 24 degrees of freedom and the column labeled $t_{0.025}$
- So the critical values are 2.064 and -2.064

Degrees of freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
⋮	⋮	⋮	⋮	⋮	⋮
23	1.320	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
⋮	⋮	⋮	⋮	⋮	⋮

Formula for confidence interval

- To estimate μ without knowing σ:
  - Collect a sample of n items, calculate $\bar{x}$ and s
  - Find critical values $t_{\alpha/2}$ and $-t_{\alpha/2}$
- Then the 100% x (1 - α) confidence interval for μ is:

$$\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Categorical variables and proportions

- So far we've been looking at numerical variables
- What about categorical variables?
- Recall that if we have a categorical variable X, a population parameter that occurs is π, the population proportion for a category in X
- Example:
  - What proportion π of people choose chocolate as their favorite ice cream out of chocolate, strawberry & vanilla?
- Just like with μ, we can estimate π with a sample!

Example

- Suppose we want to construct a 95% confidence interval for π, the proportion that choose chocolate
- What can we do?
- We could survey $n = 100$ people about it!
- Suppose 32 chose chocolate as their favorite
- So the sample proportion is $p = 0.32$
- This value will be in the center of the confidence interval for π

Example continued

- But how do we construct the confidence interval?
- Very similarly to before
- The sampling distribution of the proportion, P, approximately follows the normal distribution with:

$$\mu = \pi$$

$$\sigma = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- So: for 95% of all samples collected, the sample proportion p will be within 1.96 standard deviations of the mean π

Example continued

- We assume that our sample proportion p = 0.32 is one of the 95% within 1.96 standard deviations of π
- In other words, we assert that π is within 1.96 standard deviations of 0.32
- Since the formula for the standard deviation involves π, we use p as an estimate in the formula:

$$\sigma = \sqrt{\frac{\pi(1-\pi)}{n}}$$

$$\approx \sqrt{\frac{p(1-p)}{n}}$$

$$= \sqrt{\frac{0.32 \times 0.68}{100}}$$

$$= 0.0466$$

Example continued

- So our confidence interval is formed by asserting that π is within $1.96 \times 0.0466 = 0.0913$ of $p = 0.32$
- That is, that π is between 0.2287 and 0.4113
- This is our confidence interval!

"We are 95% confident that somewhere between 22.87% and 41.13% of people has chocolate as their favorite flavor"

Level of significance and critical values

- Just like with the mean, we can extend this method to other levels of confidence, not just 95%
- To do this, we again use the level of significance and critical values

- If we are constructing a $C\% = 100\% \times (1 - \alpha)$ confidence interval, then:
  - α is the level of significance
  - $z_{\alpha/2}$ and $-z_{\alpha/2}$ are the two critical values

Confidence interval for π

- The 100% x (1 - α) confidence interval for π is:

$$p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

lower bound of interval

upper bound of interval