Classical Inference for Means

Ryan Miller

Central Limit Theorem

- Lately, we've used results from the *Central Limit Theorem* to construct normal approximations of the sampling/null distribution of *one proportion*, or a *difference in proportions*
- CLT also can be used to construct a normal approximation of these distributions for **one mean**, or a **difference in means**
- For a single mean, CLT suggests the following normal approximation:

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- Like before, we need to replace the known population parameters μ and σ with suitable values

Note: When we first introduced the normal curve, we described its shape using parameters μ and σ, this is different from how the symbols are used here, where μ and σ refer to the *population's* mean and standard deviation

Example

- Radon gas is toxic to the human body, and is the second leading cause of lung cancer in the US
- Radon is prevalent in the Midwest and a major public health concern in many states including Iowa
- The EPA has set federal action limit of 4 pCi/L, but if the average levels in your residence are higher than 0.4 pCi/L they recommend you take measures to reduce your exposure
- Suppose you pay to have the radon level in your home tested on eight randomly selected occasions over the course of a month, the measurements are:

$$\{0.2, 0.7, 0.3, 0.9, 0.5, 0.3, 0.7, 0.6\} \text{ pCi/L}$$

  - Should you take action immediately or could the average measurement being above 0.4 pCi/L be due to random chance?
  - Also, describe a Type I and Type II error for this test? Is one error more costly than the other?

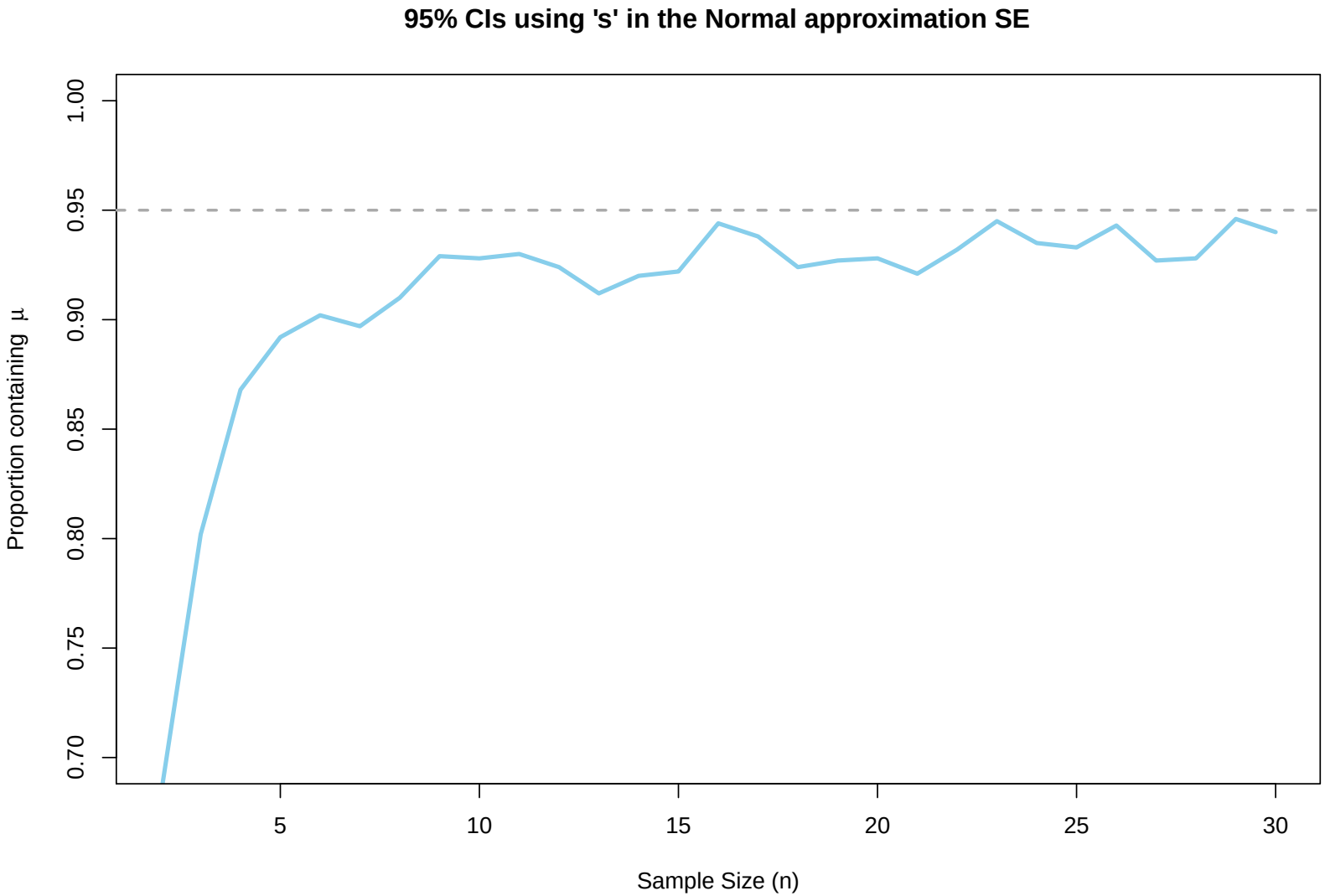
Example (Solution)

- $H_0 : \mu = 0.4$ vs. $H_A : \mu > 0.4$
- We observed a sample mean of $\bar{x} = 0.525$, and a sample standard deviation of $s = 0.243$ from our sample of size $n = 8$
- We estimate the standard error is $SE = 0.275/\sqrt{8} = 0.086$
- This suggests approximating the Null Distribution with $N(0.4, 0.086)$
- Locating 0.525 on this approximation we get a one-sided p-value of approximately 0.07
- A Type I error is home having safe radon levels but the test suggesting action should be taken
- A type II error is the home having radon levels above 0.4, but the test suggesting the home is safe

Is the Normal Approximation Flawed for Means?

- We've previously seen that the normal approximation suggested by CLT can be inaccurate when n is small
- Surprisingly, the approach is still problematic even when the sampling distribution is *perfectly normal*
- The flaw can be shown rather easily with a simple simulation:
  - Repeatedly draw random samples from a normal distribution
  - From each, construct a 95% confidence interval using the normal approximation suggested by CLT
  - Track how often these intervals contain the true mean

Simulation of the Normal Approximation



William Gosset (1876 - 1937)

- William Gosset was an English chemist who worked for Guinness Brewing in the 1890s
  - Gosset's role at Guinness was to statistically evaluate the yield of different varieties of barley
  - These experiences at Guinness prompted Gosset to investigate the statistical validity of small sample z-tests
  - In 1906, Gosset took a leave of absence from the brewery to work on the problem with Karl Pearson, shortly after he mathematically derived a modified distribution that fixed the flaw
  - His finding, the t-distribution, was published under the name "Student" because Guinness didn't want its competitors knowing that they employed statisticians!

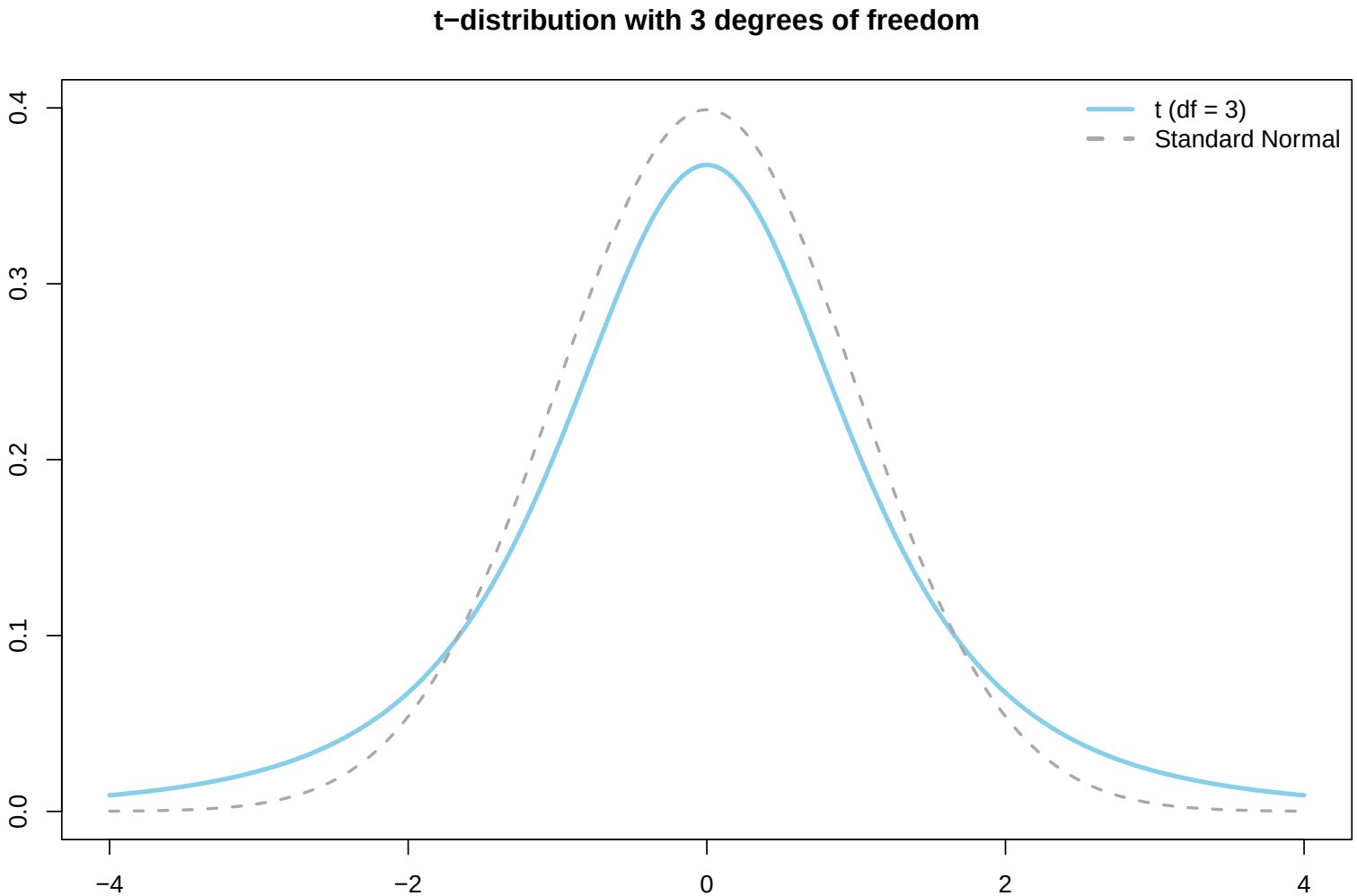
The t-distribution

- Using s as an estimate of σ produces flawed results when n is small
  - The reason for this is that s has its own variability, so treating it as a known constant in the normal approximation is overly optimistic
- Prior to modern computing, it wasn't so easy to discover this flaw, though Gosset claimed that Ronald Fisher (developer of the p-value) would have discovered if he hadn't

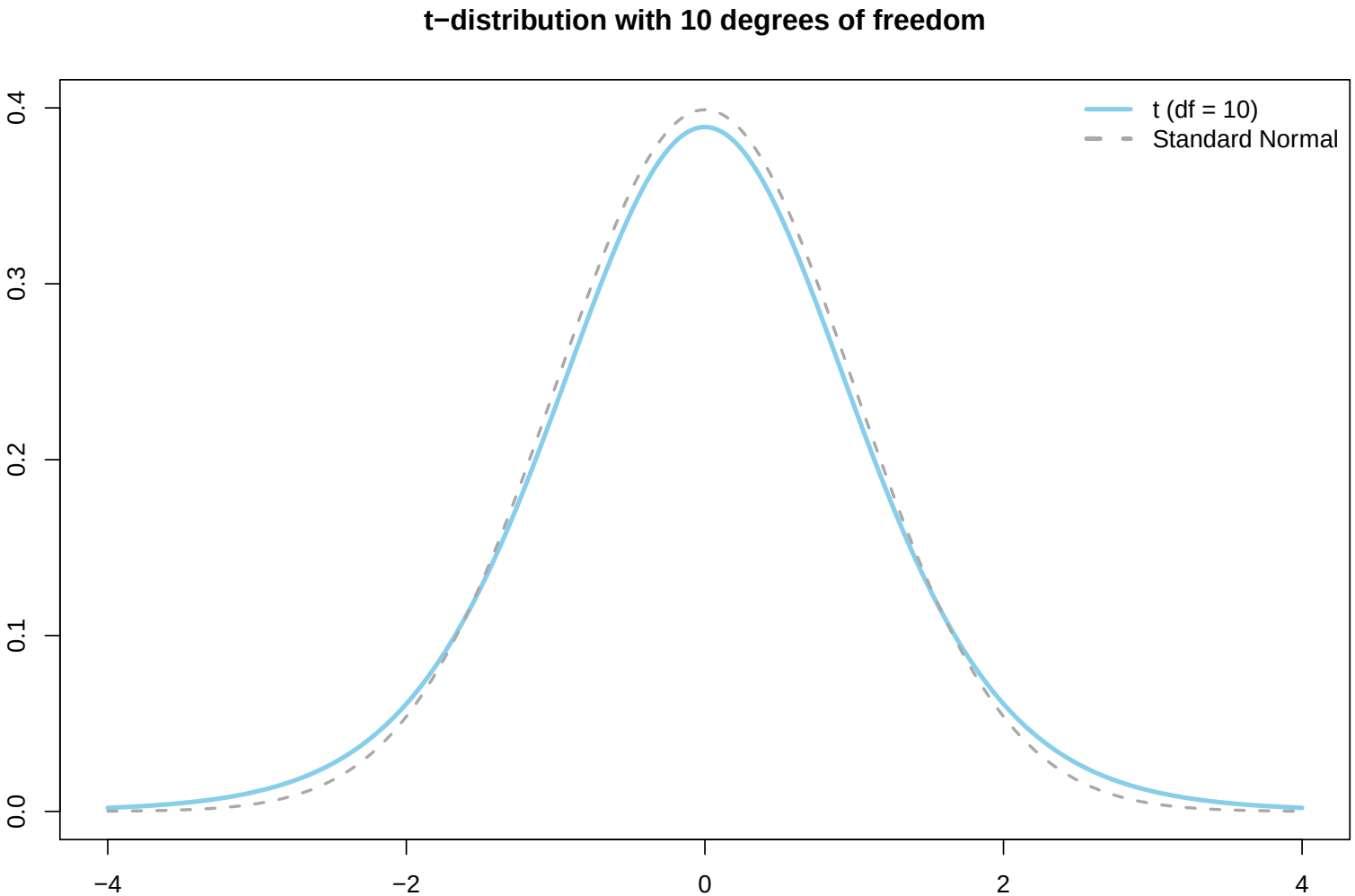
The t-distribution

- Unlike the normal distribution, the shape of the t-distribution depends upon the sample size through a parameter named **degrees of freedom** (often abbreviated as df)
  - In this context, "degrees of freedom" refers to the amount of information available for estimating the standard deviation, because the sum of the deviations, $\sum_{i=1}^{n}(x_i - \bar{x})$, must add up to zero, not all n elements can vary freely
- Thus, when applying the t-distribution to the mean of a single quantitative variable, $df = n - 1$
- The t-distribution requires the population be normally distribution
  - Generally normality is difficult to judge from a small sample, so we tend not to worry unless we observe clear outliers or substantial skew

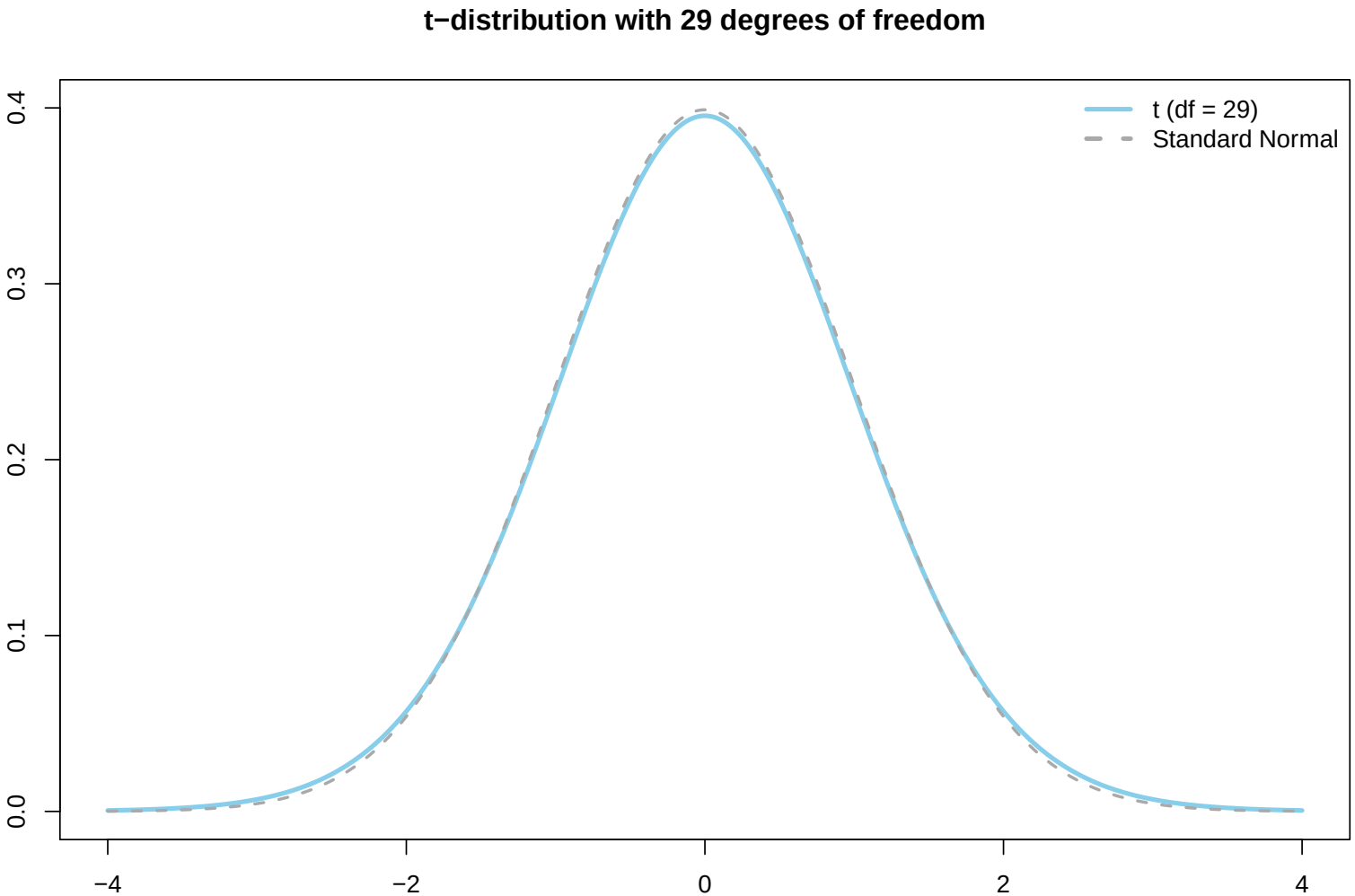
The t-distribution



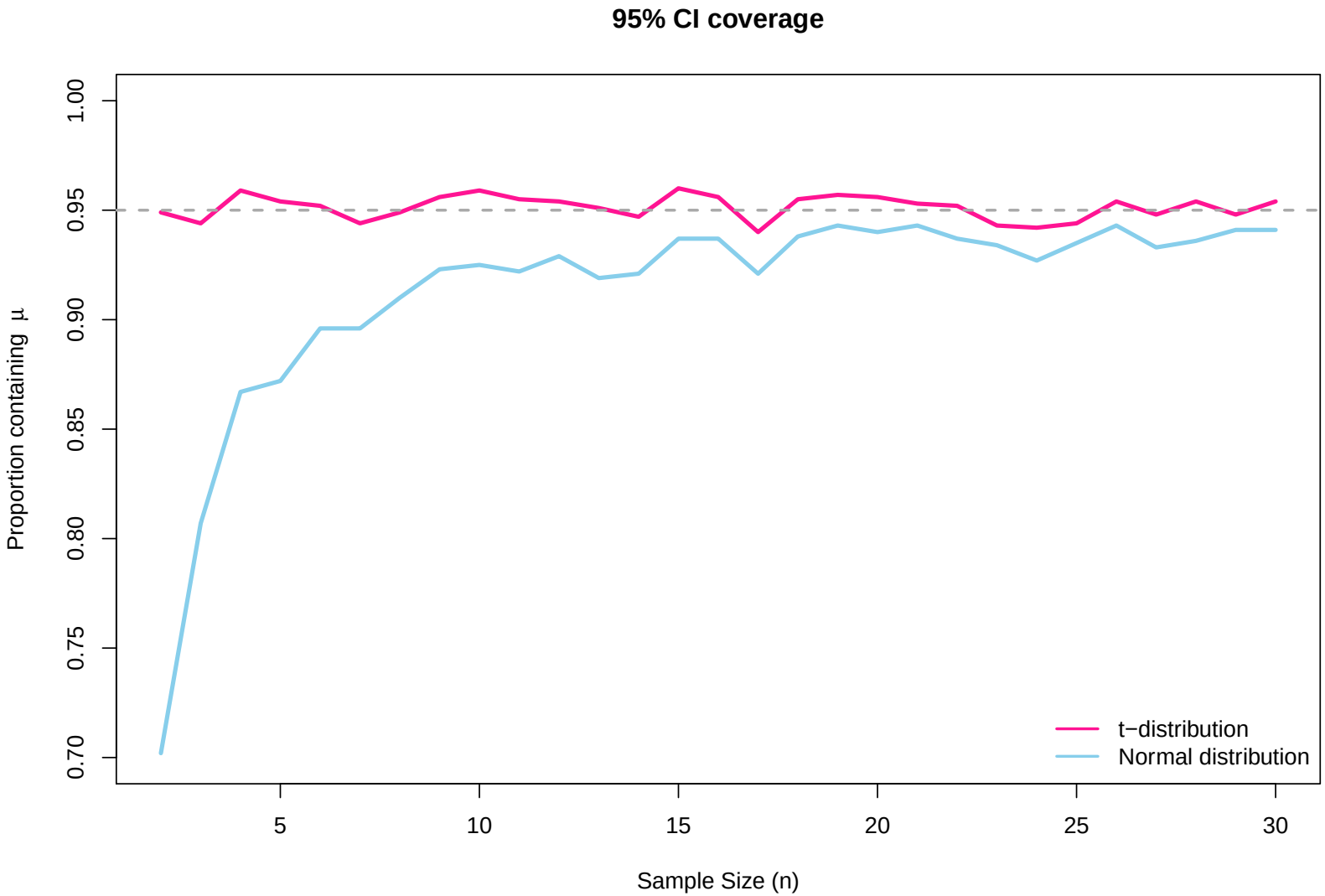
The t-distribution



The t-distribution



The t-distribution



The t-distribution

- The t-distribution has thicker tails than the normal curve, which accurately account for the uncertainty introduced by estimating σ with s
  - The difference diminishes as n increases
  - At $n = 30$, the two distributions are nearly indistinguishable
- We can calculate the area under the t-distribution, or find the percentiles needed for confidence intervals, using Minitab or StatKey

Inference for a Mean - Example #1

Using the StatKey dataset "Arsenic in Chicken", which can be found in "randomization testing for a single mean": (Note: you don't need to perform any randomization)

1. Find a 95% confidence interval estimate for the mean amount of arsenic using the t-distribution
2. Find a 95% confidence interval using the *normal approximation* instead of the t-distribution
3. Test whether the population mean differs from 80 using the t-distribution, report your p-value and conclusion
4. Suppose you had used a z-test, how would your p-value differ?

Inference for a Mean - Example #1 (solution)

1. $91 \pm 2.571(\frac{23.47}{\sqrt{6}}) = (66.37, 115.63)$
2. $91 \pm 1.96(\frac{23.47}{\sqrt{6}}) = (72.22, 109.78)$
3. $H_0 : \mu = 80$ versus $H_A : \mu \neq 80$

$$t_{test} = \frac{91 - 80}{23.47/\sqrt{6}} = 1.15$$

Using the reference distribution: $t(df = 5)$, the two-sided p-value is 0.302. We cannot reject the null hypothesis, there is insufficient evidence to claim the mean arsenic level is differs from 80 ppm

4. The test statistic is still 1.15, this leads to a p-value of 0.250 using the standard normal distribution

Inference for a Mean - Example #1 (solution)

We also could have performed this test in Minitab with the following steps:

1. Enter or copy/paste our one-sample quantitative data into a column
2. Navigate: “Stat” -> “Basic Statistics” -> “One-sample t-test”
3. Choose our variable and enter the null value

Inference for a Mean - Example #1 (lessons)

- When the sample size is small (ie: $n = 6$), it is important to account for the uncertainty in estimating σ
- The results are very different when using the $t(df = 5)$ distribution instead of the normal distribution

Inference for a Mean - Example #2

Use the StatKey dataset "Home Prices - Canton" to answer the following questions:

1. Does it appear likely that these data come from a normally distributed population? Create a randomization distribution (using $\mu_0 = 200$), does it appear skewed?
2. Conduct two one-sided hypothesis tests to determine if the mean price of homes in Canton *is less than 200k* ($H_0 : \mu \geq 200$) at the $\alpha = 0.05$ level, one using the t-distribution and another using a randomization test. How do your results compare?

Inference for a Mean - Example #2 (solution)

1. The population seems to be right skewed due to a couple of larger values appearing in the sample. The randomization distribution is right skewed
2. The randomization test yields a left-tail p-value of 0.019. The t-distribution yields a left-tail p-value of 0.055 ($t_{test} = \frac{146.8-200}{94.998/\sqrt{10}} = -1.77$). These two results are different because the t-test assumes a normally distributed population, which likely is not true.

Note: We could have done this test in Minitab by clicking "Options" on the one-sample t-test menu and specifying our one-sided alternative hypothesis

Conclusions

- The t-test assumes the data are coming from a normally distributed population
  - For n larger than 30, this assumption isn't that important
  - For small n, this can make a big difference and randomization tests should be prefered if the sample shows signs of non-normality

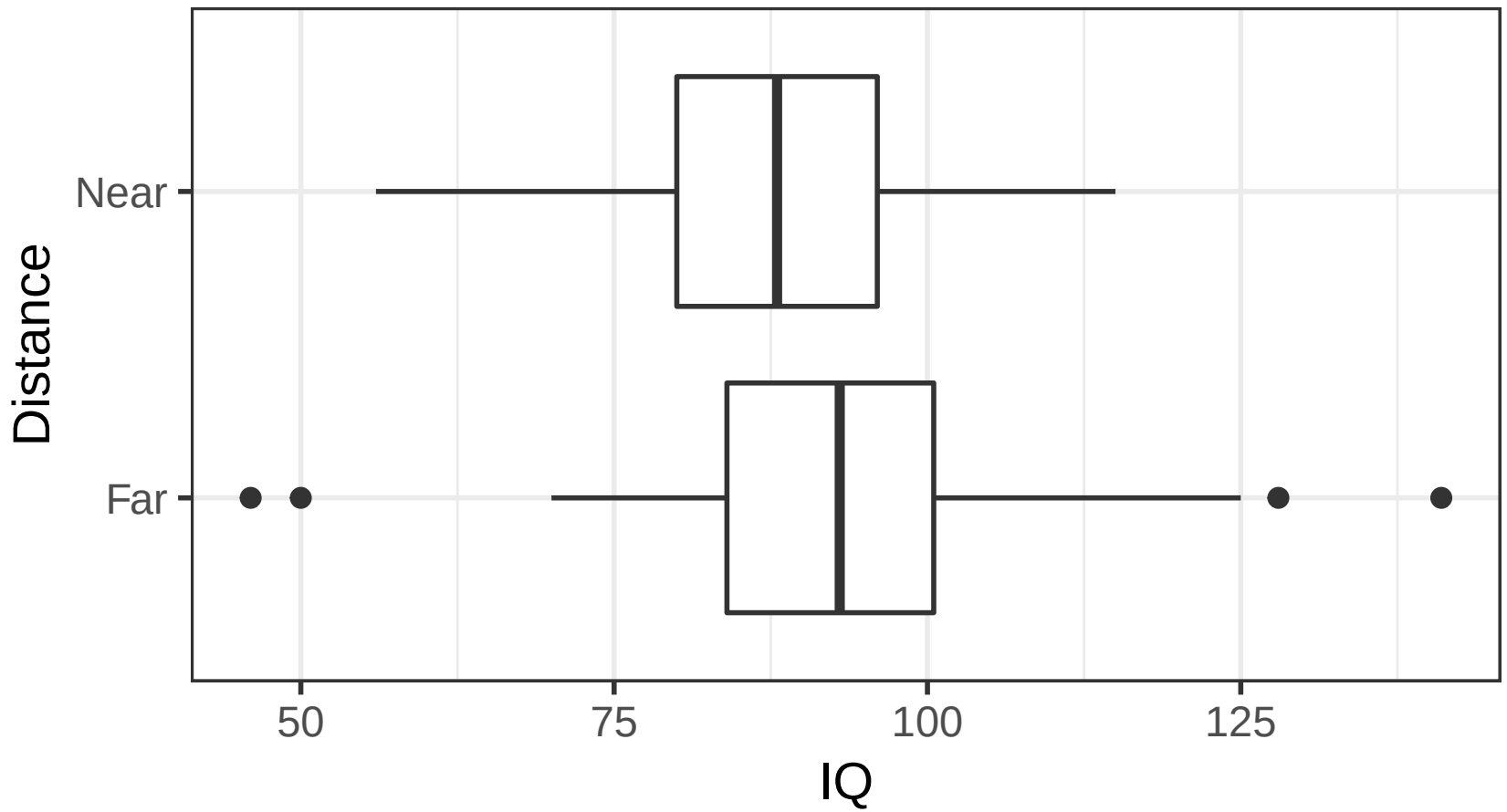
Inference for a Difference in Means

- To explore statistical approaches to analyzing a difference in means, we will use data from a study investigating relationship between lead exposure and neurological development
  - Researchers in El Paso, TX measured the IQ scores (age-adjusted) of 57 children who lived within 1 mile of a lead smelter and 67 children who lived at least 1 mile away
  - Here is a portion of Lead Exposure data:

	Distance	IQ
36	Far	104
97	Near	76
50	Far	104
107	Near	104
113	Near	92
6	Far	94

Lead Exposure and IQ

1. Do the data for each group appear to be normally distributed?
2. Does there appear to be a relationship between distance from the smelter and IQ?



Lead Exposure and IQ - Example #1

First lets see if we can use approaches for a single mean to analyze these data. With your groups:

1. Load the LeadIQ data (available here or on p-web)
2. Construct separate 95% confidence intervals for each group's mean
3. Use these reach a conclusion regarding the impact of distance from the smelter on IQ

Lead Exposure and IQ (Solution #1)

The confidence intervals are shown below:

$$\bar{x}_{\text{near}} \pm t^*_{df=56} * \frac{s_{\text{near}}}{\sqrt{n_{\text{near}}}} = (86.0, 92.4)$$

$$\bar{x}_{\text{far}} \pm t^*_{df=66} * \frac{s_{\text{far}}}{\sqrt{n_{\text{far}}}} = (88.8, 96.6)$$

These intervals seem to suggest that the group IQs are not significantly different

Lead Exposure and IQ - Revisited

The previous approach was sub-optimal, it is *more powerful* to look at the difference in means than it is to look at each mean separately. To understand why this is, we'll need a new CLT result:

$$\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

Looking at the standard error of a difference in means is always less than sum of the standard errors of each mean separately:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \sqrt{\frac{\sigma_1^2}{n_1}} + \sqrt{\frac{\sigma_2^2}{n_2}}$$

Remember: a smaller standard error means increased statistical power

Lead Exposure and IQ - Degrees of Freedom

- Since we will have to rely on estimates of σ_1 *and* σ_2 to make use of the CLT result, you might be wondering how to determine the correct degrees of freedom. The answer is quite messy...

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^2/n_1}{n_1-1} + \frac{s_2^2/n_2}{n_2-1}}$$

- Don't ever calculate this by hand, use software!
- When doing textbook problems, use the smaller of $n_1 - 1$ and $n_2 - 1$
  - This is a *conservative* approach (because it underestimates the actual degrees of freedom)

Lead Exposure and IQ - Example #2

1. Using Minitab to calculate the necessary summary statistics, then conduct a two-sample t-test (by hand) to determine whether the mean IQ differs for children living near or far from a lead smelter (use the *df* suggestion for doing textbook problems)
2. Repeat the same test using Minitab (Stat -> Basic Statistics -> Two-sample t-test)

Lead Exposure and IQ (solution #2)

$$H_0 : \mu_{\text{near}} - \mu_{\text{far}} = 0, H_A : \mu_{\text{near}} - \mu_{\text{far}} \neq 0$$

$$\bar{x}_{\text{near}} - \bar{x}_{\text{far}} = -3.49$$

$$SE_{\text{diff}} = \sqrt{\frac{12.2^2}{57} + \frac{16.0^2}{67}} = 2.54$$

$$t_{\text{test}} = \frac{-3.49 - 0}{2.54} = -1.374$$

- Using our "by hand" rule, $df = 56$, and this test statistic results in a p-value of 0.174
- We fail to reject the null hypothesis that there is no difference in IQ, though there is a trend towards children living near the smelter having slightly lower IQ levels.
- Minitab calculates $df = 120$, leading to a very similar p-value of 0.170

Paired Designs

Comparing the standard errors for one-sample (a single mean) vs. two-sample (difference in means) approaches is interesting:

$$\sqrt{\frac{\sigma_1^2}{n_1}} < \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \sqrt{\frac{\sigma_1^2}{n_1}} + \sqrt{\frac{\sigma_2^2}{n_2}}$$

- Two-sample approaches are more powerful than two one-sample approaches done separately
- But if we could collect our data such that we only needed a single one-sample test we'd have *even more* statistical power

Paired Designs - Example

- At the 2008 Olympics, several swimming world records were broken and controversy arose over new swimsuit designs providing an unfair competitive advantage
- In 2010, new international rules were implemented regulating swimsuit coverage and material
  - These rules naturally prompt the question "Do certain swimsuits really make swimmers faster?"
- Data from a study looking at the 1500m swim velocity of 12 competitive swimmers is shown in the table below:

Wetsuit	1.57	1.47	1.42	1.35	1.22	1.75	1.64	1.57	1.56	1.53	1.49	1.51
NoWetsuit	1.49	1.37	1.35	1.27	1.12	1.64	1.59	1.52	1.50	1.45	1.44	1.41

Paired Designs - Example

- The Wetsuits Data were collected using a **paired design**
  - In this design, each subject serves as their own control
  - This essentially eliminates variability between subjects, a major source of uncertainty in any experiment
- To analyze data from a paired design (with a quantitative outcome), we just use a one-sample t-test on the within subject differences, so the test statistic looks like:

$$t_{\text{test}} = \frac{\bar{x}_{\text{diff}} - \mu_0}{s_{\text{diff}}/n_{\text{pairs}}}$$

You can see that this requires us to create a new variable "difference", and then calculate that variable's mean and standard deviation.

Paired Designs - Example

1. Use Minitab conduct a two-sample t-test on the Wetsuits Data to see if swimsuit type improves 1500m swim velocity
2. Use Minitab to create a variable "difference", then conduct a one-sample paired t-test investigating the same question
3. Why is the paired test is more powerful?

Paired Designs - Example (solution)

1. The test statistic (provided by Minitab) is 1.37 and the p-value 0.186 for the two-sample t-test
2. For the paired test:

$$\bar{x}_d = 0.078, s_d = 0.022, t_{test} = \frac{0.078 - 0}{0.022/\sqrt{12}} = 12.3$$

- Using a t-distribution with 11 degrees of freedom the p-value for this test is nearly zero. There is overwhelming evidence that wearing a wetsuit improves 1500m swim velocity
- The paired test is much more powerful because the standard error is much smaller

Conclusion

Right now you should...

1. Be able to perform z and t tests
2. Know how to construct P% confidence intervals using the t-distribution
3. Know the limitations of these approaches and the assumptions involved
4. Understand why the t-distribution is necessary when σ is estimated
5. Know when to conduct a paired t-test, and the advantage it provides over a two-sample t-test

These notes cover parts of Ch 6 from the textbook, I encourage you to read through the chapter and its examples