Classical Approaches to Statistical Inference

Ryan Miller

Standard Errors

- Previously we've seen that the bootstrap and randomization distributions can be approximated by a normal curve when the sample size is sufficiently large
- But using this approximation to perform statistical inference requires us to know the standard error
- Fortunately, many brilliant statisticians have derived analytic expressions for the standard error of various sample statistics!

One Proportion (one-sample categorical data)

We will begin by looking at a single proportion, or scenarios involving *one-sample categorical data*

- Our example will be a study conducted by Johns Hopkins University where researchers studied the survival of premature babies (born at 25 weeks gestation)
- The researchers wanted to estimate the proportion of premature babies, born in hospitals similar to Johns Hopkins, that would be expected to survive premature labor
- They searched 3 years of their hospital's birth records and found 39 premature babies, 31 of whom survived at least 6 months.

Note that this research question involves a single categorical outcome: whether the baby survived or died

One Proportion (one-sample categorical data)

- In the Johns Hopkins study, 31/39 premature babies survived, resulting in a **sample statistic** of $\hat{p} = 0.795$
- The **parameter of interest** is p, the survival proportion of all such babies at similar hospitals
- We will consider 39 to be sufficiently large sample, so the normal approximation provides a 95% confidence interval for p:

$$\hat{p} \pm 1.96SE$$

- But what about the SE of $\hat{p}$?

One Proportion (one-sample categorical data)

- Statisticians have long been pre-occupied with random binary events like coin flips
- In the Johns Hopkins study, we can view the event that each baby survives as a weighted coin flip with probability p
- We won't get into the details, but in this situation the standard error of the sample proportion ($\hat{p}$) is:

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

Notice the role of sample size (n), what happens if n gets larger? Is there still a barrier to statistical inference even with this formula?

Tests and Confidence Intervals for One Proportion

We don't know p, but our best guess is $\hat{p}$, which suggests the $P\%$ confidence interval estimate of p:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

When hypothesis testing we operate in the world of null hypothesis, here our best guess at p is p_0, which suggests the test statistic:

$$z_{test} = \frac{\text{Sample Statistic} - \text{Null Value}}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

The reference distribution of this test statistic is $N(0, 1)$

One Proportion - Example #1

Recall that in Johns Hopkins study, 31/39 babies survived. With your group:

1. Calculate the 95% confidence interval estimate of p
2. Conduct a hypothesis test at the $\alpha = 0.1$ level investigating whether the survival proportion is larger than 50% (use Minitab to calculate the p-value using the standard normal distribution)
3. Compare your results for 1 and 2 to bootstrapping/randomization in StatKey

One Proportion - Example #1 (solution)

95% confidence interval:

$$\hat{p} = 0.795 \; SE = \sqrt{\frac{0.795(1-0.795)}{39}} = 0.065$$

$$0.795 \pm 1.96 * 0.065 = (0.668, 0.922)$$

Hypothesis test: $H_0 : p \leq 0.50, H_A : p > 0.50$

$$z_{test} = \frac{0.795 - 0.50}{\sqrt{\frac{0.5(1-0.5)}{39}}} = 3.685 \qquad p\text{-value} = 0.0001$$

We reject H_0 at the $\alpha = 0.05$ level, we have extremely strong evidence that more than 50% of premature babies born in hospitals similar to Johns Hopkins will survive longer than 6 months.

One Proportion - Example #2

The Johns Hopkins researchers also collected data on gestation lengths other than 25 weeks, they observed 0/29 babies born at 22 weeks survived at least 6 months

1. Calculate a 95% confidence interval estimating the survival proportion of babies born at 22 weeks
2. Do you believe that this interval actually has 95% coverage? why or why not?
3. Suppose 1/29 babies had survived, calculate the 95% confidence interval. Does this interval have any problems?

One Proportion - Example #2 (solution)

1. $\hat{p} = 0/29 = 0$ and $SE = \sqrt{\frac{0(1-0)}{29}} = 0$, so the 95% CI is (0, 0)
2. No, the interval suggests that it isn't plausible for even a single baby born at 22 weeks at comparable hospitals to survive
3. $\hat{p} = 1/29 = 0.034$ and $SE = \sqrt{\frac{0.034(1-0.034)}{29}} = 0.034$, so the 95% CI is $0.034 \pm 1.96 * 0.034$ or (-0.033, 0.101)

This interval suggests that negative values are plausible! (Obviously they aren't)

One Proportion - Example #2 (lessons)

- The standard error formulas for a single proportion rely upon:
  - The sample size being sufficiently large
  - The true proportion not being close to 0 or 1
- Statisticians have found these formulas to work well when the following conditions are met:
  - $n * p \geq 10$ and $n * (1 - p) \geq 10$
  - When checking these conditions, we use our most likely value of p ($\hat{p}$ for confidence interval and p_0 for hypothesis testing)
- When these conditions are violated statistical inference is still possible using the *binomial distribution*, a topic we won't cover in this class
  - The exact binomial confidence interval for the first part of Example 2 is (0, 0.12), it doesn't suffer from either problem we witnessed

Inference for One Proportion - Summary

- We can conduct statistical inference on a single proportion p using the sample estimate $\hat{p}$ and SE given by:

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

- We don't know p, so we use our most likely value when determining the SE
  - $\hat{p}$ is most likely in reality (ie: when we are trying to estimate p using a confidence interval)
  - p_0 is most likely in the world of the null hypothesis (ie: when testing if $p = p_0$)
- The normal approximation only works well when $n * p \geq 10$ and $n * (1-p) \geq 10$
  - When these conditions aren't met, we can still use randomization tests or bootstrapping (or the exact binomial)

Inference for a Mean (one-sample quantitative data)

- When drawing random samples of size n from a population with mean μ and population standard deviation σ, the standard error of the sample means is given by:

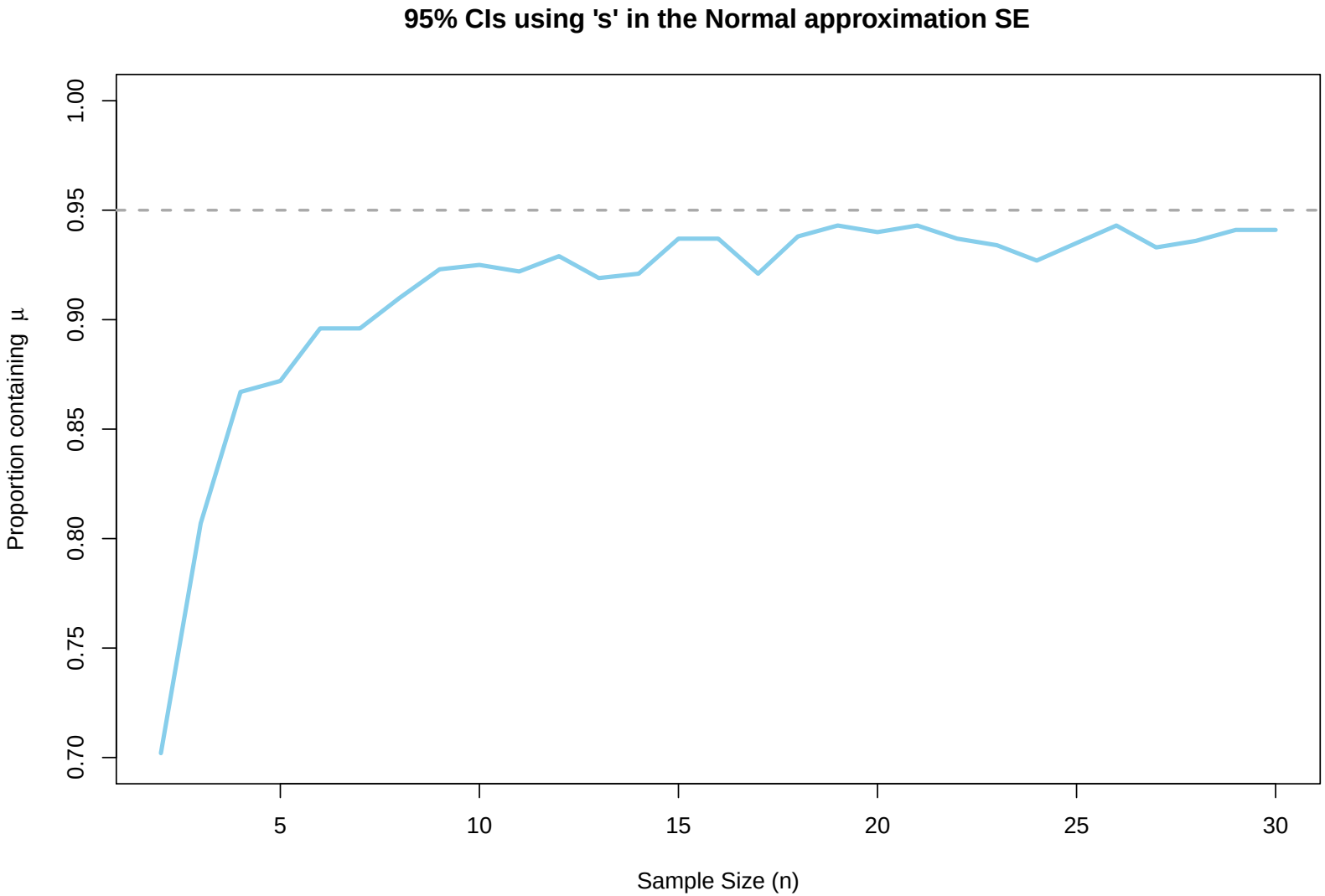
$$SE = \frac{\sigma}{\sqrt{n}}$$

- What happens to the standard error as the sample size increases?
- What makes this formula difficult to use in a real data setting?

Inference for a Mean (one-sample quantitative data)

- A seemingly natural decision is to use s, the standard deviation of the sample, in place of σ, the standard deviation of the population
  - After all, we know that s is on average the same as σ
- However, plugging in s and using the normal approximation doesn't always work...

Inference for a Mean (one-sample quantitative data)



William Gosset (1876 - 1937)

- Simply plugging in the sample standard deviation for σ produces flawed results when n is small
- Prior to modern computing, it wasn't so easy to discover this flaw
- Enter William Gosset, a chemist who worked for Guinness Brewing in the 1890s
  - His experiences at Guinness prompted Gosset to investigate the statistical validity of results from small samples
  - After taking a leave of absence from the brewery to work on the problem, Gosset derived a modified distribution fixed the flaw
  - His finding, the t-distribution, was published under the name "Student" because Guinness didn't its competitors knowing they were gaining an advantage by employing statisticians

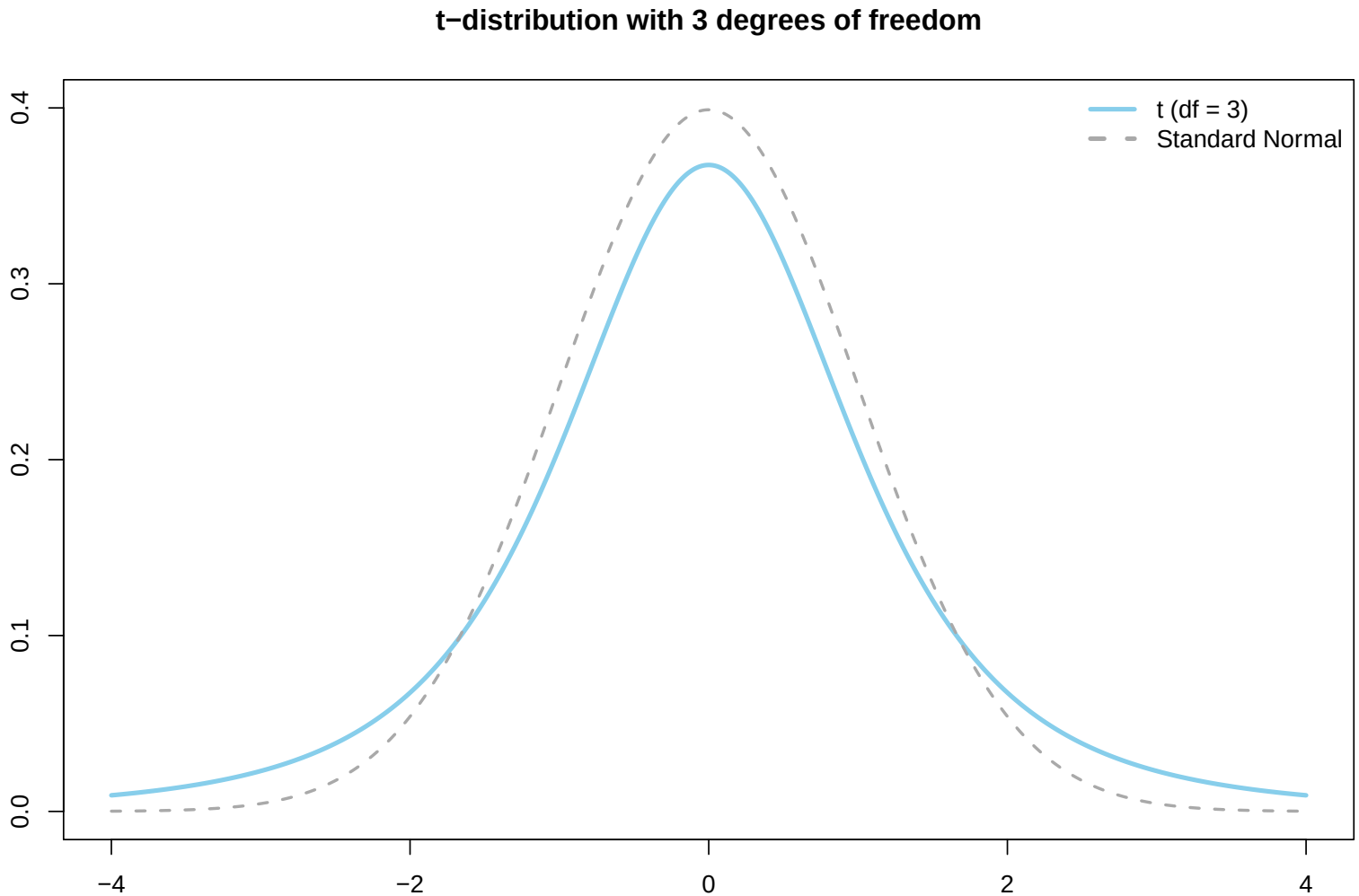
The t-distribution

- The flaw in using s in the normal approximation was that we assumed we knew the actual standard error, but really we are estimating it from the data (s comes from the sample)
  - Pretending we know SE underestimates the actual amount of uncertainty we have about the population
- Gosset showed that when the SE of a mean is estimated using the sample standard deviation, the statistic: $\frac{\bar{x}-\mu}{SE}$ no longer follows a normal curve, but something slightly different
  - This curve is known as "Student's t-distribution"

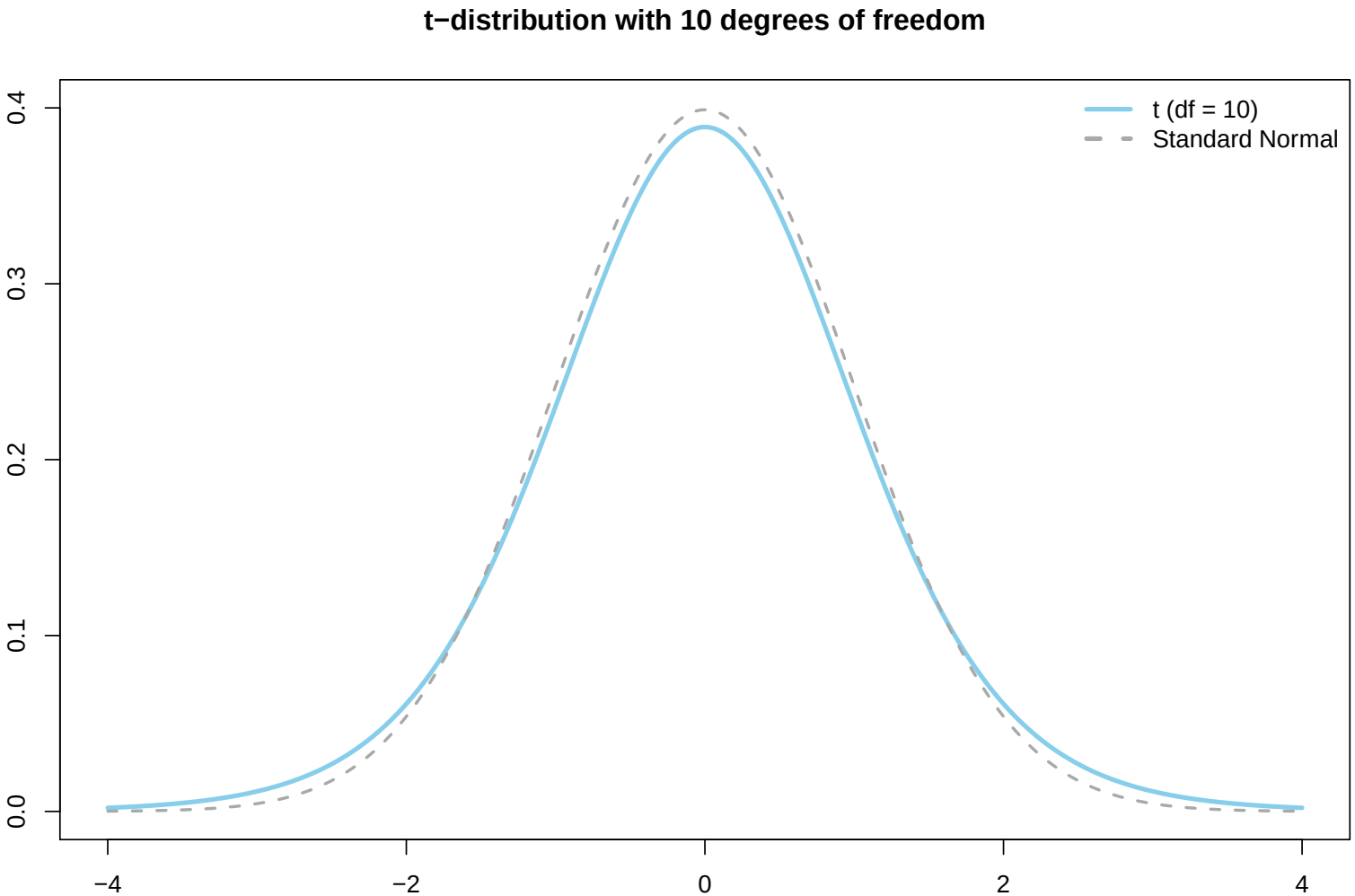
The t-distribution

- 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

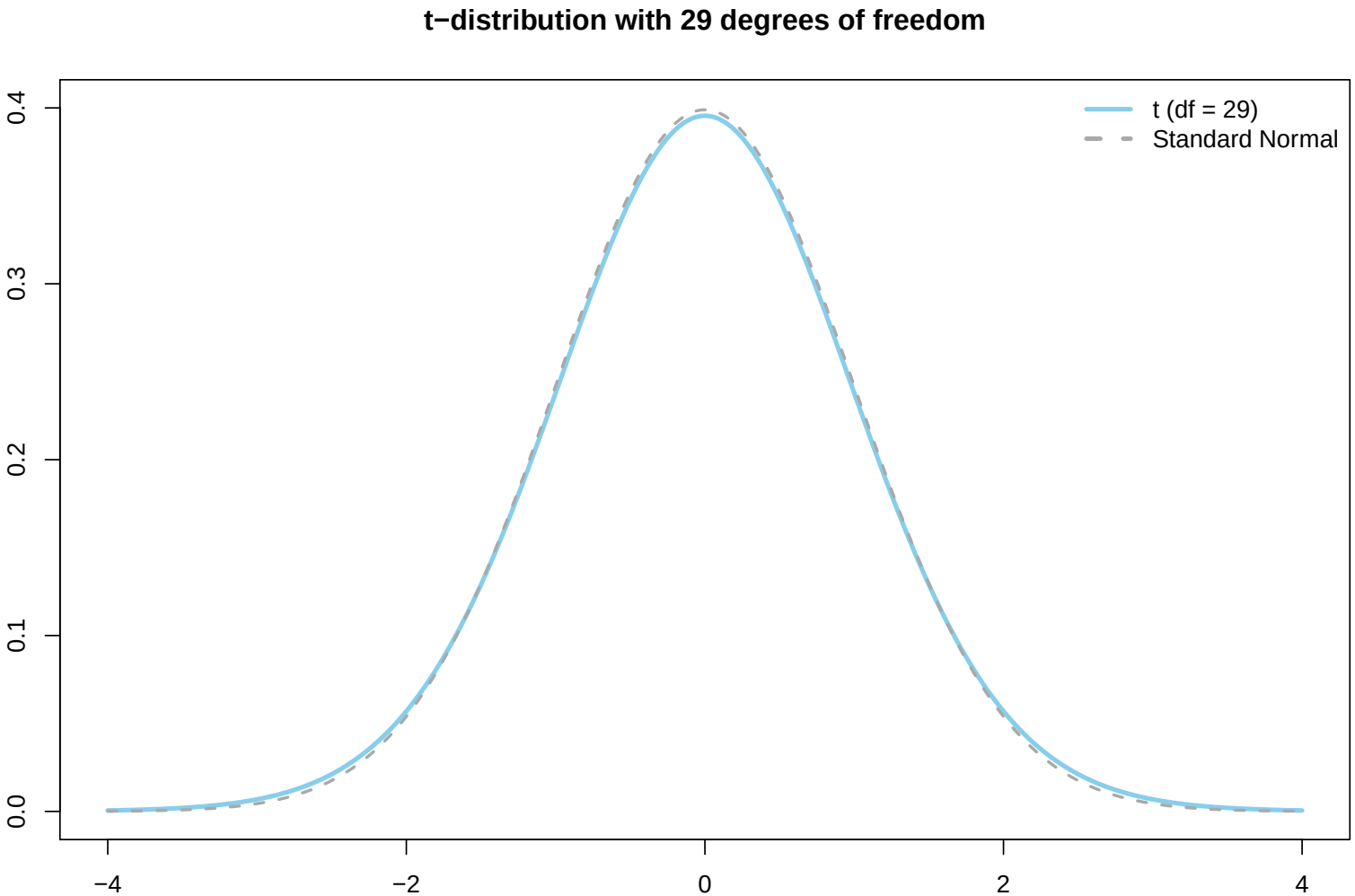
The t-distribution



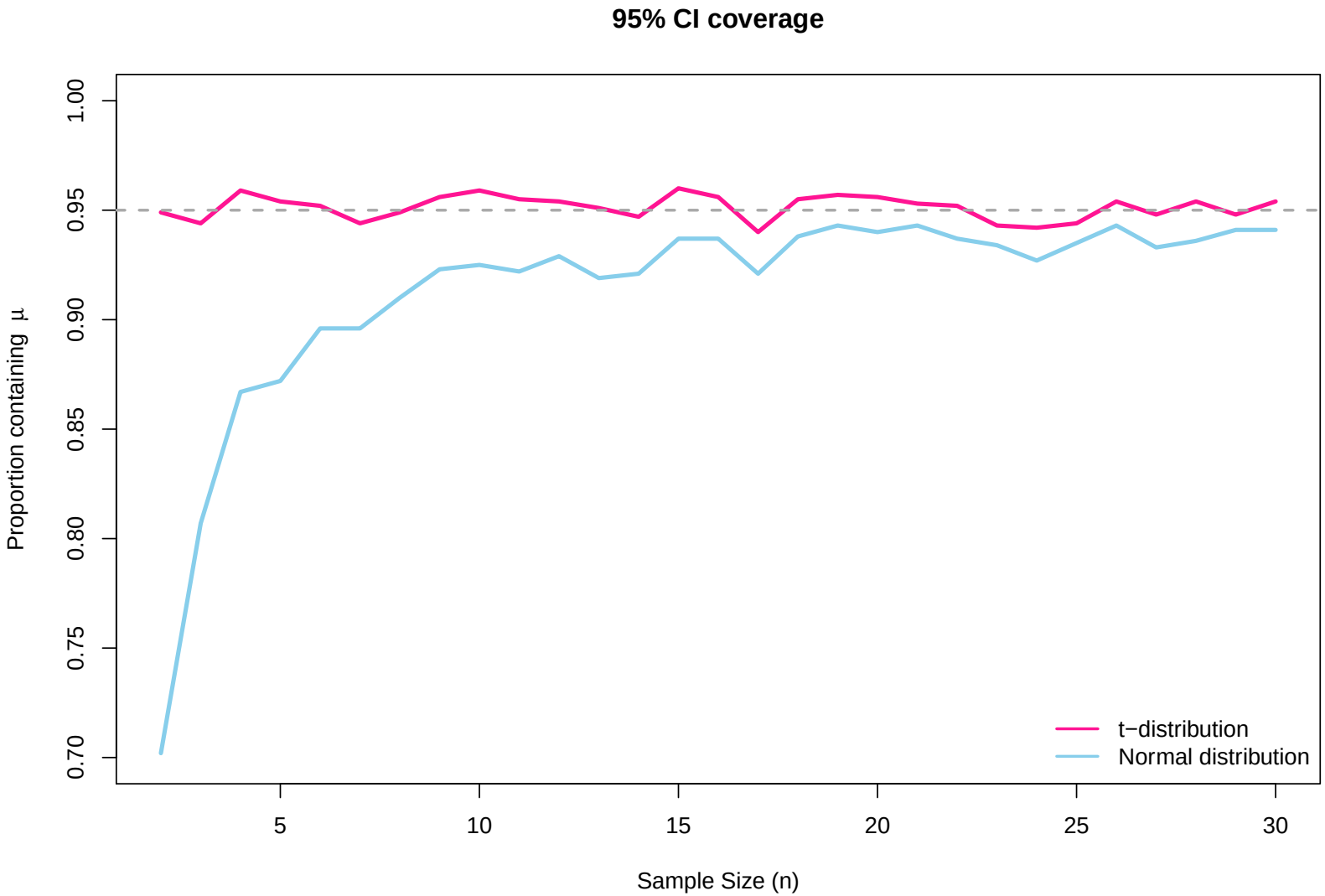
The t-distribution



The t-distribution



The t-distribution



The t-distribution

- The t-distribution has thicker tails than the normal curve, which accurately accounts for the uncertainty introduced when estimating σ using 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 critical values needed for confidence intervals, using Minitab or StatKey

Inference for a Mean - Example #1

Use the StatKey dataset "Arsenic in Chicken" 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

Summary

Scenario	Parameter	Statistic	Standard Error
One categorical variable	p	$\hat{p}$	$\sqrt{\frac{p(1-p)}{n}}$
One quantitative variable	μ	$\bar{x}$	$\frac{s}{\sqrt{n}}$
One categorical variable with groups	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	Coming soon
One quantitative variable with groups	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	Coming soon

Sample Size and Power

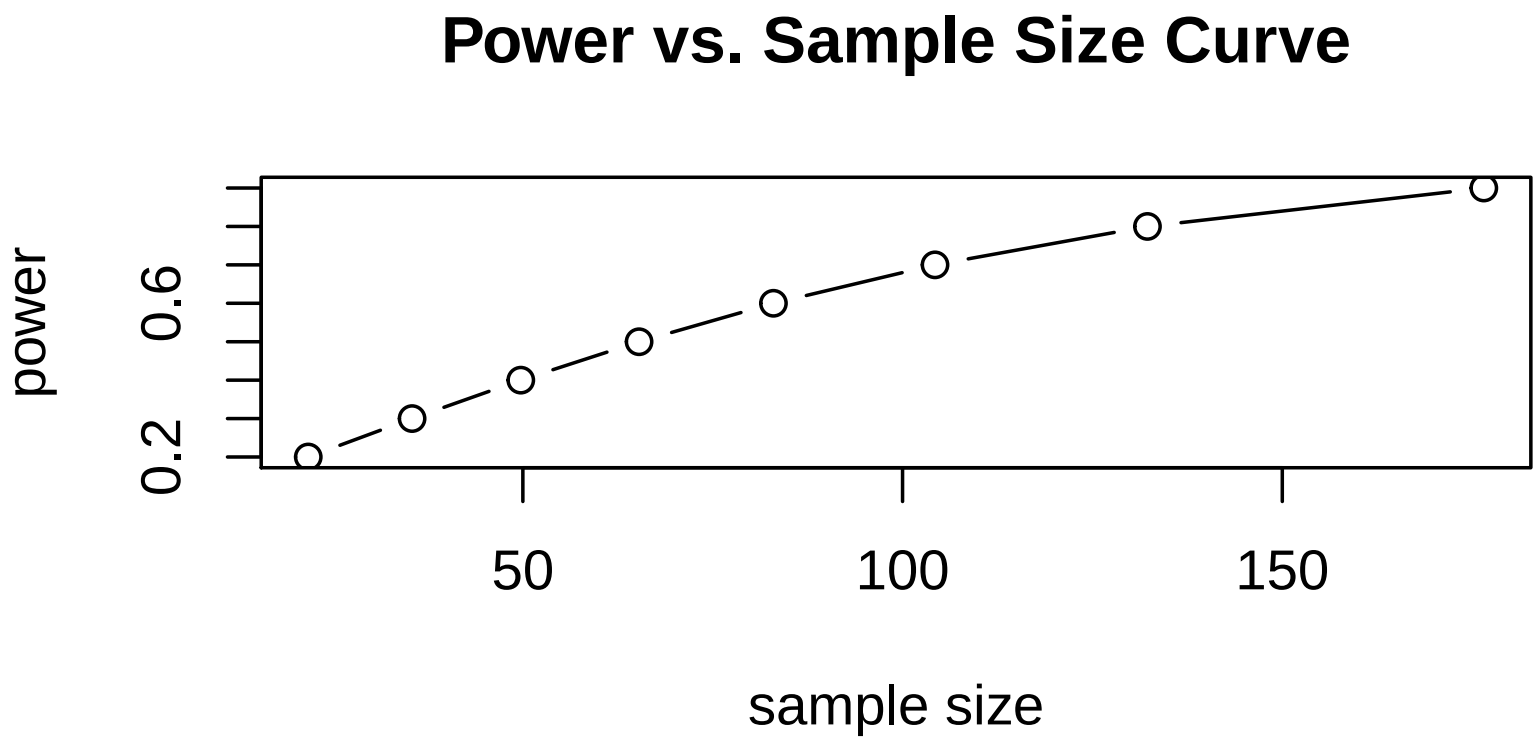
- So far we've constructed *test statistics* of the form:

$$\text{test stat} = \frac{\text{Sample Statistic} - \text{Null Value}}{SE}$$

- We've also seen that the standard error decreases as n increases
- So if we have a preliminary estimates of what the *effect size* might be, we can estimate the probability of rejecting the null hypothesis under this effect size for various different values of n

Sample Size and Power

- **Power** is the probability of rejecting a false null hypothesis
- We discussed an *underpowered* test in the Steph Curry vs. Prof Miller 3pt shooting example
- Lab #5 will look at power in detail, be sure to thoroughly read the lab



Conclusion

Right now you should...

1. Be able to perform z and t tests on single proportions and single means using the appropriate analytic standard errors
2. Know how to construct P% confidence intervals for single proportions and single means
3. Know the limitations of these approaches and the assumptions involved
4. Understand why the t-distribution is necessary when σ is estimated

These notes cover Sections 6.1 and 6.2 of the textbook, I encourage you to read through those sections and their examples