6.2, 6.4, 6.5 Inference About Means

Math 261

Outline

- Formulas for Standard Errors
- Introduction to the t distribution
- t-based Inference for Means

Central Limit Theorem

> For random samples with a *sufficiently large* sample size, the distribution of sample statistics for a mean or a proportion is approximately normal.

- For means, “sufficiently large” is often $n \geq 30$
- If the data are normal, smaller n will be sufficient
- If the data are skewed and/or have outliers, n may have to be much higher than 30

Sample Standard Error Formulas

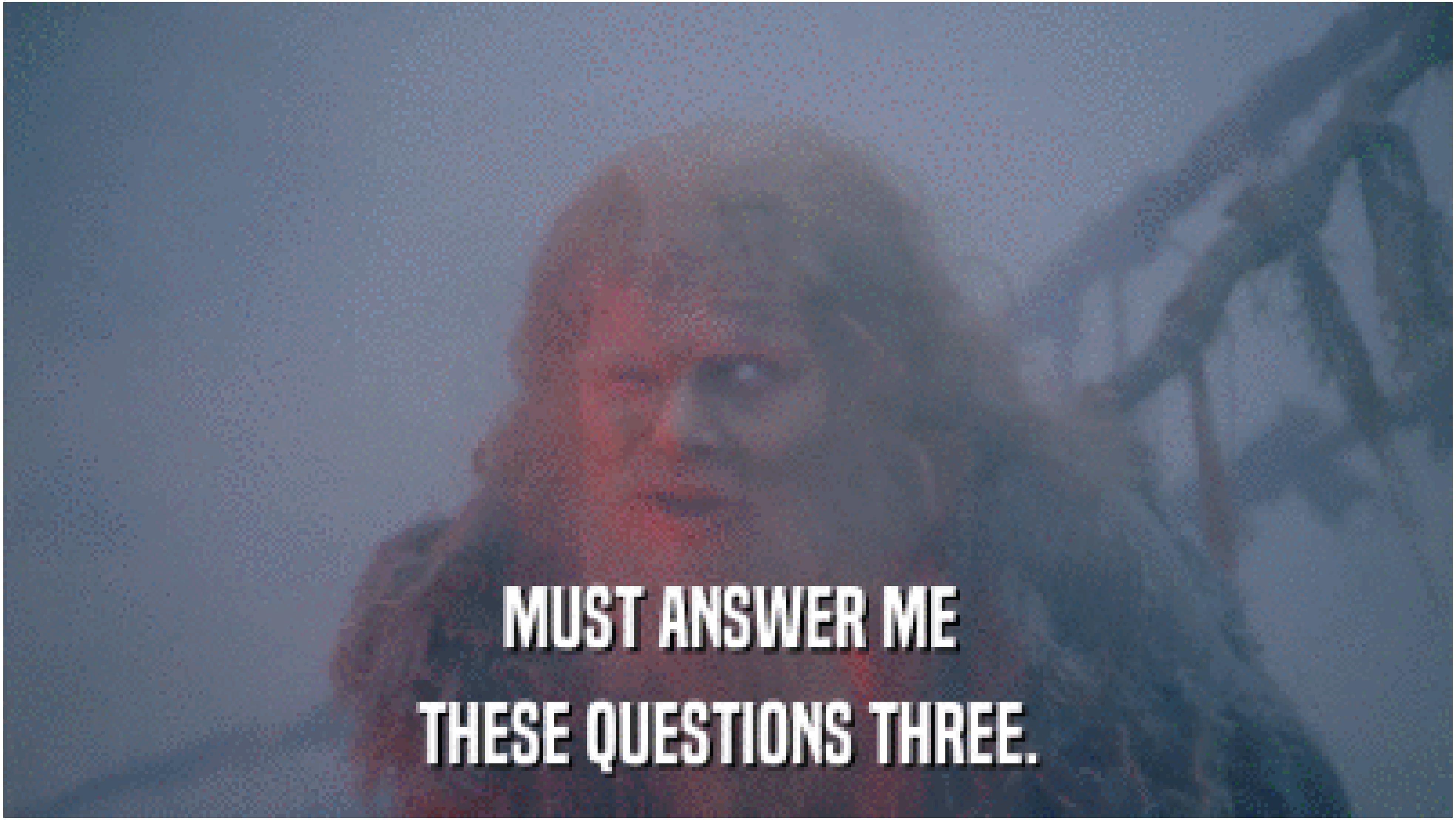
Parameter	Distribution	Standard Error
Proportion	Normal	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Difference in Proportions	Normal	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
Mean	t, df$= n-1$	$\sqrt{\frac{s^2}{n}}$
Difference in Means	t, df $= \min(n_1, n_2) - 1$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

SE of a Mean

The standard error for a sample mean can be calculated by

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

Three Questions



Three important questions

Three Questions (± 1)

- What is the standard deviation of the *population*?
- What is the standard deviation of the *sample*?
- What is the standard error of the *sample mean*?
- What is the *estimated* standard error of the *sample mean*?

The t-Distribution

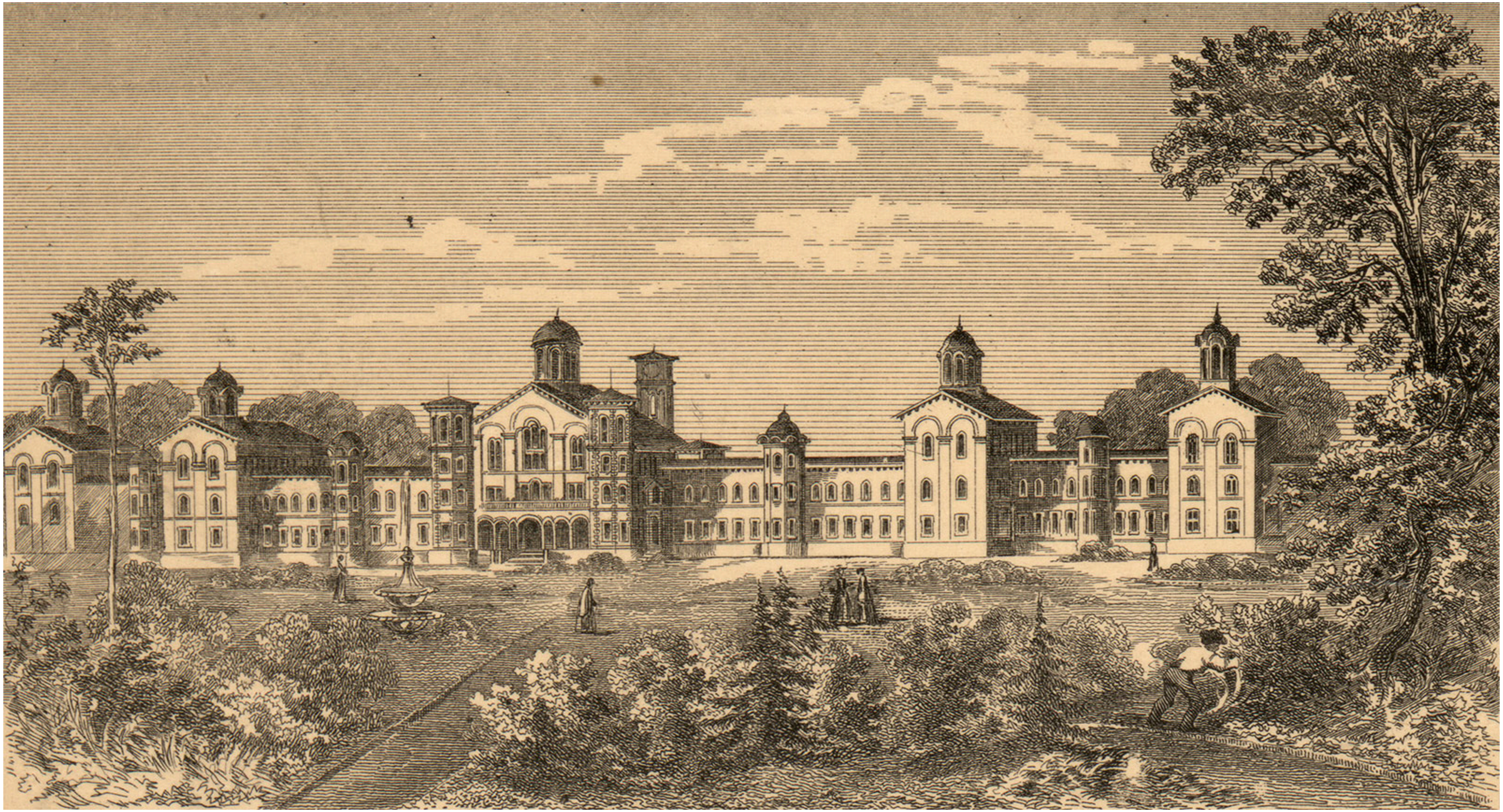
- For quantitative data, we use a t-distribution instead of the normal distribution
- Reason: Using s from the sample to estimate σ in the SE formula
- The t distribution is very similar to the standard normal, but with slightly thicker tails (to reflect the uncertainty in the sample standard deviations)
- Use StatKey to get p-values for hypothesis tests and critical values t^* for confidence intervals.

Degrees of Freedom

- The t-distribution is characterized by its *degrees of freedom (df)*
- Degrees of freedom are based on the sample size
  - Single mean: $df = n - 1$
  - Difference in means: $df = \min(n_1, n_2) - 1$
  - ANOVA: $df = n - K$
- The higher the degrees of freedom, the closer the t-distribution is to the standard normal.

t-Distribution versus Normal Distribution

Where was the first t-Test done?



Case Study: Treatments for Anorexia Nervosa

- **Anorexia nervosa** is an eating disorder characterized by weight loss (or lack of appropriate weight gain in growing children); difficulties maintaining an appropriate body weight for height, age, and stature; and, in many individuals, distorted body image. (https://www.nationaleatingdisorders.org/anorexia-nervosa)

- Randomized controlled experiment in the UK to assess effectiveness of two experimental treatments compared with the established *control* treatment (Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. eds (1993) *A Handbook of Small Data Sets. Chapman & Hall*, Data set 285 (p. 229))

t-Based Formulas

- Confidence interval

$$\text{sample statistic} \pm t^* \times \text{SE}$$

- Hypothesis test

$$t = \frac{\text{sample statistic} - \text{null parameter}}{\text{SE}}$$

- Use the t distribution for p-values and critical t^* values.

Case Study: Treatments for Anorexia Nervosa

- Today we focus on two groups:
  - Control: 26 girls
  - Family therapy: 17 girls

Case Study: Research Questions

Matched Pairs

- For a matched pairs experiment, we look at the differences for each pair, and do analysis on this one quantitative variable
- Inference for a single mean (mean difference)

Getting the Data into R

Use the following command to import the data into the data frame `Anorexia_2samp`.

```
Anorexia_2samp <-
  read.csv("http://people.kzoo.edu/enordmoe/math261/data/Anorexia_2samp.csv
```

Results for the Family Therapy Girls

- The average before weight of family therapy girls was 83.2 pounds.
- The average weight gain for girls in the family therapy group was 7.26 pounds.
- The standard deviation of these gains was 7.16 pounds.
- The sample size was 17 girls.

Q1: Weight gain for girls in Family Therapy: Hypothesis Test

1. State hypotheses
2. Check conditions
3. Calculate standard error SE
4. Calculate t-statistic
5. Compute p-value
6. Interpret in context

Q1: Weight gain for girls in Family Therapy: Hypothesis Test Calculations

Q2: Weight gain for girls in Family Therapy: Confidence Interval

1. Check conditions
2. Find t^* corresponding to desired level of confidence
3. Compute the confidence interval
4. Interpret in context

Q2: Weight gain for girls in Family Therapy: Confidence Interval Calculations

Question 3

Hypothesis Test: Do girls on family therapy gain more weight than girls on control therapy?

1. State hypotheses
    - Two independent samples
2. Check conditions
3. Calculate standard error SE
4. Calculate t-statistic
5. Compute p-value
6. Interpret in context

Summary Results for Family Therapy and Control Groups

```
favstats(Gain ~ Group, data = Anorexia_2samp)
```

```
    Group   min   Q1 median   Q3  max      mean       sd  n missing
1 Control -12.2 -7.0  -0.35  3.6 15.9 -0.450000 7.988705 26       0
2  Family  -5.3  3.9   9.00 11.4 21.5  7.264706 7.157421 17       0
```

Question 3

Do girls on family therapy gain more weight than girls on control therapy? Hypothesis Test Calculations

Question 4

How much more weight do girls on family therapy gain than girls on control therapy? Confidence interval

1. Check conditions
2. Find t^* corresponding to desired level of confidence
3. Compute the confidence interval
4. Interpret in context

Question 4

How much more weight do girls on family therapy gain than girls on control therapy? Confidence interval calculations

Inference formulas for means

Parameter of Interest	Confidence Interval	Test of Significance
μ	$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
$\mu_1 - \mu_2$	$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{2_2^2}{n_2}}}$

A Few Words about Conditions

Normality Conditions

- Using the t-distribution requires that the data comes from a *normal distribution*
- Note: this assumption is about the population data, *not* the distribution of the statistic.
- For large sample sizes we do not need to worry about this, because s will be a very good estimate of σ and t will be very close to $N(0, 1)$.

Small Samples

- For **small sample sizes** ($n < 30$), we can only use the t-distribution if the distribution of the data is approximately **normal**.
  - Problem: Hard to assess normality for small samples.

Summary