Announcements

Exam 2 is FRIDAY

- Office Hours This Week
  - MONDAY 2pm-4pm
  - WEDNESDAY 8am-9:30am
  - WEDNESDAY 4pm-6pm (extra)
  - E-mail will be guaranteed answered if sent before Thursday 6pm.
- Unclaimed exams & homeworks (1, 2, & 3) are up front
- Did you accidently pick up an exam that wasn't yours? Please return it to the front table, no questions asked.
- Please Remove Chapter 22 from your outlines, we will skip 22 and move straight to Chapter 23

Chapter 23. Inferences about Means

- Recall: Proportions (0,1; success/failure) vs. Quantitative Data (potentially has different values for each individual, wider range)
- Quantitative Data is summarized with means and standard deviations (like Chapter 5)
- Chapter 23 involves INFERENCE (making generalizations from samples back to the pouplation) so it involves thinking about Chapter 18 (sampling distribution/distribution of all possible samples)

Recall: CLT from Chapter 18

- No matter what the shape (distribution) of the population appears to be or even if it is unknown
- The sampling distribution (all possible samples of the same size taken from the population) will be NORMAL as long as the following conditions are met
  - The samples are random
  - The observations are independent (10% condition is the best way to check this one)
  - Samples should be a reasonable size (> 50) if working with sample means from quantitative data or should satisfy the np, nq conditions if working with proportions from binary data

Standard Deviation vs. Standard Error for a sampling distribution

- All kinds of distributions have a standard deviation, recall we are working with sampling distribution
- Formula for SD depends on the type of variable (quantitative vs. binary)
- For proportions (binary): $SD_{(p)} = \sqrt{\frac{pq}{n}}$
- For quantitative data: $SD_{(\bar{y})} = \frac{\sigma}{\sqrt{n}}$
- The Standard Error is a standard deviation that has been constructed using sample information as opposed to population information.

$\hat{p}$ versus $\bar{y}$

- In real life, for proportions, we have p-hat (sample proportion) and from that, we can construct an SE because once we know p-hat, we know q-hat and we of course know the sample size.
- For the sample mean, y-bar, of a quantitative variable, there is no simple direction relationship with s, the sample standard deviation. While we can calculate an SE for y-bar, this situation creates some problems that we do not have with p-hat.

SD vs. SE for a sampling distribution

	Parameters Known (SD)	Parameters Unknown (SE)
Binary outcomes (0,1)	$SD_{(p)} = \sqrt{\frac{pq}{n}}$	$SE_{(\hat{p})} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$
Quantitative Outcomes (has a range of values)	$SD_{(\bar{y})} = \frac{\sigma}{\sqrt{n}}$	$SE_{(\bar{y})} = \frac{s}{\sqrt{n}}$

The problem with only knowing the SE for quantitative data

- For small samples (< 50) statisticians noticed that the SE of y-bar varied quite a bit and the sampling distribution was not really quite normal.
- Gosset, a quality control engineer at Guiness Brewery in Ireland solved the problem.
- He developed a new distribution, it looks normal, but it is not, it is "t".

The t-distribution: Properties

- Has mean "mu"
- Has standard error
- Is bell-shaped
- Is symmetric around mu
- Has DEGREES OF FREEDOM (df) n-1
- Has fatter "tails" than the normal
- See Table A-84 (back of textbook)
- Basically used when sigma is unknown and you must use s from the sample.

Standard Normal Distribution (Z) vs. t distribution (t)

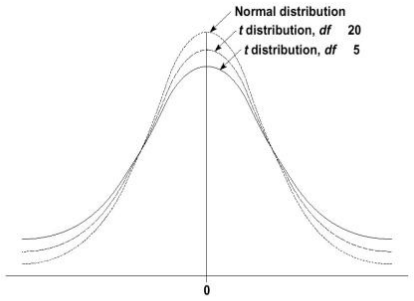

Fig. 9.1 Degrees of Freedom and the Shape of the t-Distribution

Notice there is only one standard normal but the distribution of t is determined by its "degrees of freedom" (df) which is n-1

The t-distribution: formula

- For the same distance between the mean y-bar and mu and same size denominator (SD for Z distribution, SE for t-distribution), a t-distribution will produce larger p-values and wider confidence intervals
- Formula

$$t = \frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}} \quad \text{VS.} \quad Z = \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Upshot – having to use t instead of normal (Z) has a "cost" – i.e. a larger margin of error.

	PROPORTION IN ONE TAIL					
	0.25	0.10	0.05	0.025	0.01	0.005
	PROPORTION IN TWO TAILS COMBINED					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707

Suppose n=16 and sigma is unknown, we would need to use t with df=15. Notice t=2.131

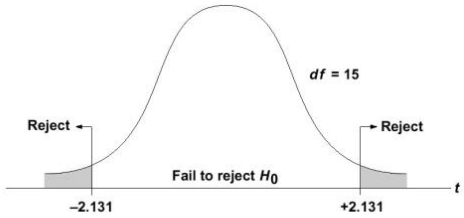

Fig. 9.5 The Critical Region in the t Distribution for $\alpha = .05$ and df = 15

Z vs. t (p. 433)

- If sigma is known, use the standard normal Z distribution and it's formulas.
- When sigma is unknown and your sample is small (n size 50 or less) you should probably use t.
- Your textbook has various criteria for sample size, they basically use t for all cases where sigma is unknown. When n is large enough, t and Z are effectively the same.

Assumptions and Conditions

- Random Samples
- Independence – use check to make sure the sample is no larger than 10% of the population
- (p. 435) for very small samples, $n < 15$ or so your data should be relatively normal for this to work.
- (p. 435) for samples between 15- about 50, t works well as long as there is symmetry (no skewness)
- (p. 435) for samples over size 50 you can use t even if it's non-normal or skewed (you could probably use the normal after size 50)