More Confidence Intervals: Proportions and Small Samples

Keegan Korthauer

Department of Statistics

UW Madison

Exam 1

- Mean: 34.5 (69%)
- Median: 36.5 (73%)
- Standard deviation: 8.6
- Most missed questions:
  - **Problem 4**: Mutually Exclusive ≠ Independent
  - **Problem 9b**: Waiting time between events in Poisson process is Exponential -> need to convert rate parameter to appropriate units (here from minutes to seconds)
  - **Problem 10a**: for two normal RVs X and Y

$$X - Y \sim N(\mu_X - \mu_Y,\ \sigma_X^2 + \sigma_Y^2)$$

- Solutions posted

Exam 1 – Unofficial Letter Grades

Score (Out of 50 Points)	Tentative Letter Grade
[42 – 50]	A
[39 – 42)	AB
[34.5 – 39)	B
[30 – 34.5)	BC
[22 – 30)	C
[19 – 22)	D
[15 – 19)	F

Large-Sample CI for a Population Mean

- CLT says that for **large samples**: $\bar{X} \sim N\left(\mu,\ \dfrac{\sigma^2}{n}\right)$
- General form of CIs: **point estimate** ± **critical value** × **standard deviation**
- Point Estimate: $\bar{X}$
- Critical Value of Normal Distribution: $z_{\alpha/2}$
- Standard Deviation of the point Estimate: **σ/√n**
- Level 100(1-α)% CI for μ is: $\bar{X} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$

CONFIDENCE INTERVALS FOR PROPORTIONS

Normal Approximation to Binomial

- Recall that if X ~ Bin(n, p), then we can write X as a sum of independent and identically distributed RVs from a Bernoulli(p) population:

$$X = Y_1 + \ldots + Y_n$$

where $Y_1, \ldots, Y_n$ ~ Bern(p) (with mean p and variance p(1-p))

- Also note that $\hat{p} = \frac{X}{n} = \frac{Y_1 + \ldots + Y_n}{n} = \bar{Y}$

- Then by the CLT if n is large enough,

$$\hat{p} \sim N(p,\ p(1-p)/n) \text{ and } X \sim N(np,\ np(1-p))$$

(approximately)

Deriving the CI for a Proportion

- Let *p* represent the proportion of successes in the population
- We sample *n* members of the population and count *X* 'successes'

$$\hat{p} = \frac{X}{n} \sim N\left(p,\ \frac{p(1-p)}{n}\right)$$

- Then for 95% of all possible samples,

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$$

- Looks like a CI, but *p* is unknown!
  - Traditional approach: replace *p* with X/n
  - Modern approach: replace *p* with (X+2)/(n+4) (see next slide)

Summary of CIs for a Proportion

Let X be the number of successes in *n* independent Bernoulli trials with success probability *p*, so that X ~ Bin(*n*, *p*)

Let $\tilde{n} = n + 4$ and $\tilde{p} = (X + 2) / \tilde{n}$

- A level 100(1-α)% CI for *p* is $\tilde{p} \pm z_{\alpha/2}\sqrt{\dfrac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$

- The level 100(1-α)% upper & lower confidence bounds for *p* are

$\tilde{p} + z_{\alpha}\sqrt{\dfrac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$ (upper) $\tilde{p} - z_{\alpha}\sqrt{\dfrac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$ (lower)

Notes

- If the lower limit of the CI or lower confidence bound is less than 0, replace it with 0
- If the upper limit of the CI or upper confidence bound is greater than 1, replace it with 1
- Although derived from CLT, these CIs for proportions work well for any sample size (even small samples)

Example

In a random sample of 50 steel rods used in optical storage devices, 4 or them were defective.

Find a 99% CI for the proportion of defective rods in the entire sample

[0.000765, 0.222]

CONFIDENCE INTERVAL FOR POPULATION MEAN (SMALL RANDOM SAMPLE)

Sample Size Problem

- If we have a small sample ($n < 30$):
  - the Central Limit Theorem does not apply
  - the sample mean may not be approximately normal
  - the sample standard deviation may not be close to σ
  - how to construct a confidence interval?
- If we know the population is approximately normal:
  - the sample mean will be approximately normal, even for a small sample
  - the sample standard deviation still may not be close to σ
  - we can use a different distribution to account for our lack of knowledge of σ: the **Student's *t* distribution**

Student's *t* Distribution

Let $X_1,\ldots,X_n$ be a **small** (n < 30) sample from a **normal** population with mean μ. Then

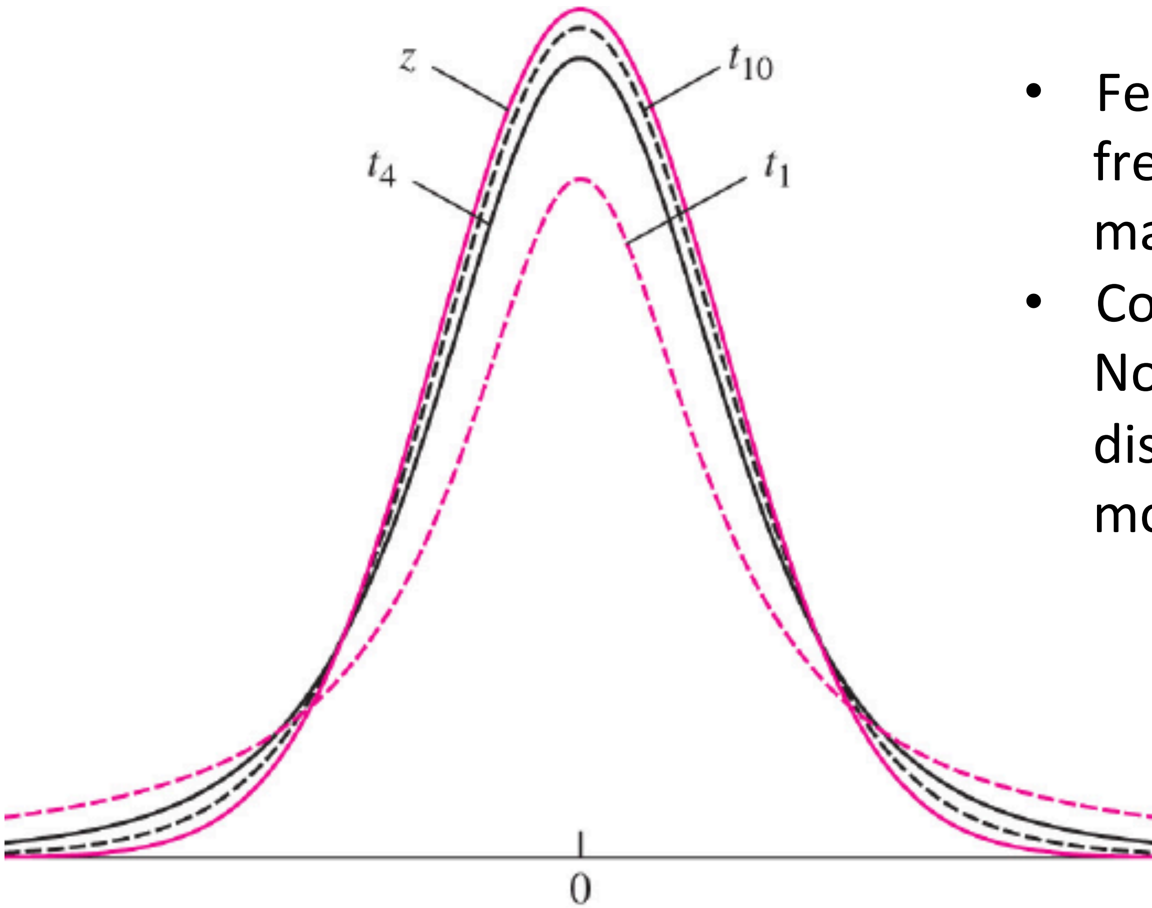
$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} is the Student's *t* distribution with *n-1* degrees of freedom

The distribution is always centered at zero and has only one parameter (degrees of freedom *n-1*) that determines its shape

As *n* gets very large, t_{n-1} approaches N(0, 1)

Student's *t* distribution



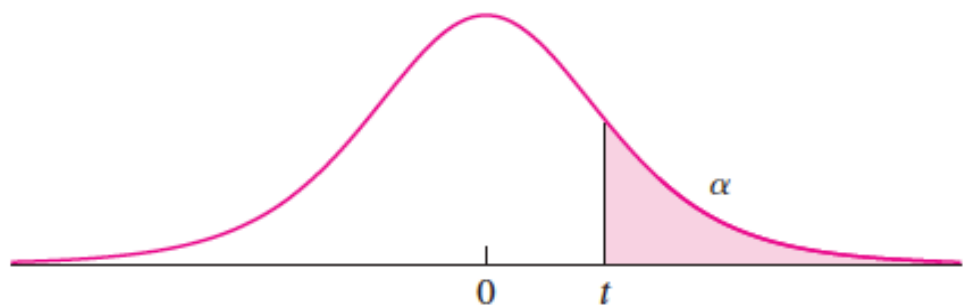
- Fewer degrees of freedom means more mass in the tails
- Compared to the Normal, does the *t* distribution have more or less spread?

For a demo, see: http://www.math.uah.edu/stat/applets/SpecialCalculator.html
(choose "Student t" from the dropdown menu, see how the shape changes with n)

How to use the *t*-distribution table

TABLE A.3 Upper percentage points for the Student's *t* distribution

Also in the back cover of the text



Degrees of freedom (n-1)

Tail area (upper)

	α								
ν	**0.40**	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144	4.587

t statistic

Finding *t* Probabilities Using R

- What if we want to find the probability that a t statistic with 1 df is greater than 1.96?
- From the table, we can only tell it is between 0.10 and 0.25
- Use the R function for the *t* distribution:

  `pt(q, df, lower.tail = TRUE)`

  where `q` is the *t* statistic, `df` is the degrees of freedom. Change `lower.tail = FALSE` to get upper (right-tail) probabilities.
- Example:
  ```
  > pt(1.96, 1, lower.tail=FALSE)
  [1] 0.1501714
  ```
- There is also a `qt()` function for finding the t statistic for a given tail area

Confidence Intervals and Bounds

Let $X_1,\ldots,X_n$ be a **small** (n < 30) sample from a **normal** population with mean μ. Then

- A level 100(1-α)% confidence interval for μ is

$$\bar{X} \pm t_{n-1,\,\alpha/2}\frac{s}{\sqrt{n}}$$

- A level 100(1-α)% upper confidence bound for μ is

$$\bar{X} + t_{n-1,\,\alpha}\frac{s}{\sqrt{n}}$$

- A level 100(1-α)% lower confidence bound for μ is

$$\bar{X} - t_{n-1,\,\alpha}\frac{s}{\sqrt{n}}$$

Notes

- Must have knowledge that the sample comes from a population that is **approximately normal** to use the *t* distribution

- Samples containing outliers or highly skewed samples should not be used (outliers and skew are evidence that the sample is **not** normal)

- When we know the population is approximately normal **and** we know the population standard deviation σ we can use the method for calculating CIs from large samples (last lecture)

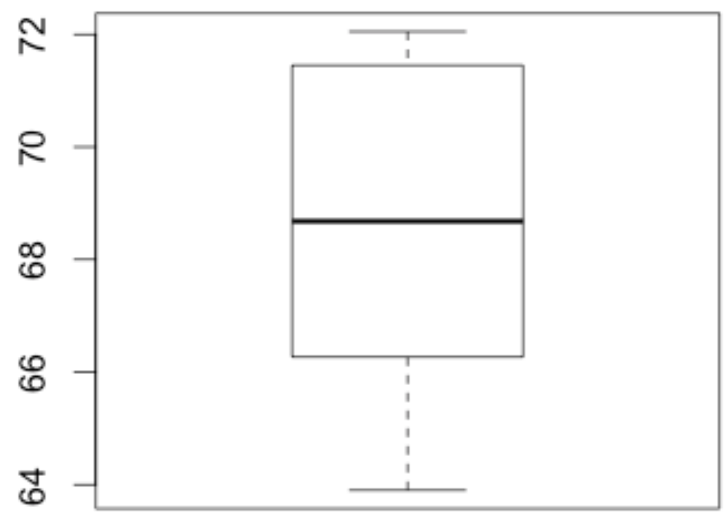
- 'Toolbox' analogy

Example – Small Sample of Heights

- Example – a random sample of the heights (in inches) of 5 UW students:

63.90, 71.45, 68.68, 72.05, 66.27

- Say we know that the heights of all students at UW are approximately normally distributed
- Construct a 95% confidence interval for the mean height



[64.193, 72.747]

Next

- Confidence intervals for the difference in means and proportions (guest lecturer)

- Homework 5 due on Friday

- Check Learn@UW for Homework 6 over the weekend