Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. When to use which formula.
2. Review list.
3. Example problems.

Read ch6. The midterm will be on ch 1-6.
http://www.stat.ucla.edu/~frederic/13/W23 .
Bring a PENCIL and CALCULATOR and any books or notes you want to the midterm and final. You cannot use a computer, laptop, ipad, or phone on the exams though.

No class Mon Feb20, President's Day!

t versus normal and assumptions.

Suppose $X_1, \ldots, X_n$ are iid with mean μ and SD σ.

CLT: $(\bar{x} - \mu) \div (\sigma/\sqrt{n}) \sim$ std. normal.

If $X_1, \ldots, X_n$ are normal, then $(\bar{x} - \mu) \div (\sigma/\sqrt{n})$ is std. normal.

σ is the SD of the population from which $X_1, \ldots, X_n$ are drawn. s is the SD of the sample, $X_1, \ldots, X_n$.

Gosset (1908) showed that replacing σ with s,

if $X_1, \ldots, X_n$ are normal, then $(\bar{x} - \mu) \div (s/\sqrt{n})$ is t distributed.

So we need the multiplier from the t distribution.

t versus normal and assumptions.

To sum up,

if the observations are iid and n is large, then

P(μ is in the range $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$) ~ 95%.

If the observations are iid and normal, then

P(μ is in the range $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$) ~ 95%.

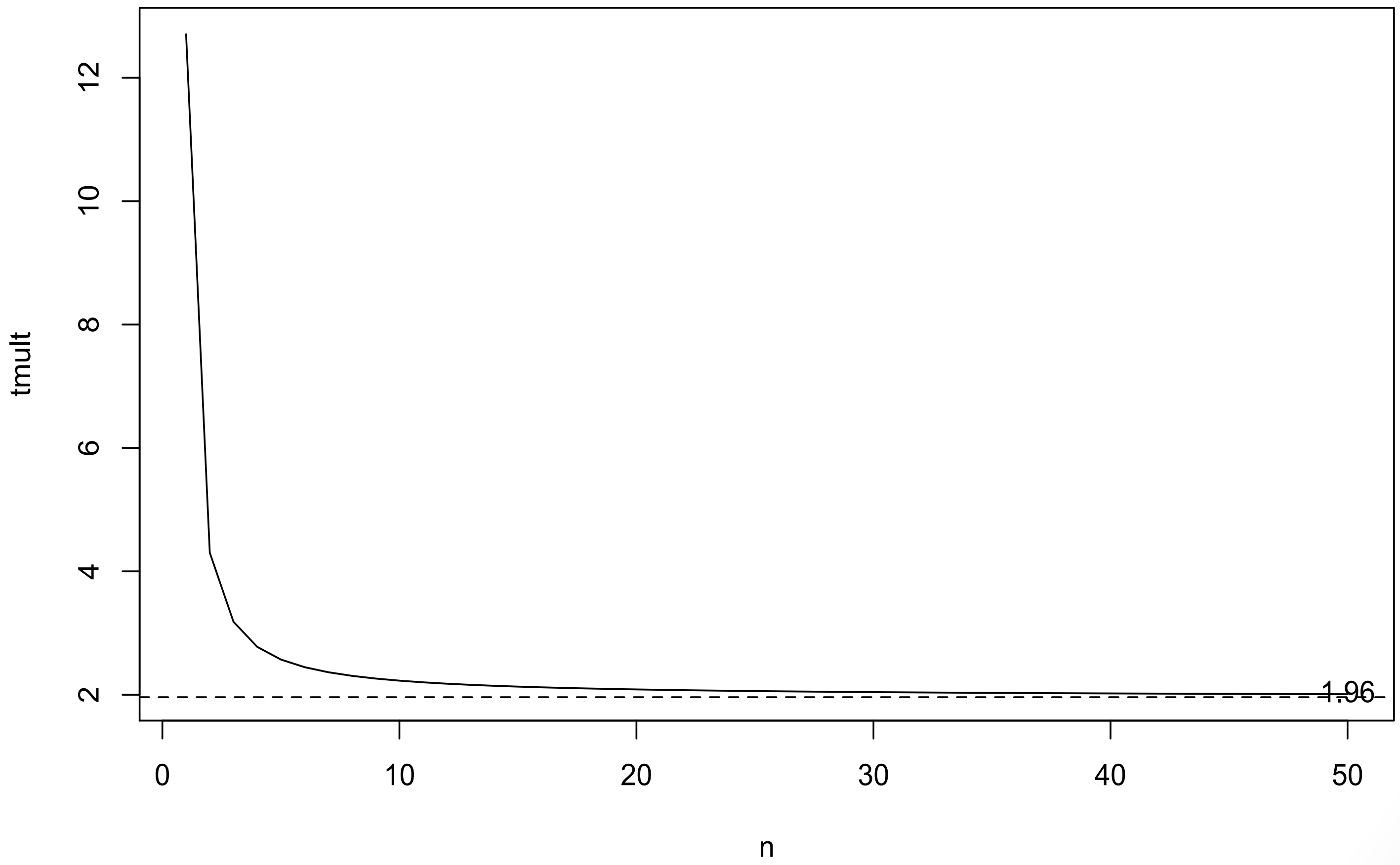
If the obs. are iid and normal and σ is unknown, then

P(μ is in the range $\bar{x}$ +/- t_{mult} $s/\sqrt{n}$) ~ 95%.

where t_{mult} is the multiplier from the t distribution.

This multiplier depends on n.

t versus normal and assumptions.



When to use which formula.

a. 1 sample numerical data, iid observations, want a 95% CI for μ.

- If n is large and σ is known, use $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$.
- If n is small, draws are normal, and σ is known, use $\bar{x}$ +/- 1.96 $\sigma/\sqrt{n}$.
- If n is small, draws are normal, and σ is unknown, use $\bar{x}$ +/- t_{mult} $s/\sqrt{n}$.
- If n is large and σ is unknown, t_{mult} ~ 1.96, so we can use $\bar{x}$ +/- 1.96 $s/\sqrt{n}$.

$n \geq 30$ is often considered large enough to use 1.96.

In practice, we typically do not know the draws are normal, but if the distribution looks roughly symmetrical without enormous outliers, the t formula may be reasonable.

b. 1 sample binary data, iid observations, want a 95% CI for π.

View the data as 0 or 1, so sample percentage p = $\bar{x}$, and

$s = \sqrt{[p(1-p)]}$, $\sigma = \sqrt{[\pi(1-\pi)]}$.

When to use which formula.

a. 1 sample numerical data, iid observations, want a 95% CI for μ.

- If n is large and σ is known, use $\bar{x}$ +/- 1.96 σ/√n.
- If n is small, draws are normal, and σ is known, use $\bar{x}$ +/- 1.96 σ/√n.
- If n is small, draws ~ normal, and σ is unknown, use $\bar{x}$ +/- t_{mult} s/√n.
- If n is large and σ is unknown, t_{mult} ~ 1.96, so we can use $\bar{x}$ +/- 1.96 s/√n.

b. 1 sample binary data, iid observations, want a 95% CI for π.

View the data as 0 or 1, so sample percentage p = $\bar{x}$, and

s = √[p(1-p)], σ = √[π(1−π)].

If n is large and π is unknown, use $\bar{x}$ +/- 1.96 s/√n.

Here large n means ≥ 10 of each type in the sample.

When to use which formula.

What if n is small and the draws are not normal, and you want a theory-based test or CI?

How should you find the t multiplier for a CI or a p-value using the t-statistic, when n is small?

These are questions outside the scope of this course, but some techniques have been developed, such as the bootstrap, which are sometimes useful in these situations.

When to use which formula.

c. Numerical data from 2 samples, iid observations, want a 95% CI for μ_1 - μ_2.

If n is large and σ is unknown, use $\bar{x}_1$ - $\bar{x}_2$+/- 1.96 $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$.

As with one sample, if σ_1 is known, replace s_1 with σ_1, and the same for σ_2. And as with one sample, if σ_1 and σ_2 are unknown, the sample sizes are small, and the distributions are roughly normal, then use t_{mult} instead of 1.96. If the sample sizes are small, the distributions are normal, and σ_1 and σ_2 are known, then use 1.96.

d. Binary data from 2 samples, iid observations, want a 95% CI for π_1 - π_2.

same as in c above, with $p_1 = \bar{x}_1$, $s_1 = \sqrt{[p_1(1-p_1)]}$, $\sigma_1 = \sqrt{[\pi_1(1-\pi_1)]}$.

Large for binary data means sample has ≥ 10 of each type.

For testing, use pooled estimate of p for the SE.

For CIs for the difference in proportions,

$$\text{SE} = \sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$$

In testing the difference in proportions,

$$\text{SE} = \sqrt{\left(\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}\right)}$$

where $\hat{p}$ is the proportion in both groups combined.

2. Review list.

1. Meaning of SD.
2. Parameters and statistics.
3. Z statistic for proportions.
4. Simulation and meaning of pvalues.
5. SE for proportions.
6. What influences pvalues.
7. CLT and validity conditions for tests.
8. 1-sided and 2-sided tests.
9. Reject the null vs. accept the alternative.
10. Sampling and bias.
11. Significance level.
12. Type I, type II errors, and power.
13. CIs for a proportion.
14. CIs for a mean.
15. Margin of error.
16. Practical significance.
17. Confounding.
18. Observational studies and experiments.
19. Random sampling and random assignment.
20. Two proportion CIs and testing.
21. IQR and 5 number summaries.
22. Testing and CIs for 2 means.
23. Placebo effect, adherer bias, and nonresponse bias.
24. Prediction and causation.

3. Example problems.

NCIS was the top-rated tv show in 2014. It was 3rd in 2016 and is now 5th in 2017.

A study finds that in a certain city, people who watch NCIS are much more likely to die than people who do not watch NCIS. Can we conclude that NCIS is a dangerous tv show to watch?

Example problems.

NCIS was the top-rated tv show in 2022.

A study finds that in a certain city, people who watch NCIS are much more likely to die than people who do not watch NCIS. Can we conclude that NCIS is a dangerous tv show to watch?

No. Age is a confounding factor. The median age of a viewer is 61 years old.

- 1. Suppose the population of American adults has a mean systolic blood pressure of 120 mm Hg and an SD of 20 mm Hg. You take a simple random sample of 100 American adults. Which of the following is true?
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **2** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **20** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **2** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **0.2** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **0.2** mm Hg.

- 1. Suppose the population of American adults has a mean systolic blood pressure of 120 mm Hg and an SD of 20 mm Hg. You take a simple random sample of 100 American adults. Which of the following is true?
- **A typical adult's blood pressure would differ from 120 by about 20 mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about 2 mm Hg.**
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **20** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **2** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **0.2** mm Hg.
- A typical adult's blood pressure would differ from 120 by about **20** mm Hg, and a typical sample of size 100 would have a sample mean that differs from 120 by about **0.2** mm Hg.