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

1. Z versus t.
2. Example problems.

http://www.stat.ucla.edu/~frederic/13/W23 .

The midterm if Fri Feb24 in class and will be on ch 1-6.
It will be 15 questions, all multiple choice.
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.

# T-test versus Z-test.

For 1 sample numerical data, iid observations.

$z = (\bar{x} - \mu) \div (s/\sqrt{n})$ . If you know σ, use σ in place of s.

$t = (\bar{x} - \mu) \div (s/\sqrt{n})$ . Same formula.

$(s/\sqrt{n})$ = SE for the mean.

The only difference is how you convert it to a p-value. With z, you use the normal distribution to find a p-value, and with t, you use the t distribution.

If n is small, population is normal and σ is unknown, should call it t.

If n is large, calling it z is correct.

If n is large and population is normal and σ is unknown, calling it z is correct and calling it t is also correct. Technically, if you really know population is normal, then calling it t would be preferable. But when n is large p-value will be essentially the same anyway.

$n \geq 30$ is often considered large sample size for quantitative data.

For 0-1 data, must have ≥10 of each type in your sample.

# T-test versus Z-test.

For 1 sample categorical data, iid observations.

z = $(\bar{x} - \mu) \div (s/\sqrt{n})$ . If you know σ, use σ in place of s.

Never use t for categorical data because the population cannot be normal.

For 0-1 data,

**p = $\bar{x}$, and**

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

For 0-1 data, must have ≥10 of each type in your sample.

For testing the difference between means of 2 groups for quantitative data, still use

(observed difference - expected difference under Ho) / SE,

where now

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

$s_1$ = standard deviation of group 1,

$s_2$ = standard deviation of group 2.

Here expected difference under Ho is always 0.

For testing the difference in proportions for 2 groups, still use (observed difference - expected difference under Ho) / SE, where now

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

$\hat{p}$ is the pooled proportion.

It is the proportion of 1's in both groups combined.

Again, with 2 groups, expected difference under Ho is always 0.

- 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.

EXAMPLE PROBLEMS.

- In the study on echinacea by O'Neil et al. (2008), with 58 volunteers, which of the following is a valid conclusion based on the data?
- a. This was an observational study, so confounding factors such as the overall health and wealth of the volunteers are a likely explanation for the results.
- b. In this experiment, the echinacea group got sick less than the placebo group, but the difference was not statistically significant, in part because the sample size was so small.
- c. This study showed that echinacea works to prevent colds, but its effect is very minimal.
- d. The explanatory variable is a confounding factor t-test with 95% central limit theorem.
- e. None of the above.

EXAMPLE PROBLEMS.

- In the study on echinacea by O'Neil et al. (2008), with 58 volunteers, which of the following is a valid conclusion based on the data?
- a. This was an observational study, so confounding factors such as the overall health and wealth of the volunteers are a likely explanation for the results.
- **b. In this experiment, the echinacea group got sick less than the placebo group, but the difference was not statistically significant, in part because the sample size was so small.**
- c. This study showed that echinacea works to prevent colds, but its effect is very minimal.
- d.  The explanatory variable is a confounding factor t-test with 95% central limit theorem.
- e. None of the above.

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels, to see if going to UCLA is associated with higher levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

a. Find a 95%-CI for how much less an average UCLA student's blood glucose level is than an average 2nd grader.

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

a. Find a 95%-CI for how much less an average UCLA student's blood glucose level is than an average 2nd grader.

$2.0 \pm 1.96 \sqrt{(1.5^2/100 + 2.2^2/80)} = 2.0 \pm 0.564.$

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

b. Is the difference observed between the mean blood glucose at UCLA and in 2nd grade statistically significant?

# Example problems.

Suppose you sample 100 UCLA students and 80 $2^{nd}$ graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in $2^{nd}$ grade is 7.5 mmol/L, with an SD of 2.2.

b. Is the difference observed between the mean blood glucose at UCLA and in $2^{nd}$ grade statistically significant?

Yes. The 95%-CI does not come close to containing 0.

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

c. Is this an observational study or an experiment?

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

c. Is this an observational study or an experiment?

Observational study.

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

d. Does going to UCLA cause your blood glucose level to drop?

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

d. Does going to UCLA cause your blood glucose level to drop?

No. Age is a confounding factor. Young kids eat more candy.

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

e. The mean blood glucose level of all 43,301 UCLA students is a

parameter random variable t-test

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

e. The mean blood glucose level of all 43,301 UCLA students is a

parameter

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

f. If we took another sample of 100 UCLA students and 80 2nd graders, and used the difference in sample means to estimate the difference in population means, how much would it typically be off by?

# Example problems.

Suppose you sample 100 UCLA students and 80 2nd graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in 2nd grade is 7.5 mmol/L, with an SD of 2.2.

f. If we took another sample of 100 UCLA students and 80 2nd graders, and used the difference in sample means to estimate the difference in population means, how much would it typically be off by? $SE = \sqrt{(1.5^2/100 + 2.2^2/80)} = .288$ mmol/L

# Example problems.

Suppose you sample 100 UCLA students and 80 $2^{nd}$ graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in $2^{nd}$ grade is 7.5 mmol/L, with an SD of 2.2.

g. How much does one UCLA student's blood glucose level typically differ from the mean of UCLA students?

# Example problems.

Suppose you sample 100 UCLA students and 80 $2^{nd}$ graders and record their blood glucose levels. The mean at UCLA is 5.5 mmol/L, with an SD of 1.5, and the mean in $2^{nd}$ grade is 7.5 mmol/L, with an SD of 2.2.

g. How much does one UCLA student's blood glucose level typically differ from the mean of UCLA students?

1.5 mmoL/L.

# Example problems.

- Researchers take a simple random sample of Californians and a simple random sample of Texans to see who does more exercise. They find that the Californians spend 2.5 hours per week exercising on average and the Texans spend 2.0 hours per week exercising on average. The researchers do a 2-sided test on the difference between the two means and find a p-value of 2.3%. Which of the following would be true of 90% and 95% confidence intervals for the weekly mean exercising time for Californians minus the mean exercising time for Texans?
- a. Both the 90% CI and the 95% CI will contain zero.
- b. Neither the 90% CI nor the 95% CI will contain zero.
- c. The 95% CI will not contain zero, but the 90% CI might contain zero.
- d. The 95% CI will contain zero, but the 90% CI might not contain zero.
- e. The 95% CI has a non-response bias in the margin of error due to confounding factors from the observation study on the null hypothesis.

# Example problems.

- Researchers take a simple random sample of Californians and a simple random sample of Texans to see who does more exercise. They find that the Californians spend 2.5 hours per week exercising on average and the Texans spend 2.0 hours per week exercising on average. The researchers do a 2-sided test on the difference between the two means and find a p-value of 2.3%. Which of the following would be true of 90% and 95% confidence intervals for the weekly mean exercising time for Californians minus the mean exercising time for Texans?
- a. Both the 90% CI and the 95% CI will contain zero.
- **b. Neither the 90% CI nor the 95% CI will contain zero.**
- c. The 95% CI will not contain zero, but the 90% CI might contain zero.
- d. The 95% CI will contain zero, but the 90% CI might not contain zero.
- e. The 95% CI has a non-response bias in the margin of error due to confounding factors from the observation study on the null hypothesis.

# Example problems.

- The Physician's Health Study I studied aspirin's effect on reducing the risk of heart attacks. Which of the following was **not** a reason for randomly assigning people to treatment or control in this experiment?
- a. To ensure that the sample is more representative of the overall population.
- b. To ensure that the treatment and control groups are similar with respect to known potential confounders such as diet and exercise.
- c. To ensure that the treatment and control groups are similar with respect to unknown confounding factors.

# Example problems.

- The Physician's Health Study I studied aspirin's effect on reducing the risk of heart attacks. Which of the following was **not** a reason for randomly assigning people to treatment or control in this experiment?
- **a. To ensure that the sample is more representative of the overall population.**
- b. To ensure that the treatment and control groups are similar with respect to known potential confounders such as diet and exercise.
- c. To ensure that the treatment and control groups are similar with respect to unknown confounding factors.