

Friday, March 22, 2019

- Warm-up

- **Big Town Fisheries recently stocked a new lake in a city park with 2,000 fish of various sizes. The distribution of the lengths of these fish is approximately normal. The Fishery claims that the mean length of the fish is 8 inches. If the claim is true, which of the following would be more likely?**

A random sample of 15 fish having a mean length greater than 10 inches

OR

A random sample of 50 fish having a mean length that is greater than 10 inches

Justify your answer.

- Check Homework
- Extend warm-up
- Hypothesis Tests & Confidence Intervals





Big Town Fisheries recently stocked a new lake in a city park with 2,000 fish of various sizes. The distribution of the lengths of these fish is approximately normal. The Fishery claims that the mean length of the fish is 8 inches. If the claim is true, which of the following would be more likely?

A random sample of 15 fish having a mean length greater than 10 inches

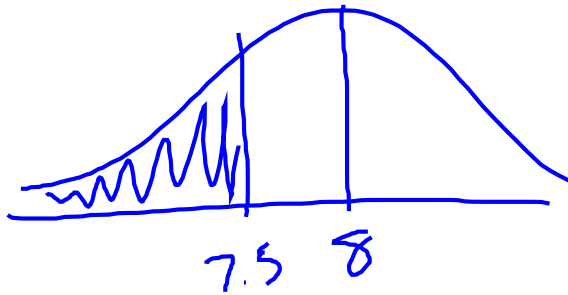
OR

A random sample of 50 fish having a mean length that is greater than 10 inches

Justify your answer.

Extend Warm-up

- Suppose the standard deviation of the sampling distribution of the sample mean for random samples of size 50 is 0.3 inch. If the mean length of the fish is 8 inches, use the normal distribution to compute the probability that a random sample of 50 fish will have a mean length less than 7.5 inches.

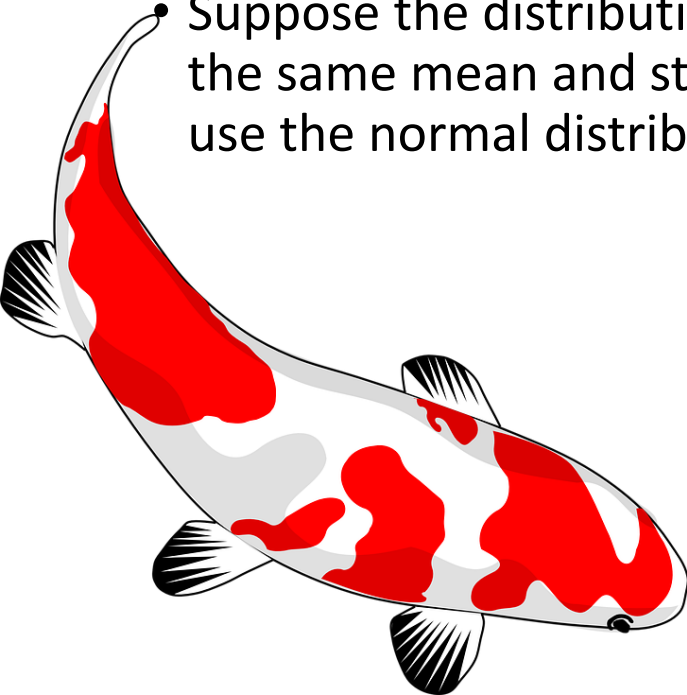


$$s: \frac{0.3}{\sqrt{50}}$$

$$z: \frac{7.5 - 8}{\frac{0.3}{\sqrt{50}}}$$

- Suppose the distribution of fish lengths in this lake was nonnormal but had the same mean and standard deviation. Would it still be appropriate to use the normal distribution to compute the probability?

Yes CLT





Objectives

- Content Objective: I will use the t-distribution to compare means of different samples.
- Social Objective: I will listen and not cause distractions for myself or others.
- Language Objective: I will take clear notes that I can understand when I refer to them later.

Assumptions and Conditions

- **Independence Assumption** (Each condition needs to be checked for both groups):
 - **Randomization Condition**: Were the data collected with suitable randomization (representative random samples or a randomized experiment)?
 - **10% Condition**: We don't usually check this condition for differences of means. We will check it for means only if we have a very small population or an extremely large sample.



Assumptions and Conditions

- **Normal Population Assumption:**
 - **Nearly Normal Condition:** This must be checked for *both* groups. A violation by either one violates the condition.
- **Independent Groups Assumption:** The two groups we are comparing must be independent of each other.



Formulas

Remember that, for independent random quantities, variances add.

So, the standard deviation of the difference between two sample means is

$$SD(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

We still don't know the true standard deviations of the two groups, so we need to estimate and use the standard error

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Two-Sample t -Interval

When the conditions are met, we are ready to find the confidence interval for the difference between means of two independent groups, $\mu_1 - \mu_2$.

The confidence interval is $(\bar{y}_1 - \bar{y}_2) \pm t_{df}^* \times SE(\bar{y}_1 - \bar{y}_2)$

where the standard error of the difference of the means is

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The critical value t_{df}^* depends on the particular confidence level, C , that you specify and on the number of degrees of freedom, which we get from the sample sizes and a special formula.

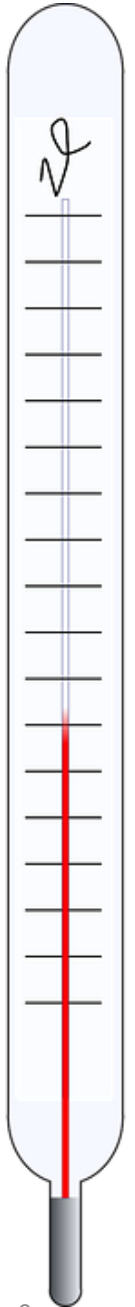


Degrees of Freedom

- The special formula for the degrees of freedom for our t critical value is a bear:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2} \right)^2}$$

- Because of this, we will let technology calculate degrees of freedom for us!



Testing the Difference Between Two Means

- The hypothesis test we use is the two-sample t -test for means.
- The conditions for the two-sample t -test for the difference between the means of two independent groups are the same as for the two-sample t -interval.



Sampling Distribution for the Difference Between Two Means

- When the conditions are met, the standardized sample difference between the means of two independent groups

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)}$$

can be modeled by a Student's t -model with a number of degrees of freedom found with a special formula.

- We estimate the standard error with

$$H_0: \mu_1 - \mu_2 = 0$$
$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



The paper “The Truth About Lying in Online Dating Profiles” (Proceedings, Computer-Human Interactions [2007]: 1-4) describes an investigation in which 40 men and 40 women with online dating profiles agreed to participate in a study. Each participant’s height (in inches) was measured and the actual height was compared to the height given in that person’s online profile. The difference between the online profile height and the actual height (profile – actual) were used to compute the values in the accompanying table.

Men	Women
$\bar{x}_d = 0.57$	$\bar{x}_d = 0.03$
$s_d = 0.81$	$s_d = 0.75$
$n=40$	$n=40$

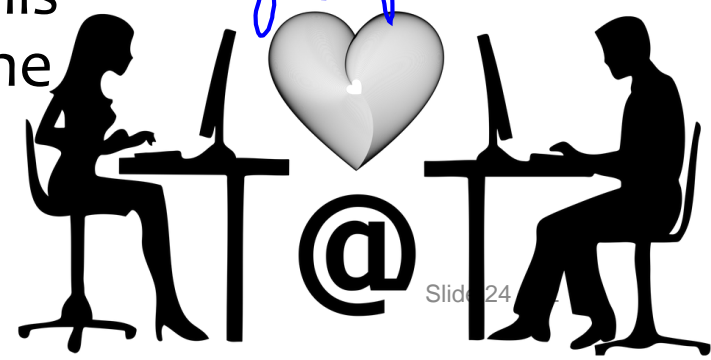
Check conditions...

Men
 Assume representative
 $40 < 10\%$ of all men w/online dating
 $40 > 30$ CLT means nearly normal

Women
 Assume represent.
 $40 < 10\%$ of all women w/online dating
 $40 > 30$ CLT means nearly normal

For purposes of this exercise, assume it is reasonable to regard the two samples in this study as being representative of male online daters and female online daters.

Assume independent groups



Use the two-sample t test to test if males overestimate their height significantly more than females.

2 sample t -test

$$t_{77.5} = 3.09$$

$$p\text{-value} = 0.001$$

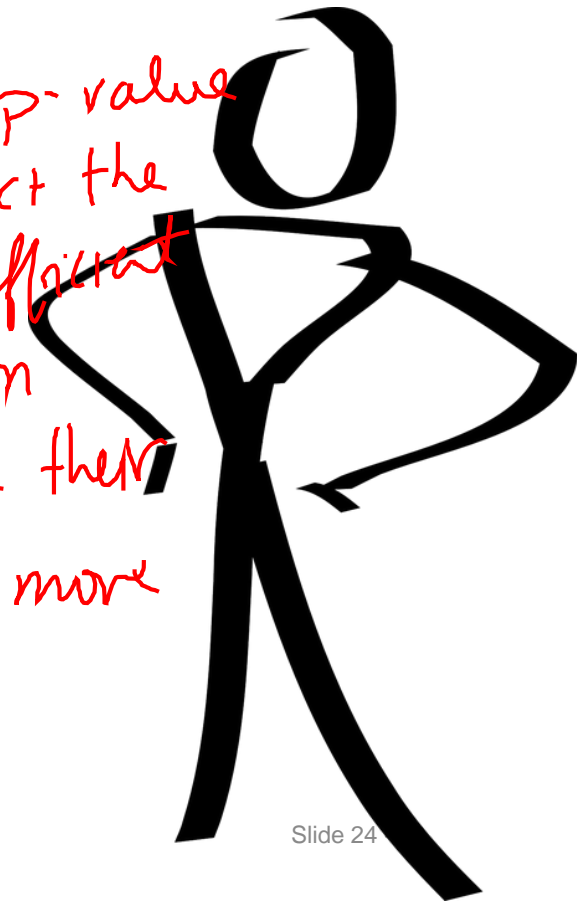
$$df = 77.5$$

$$H_0: \mu_m = \mu_f$$

$$H_A: \mu_m > \mu_f$$

Due to a low p -value of 0.001 we reject the null. There is sufficient evidence that males on average overestimate their height significantly more than females.

Men	Women
$\bar{x}_d = 0.57$	$\bar{x}_d = 0.03$
$s_d = 0.81$	$s_d = 0.75$
$n=40$	$n=40$



Create a 95% confidence interval to approximate the average overestimation of height difference between men and women.

2 sample confidence interval


Men	Women
$\bar{x}_d = 0.57$	$\bar{x}_d = 0.03$
$s_d = 0.81$	$s_d = 0.75$
$n=40$	$n=40$

$$df = 77.5$$

$$0.54 \pm 0.34$$

$$(0.192, 0.887)$$

I am 95% confident that the true mean difference in height estimation is between 0.192 and 0.887.



Homework
P 584
(27-30)