KEY

Notes 10.2a: Comparing Two Means

1. **If we want to compare the mean of some quantitative variable for the individuals in Population 1 and Population 2?**
   - The best approach is to take separate random samples from each population and to compare the sample means.
   - Our parameters of interest are the population means μ_1 and μ_2.
2. **Suppose we want to compare the average effectiveness of two treatments in a completely randomized experiment.**
   - The parameters $\mu 1$ and $\mu 2$ are the true mean responses for Treatment 1 and Treatment 2. We use the mean response in the two groups to make the comparison.

Here's a table that summarizes these two situations:

Population or treatment	Parameter	Statistic	Sample size
1	μ_1	$\bar{x}_1$	n_1
2	μ_2	$\bar{x}_2$	n_2

The Sampling Distribution of the Difference Between Sample Means

Choose
- an SRS of size n_1 from Population 1 with mean μ_1 and std. dev. σ_1
- an SRS of size n_2 from Population 2 with mean μ_2 and std. dev. σ_2.
- **The samples MUST be independent**

* ADDITIONAL CONDITION FOR 2 SAMPLE CI's & TESTS.

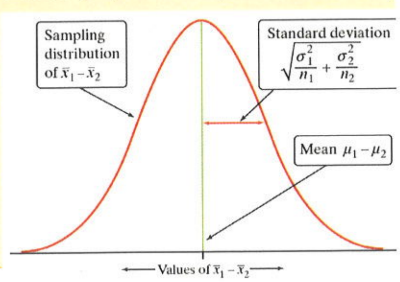
Center The mean of the sampling distribution is an **unbiased estimator** of the difference in population means $\mu_{\bar{x}_1-\bar{x}_2}=\mu_{\bar{x}_1}-\mu_{\bar{x}_2}=\mu_1-\mu_2$

Spread $\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_2}}$

This formula on Green sheet!

As long as each sample is no more than 10% of its population (10% condition,

Shape of the sampling distribution of is approximately normal when…
1) Both population distributions are Normal
2) When both sample sizes are large enough ($n_1 \geq 30$ and $n_2 \geq 30$)



3) Small samples - You must graph (histograms) BOTH samples to look for skewness and/or severe outliers

The Two-Sample t Statistic

Important Formula:

When data come from two random samples or two groups in a randomized experiment, the statistic $\bar{x}_1 - \bar{x}_2$ is our best guess for the value of $\mu_1 - \mu_2$.

When the Independent condition is met, the standard deviation of the statistic $\bar{x}_1 - \bar{x}_2$ is :

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Since we don't know the values of the parameters σ_1 and σ_2, we replace them in the standard deviation formula with the sample standard deviations. The result is the **standard error** of the statistic $\bar{x}_1 - \bar{x}_2$: $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

If the Normal condition is met, we standardize the observed difference to obtain
a t statistic th ... served difference is from its mean in standard devia

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

TEST STATISTIC FOR 2 SAMPLE MEANS is "t".

The two-sample t statistic has approximately a t distribution. We can use technology to determine degrees of freedom OR we can use a conservative approach, using the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom.

DF

2 WAYS TO DETERMINE "DF":
1. TECHNOLOGY - Calculator DF
2. CONSERVATIVE - DF is based on the smaller sample size

* MUST STATE METHOD USED!!

Confidence Intervals for $\mu_1 - \mu_2$

Two-Sample t Interval for a Difference Between Means

When the Random, Normal, and Independent conditions are met, an approximate level C confidence interval for $(\bar{x}_1 - \bar{x}_2)$ is

Formula FOR 2 Sample CI.

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^* is the critical value for confidence level C for the t distribution with degrees of freedom from either technology or the smaller of $n_1 - 1$ and $n_2 - 1$.

Significance Tests for $\mu_1 - \mu_2$

Two-Sample t Test for the Difference Between Two Means

If the Random, Normal, and Independent conditions are met, we can proceed:

To do a test, standardize $\bar{x}_1 - \bar{x}_2$ to get a two-sample t statistic:

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

To find the P-value, use the t distribution with degrees of freedom given by technology or by the conservative approach (df = smaller of n_1 - 1 and n_2 - 1).

2 Sample t statistic for T.O.H.

- **Example:** Who's Taller at Ten, Boys or Girls?
- ***Describing the Sampling Distribution of a Difference Between 2 Means***

EXAMPLE: Based on information from the U.S. National Health and Nutrition Examination Survey (NHANES), the heights (in inches) of ten-year-old girls follow a Normal distribution *N*(56.4, 2.7). The heights (in inches) of ten-year-old boys follow a Normal distribution *N*(55.7, 3.8). A researcher takes independent SRSs of 12 girls and 8 boys of this age and measures their heights. After analyzing the data, the researcher reports that the sample mean height of the boys is larger than the sample mean height of the girls

a) Describe the shape, center, and spread of the sampling distribution of $\bar{x}_f - \bar{x}_m$.

b) Find the probability of getting a difference in sample means $\bar{x}_f - \bar{x}_m$ that is less than 0.

c) Does the result in part (b) give us reason to doubt the researchers' stated results?

2 POPULATIONS

FEMALE HEIGHTS $\mu_F = 56.4$in $N(56.4, 2.7)$

MALE HEIGHTS $\mu_M = 55.7$ $N(55.7, 3.8)$

SRS
$n_1 = 12$
$n_2 = 8$

(A) THE SAMPLING DISTRIBUTION OF $\bar{X}_F - \bar{X}_M$

(1) SHAPE: Since both population distributions are normal, THE SAMPLING DISTRIBUTION OF $\bar{X}_F - \bar{X}_M$ IS APPROXIMATELY NORMAL

(2) CENTER: $\underbrace{\mu_{\bar{X}_F - \bar{X}_M}}_{\text{Sampling Distribution}} = \underbrace{\mu_F - \mu_M}_{\text{Pop. Parameters}} = 56.4 - 55.7$

$= \boxed{0.7 \text{ INCHES}}$

(3) SPREAD:

$$\sigma_{\bar{X}_F - \bar{X}_M} = \sqrt{\frac{\sigma_F^2}{n_F} + \frac{\sigma_M^2}{n_M}}$$

$$= \sqrt{\frac{2.7^2}{12} + \frac{3.8^2}{8}} = \boxed{1.55 \text{ inches}}$$

(B) FIND THE PROBABILITY OF GETTING DIFFERENCE $(\bar{X}_F - \bar{X}_M)$ LESS THAN 0.

WRITE A PROBABILITY STATEMENT →

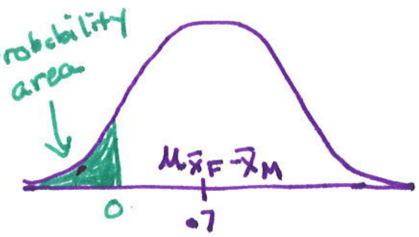
$$P(\bar{X}_F - \bar{X}_M < 0) = .3258$$

THIS IS EQUIVALENT TO $P(\bar{X}_F < \bar{X}_M)$

THAT IS "MEAN HT OF 10 YEAR OLD GIRL IS SHORTER THAN A BOY"

OR "mean HT OF A BOY IS TALLER THAN A GIRL" $P(\bar{X}_M > \bar{X}_F)$

Graph the Sampling distribution →



TO FIND THE PROBABILITY:

normalcdf(-E99, 0, .7, 1.55) = .3258

LB, UB, μ, σ

OR FIND PROBABILITY BY STANDARDIZING Z:

$N(0,1) \longrightarrow Z = \frac{0 - .7}{1.55} = -.45$

$P(Z \leq -.45) = .3264$ $\boxed{\text{normalcdf}(-E99, -.45, 0, 1)}$

(C) Researcher claims boys are taller than girls (at 10 years old)

Based on these results, there is about a 33% chance of getting a sample mean difference less than zero (0) due to sampling variability.

THIS MEANS THAT WE WOULD EXPECT 1 OUT 3 10 year old boys to be taller than girls.

Since this is not an unusual result, we should not doubt the researchers claim that boys are taller than girls.

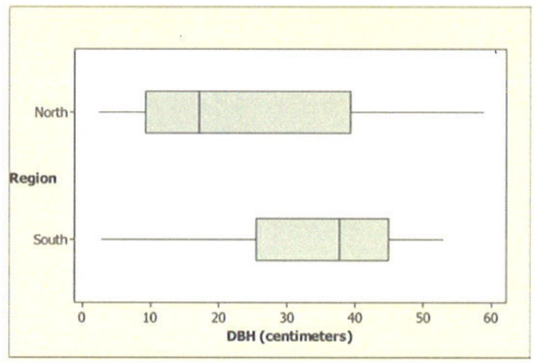
Two-Sample *t* Interval for a Difference Between Means

- **Example:** Big Trees, Small Trees, Short Trees, Tall Trees

The Wade Tract Preserve in Georgia is an old-growth forest of longleaf pines that has survived in a relatively undisturbed state for hundreds of years. One question of interest to foresters who study the area is "How do the sizes of longleaf pine trees in the northern and southern halves of the forest compare?" To find out, researchers took random samples of 30 trees from each half and measured the diameter at breast height (DBH) in centimeters. Comparative boxplots of the data and summary statistics from Minitab are shown below. Construct and interpret a 90% confidence interval for the difference in the mean DBH for longleaf pines in the northern and southern halves of the Wade Tract Preserve.

Descriptive Statistics: North, South

	Variable	N	Mean	StDev
μ_2	North	30	23.70 ✓	17.50
μ_1	South	30	34.53 ✓	14.26



NAME OF INTERVAL:
2 sample t-interval for the difference of means $(\mu_1 - \mu_2)$

TIP: look at the sample means and make μ_1 the larger mean to have a positive difference.

DEFINE PARAMETERS:

μ_1 = TRUE mean diameter of trees in the SOUTH $\bar{X}_1 = 34.53$

μ_2 = TRUE mean diameter of trees in the NORTH $\bar{X}_2 = 23.70$

SIGNIFICANCE LEVEL:
Want to estimate the difference $(\mu_1 - \mu_2)$ at the 90% C.L. $\alpha = .10$

CONDITIONS:

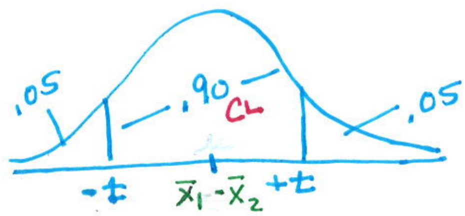
σ IS UNKNOWN (t-interval)

RANDOM – Random samples from both the NORTH and SOUTH

INDEPENDENT – ① The samples from the North and South are independent.
② IT IS FAIR TO ASSUME THAT THERE ARE AT LEAST 10(30) = 300 TREES IN EACH REGION.

NORMAL – SINCE BOTH SAMPLE SIZES ARE RELATIVELY LARGE WITH SAMPLE SIZES OF 30, IT IS REASONABLE BOTH DISTRIBUTIONS ARE APPROXIMATELY NORMAL.

Construct a 2 Sample TInterval for $\mu_1 - \mu_2$



$\boxed{\alpha = .10}$

Hand Calculate

Formula:

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Conservative df = n−1 of smaller of the 2 samples

df = 29

$t^* = \text{invT}(.05, \underset{\uparrow DF}{29})$ $\quad t^* = \pm 1.70$

$$34.53 - 23.70 \pm (1.70)\sqrt{\frac{14.26^2}{30} + \frac{17.50^2}{30}}$$

$10.83 \pm (1.70)\underset{\text{SE}}{(4.12)}$

$10.83 \pm \underset{\text{ME}}{7.01}$ $\quad (3.82, 17.84)$

Calculator

2-Samp TInterval

South	North
$\bar{x}_1 = 34.53$	$\bar{x}_2 = 23.70$
$Sx_1 = 14.26$	$Sx_2 = 17.50$
$n_1 = 30$	$n_2 = 30$

Pooled No ← Always

(3.9362, 17.724)

df = 55.7

Technology. Complex Formula do not need to know

* You must state what df you used (Conservative or Technology)!!!

Conclude:

We are 90% confident that the interval 3.83 to 17.83cm captures the true difference in the actual mean DBH between the southern and northern trees.

This suggests that the mean diameter of southern trees is between 3.83 and 17.83cm larger than the mean diameter of the northern trees, on average.

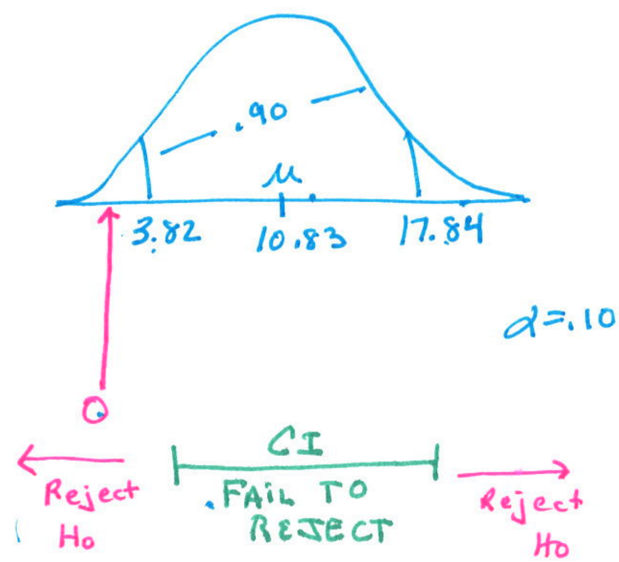
TREE EXAMPLE

DISCUSSION

BASED ON THE CI, IS THERE CONVINCING EVIDENCE THAT THE TREE DIAMETER IS DIFFERENT BETWEEN THE NORTH AND SOUTH?

$H_0: \mu_N = \mu_S$

$H_A: \mu_N \neq \mu_S$



SIGNIFICANCE LEVEL
$\alpha = .10$ 10% Chance of making a TYPE 1 ERROR

SINCE 0 IS NOT IN THE CI, We have convincing evidence to reject Ho and believe the diameters in North + South are different.

■ **Example:** Calcium and Blood Pressure

Does increasing the amount of calcium in our diet reduce blood pressure? Examination of a large sample of people revealed a relationship between calcium intake and blood pressure. The relationship was strongest for black men. Such observational studies do not establish causation. Researchers therefore designed a randomized comparative experiment. The subjects were 21 healthy black men who volunteered to take part in the experiment. They were randomly assigned to two groups: 10 of the men received a calcium supplement for 12 weeks, while the control group of 11 men received a placebo pill that looked identical. The experiment was double-blind. The response variable is the decrease in systolic (top number) blood pressure for a subject after 12 weeks, in millimeters of mercury. An increase appears as a negative response Here are the data:

Group 1 (calcium):	7	−4	18	17	−3	−5	1	10	11	−2	
Group 2 (placebo):	−1	12	−1	−3	3	−5	5	2	−11	−1	−3

EXAMPLE CALCIUM + BLOOD PRESSURE

2 SAMPLE t-test for $\mu_1 - \mu_2$

GRAPHS

GROUP 1 (calcium) $\bar{X}_1 = 5.0$ $S_1 = 8.74$ $n_1 = 10$

GROUP 2 (placebo) $\bar{X}_2 = -.27$ $S_2 = 5.90$ $n_2 = 11$

PARAMETERS: μ_1 = true mean decrease in blood pressure (Calcium Supplements)

μ_2 = true mean decrease in blood pressure (PLACEBO)

HYPOTHESIS

$H_0: \mu_1 - \mu_2 = 0$ $H_0: \mu_1 = \mu_2$

$H_A: \mu_1 - \mu_2 > 0$ $H_A: \mu_1 > \mu_2$

SIGNIFICANCE LEVEL: $\alpha = .05$

SKETCH GRAPH:

CONDITIONS: Random, Normal, Independent, σ

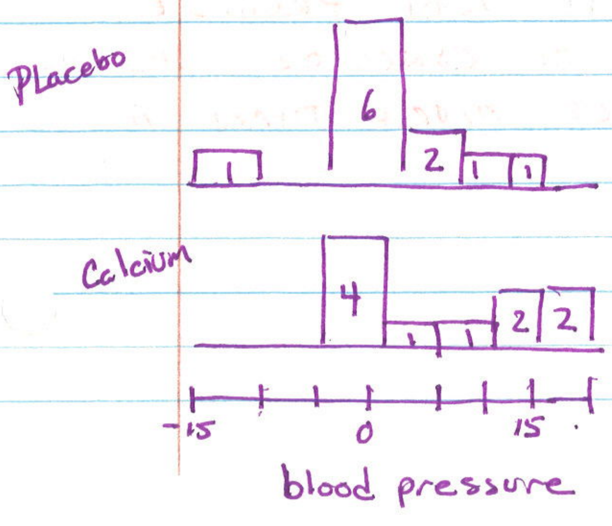
• Random – 21 subjects were randomly assigned TO THE 2 TREATMENTS.

INDEPENDENT: ① Due to Random assignment, these 2 groups can be viewed as independent.

② Reasonable individual observations are independent

Normal: Since samples are both under 30, we looked at graphs (above) and do not show clear evidence of skewness & no outliers

σ UNKNOWN (t inference)



STATE TEST BY NAME OR FORMULA

NAME: $\boxed{\text{2 Sample T test for } \mu_1 - \mu_2}$

TEST STATISTIC $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$

STATE DF: DF = 9 (Conservative)

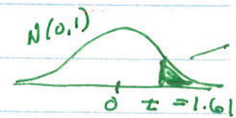
$$t = \frac{(5 - (-.27)) - 0}{\sqrt{\frac{8.74^2}{10} + \frac{5.90^2}{11}}} = \frac{5.27}{3.28} = 1.61 \qquad t = 1.61$$

SE

PVALUE

Pvalue =

State probability: $P(t \geq 1.61) = .071$



tcdf(1.61, E99, 9)

find pvalue

write decision in conclusion

Check w/ CALC: STAT TEST 2 SAMPTTEST

* ALWAYS USE Pooled $\boxed{NO}$ ← we are not going to pool variances

What you need to write using calc.

$t =$ 1.60 df = 15.6 (technology) pvalue = $P(t \geq 1.60) = .064$

CONCLUSION BECAUSE THE PVALUE IS GREATER THAN $\alpha = .05$, WE FAIL TO REJECT H_0

THE EXPERIMENT DID NOT PROVIDE CONVINCING EVIDENCE TO CONCLUDE CALCIUM REDUCES BLOOD PRESSURE MORE THAN A PLACEBO

STATE TEST BY NAME OR FORMULA

NAME: 2 Sample T test for $\mu_1 - \mu_2$

TEST STATISTIC $$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

STATE DF:

PVALUE

State probability:
find pvalue
write decision in conclusion

Check w/ CALC: STAT TEST 2 SAMPTTEST
* ALWAYS USE Pooled $\boxed{\text{NO}}$ ← we are not going to pool variances

$t =$ ______ $df =$ __ $p =$ ______

CONCLUSION BECAUSE THE PVALUE IS GREATER THAN $\alpha = .05$, WE FAIL TO REJECT Ho

THE EXPERIMENT DID NOT PROVIDE CONVINCING EVIDENCE TO CONCLUDE CALCIUM REDUCES BLOOD PRESSURE MORE THAN A PLACEBO

EXAMPLE — CALCIUM + BLOOD PRESSURE

2 SAMPLE t-test for $\mu_1 - \mu_2$

GRAPHS

GROUP 1 (calcium) $\bar{x}_1 =$ $s_1 =$ $n_1 =$

GROUP 2 (placebo) $\bar{x}_2 =$ $s_2 =$ $n_2 =$

PARAMETERS: $\mu_1 =$

$\mu_2 =$

HYPOTHESIS

SIGNIFICANCE LEVEL:

SKETCH GRAPH:

CONDITIONS: Random, Normal, Independent, σ

- Random – 21 subjects were randomly assigned to the 2 treatments.

INDEPENDENT: ① Due to Random assignment, these 2 groups can be viewed as independent.
② Reasonable individual observations are independent

Normal: Since samples are both under 30, we looked at graphs (above) and do not show clear evidence of skewness & no outliers

σ UNKNOWN (t inference)