AP Statistics
Unit 07 – Day 03 Notes
Confidence Intervals: 2-Sample Means

Name Key
Period______

If we choose independent SRSs of size n_1 from Population 1 and size n_2 from Population 2, then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ has the following properties:

- **Shape:** Normal if both population distributions are Normal; approximately Normal otherwise if both samples are large enough ($n_1 \geq 30$ and $n_2 \geq 30$) by the central limit theorem.
- **Center:** Its mean is $\mu_1 - \mu_2$.
- **Spread:** As long as each sample is no more than 10% of its population, its standard deviation is:

$$\boxed{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Before estimating or testing a claim about $\mu_1 - \mu_2$, check that the **conditions** are met:

- **Random:** The data are produced by independent random samples of size n_1 from Population 1 and of size n_2 from Population 2 or by two groups of size n_1 and n_2 in a randomized experiment.
- **10%:** When sampling without replacement, check that the two populations are at least 10 times as large as the corresponding samples.
- **Normal/Large:** Both population distributions (or true distributions of responses to the two treatments) are Normal or both sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$). If either population (treatment) distribution has unknown shape and the corresponding sample size is less than 30, use a graph of the sample data to assess the Normality of the population (treatment) distribution. Do not use two-sample t-procedures if the graph shows strong skewness or outliers.

*Also, be sure not to use a two-sample t-procedure to compare means for **PAIRED DATA!** We use **paired t procedures** for that.*

TWO-SAMPLE T-INTERVAL FOR THE DIFFERENCE BETWEEN MEANS

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

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df}^* \cdot SE(\bar{x}_1 - \bar{x}_2)$$

where the standard error of the difference of means

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

The critical value t^*_{df} depends on the particular confidence level, C, that you specify. It also depends on the number of degrees of freedom. **You have two options here for finding degrees of freedom:**

- **Option 1 (Technology):** Use the t distribution with degrees of freedom calculated from the data by the formula below:

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

← nope! ew. (but the calc does it)

- **Option 2 (Conservative):** Use the t distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$. With this option the resulting confidence interval has a margin of error as *large as or larger than* is needed for the desired confidence level. The significance test using this option (discussed later) gives a P-value *equal to or greater than* the true P-value. As the sample sizes increase, confidence levels and P-values from this option become more accurate. *we use this one :) when doing it by hand

ON YOUR CALCULATOR: option 1

pooled?

CI p	CI u	ST p	ST u

1. STAT → TESTS → 2-SampTInt
2. Choose Stats as the input method and enter the su
   (or choose data & choose your lists)
3. Choose "No" for pooling (we will discuss this later)
4. Press Enter!

*You must NAME THE TEST in **PLAN**, as well as directly st

Example 1: Medium or Large Drink?

A fast-food restaurant uses an automated filling machi ... different settings for small, medium, and large drink cu ... turer, when the large setting is chosen, the amount of liquid L ... mal distribution with mean 27 ounces and standard deviation 0.8 ounces. When the medium setting is chosen, the amount of liquid M dispensed follows a Normal distribution with mean 17 ounces and standard deviation 0.5 ounces. To test the manufacturer's claim, the restaurant manager measures the amount of liquid in each of 20 random cups filled with the large setting and 25 random cups filled with the medium setting. A graph of each set of data shows no outliers and no strong skewness. Let $\bar{x}_L - \bar{x}_M$ be the difference in the sample mean amount of liquid under the two settings.

a. What is the shape of the sampling distribution of $\bar{x}_L - \bar{x}_M$? Why?

The sampling distribution is approximately Normal because both population distributions are Normal.

b. Construct and interpret a 90% confidence interval for the true mean difference amount of liquid between large cups and medium cups.

State: μ_1 = true mean amount of liquid in large cups (oz)
μ_2 = true mean amount of liquid in medium cups (oz)
$\bar{x}_1 = 27$ oz   we want to find the difference in means
$\bar{x}_2 = 17$ oz   (in liquid between large and medium cups) with 90% confidence $(\mu_1 - \mu_2)$.

Plan: Random: our data come from 2 independent Random samples ✓
10% condition: $200 <$ all large cups filled ✓
$n_1 = 20$
$n_2 = 25$ $250 <$ all medium cups filled ✓
Normal/Large: $n_1 = 20 \cancel{\geq} 30$ $n_2 = 25 \cancel{\geq} 30$
but no skew and no outliers stated ✓   but no skew and no outliers stated ✓

because our conditions are met, we will use a 2-sample t-interval for difference of means $\mu_1 - \mu_2$.

Do: $(27-17) \pm 1.729\sqrt{\frac{0.8^2}{20} + \frac{0.5^2}{25}} = (9.646, 10.354)$

$df = 20 - 1 = 19$
$t^*_{19} = 1.729$

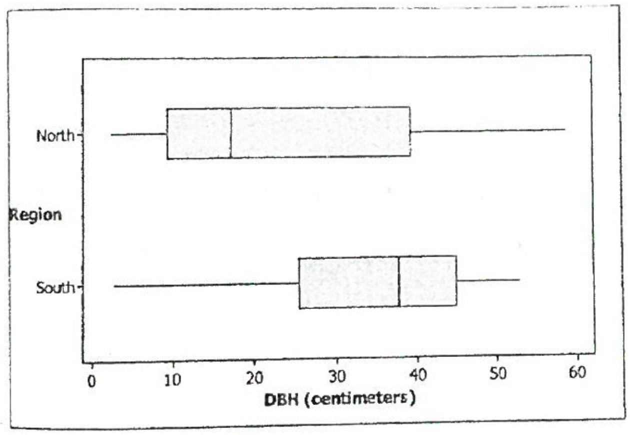
Conclude: we are 90% confident that the interval from 9.646 to 10.354 oz contains the mean difference in liquid between the large and medium cups $(\mu_1 - \mu_2)$.

Example 2: 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. Here are comparative boxplots of the data and summary statistics:

Descriptive Statistics: North, South

Variable	N	Mean	StDev
North	30	23.70	17.50
South	30	34.53	14.26



Construct and interpret a 90% confidence interval for the difference in the mean DBH of longleaf pines in the northern and southern halves of the Wade Tract Preserve.

State: μ_1 = true mean DBH of long leaf pines in the southern half of the Wade Tract Preserve.
μ_2 = true mean DBH of long leaf pines in the Northern half of the Wade Tract Preserve. in cm (WTP) in cm

We want to find the difference in means ($\mu_1 - \mu_2$) (of DBH between the Northern and Southern long leaf pines in the Wade Tract presere) with 90% confidence.

$\bar{x}_1 = 34.53$ cm $\quad \bar{x}_2 = 23.70$ cm

Plan: Random: random samples from each half ✓
10% condition: 300 < all Northern long leaf pines in WTP ✓
300 < all Southern long leaf pines in WTP ✓

Normal/Large: $n_1 = 30 \geq 30$ ✓
$n_2 = 30 \geq 30$ ✓

Because our conditions are met, we will use a 2-sample t-interval for difference in means $\mu_1 - \mu_2$

DO: $(34.53, 23.70) + 1.699\sqrt{\dfrac{14.26^2}{30} + \dfrac{17.50^2}{30}}$

df = 30 − 1 = 29

$t^*_{29} = 1.699$

$= (3.83, 17.83)$

Conclude: We are 90% confident that the interval from 3.83 cm to 17.83 cm captures the true difference in mean DBH between Northern and Southern long leaf pines in WTP ($\mu_1 - \mu_2$).