10.2 Comparing 2 means

When we compare 2 population means, we use their difference, $\mu_1 - \mu_2$. The statistic used to estimate this difference is the difference of sample means $\bar{x}_1 - \bar{x}_2$.

Sampling distribution of $\bar{x}_1 - \bar{x}_2$

Choose an SRS of n_1 from the first population with mean μ_1 and standard deviation σ_1 and an SRS of size n_2 from the second population with mean μ_2 and standard deviation σ_2.

Shape: When the population distributions are Normal, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is also Normal. If they are not Normal the sampling will be approximately Normal if $n_1 \geq 30$ and $n_2 \geq 30$.

Center: mean of the sampling distribution

$\mu_{\bar{x}_1 - \bar{x}_2}$ is $\mu_1 - \mu_2$

Spread: standard deviation of the sampling distribution, $\sigma_{\bar{x}_1 - \bar{x}_2}$ is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Two sample z - statistic

If the population standard deviations are known, the two sample z statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

This is very unlikely to happen since we hardly ever know the population standard deviations

Two sample t statistic

Since we typically do not know σ, we will estimate using sample standard deviations s_1 & s_2.

This gives us a standard error of:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Therefore the t statistic is:

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

Conditions for inference (CI + significance tests) about $\mu_1 - \mu_2$

1. Random: Data comes from 2 independent random samples or randomized experiments.
2. 10% condition: $n_1 \leq \frac{1}{10} N_1$ and $n_2 \leq \frac{1}{10} N_2$
3. Large counts: both populations are Normal
   - If not Normal, $n_1 \geq 30$ AND $n_2 \geq 30$.
   - If not normal and not greater than 30, graph and assess for skewness & outliers.

Degrees of Freedom to find t^*

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

OR

② use the smaller of n_1-1 and n_2-1.

2 sample t-interval for $\mu_1 - \mu_2$

When conditions are met, an approximate C% CI for $\mu_1 - \mu_2$ is

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

where t^* is the critical value with C% of its area between $-t^*$ and t^* for the t distribution with appropriate df.

Ex: Check your understanding on p. 644

State: We will perform a 99% CI for $\mu_1 - \mu_2$ where μ_1 = true mean wheat price in July and μ_2 = true mean wheat price in September.

Plan:

Random: independent random samples

10%: $90 \leq 10\%$ of all wheat producers in July
$45 \leq 10\%$ of all wheat producers in Sept.

Large counts: $90 \geq 30$ and $45 \geq 30$.

We will use the 2 sample t interval for $\mu_1 - \mu_2$

Do: Using 2 sample t interval on calculator
with $\bar{x}_1 = 2.95$, $s_1 = .22$, $n_1 = 90$ and
$\bar{x}_2 = 3.61$, $s_2 = .19$, $n_2 = 45$

when get the CI of $(-.7561, -.5639)$ with 100.4 df.

Conclude: We are 99% confident the interval of $-.7561$ to $-.5639$ capture the true difference in mean wheat price in July and September.

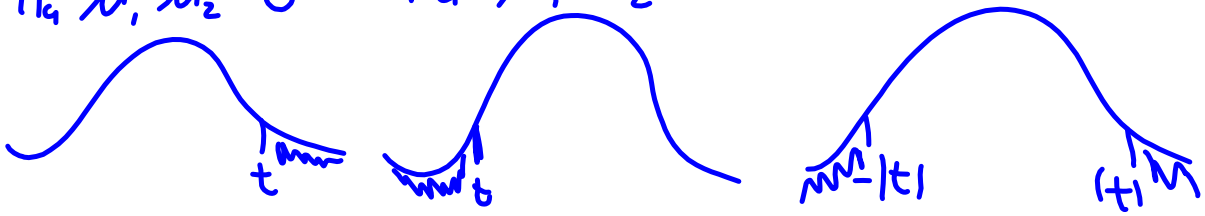
2 sample t-test for $\mu_1 - \mu_2$

Suppose the conditions are met. To test $H_0: \mu_1 - \mu_2 = 0$, compute the t statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Find the probability of getting a t statistic this large or larger in the direction of H_a.

$H_a: \mu_1 - \mu_2 > 0$ $H_a: \mu_1 - \mu_2 < 0$ $H_a: \mu_1 - \mu_2 \neq 0$



check your understanding on p. 649

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

$H_a: \mu_1 - \mu_2 > 0$

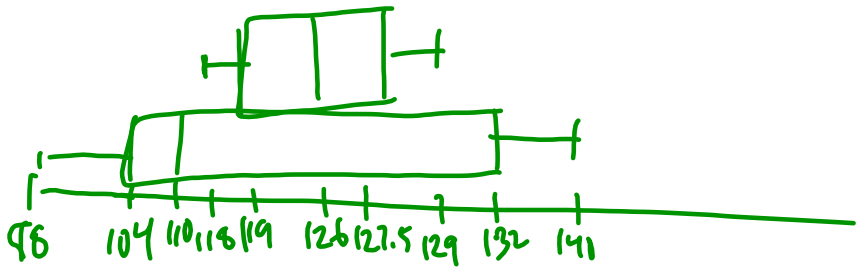
where μ_1 = true mean breaking strength of polyester after 2 weeks.

μ_2 = true mean breaking strength of polyester after 16 weeks

We will use a $\alpha = .05$ for a significance level.

Plan: Random: it is a randomly designed experiment

Large counts: The graphs show no strong skew or outliers.



We will perform a 2 sample t test for $\mu_1 - \mu_2$

Do: on calc, use 2 sample t test,

with $\bar{x}_1 = 123.8$, $s_1 = 4.60$ $n_1 = 5$

$\bar{x}_2 = 116.4$, $s_2 = 16.09$ $n_2 = 5$

This gives us a t statistic of .9894 with df = 4.65 and a p value of .186.

Conclude: Fail to reject H_0 since .186 > .05.

meaning there is not convincing evidence that the polyester decays more in 16 weeks than 2 weeks in _____.

Using t procedures wisely

1. use two sample t procedures ($\mu_1 - \mu_2$) if we do inference on 2 distinct groups of subjects.
2. use paired t procedures (μ_D) if inference is done on 1 group of subjects receiving both treatments.