Sets 28 and 29

Instead of drawing samples from one population, we may take random samples from two populations so that we can carry out some sort of comparison.

We may wish to compare μ_1 and μ_2, the population means for populations 1 and 2. We do so by examining the difference, $\mu_1 - \mu_2$.

Example: Suppose we wish to compare the μ_1, the mean lead content (in ppm) per mL in Victoria tap water with μ_2, the mean lead content (in ppm) per mL in Vancouver tap water.

- If the means are equal, then $\mu_1 - \mu_2 = 0$.
- If the means are different, then $\mu_1 - \mu_2 \neq 0$.
- If the lead content is higher in Victoria, then $\mu_1 - \mu_2 > 0$.
- If the lead content is higher in Vancouver, then $\mu_1 - \mu_2 < 0$.
- If the lead content is higher in Victoria by at least $4\ ppm$ than in Vancouver, then $\mu_1 - \mu_2 > 4$
- If the lead content is higher in Vancouver by at least $2\ ppm$ than in Victoria, then $\mu_1 - \mu_2 < -2$

The point estimate for $\mu_1 - \mu_2$ we will use is $\overline{x}_1 - \overline{x}_2$.

The following pivotal quantities apply to a variety of cases where two samples are drawn **independently** from *two* populations.

As before, any confidence interval we construct will have the form *estimate* $\pm$ *(c.v.)(e.s.e.)*

All of the pivotal quantities below have the form $\dfrac{\text{estimate} - \text{parameter}}{\text{e.s.e.}}$.

Large Sample Size Procedures:

$$Z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)}}$$

Assumptions:

- Independent random samples from two populations.
- **Both** sample sizes are large ($n_1 \geq 40, n_2 \geq 40$) and the population standard deviations are unknown.
- Populations may have any distribution.

Example: We wish to compare the cube compressive strength (in N/mm^2) of two types of concrete. The summary statistics are as follows:

	sample size	sample mean	sample sd
Type A	70	31.9	1.4
Type B	50	35.6	2.1

Test the research hypothesis that the two types of concrete have different mean cube compressive strengths.

Small Sample Size Procedures: In the case where n_1 or n_2 (or both) are small, there are two choices of test statistic. This choice depends on whether or not we assume that $\sigma_1 = \sigma_2$.

While there are hypothesis tests that we could carry out to investigate whether or not $\sigma_1 = \sigma_2$, we will use the following rule for our class:

- Using s_1, s_2, calculate the *larger* standard deviation divided by the *smaller*
- If this value is **less than 1.4**, we assume $\sigma_1 = \sigma_2$. If it is **greater than 1.4**, we assume $\sigma_1 \neq \sigma_2$.

Note that we only need to decide whether or not $\sigma_1 = \sigma_2$ when the sample size is **not** large.

Pooled Procedures:

$$t_{n_1+n_2-2} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

Assumptions:

- Independent random samples from two populations. At least one of the sample sizes is small, and the population standard deviations are unknown.
- We know that (or assume that) $\sigma_1 = \sigma_2$
- Both populations have normal (or approximately normal) distribution.

Comments: The value $\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ is sometimes denoted by s_p^2, and is called the **pooled variance estimate**.

Recall that for pooled procedures, we assumed that $\sigma_1 = \sigma_2$. The value of s_p^2 is the estimate for both σ_1^2 and σ_2^2.

Unpooled Procedures:

$$t_\gamma = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

where γ, the number of degrees of freedom, is the integer part of

$$\gamma = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Assumptions:

- Independent random samples from two populations. At least one of the sample sizes is small, and the population standard deviations are unknown.
- We know that (or assume that) $\sigma_1 \neq \sigma_2$
- Both populations have normal (or approximately normal) distribution.

Comments: Unsurprisingly, the value of γ is almost never an integer. We always round *down* (i.e. take the integer part) as the number of degrees of freedom when we are using our tables.

If you are carrying out an unpooled test on $\mu_1 - \mu_2$ using R (or other software), the p-value will be calculated using the unrounded value of γ as the degrees of freedom.

Example: We wish to compare the lifespans of smart-phones produces by two companies. Let μ_1, μ_2 be the mean lifespan (in weeks) of smart-phones produced by Company A, B (respectively). The summary of our study's observations are as follows:

Company A	Company B
$\overline{x}_1 = 148$	$\overline{x}_2 = 153$
$s_1 = 8.3$	$s_2 = 5.1$
$n_1 = 15$	$n_2 = 6$

(a) What is the estimated standard error of $\overline{x}_1 - \overline{x}_2$?

(b) What probability distribution is used to calculate the p-value in a hypothesis test on $\mu_1 - \mu_2$? (**Note:** It is **not** enough to just say "t-distribution"; you must also specify the number of degrees of freedom)

(c) Test $H_0 : \mu_1 - \mu_2 = 0, H_a : \mu_1 - \mu_2 < 0$, at the significance level $\alpha = 0.1$.

Example: A study is conducted to compare the lead concentration in the water from sources 1 and 2. Let μ_1 be the mean concentration (in ppb) at source 1, and let μ_2 be the mean concentration (in ppb) at source 2. The summary of our observations are as follows.

Source 1	Source 2
$\overline{x}_1 = 100$	$\overline{x}_2 = 80$
$s_1 = 20$	$s_2 = 25$
$n_1 = 15$	$n_2 = 15$

(a) Find a 95% confidence interval for the difference in the mean lead concentrations.

(b) What assumptions (not given in the question) did we need to make in constructing this confidence interval?

(c) Using only the confidence interval, determine whether or not it is reasonable that the two mean concentration levels are the same.

Example: Suppose that we randomly select 10 Belgian Blue bulls and 6 Longhorn bulls and measure their lifespans. Let μ_1 be the mean lifespan of Belgian Blue bulls and μ_2 be the mean lifespan of Longhorn bulls. The lifespans of both types of bulls is known to be normally distributed. We collect the following data:

$$\overline{x}_1 = 17.2, s_1 = 2.9, \overline{x}_2 = 18.5, s_2 = 3.6$$

(a) What is the estimated standard error of $\overline{x}_1 - \overline{x}_2$?

(b) If we were to perform a hypothesis test on $\mu_1 - \mu_2$, what distribution (including degrees of freedom) would we use to calculate the p-value?

(c) Test the hypotheses $H_0 : \mu_1 - \mu_2 = 0$, $H_1 : \mu_1 - \mu_2 \neq 0$.

(d) Create a 95% confidence interval for $\mu_1 - \mu_2$, and interpret.

Example: We are comparing the fat content of chicken eggs where the chickens have been assigned to one of two different diets. We have 12 eggs from chickens on the first diet, and 6 on the second diet. Let μ_1, μ_2 be the mean fat content (in g) of eggs from chickens on diets one and two (respectively). Our summary statistics:

$$\overline{x}_1 = 5.5, s_1 = 0.4, \overline{x}_2 = 6.5, s_2 = 0.7$$

(a) What is the estimated standard error of $\overline{x}_1 - \overline{x}_2$?

(b) If we were to perform a hypothesis test on $\mu_1 - \mu_2$, what distribution (including degrees of freedom) would we use to calculate the p-value?

(c) Create a 95% confidence interval for $\mu_1 - \mu_2$.

(d) Using our data as a pilot study, estimate the sample size we would need in a future study to create a 95% confidence interval for $\mu_1 - \mu_2$ with a margin of error of 0.25. Assume that $n_1 = n_2$ in this future study.

Set 30

Again, we consider the scenario where we take samples independently from two populations. Instead of considering the case where we are comparing means, today we compare p_1 and p_2, the population proportions for the two populations.

As before, in order to compare the two proportions, we need only examine $p_1 - p_2$.

We use $\widehat{p_1} - \widehat{p_2}$ (the difference in sample proportions) as the point estimate for $p_1 - p_2$, where $\widehat{p_1} = x_1/n_1$ (number of successes out of number of trials for the first sample) and $\widehat{p_2} = x_2/n_2$ (number of successes out of number of trials for the second sample.

When there are not too few successes or failures in each group (at least 5 of each for both samples) then the following pivotal quantity has approximately standard normal distribution.

$$Z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$$

Here, the denominator is the standard error of $\widehat{p_1} - \widehat{p_2}$. From this, we can gather that the estimated standard error of $\widehat{p_1} - \widehat{p_2}$ is:

$$\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

- In cases where we are constructing a confidence interval, it is clear that we should use the estimated standard error, as we did for a single proportion:

$$\widehat{p_1} - \widehat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

- In all hypothesis tests on two proportions, the numerator of the test statistic will be $(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)$.
- For the denominator of the test statistic, there are a couple of cases to consider:
- In our class, we will only be hypothesizing a difference between the proportions. That is, our null hypothesis will have the form $H_0 : p_1 - p_2 = k$, for some real number k.
- We may choose to use the *unpooled* estimated standard error, $\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$ as the denominator of the test statistic.
- Our test statistic would be:

$$Z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}}$$

  and it has (approximately) standard normal distribution.

- Frequently, we will be testing the null hypothesis is $H_0 : p_1 - p_2 = 0$. Of course, this is equivalent to $H_0 : p_1 = p_2$.
- In this case, if the null hypothesis were true and $p_1 = p_2$, then it would make sense to use a **pooled estimate** for the sample proportions:

$$\widehat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

(total number of successes from both samples over the total of both sample sizes).

- In this case, *pooled* estimated standard error is:

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

- Our test statistic would be:

$$Z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

and it has (approximately) standard normal distribution.

- Again, note that the numerator of the test statistic is the same, whether we choose the pooled or the unpooled estimated standard error
- Even for hypothesis tests on two proportions where you could potentially use the pooled estimated standard error, it would still be correct for you to use the unpooled estimated standard error instead.

Example: Motherboards are made by one of two manufacturing processes. 300 motherboards made by the first process and 500 motherboards made by the second process are sampled at random. From the first process, 15 have flaws. From those made by the second process, 30 have flaws. Let p_1, p_2 denote the proportion of motherboards made by process one, two (respectively) which are defective.

(a) What is the estimate for $p_1 - p_2$?

(b) What is the unpooled estimated standard error of $\widehat{p_1} - \widehat{p_2}$?

(c) Test the research hypothesis that the first process makes a smaller proportion of defective items than the second process, using the e.s.e. from part (b).

(d) Test the same hypotheses in (c), this time using the pooled estimated standard error.

(e) Create a 93% confidence interval for $p_1 - p_2$.

(f) What does the confidence interval tell you about $p_1 - p_2$?

(g) Suppose we wish to use these data as a pilot study to estimate the sample size we would need in the future to create a 95% confidence interval with a margin of error of 0.01. What sample size is needed (assuming that $n_1 = n_2$).

Set 31

Sometimes, the data we have collected forms a set of **matched pairs**. Rather than having two *independent* samples $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ we have *pairs* of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Some examples:

- We have a sample of cancer patients who will receive a new drug, we first measure the size of their tumor before receiving the medication, and again after receiving the medication.
- We have two appraisers working for an insurance company. The first appraiser examines ten works of art, and then the second appraiser examines the same ten works of art.

Note: We need to read the scenario carefully; just because $n_1 = n_2$, this does not necessarily mean that the data is paired.

For example, if the first appraiser examined ten works of art, and the second appraiser examined ten *different* works of art, this would be two **independent** samples rather than paired data, and we would be looking at $\mu_1 - \mu_2$.

Notation: If x_i is the i^{th} observation from the first sample, and y_i is the i^{th} observation from the second sample, then D_i is the difference between these two observations:

$$D_i = x_i - y_i$$

We will be working under the assumption that these *differences* are normally distributed.

The parameter of interest in a hypothesis test will be μ_D, the mean difference between paired samples.

The estimate we will use will be $\overline{x}_D$, the average of the observed differences.

Suppose we have n_D pairs of data, and have computed the n_D differences. The test statistic for hypothesis testing we will use is:

$$t = \frac{\overline{x}_D - \mu_D}{s_D/\sqrt{n_D}}$$

(where s_D is the sample standard deviation of the n_D differences.)

This test statistic has t distribution with $n_D - 1$ degrees of freedom.

Note: As before, if the sample size is large enough ($n_D \geq 40$) we can use standard normal distribution.

Example: An insurance company is worried about differences in the values of art objects as estimated by two appraisers. The company selects 5 works of art, and asks both appraisers to determine a value. The following are the appraised values (in millions of dollars). Is there a significant difference in appraised values?

	Object 1	Object 2	Object 3	Object 4	Object 5
Appraiser 1	22.10	92.70	2.76	75.60	4.13
Appraiser 2	21.30	92.10	1.54	78.90	4.78

Example: Find a 99% confidence interval for μ_D, the mean difference in the appraisal values.