9.3 Analysis of Paired Data:

In the last two sections, we compared two means μ_1 and μ_2 by using a random sample $X_1, \ldots, X_m$ from the first population, and a completely independent (of the X's) sample $Y_1, \ldots, Y_n$ from the second population.

Sometimes, we may have n people or objects and we make observations on each twice. This creates a pairing between the first observations $X_1, \ldots, X_n$ and the second observations $Y_1, \ldots, Y_n$.

Ex: Investigate cars with two sets of tires, using the same cars each time.

The Paired t-Test and CI:

Assumptions: The data consists of n independently selected pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $E[X_i] = \mu_1$ and $E[Y_i] = \mu_2$. Define the differences

$$D_1 = X_1 - Y_1,\ D_2 = X_2 - Y_2,\ \ldots,\ D_n = X_n - Y_n.$$

We assume the D_i's are normally distributed with mean value μ_D and variance σ_D^2. (Usually this is because the X_i's and Y_i's are both normally distributed.)

$$\mu_D = E[X_1 - Y_1] = E[X_1] - E[Y_2] = \mu_1 - \mu_2$$

So to perform a hypothesis test about $\mu_1 - \mu_2$ with paired data, we perform a one-sample t-Test on the differences $D_1, \ldots, D_n$ with mean μ_D.

Hypotheses: $H_0: \mu_D = \Delta_0$ vs $H_a: \mu_D \underset{\neq}{\overset{<}{>}} \Delta_0$.

Test Statistic: $t = \dfrac{\bar{d} - \Delta_0}{\frac{s_D}{\sqrt{n}}}$ ($\bar{d}$ and s_D are the sample mean and SD of the differences $d_1, \ldots, d_n$.)

P-value: Find the area under the t_{n-1} curve according to H_0.

Conclusion: Reject H_0 when P-value $\leq \alpha$, otherwise fail to reject H_0.

CI: The paired CI for $\mu_1 - \mu_2$ is the same as the one-sample t CI for μ_D using the differences.

A $100(1-\alpha)\%$ CI is $\bar{d} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s_D}{\sqrt{n}}$

Example 1: The zinc level in the water of a certain river is measured in 6 locations at the surface of the water and at the bottom of the water.

	Location 1	2	3	4	5	6
Zinc concentration in bottom water (mg/L)	0.5	0.6	0.4	0.4	0.7	0.6
Zinc concentration in surface water (mg/L)	0.4	0.6	0.3	0.2	0.5	0.5
Differences	0.1	0	0.1	0.2	0.2	0.1

Conduct a hypothesis test with $\alpha = 0.05$ to see if there is more zinc on average in the bottom water than on the surface. Find a 95% CI for the average difference in zinc concentration.

Hypotheses: $H_0: \mu_B - \mu_S = 0$ vs $H_a: \mu_B - \mu_S > 0$

$H_0: \mu_D = 0$ vs $H_a: \mu_D > 0$

Test Statistic: $t = \dfrac{\bar{d} - \Delta_0}{s_D/\sqrt{n}}$ $\quad \bar{d} = 0.117,\ s_D = 0.0753$

$$t = \frac{0.117 - 0}{0.0753/\sqrt{6}} \approx 3.806$$

P-value: $1 - \text{pt}(3.806, df = 5) \approx 0.0063$

Conclusion: We reject H_0 since P-value $\leq \alpha$. There is statistically significant evidence that the zinc level in bottom water is higher than the surface.

CI: $\bar{d} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s_D}{\sqrt{n}}$, $\quad t_{0.025, 5} = \text{qt}(0.975, df = 5) \approx 2.571$

$0.117 \pm 2.571 \cdot \frac{0.0753}{\sqrt{6}} = (0.038, 0.196)$ is a 95% CI for the difference of the means.

9.4 Inferences Concerning a Difference Between Population Proportions:

In this section, we develop a hypothesis test and CI for $p_1 - p_2$, the difference between two population proportions.

Prop: Let $X \sim Bin(m, p_1)$ and $Y \sim Bin(n, p_2)$ be independent.

Let $\hat{p}_1 = \frac{X}{m}$ and $\hat{p}_2 = \frac{Y}{n}$.

$$E[\hat{p}_1 - \hat{p}_2] = E[\hat{p}_1] - E[\hat{p}_2] = E[\tfrac{X}{m}] - E[\tfrac{Y}{n}] = \tfrac{1}{m}E[X] - \tfrac{1}{n}E[Y]$$
$$= \tfrac{1}{m}\cdot mp_1 - \tfrac{1}{n}\cdot np_2 = p_1 - p_2.$$

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + (-1)^2 Var(\hat{p}_2) = Var(\tfrac{X}{m}) + Var(\tfrac{Y}{n})$$
$$= \tfrac{1}{m^2}Var(X) + \tfrac{1}{n^2}Var(Y)$$
$$= \tfrac{1}{m^2}\cdot mp_1(1-p_1) + \tfrac{1}{n^2}\cdot np_2(1-p_2) = \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}$$

When both m and n are large, $\hat{p}_1$ and $\hat{p}_2$ are approximately normal, so $\hat{p}_1 - \hat{p}_2$ is approximately normal. Then

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{m} + \frac{p_2 q_2}{n}}} \approx N(0,1). \quad (q_1 = 1-p_1 \text{ and } q_2 = 1-p_2)$$

Large Sample CI for $p_1 - p_2$:

A $100(1-\alpha)\%$ CI for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}$$

This is safe to use as long as we have at least 10 successes and 10 failures from each sample:

$m\hat{p}_1$, $m\hat{q}_1$, $n\hat{p}_2$, $m\hat{q}_2$ are all at least 10.

Example 1: Say we have two coins. We flip the first coin 80 times and get 44 heads. We flip the second coin 100 times and get 34 heads. Find a 99% CI for $p_1 - p_2$, the difference of the probability the first coin comes up heads and the probability the second coin comes up heads.

$m = 80$ $\quad n = 100$

$\hat{p}_1 = \frac{44}{80} = .55 \quad \hat{p}_2 = \frac{34}{100} = .34$

$\hat{q}_1 = 1 - .55 = .45 \quad \hat{q}_2 = .66$

$z_{\frac{\alpha}{2}} = \text{qnorm}(.995) \approx 2.576$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}.$$

$$.55 - .34 \pm 2.576\sqrt{\frac{.55(.45)}{80} + \frac{.34(.66)}{100}}$$

$\boxed{(.022, .398)}$ is our 99% CI for $p_1 - p_2$.

Large Sample Z-test for $H_0: p_1 - p_2 = 0$:

In general we could look at $H_0: p_1 - p_2 = \Delta_0$, but it turns out the test is different when $\Delta_0 = 0$ and when $\Delta_0 \neq 0$. So we will look at the common case $H_0: p_1 - p_2 = 0$.

sample 1		sample 2
m	sample size	n
$\hat{p}_1 = \frac{X}{m}$ $\frac{5}{10}$		$\hat{p}_2 = \frac{Y}{n}$ $\frac{5}{10}$
$m\hat{p}_1$	successes	$n\hat{p}_2$

$$\hat{p} = \frac{X+Y}{m+n}$$

When H_0 is true, $p_1 - p_2 = 0$, so $p_1 = p_2$. Call this $p = p_1 = p_2$.

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{m} + \frac{p_2 q_2}{n}}} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{pq}{m} + \frac{pq}{n}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{m} + \frac{1}{n}\right)}} \approx N(0,1)$$

But we don't know p. We can use

$$\hat{p} = \frac{X+Y}{m+n} = \frac{m\hat{p}_1 + n\hat{p}_2}{m+n}$$

Test Procedure

Hypotheses:

$H_0: p_1 - p_2 = 0$ vs $H_a: p_1 - p_2 \; >, <, \neq \; 0$.

Test Statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}}$ where $\hat{p} = \dfrac{\text{total succes in two samples}}{m+n}$

$Z \approx N(0,1)$

P-Value: Use standard normal distribution to find the P-value according to the alternative hypothesis.

Conclusion: Reject H_0 if P-value $\leq \alpha$

Fail to reject H_0 if P-value $> \alpha$.

Test is safe to use if there are at least 10 successes and 10 failures in each sample: $m\hat{p}_1$, $m\hat{q}_1$, $n\hat{p}_2$, $n\hat{q}_2$ are all at least 10.

Example 2: SSD drives from two factories are checked for defects. Out of 200 drives from the first factory, 20 are defective. Out of 300 drives from the second factory, 15 are defective. Perform a hypothesis test to see if the true proportion of defective drives is the same for both factories. Use $\alpha = 0.05$.

$\hat{p}_1 = \frac{20}{200} = 0.1$ $\quad \hat{p}_2 = \frac{15}{300} = 0.05$ $\quad \hat{p} = \frac{35}{500} = 0.07$

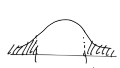
$\hat{q} = 1 - 0.07 = 0.93$

Hypotheses: $H_0: p_1 - p_2 = 0$

$H_a: p_1 - p_2 \neq 0$

Test Statistic: $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m} + \frac{1}{n}\right)}} = \dfrac{0.1 - 0.05}{\sqrt{.07(.93)\left(\frac{1}{200} + \frac{1}{300}\right)}} \approx 2.147$

P-value: 2 · pnorm(2.147, lower.tail = FALSE)



2.147

P-value ≈ 0.0318

Conclusion: Since P-value $\leq \alpha$, we reject H_0. We have statistically significant evidence that there is a difference between the proportion of defective drives at the two factories.

9.5 Inferences Concerning Two Population Variances

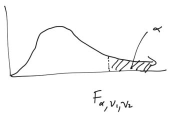
Sometimes we want to test if two populations have the same variance,

$H_0: \sigma_1^2 = \sigma_2^2$.

F-distribution: If $X_1 \sim \chi^2(\nu_1)$ and $X_2 \sim \chi^2(\nu_2)$, then

$$F = \frac{X_1/\nu_1}{X_2/\nu_2}$$ has the F-distribution with ν_1 and ν_2 df.

$F \sim F(\nu_1, \nu_2)$.



The F-test for Equality of Variances:

Let $X_1, \ldots, X_m$ be a random sample from a normal distribution with variance σ_1^2, and let $Y_1, \ldots, Y_n$ be a random sample from a normal distribution with variance σ_2^2, independent of the X_i's.

Let S_1^2 and S_2^2 be the two sample variances. Then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$ has an $F(m-1, n-1)$ distribution.

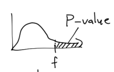
Test Procedure:

Hypotheses: $H_0: \sigma_1^2 = \sigma_2^2$ vs $H_a: \sigma_1^2 \; \substack{> \\ < \\ \neq} \; \sigma_2^2$.

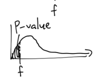
Test Statistic: $f = \frac{S_1^2}{S_2^2}$

P-value:

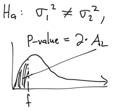
if $H_a: \sigma_1^2 > \sigma_2^2$, P-value $= A_R = 1 - \text{pf}(f, m-1, n-1)$

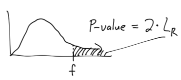


if $H_a: \sigma_1^2 < \sigma_2^2$, P-value $= A_L = \text{pf}(f, m-1, n-1)$



if $H_a: \sigma_1^2 \neq \sigma_2^2$, P-value $= 2 \cdot \min(A_L, A_R)$





Conclusion: If P-value $\le \alpha$, we reject H_0.
If P-value $> \alpha$, fail to reject H_0.

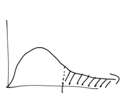
Test requires both populations are normal, and the two samples are independent of each other.

Example 1: The test scores on two standardized tests are normally distributed. A random sample of 20 people take the first test and the sample variance is $S_1^2 = 330$. A independent random sample of 25 people take the second test and the sample variance is $S_2^2 = 280$. Conduct a test with $\alpha = 0.05$ to determine if the test scores have the same variance $(\sigma_1^2 = \sigma_2^2)$.

Hypotheses: $H_0: \sigma_1^2 = \sigma_2^2$, $H_a: \sigma_1^2 \neq \sigma_2^2$

Test Statistic: $f = \frac{S_1^2}{S_2^2} = \frac{330}{280} \approx 1.179$

P-value: f comes from an $F(20-1, 25-1) = F(19, 24)$ distribution



In this case the probability to the right of 1.179 is smaller. So,

P-value $= 2 \cdot \text{pf}(1.179, 19, 24, \text{lower.tail} = \text{FALSE}) \approx 0.695$.

Conclusion: Since P-value $> \alpha$, we fail to reject H_0. There is insufficient evidence to show the tests have different variances.