Inferences on Two Groups or Populations

Links: Math 3339

Matched Pairs T-Test

A matched pair is when two subject units are exactly the same for both responses. There are two populations, but they depend on each other because they share the same subject. We calculate the differences first and find the mean and standard deviation of the differences. Then this problem is the same as a one-sample confidence interval.
Matched pairs is a special test that is used when we are comparing corresponding values in data, it is only used when the data samples are *dependent* upon one another (like before and after results).

- We first find the differences from each observation
- The point estimate is $\overline{x}_d$ = mean of the differences
- The standard deviation is s_d = the standard deviation of the differences
- the margin of error is $m = t^*(\frac{s_d}{\sqrt{n}})$ (critical value multiplied by standard error)
- the confidence interval is $\overline{x}_d \pm t^*(\frac{s_d}{\sqrt{n}})$ (point estimate plus or minus the margin of error)

If we want a hypothesis test, the test statistic is: $t = \frac{\overline{x}_d - \mu_d}{s_d/\sqrt{n}}$

Assumptions of the Matched pairs t-test:

1. Each sample is a simple random sample of size *n* from the same population
2. The test is conducted on paired data (samples are not independent).
3. *unknown* population standard deviation
4. either a Normal population or large samples ($n \geq 30$)

Hypotheses for the matched pairs t-test:

- Null Hypothesis: $H_0 : \mu_d = 0$
- Alternative Hypotheses:
    - (two-tailed) $H_a : \mu_d \neq 0$
    - (left-tailed) $H_a : \mu_d < 0$
    - (right-tailed) $H_a : \mu_d > 0$
- (where μ_d is the mean of the differences)

(see slides 7 to 10 of lecture 16 slides for example)

Two Population Inference

The goal of two-population inference is to compares the responses in two groups. Each group is considered to be a sample from a distinct population, and the responses in each group are *independent* of those in the other group.

Assumptions for the difference of the two means:

1. Both samples must be independent simple random samples from the populations of interest
2. Both sets of data must come from normally distributed populations.

Two samples (where both follow normal distribution)
$x_1 \sim N(\mu_1, \sigma_1) \qquad x_2 \sim N(\mu_2, \sigma_2)$
$E(\overline{x}_1 - \overline{x}_2) = E(\overline{x}_1) - E(\overline{x}_2) = \mu_1 - \mu_2$
$Var(\overline{x}_1 - \overline{x}_2) = Var(\overline{x}_1) + Var(\overline{x}_2) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$
$Sd(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Two-sample t

if the population standard deviations σ_1 and σ_2 is unknown, the sample standard deviations s_1 and s_2 is used. When we use the sample standard deviations we use the *two-sample t statistic*

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

with *k* degrees of freedom approximated by software or the smaller value of $n_1 - 1$ or $n_2 - 1$.
(calculators and software such as R use an approximate degrees of freedom called *Satterthwaite* degrees of freedom calculated using a big formula that is on slide 16 of lecture 16 slides, if doing by hand use the smaller of the two sample sizes minus one as shown above).

Interval Estimation of mu1 - mu2

1. Post Estimate: $\overline{x}_1 - \overline{x}_2$
2. Confidence level: $C = 1 - \alpha$
3. Critical value: t^* with degrees of freedom $n_1 - 1$ or $n_2 - 1$ (whichever is smaller). In R the critical value is: `qt(1+C/2,df)`
4. Margin of Error: $E = t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
5. Confidence interval: point estimate ± margin of error
    - $(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2,df} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Hypothesis Test for two population means (Two-Sample t-Test)

Used to compare the responses to two treatments or characteristics of two populations.
These tests are different than the matched pairs t-test.

Hypotheses:

- Null: $H_0 : \mu_1 = \mu_2$ (first mean minus second mean is 0)
- Alternative: $H_a : \mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$

Assumptions for a two-sample t-test

1. Each group is considered to be a simple random sample from two distinct populations
2. The responses in each group are independent of those in the other group
3. The distribution of the variables are Normal or have a large sample $n_1 > 30$ and $n_2 > 30$.

Test statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
With degrees of freedom $n_1 - 1$ or $n_2 - 1$ (whichever is smaller).

Comparing Two Proportions

1. Each group is considered to be a simple random sample from two distinct populations
2. the population sizes are both at least ten times the sizes of the samples
3. the number of successes and failures in **both** samples must all be ≥ 10 (i.e. both np and $n(1-p)$ must be ≥ 10 for both samples).

$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$
$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{Var(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Confidence intervals for comparing two proportions

Choose a simple random sample of n_1 from a large population having proportion p_1 of successes and an independent simple random sample of size n_2 from another population having proportion p_2 of successes.

1. Point estimate: $D = \hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$
2. Confidence level: *C* a percent predetermined in the problem if not use 95%
3. Critical value z^* is the value for standard Normal density curve with area *C* between $-z^*$ and z^*. (calculate z^* in R with `qnorm((1+C)/2)`).
4. Confidence interval:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

5. Interpret.

Assumptions for Two-Sample Proportion Test

1. Both samples must be independent simple random samples from the populations of interest

2. The population sizes are both at least ten times the sizes of the samples
3. The number of successes and failures in both samples must all be at least 10

Test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

(see lecture 16 slides for examples)