Formal definition of paired sampling distribution

Let X_1 and X_2 represent two measurements' distributions,
$X_{1,i}$ represent the ith individual's (unknown) first measurement,
and $X_{2,i}$ represent the ith individual's second measurement.
Define $X_{\text{diff,i}} = X_{2,i} - X_{1,i}$
Our statistic is a mean of differences.

$$\overline{X_{\text{diff}}} = \frac{\sum_{i=1}^{n}(X_{2,i} - X_{1,i})}{n}$$

Usually, $\overline{X_{\text{diff}}}$ approximately follows a normal distribution.

$$\overline{X_{\text{diff}}} \sim \mathcal{N}(\mu_{\text{diff}}, SE)$$

$$\frac{\overline{X_{\text{diff}}} - \mu_{\text{diff}}}{SE} \sim \mathcal{N}(0,\, 1)$$

$$SE = \frac{\sigma_{\text{diff}}}{\sqrt{n}}$$

Inference from paired data

- Now, imagine μ_{diff} and σ_{diff} are unknown, but we want to infer about our parameter of interest:
- We obtain a sample of differences, which has mean $\overline{x_{\text{diff}}}$ and standard deviation s_{diff}. We now have a point estimate:
- Due to our uncertainty in both μ_{diff} and σ_{diff}, we use a t distribution for inference.

Standard Error:

$$SE \approx \frac{s_{\text{diff}}}{\sqrt{n}}$$

Degrees of freedom:

$$df = n - 1$$

Confidence interval:

$$\mu_{\text{diff}} \approx \overline{x_{\text{diff}}} \pm t^{\star} SE$$

Hypothesis testing:

$$t_0 = \frac{\overline{x_{\text{diff}}} - (\mu_{\text{diff}})_0}{SE}$$

Formal definition of unpaired sampling distribution

Let X_1 and X_2 represent two distributions.

Let $X_{1,i}$ represent the ith (out of n_1) value from the first distribution.

Let $X_{2,j}$ represent the jth (out of n_2) value from the second distribution.

We are interested in a difference of means.

$$\overline{X_2} - \overline{X_1} = \frac{\sum_{j=1}^{n_2} X_{2,j}}{n_2} - \frac{\sum_{i=1}^{n_1} X_{1,i}}{n_1}$$

Usually, $\overline{X_2} - \overline{X_1}$ approximately follows a normal distribution.

$$\overline{X_2} - \overline{X_1} \sim \mathcal{N}(\mu_2 - \mu_1, SE)$$

$$SE = \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$$

Inference from unpaired data

- Now, imagine μ_1, μ_2, σ_1 and σ_2 are unknown. From each population, we take a random sample, and then we wish to infer a confidence interval for $\mu_2 - \mu_1$ or test whether there is evidence to disprove $\mu_2 - \mu_1 = 0$.
- How best to determine (from 2 samples) a confidence interval for $\mu_2 - \mu_1$ and test whether $\mu_2 - \mu_1 = 0$ is an open question, called the Behrens-Fisher problem.
- Different people use different strategies. Old people will probably be more familiar with Student's approach, which assumes $\sigma_1 = \sigma_2$.
- The modern approach (used in our text) is Welch's t-test. Along with randomization techniques (like we simulated with shuffling cards), this is the current standard approach.
- Welch test's main drawback is the annoyingly complicated formula for determining degrees of freedom.

Inference from unpaired data

Now, imagine μ_1, μ_2, σ_1 and σ_2 are unknown.
Let $\overline{x_1}$ represent the (known) mean of first measurements.
Let $\overline{x_2}$ represent the (known) mean of second measurements.
Due to our uncertainty in $\mu_2 - \mu_1$ and σ_1 and σ_2, we use a t distribution.

Standard error:

$$SE = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$$

Confidence interval:

$$\mu_2 - \mu_1 \quad \approx \quad (\overline{x_2} - \overline{x_1}) \pm t^\star SE$$

Hypothesis testing:

$$t_0 = \frac{(\overline{x_2} - \overline{x_1}) - (\mu_2 - \mu_1)_0}{SE}$$

Degrees of freedom:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\frac{(s_1)^4}{(n_1)^3 - (n_1)^2} + \frac{(s_2)^4}{(n_2)^3 - (n_2)^2}}$$

Approximation for calculations by hand

Welch's t test has a gnarly formula for df.

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\frac{(s_1)^4}{(n_1)^3-(n_1)^2} + \frac{(s_2)^4}{(n_2)^3-(n_2)^2}}$$

The formula for degrees of freedom is annoying to evaluate for mere mortals. So, unless otherwise instructed, we will use a conservative estimate (conservative w.r.t. type I error).

$$\boxed{\text{df} \approx \min(n_1,\, n_2) - 1}$$

Don't be surprised if other texts (or people) tell you to use $df = n_1 + n_2 - 2$. We only use this if we have a strong argument for why we believe $\sigma_1 = \sigma_2$.

Hypotheses under paired and unpaired

- With paired data, the statistic is a mean of differences. Usually we are wondering whether the population mean of differences is 0.

$$H_0: \quad \mu_{diff} = 0$$

$$H_A: \quad \mu_{diff} \neq 0$$

- With unpaired data, the statistic is a difference of means. Usually we are wondering whether the difference of population means is 0.

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

$$H_A: \quad \mu_2 - \mu_1 \neq 0$$

Hypotheses under paired and unpaired (other notation)

- With paired data, the statistic is a mean of differences. Usually we are wondering whether the population mean of differences is 0.

$$H_0: \; E(X_2 - X_1) = 0$$
$$H_A: \; E(X_2 - X_1) \neq 0$$

- With unpaired data, the statistic is a difference of means. Usually we are wondering whether the difference of population means is 0.

$$H_0: \;\; E(X_2) - E(X_1) = 0$$
$$H_A: \;\; E(X_2) - E(X_1) \neq 0$$

Example problem

An experiment has $n_1 = 4$ plants in the treatment group and $n_2 = 6$ plants in the control group. After some time, the plants' heights (in cm) are measured, resulting in the following data:

	value1	value2	value3	value4	value5	value6
sample 1:	16.4	14.2	19.4	17.3		
sample 2:	10.3	9.9	9.4	11	10.4	10.7

1. Determine degrees of freedom.
2. Determine $t^\star$ for a 98% confidence interval.
3. Determine SE.
4. Determine a lower bound of the 98% confidence interval of $\mu_2 - \mu_1$.
5. Determine an upper bound of the 98% confidence interval of $\mu_2 - \mu_1$.
6. Determine $|t_{\text{obs}}|$ under the null hypothesis $\mu_2 - \mu_1 = 0$.
7. Determine a lower bound of the two-tail p-value.
8. Determine an upper bound of two-tail p-value.
9. Do you reject the null hypothesis with a two-tail test using a significance level $\alpha = 0.02$? (yes or no)

Solution

These data are unpaired. We might as well find the sample means and sample standard deviations (use a calculator's built-in function for standard deviation).

$$\overline{x_1} = 16.8$$
$$\overline{x_2} = 10.3$$
$$s_1 = 2.15$$
$$s_2 = 0.571$$

We make a conservative estimate of the degrees of freedom using the appropriate formula.

$$df \;=\; \min(n_1,\, n_2) - 1 \;=\; \min(4, 6) - 1 \;=\; 3$$

We use the t table to find $t^\star$ such that $P(|T| < t^\star) = 0.98$

$$t^\star = 4.54$$

We use the SE formula for unpaired data.

$$SE = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} = \sqrt{\frac{(2.15)^2}{4} + \frac{(0.571)^2}{6}} = 1.1$$

We find the bounds of the confidence interval.

$$CI = (\overline{x_2} - \overline{x_1}) \pm t^\star SE$$

$$CI = (-11.494, -1.506)$$

We find t_{obs}.

$$t_{\text{obs}} = \frac{(\overline{x_2} - \overline{x_1}) - (\mu_2 - \mu_1)_0}{SE} = \frac{(10.3 - 16.8) - 0}{1.1} = -5.91$$

We find $|t_{\text{obs}}|$.

$$|t_{\text{obs}}| = 5.91$$

We use the table to determine bounds on p-value. Remember, $df = 3$ and p-value $= P(|T| > |t_{\text{obs}}|)$.

$$0.005 < p\text{-value} < 0.01$$

We should consider both comparisons to make our decision.

$$|t_{\text{obs}}| > t^\star$$

$$p\text{-value} < \alpha$$

Thus, we reject the null hypothesis. Also notice the confidence interval does not contain 0.

Answer list

1. 3
2. 4.54
3. 1.1
4. -11.494
5. -1.506
6. 5.909
7. 0.005
8. 0.01
9. yes

Example problem 2

An experiment has $n_1 = 6$ plants in the treatment group and $n_2 = 8$ plants in the control group. After some time, the plants' heights (in cm) are measured, resulting in the following data:

	value1	value2	value3	value4	value5	value6	value7	value8
sample 1:	0.81	0.98	1.39	1.34	0.78	1.11		
sample 2:	1.31	1.3	1.45	1.42	1.22	1.37	1.34	1.31

1. Determine degrees of freedom.
2. Determine $t^\star$ for a 98% confidence interval.
3. Determine SE.
4. Determine a lower bound of the 98% confidence interval of $\mu_2 - \mu_1$.
5. Determine an upper bound of the 98% confidence interval of $\mu_2 - \mu_1$.
6. Determine $|t_{\text{obs}}|$ under the null hypothesis $\mu_2 - \mu_1 = 0$.
7. Determine a lower bound of the two-tail p-value.
8. Determine an upper bound of two-tail p-value.
9. Do you reject the null hypothesis with a two-tail test using a significance level $\alpha = 0.02$? (yes or no)

Solution 2

These data are unpaired. We might as well find the sample means and sample standard deviations (use a calculator's built-in function for standard deviation).

$$\overline{x_1} = 1.07$$

$$\overline{x_2} = 1.34$$

$$s_1 = 0.259$$

$$s_2 = 0.0729$$

We make a conservative estimate of the degrees of freedom using the appropriate formula.

$$df \;=\; \min(n_1,\, n_2) - 1 \;=\; \min(6, 8) - 1 \;=\; 5$$

We use the t table to find $t^\star$ such that $P(|T| < t^\star) = 0.98$

$$t^\star = 3.36$$

We use the SE formula for unpaired data.

$$SE = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} = \sqrt{\frac{(0.259)^2}{6} + \frac{(0.0729)^2}{8}} = 0.109$$

We find the bounds of the confidence interval.

$$CI = (\overline{x_2} - \overline{x_1}) \pm t^\star SE$$

$$CI = (-0.096, 0.636)$$

We find t_{obs}.

$$t_{\text{obs}} = \frac{(\overline{x_2} - \overline{x_1}) - (\mu_2 - \mu_1)_0}{SE} = \frac{(1.34 - 1.07) - 0}{0.109} = 2.48$$

We find $|t_{\text{obs}}|$.

$$|t_{\text{obs}}| = 2.48$$

We use the table to determine bounds on p-value. Remember, $df = 5$ and $p\text{-value} = P(|T| > |t_{\text{obs}}|)$.

$$0.05 < p\text{-value} < 0.1$$

We should consider both comparisons to make our decision.

$$|t_{\text{obs}}| < t^\star$$

$$p\text{-value} > \alpha$$

Thus, we retain the null hypothesis. Also notice the confidence interval does contain 0.

Answer list

1. 5
2. 3.36
3. 0.109
4. -0.096
5. 0.636
6. 2.481
7. 0.05
8. 0.1
9. no