Comparison of Two Independent Samples

Many times in the sciences it is useful to compare two groups.
- Male vs. Female
- Drug vs. Placebo
- NC vs. Disease

Q: Different?

Population 1
- Mean: \( \mu_1 \)
- Sample: \( \bar{y}_1 \)
- Sample size: \( n_1 \)

Population 2
- Mean: \( \mu_2 \)
- Sample: \( \bar{y}_2 \)
- Sample size: \( n_2 \)

What seems like appropriate estimates for these quantities?

\[
\text{Standard Error of } \bar{y}_1 - \bar{y}_2 = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

We know \( \bar{y}_1 - \bar{y}_2 \) estimates \( \mu_1 - \mu_2 \).

What we need to describe next is the precision of our estimate:

\[
SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{SE_1^2 + SE_2^2}
\]
Standard Error of $\bar{y}_1 - \bar{y}_2$

**Example:** A study is conducted to quantify the benefits of a new cholesterol lowering medication. Two groups of subjects are compared, those who took the medication twice a day for 3 years, and those who took a placebo. Assume subjects were randomly assigned to either group and that both groups data are normally distributed. Results from the study are shown below:

<table>
<thead>
<tr>
<th></th>
<th>Medication</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{y}$</td>
<td>209.8</td>
<td>224.3</td>
</tr>
<tr>
<td>$n$</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>$s$</td>
<td>44.3</td>
<td>46.2</td>
</tr>
<tr>
<td>SE</td>
<td>14.0</td>
<td>14.6</td>
</tr>
</tbody>
</table>

**Example:** Cholesterol medicine (cont')

(e.g., ftp://ftp.nist.gov/pub/dataplot/other/reference/CHOLEST2.DAT)

Calculate an estimate of the true mean difference between treatment groups and this estimate’s precision.

First, denote medication as group 1 and placebo as group 2.

$$
\bar{y}_1 - \bar{y}_2 = 209.8 - 224.3 = -14.5
$$

Then we use the pooled variance to calculate the pooled version of the standard error.

$$
SE_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{SSE_{pooled}}{n_1 + n_2}}
$$

NOTE: If $n_1 = n_2$ and $s_1 = s_2$ the pooled and unpoled will give the same answer for $SE_{\bar{y}_1 - \bar{y}_2}$.

It is when $n_1 \neq n_2$ that we need to decide whether to use pooled or unpoled:

- If $s_1 = s_2$ then use pooled (unpoled will give similar answer)
- If $s_1 \neq s_2$ then use unpoled (pooled will NOT give similar answer)

**CI for $\mu_1 - \mu_2$**

**Recall:** We described a CI earlier as:

the estimate $\pm$ (an appropriate multiplier x SE)

A 100(1- $\alpha$)% confidence interval for $\mu_1 - \mu_2$

$$
(\bar{y}_1 - \bar{y}_2) \pm t(df) \cdot SE_{\bar{y}_1 - \bar{y}_2}
$$

where $df = \frac{(SE_{\bar{y}_1}^2 + SE_{\bar{y}_2}^2)(n_1 - 1 + n_2 - 1)}{SE_{\bar{y}_1}^2/(n_1 - 1) + SE_{\bar{y}_2}^2/(n_2 - 1)}$
Example: Cholesterol medication (cont')

Calculate a 95% confidence interval for \( \mu_1 - \mu_2 \)

We know \( \bar{x}_1 - \bar{x}_2 \) and \( SE(\bar{x}_1 - \bar{x}_2) \) from the previous slides. Now we need to find the t multiplier.

\[
df = \frac{(14^2 + 14.6^2)}{(10-1) + (10-1)} \approx 17.97 \approx 17
\]

Round down to be conservative

\( t \) multiplier

\[
\frac{(\bar{x}_1 - \bar{x}_2) \pm t(df) \cdot SE(\bar{x}_1 - \bar{x}_2)}{21}
\]

\( -14.5 \pm 2.1010(20.24) \)

\( (-57.21, 28.21) \)

CONCLUSION: We are highly confident at the 0.05 level. that the true mean difference in cholesterol between the medication and placebo groups is between -57.02 and 28.02 mg/dL.

Note the change in the conclusion of the parameter that we are estimating. Still looking for the 5 basic parts of a CI conclusion (see slide 38 of lecture set 5).

---

Hypothesis Testing: The independent t test

The idea of a hypothesis test is to formulate a hypothesis that nothing is going on and then to see if collected data is consistent with this hypothesis (or if the data shows something different)

Like innocent until proven guilty

There are four main parts to a hypothesis test:

- Hypotheses
- Test statistic
- p-value
- Conclusion

---

Hypothesis Testing: #1 The Hypotheses

There are two hypotheses:

- Null hypothesis (aka the "status quo" hypothesis)
  - Denoted by \( H_0 \)
- Alternative hypothesis (aka the research hypothesis)
  - Denoted by \( H_a \)

---

Hypothesis Testing: #1 The Hypotheses

- If we are comparing two group means nothing going on would imply no difference
  - The means are "the same"
    \[ (\mu_1 - \mu_2) = 0 \]
- For the independent t-test the hypotheses are:
  \[ H_0: (\mu_1 - \mu_2) = 0 \] (no statistical difference in the population means)
  \[ H_a: (\mu_1 - \mu_2) \neq 0 \] (a statistical difference in the population means)
Hypothesis Testing: #1 The Hypotheses

Example: Cholesterol medication (cont')
Suppose we want to carry out a hypothesis test to see if the data show that there is enough evidence to support a difference in treatment means.
Find the appropriate null and alternative hypotheses.

\[ H_0: \mu_1 - \mu_2 = 0 \]  
(no statistical difference in the true means of the medication and placebo groups)

\[ H_1: \mu_1 - \mu_2 \neq 0 \]  
(a statistical difference in the true means of the medication and placebo groups, medication has an effect on cholesterol)

Hypothesis Testing: #2 Test Statistic

A test statistic is calculated from the sample data:
\[ t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}} \]

On a t distribution, \( t_s \) could fall anywhere:
- If the test statistic is close to 0, this shows that the data are compatible with \( H_0 \) (no difference)
- If the test statistic is far from 0 (in the tails, upper or lower), this shows that the data are incompatible to \( H_0 \) (there is a difference)

\[ t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}} \]

Hypothesis Testing: #3 P-value

How far is far?
- For a two-tailed test (i.e. \( H_1: \mu_1 - \mu_2 \neq 0 \)) The p-value of the test is the area under the Student’s T distribution in the double tails beyond \( -t_s \) and \( t_s \).

Definition (p. 238): The p-value for a hypothesis test is the probability, computed under the condition that the null hypothesis is true, of the test statistic being at least as extreme or more extreme as the value of the test statistic that was actually obtained (from the data).
Hypothesis Testing: #3 P-value

- What this means is that we can think of the p-value as a measure of compatibility between the data and $H_0$.
  - A large p-value (close to 1) indicates that $t_0$ is near the center (data support $H_0$).
  - A small p-value (close to 0) indicates that $t_0$ is in the tail (data do not support $H_0$).

Where do we draw the line?

- How small is small for a p-value?

The threshold value on the p-value scale is called the significance level, and is denoted by $\alpha$.

- The significance level is chosen by whomever is making the decision (BEFORE THE DATA ARE COLLECTED!)
- Common values for include 0.1, 0.05 and 0.01.

Rules for making a decision:

1. If $p < \alpha$ then reject $H_0$ (statistical significance).
2. If $p > \alpha$ then fail to reject $H_0$ (no statistical significance).

Example: Cholesterol medication (cont’)

Find the p-value that corresponds to the results of the cholesterol lowering medication experiment.

We know from the previous slides that $t = -0.716$ (which is close to 0).

This means that the p-value is the area under the curve beyond ±0.716 with 18 df.

Example: Cholesterol medication (cont’)

Using SOCR we can find the area under the curve beyond ±0.716 with 18 df to be:

$p > 2(0.2) = 0.4$

NOTE: when $H_0$ is $\neq$, the p-value is the area beyond the test statistic in BOTH tails.

Example: Cholesterol medication (cont’)

Suppose the researchers had set $\alpha = 0.05$.

Our decision would be to fail to reject $H_0$ because $p > 0.4$ which is $> 0.05$.

#4 CONCLUSION: Based on this data there is no statistically significant difference between true mean cholesterol of the medication and placebo groups $(p > 0.4)$.

In other words the cholesterol lowering medication does not seem to have a significant effect on cholesterol.

Keep in mind, we are saying that we couldn’t provide sufficient evidence to show that there is a significant difference between the two population means.

Hypothesis Testing Summary

- Important parts of Hypothesis test conclusions:
  1. Decision (significance or no significance)
  2. Parameter of interest
  3. Variable of interest
  4. Population under study
  5. (optional but preferred) P-value
Was Cavendish’s experiment biased?

A number of famous early experiments of measuring physical constants have later been shown to be biased.

Mean density of the earth

True value = 5.517

Cavendish’s data: (from previous Example 7.2.2)


n = 23, sample mean = 5.483, sample SD = 0.1904

Simulate taking 400 sets of 23 measurements from \(N(5.517,0.1904)\). Plotted are the results of the sample means.

Are the Cavendish values unusually diff. from true mean?

SD=0.1904

N(5.517,0.1904)

Comparing CI’s and significance tests

- These are different methods for coping with the uncertainty about the true value of a parameter caused by the sampling variation in estimates.
- **Confidence interval**: A fixed level of confidence is chosen. We determine a range of possible values for the parameter that are consistent with the data (at the chosen confidence level).
- **Significance test**: Only one possible value for the parameter, called the hypothesized value, is tested. We determine the strength of the evidence (confidence) provided by the data against the proposition that the hypothesized value is the true value.

Measuring the distance between the true-value and the estimate in terms of the SE

- Intuitive criterion: Estimate is credible if it’s not far away from its hypothesized true-value!
- But how far is far-away?
- Compute the distance in standard-terms:
  \[
  T = \frac{\text{Estimator} - \text{True Parameter Value}}{\text{SE}}
  \]
- Reason is that the distribution of \(T\) is known in some cases (Student’s \(t\), or \(N(0,1)\)). The estimator (obs-value) is typical/atypical if it is close to the center/tail of the distribution.
Review

What intuitive criterion did we use to determine whether the hypothesized parameter value ($p=0.2$ in the ESP Example 9.1.1, and $\mu=5.517$ in Example 9.1.2) was credible in the light of the data? (Determine if the data-driven parameter estimate is consistent with the pattern of variation we’d expect get if hypothesis was true. If hypothesized value is correct, our estimate should not be far from its hypothesized true value.)

Why was it that $\mu=5.517$ was credible in Example 9.1.2, whereas $p=0.2$ was not credible in Example 9.1.1? (The first estimate is consistent, and the second one is not, with the pattern of variation of the hypothesized true process.)

Hypotheses

Guiding principles

We cannot rule in a hypothesized value for a parameter, we can only determine whether there is evidence to rule out a hypothesized value.

The null hypothesis tested is typically a skeptical reaction to a research hypothesis.

Comments

How can researchers try to demonstrate that effects or differences seen in their data are real? (Reject the hypothesis that there are no effects)

How does the alternative hypothesis typically relate to a belief, hunch, or research hypothesis that initiates a study? ($H_0$:$H_1$ specifies the type of departure from the null-hypothesis, $H_0$ (skeptical reaction), which we are expecting (research hypothesis itself).

In the Cavendish’s mean Earth density data, null hypothesis was $H_0$: $\mu=5.517$. We suspected bias, but not bias in any specific direction, hence $H_0: \mu=5.517$.

The $t$-test

Using $\hat{\theta}$ to test $H_0: \theta=\theta_0$ versus some alternative $H_1$.

STEP 1 Calculate the test statistic,

$$t_0 = \frac{\hat{\theta} - \theta_0}{\text{standard error}}$$

[This tells us how many standard errors the estimate is above the hypothesized value ($t_0$ positive) or below the hypothesized value ($t_0$ negative).]

STEP 2 Calculate the $P$-value using the following table.

STEP 3 Interpret the $P$-value in the context of the data.
## Alternative Evidence against $H_0$: $θ > θ_0$

- $H_1: θ > θ_0$ (too much bigger than $θ_0$)
  - $P = pr(T ≥ t_0)$

- $H_1: θ < θ_0$ (too much smaller than $θ_0$)
  - $P = pr(T ≤ t_0)$

- $H_1: θ ≠ θ_0$ (too far from $θ_0$)
  - $P = 2 pr(|T| ≥ |t_0|)$

where $T = \text{Student}(df)$

## Interpretation of the p-value

<table>
<thead>
<tr>
<th>Approximate Size of $P$-Value</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&gt; 0.12$ (12%)</td>
<td>No evidence against $H_0$</td>
</tr>
<tr>
<td>$0.10$ (10%)</td>
<td>Weak evidence against $H_0$</td>
</tr>
<tr>
<td>$0.05$ (5%)</td>
<td>Some evidence against $H_0$</td>
</tr>
<tr>
<td>$0.01$ (1%)</td>
<td>Strong evidence against $H_0$</td>
</tr>
<tr>
<td>$0.001$ (0.1%)</td>
<td>Very Strong evidence against $H_0$</td>
</tr>
</tbody>
</table>

## Is a second child gender influenced by the gender of the first child, in families with >1 kid?

### First and Second Births by Sex

<table>
<thead>
<tr>
<th>Second Child</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>3,202</td>
<td>2,776</td>
<td>5,978</td>
</tr>
<tr>
<td>Female</td>
<td>2,620</td>
<td>2,792</td>
<td>5,412</td>
</tr>
<tr>
<td>Total</td>
<td>5,822</td>
<td>5,568</td>
<td>11,390</td>
</tr>
</tbody>
</table>

- Research hypothesis needs to be formulated first before collecting/looking/interpreting the data that will be used to address it. Mothers whose 1st child is a girl are more likely to have a girl, as a second child, compared to mothers with boys as 1st children.
- Data: 20 yrs of birth records of 1 Hospital in Auckland, NZ.

## Analysis of the birth-gender data – data summary

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of births</th>
<th>Number of girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Previous child was girl)</td>
<td>5412</td>
<td>2792 (approx. 51.6%)</td>
</tr>
<tr>
<td>2 (Previous child was boy)</td>
<td>5978</td>
<td>2776 (approx. 46.4%)</td>
</tr>
</tbody>
</table>

- Let $p_1$=true proportion of girls in mothers with girl as first child, $p_2$=true proportion of girls in mothers with boy as first child. Parameter of interest is $p_1 - p_2$.
- $H_0: p_1 - p_2 = 0$ (skeptical reaction). $H_1: p_1 - p_2 > 0$ (research hypothesis)

## Hypothesis testing as decision making

<table>
<thead>
<tr>
<th>Decision Making</th>
<th>Actual situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision made</td>
<td>Accept $H_0$ is true</td>
</tr>
<tr>
<td>Type I error</td>
<td>OK</td>
</tr>
</tbody>
</table>

- Sample sizes: $n_1=5412$, $n_2=5978$, Sample proportions (estimates) $\hat{p}_1 = 2792/5412 = 0.5159$, $\hat{p}_2 = 2776/5978 = 0.4644$.
- $H_0: p_1 - p_2 = 0$ (skeptical reaction). $H_1: p_1 - p_2 > 0$ (research hypothesis)

## Analysis of the birth-gender data

- Samples are large enough to use Normal-approx. Since the two proportions come from totally different mothers they are independent → use formula 8.5.5.a

\[
t_0 = \frac{\text{Estimate - Hypothesized Value}}{SE} = 5.49986 =
\]

\[
\frac{\hat{p}_1 - \hat{p}_2 - 0}{SE} = \frac{\hat{p}_1 - \hat{p}_2}{SE}
\]

\[
SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}}
\]

\[
P = Pr(T > t_0) = 1.9 \times 10^{-8}
\]
Analysis of the birth-gender data

- We have strong evidence to reject the $H_0$, and hence conclude mothers with first child a girl are more likely to have a second child.
- How much more likely? A 95% CI: $CI (\hat{p}_1 - \hat{p}_2) = [0.033; 0.070]$. And computed by:

\[
\hat{p} = \frac{\hat{p}_1 \hat{p}_2 (1-\hat{p}_1)(1-\hat{p}_2)}{\hat{p}_1 n_1 + \hat{p}_2 n_2} = 0.0515 \pm 1.96 \times 0.0093677 = [3\%; 7\%]
\]

Example – 7.51

- A study was undertaken to compare the respiratory responses of hypnotized and non-hypnotized subjects to certain instructions.
- 16 male volunteers were allocated at random to an experimental group to be hypnotized or to a control group. Baseline measurements were taken at the start of the experiment.
- In analyzing the data, the researchers noticed that the baseline breathing patterns of the two groups were different; this was surprising, since all the subjects had been treated the same up to that time.
- One explanation proposed for this unexpected difference was that the experimental group were more excited in anticipation of the experience of being hypnotized.
- The summary of the baseline measurements of total ventilation is provided (liters of air per minute per square meter of body area).

<table>
<thead>
<tr>
<th>Control</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.21</td>
<td>6.08</td>
</tr>
<tr>
<td>7.18</td>
<td>6.797</td>
</tr>
<tr>
<td>5.7</td>
<td>6.346</td>
</tr>
<tr>
<td>6.325</td>
<td>5.41</td>
</tr>
<tr>
<td>4.86</td>
<td>5.321</td>
</tr>
<tr>
<td>4.79</td>
<td>5.743</td>
</tr>
<tr>
<td>4.78</td>
<td>5.62</td>
</tr>
<tr>
<td>4.5</td>
<td>5.321</td>
</tr>
</tbody>
</table>

Example – 7.51

Let 1 denote experimental (to be hypnotized) and 2 denote control.

Welch’s formula (Eqn 7.1) yields $13.97$, so $df = 13$.

(a) $H_0$: Mean ventilation is the same in the “to be hypnotized” condition and in the “control” condition ($\mu_1 = \mu_2$).

$H_A$: Mean ventilation is different in the “to be hypnotized” condition than in the “control” condition ($\mu_1 \neq \mu_2$).

$H_0$ is rejected. There is sufficient evidence ($0.01 < P-value < 0.02$) to conclude that mean ventilation is different in the “to be hypnotized” condition than in the “control” condition.

(b) $H_0$: Mean ventilation is the same in the “to be hypnotized” condition and in the “control” condition ($\mu_1 = \mu_2$).

$H_A$: Mean ventilation is higher in the “to be hypnotized” condition than in the “control” condition ($\mu_1 > \mu_2$).

$H_0$ is rejected. There is sufficient evidence ($0.005 < P-value < 0.01$) to conclude that mean ventilation is higher in the “to be hypnotized” condition than in the “control” condition.

(c) The non-directional alternative (part (a)) is more appropriate. According to the narrative, the researchers formulated the directional alternative in part (b) after they had seen the data. Thus, it would not be legitimate for them (or us) to use a directional alternative.