Hypothesis testing for two means

Why do we let people die?

HYPOTHESIS TESTING

\[ H_0 : \mu = a \]
\[ H_a : \mu \neq a \]

always compare **one-sample** to a hypothesized population parameter

sometimes we want to compare two (or more) population parameters

\[ H_0 : \mu_1 = \mu_2 \]
\[ H_a : \mu_1 \neq \mu_2 \]

STATISTIC

since we want to compare the difference of two population means

our statistic should be the difference of two sample means

\[ \overline{X}_1 - \overline{X}_2 \]

and we will compare that statistic to the hypothesized value of the parameter

\[ H_0 : \mu_1 - \mu_2 = 0 \]

if the statistic is much different from the hypothesized parameter, we will reject \( H_0 \)

same approach as before, different sampling distribution

SAMPLING DISTRIBUTION

We will compute a \( t \) test statistic that has a sampling distribution as a \( t \) distribution with \( df = n_1 + n_2 - 2 \)

two restrictions

1. the two samples drawn from the respective populations are independent
2. the variances of the two populations are equal

TWO-SAMPLE CASE FOR THE MEAN

useful when you want to compare means of two groups

- different teaching methods
- survival with and without drug
- depression with and without treatment
- height of males and females

the null hypothesis is that there is no difference between the means

\[ H_0 : \mu_1 = \mu_2 \]
\[ H_a : \mu_1 \neq \mu_2 \]

or another way to say the same thing

\[ H_0 : \mu_1 - \mu_2 = 0 \]
\[ H_a : \mu_1 - \mu_2 \neq 0 \]

INDEPENDENCE

drawing a sample with a particular value of \( \overline{X}_1 \) should not affect the probability of drawing a sample with any other particular value of \( \overline{X}_2 \)

remember statistical independence

\[ P(X \text{ and } Y) = P(X) \times P(Y) \]

same idea here
INDEPENDENCE

in practice this means we need to be careful about how we sample if comparing treatments, randomly divide a random sample into an experimental group and a control group

Thus, even if you hope your new treatment will save lives, you have to have one group of patients without the treatment (maybe even a “sham” treatment). It seems cruel, but you cannot assume the treatment works, you have to demonstrate it.

take random samples from each population (no overlap, so no risk of dependence)

avoid situations like repeating subjects: e.g. comparing depression for the same subjects before and after treatment (there are ways to test this situation, but not with these techniques)

HOMOGENEITY OF VARIANCE

to carry out hypothesis testing we need to calculate standard error to get standard error we need to estimate (or know) the standard deviation since we sample two groups, we need a pooled estimate of $\sigma^2$ to get a pooled estimate we need to be certain that $\sigma_1^2 = \sigma_2^2$

note this is a statement about the populations we would not expect the sample variances to be identical

FORMULAS

deviation formula
$$s^2 = \frac{\Sigma (X_{1i} - \overline{X}_1)^2 + \Sigma (X_{2i} - \overline{X}_2)^2}{n_1 + n_2 - 2}$$

deviations relative to the sample mean of each sample!

raw score form:
$$s^2 = \frac{\Sigma X_{3i}^2 - (\Sigma X_{3i})^2/n_1 + (\Sigma X_{3i})^2/n_2}{n_1 + n_2 - 2}$$

• $X_{3i}$ refers to the $i$th score from sample 1
• $X_{3i}$ refers to the $i$th score from sample 2
• $n_1$ refers to the number of scores in sample 1
• $n_2$ refers to the number of scores in sample 2

HYPOTHESIS TESTING

we want to compare population means from two populations we have
• $H_0: \mu_1 = \mu_2$
• $\sigma_1^2 = \sigma_2^2$
• Independent samples of size $n_1$ and $n_2$

although we draw two random samples (one from each population), we are only interested in one statistic $\overline{X}_1 - \overline{X}_2$

but we need to know the sampling distribution for this statistic

FORMULAS

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

where
• $s_1^2$ is the variance among scores in sample 1
• $s_2^2$ is the variance among scores in sample 2

SAMPLING DISTRIBUTION OF DIFFERENCES

it turns out that the sampling distribution is familiar

1. Shape: As sample sizes get large, distribution becomes normal.
2. Central tendency: The mean of the sampling distribution equals $\mu_1 - \mu_2$.
3. Variability: The standard deviation of the sampling distribution (standard error of the difference between means) is

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$$

We have to estimate $\sigma$ from our data our estimate is called the pooled estimate because we use scores from both samples
STANDARD ERROR
we use the pooled \( s \) to calculate an estimate of standard error for the sampling distribution of differences
\[
\hat{s}_{X_1 - X_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]
this gives us an estimate of the standard deviation of the sampling distribution of the difference of sample means
we need to know one more thing

DEGREES OF FREEDOM
we have two samples with (possibly) different numbers of scores
the degrees of freedom in sample 1
\[ df = n_1 - 1 \]
from sample 2
\[ df = n_2 - 1 \]
added together gives the result
(depends on independence!)
\[ df = n_1 + n_2 - 2 \]
(same as in denominator of standard deviation estimate)

EXAMPLE
A neurosurgeon believes that lesions in a particular area of the brain, called the thalamus, will decrease pain perception. If so, this could be important in the treatment of terminal illness accompanied by intense pain. As a first attempt to test this hypothesis, he conducts an experiment in which 16 rats are randomly divided into two groups of 8 each. Animals in the experimental group receive a small lesion in the part of the thalamus thought to be involved in pain perception. Animals in the control group receive a comparable lesion in a brain area believed to be unrelated to pain. Two weeks after surgery each animal is given a brief electrical shock to the paws. The shock is administered with a very low intensity level and increased until the animal first flinches. In this manner, the pain threshold to electric shock is determined for each rat. The following data are obtained. Each score represents the current level (milliamperes) at which flinching is first observed. The higher the current level, the higher is the pain threshold.

HYPOTHESIS
Step 1. State the hypotheses and the criterion.
Directional hypothesis because we expect the lesion will increase the threshold.
\[ H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0 \]
(lesion makes no difference)
\[ H_a : \mu_1 < \mu_2 \text{ or } \mu_1 - \mu_2 < 0 \]
(lesion increases pain threshold, less sensitivity)
we will set \( \alpha = 0.05 \) for a one-tailed test
We expect a negative \( t \) value (see \( H_a \))

HYPOTHESIS TESTING
now we have everything we need to apply the techniques of hypothesis testing
1. State the hypothesis and the criterion.
2. Compute the test statistic.
3. Compute the \( p \)-value.
4. Make a decision.

DATA
now we consider the data from the experiment the researcher gets the following

<table>
<thead>
<tr>
<th>Control Group (False lesion)</th>
<th>Experimental Group (Thalamic lesion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>( X_2 )</td>
</tr>
<tr>
<td>0.8</td>
<td>1.9</td>
</tr>
<tr>
<td>0.7</td>
<td>1.8</td>
</tr>
<tr>
<td>1.2</td>
<td>1.6</td>
</tr>
<tr>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td>0.4</td>
<td>1.0</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>1.4</td>
<td>1.7</td>
</tr>
<tr>
<td>1.1</td>
<td>0.7</td>
</tr>
</tbody>
</table>
**Computing Test Statistic**

Step 2. we have $n_1 = 8$, $n_2 = 8$

from the data we calculate

$X_1 = 0.875$

$X_2 = 1.3625$

$X_1 - X_2 = -0.4875$

$s^2 = 0.403$

(using any formula you want), so that

the estimate of standard error is

$s_{X_1 - X_2} = \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$

$s_{X_1 - X_2} = \sqrt{0.403 \left( \frac{1}{8} + \frac{1}{8} \right)} = 0.2015$

**Computing the Test Statistic**

Test statistic $t = \frac{X_1 - X_2 - (\mu_1 - \mu_2)}{s_{X_1 - X_2}}$

$t = \frac{(0.875 - 1.3625) - 0}{0.2015} = -2.419$

Step 3. Compute the $p$-value.

we need to calculate the degrees of freedom

$df = n_1 + n_2 - 2 = 16 - 2 = 14$

We use the $t$ Distribution Calculator to compute

$p = 0.015$

**Interpret Results**

Step 4. Make a decision.

our interpretation of the test is that the difference between the calculated sample means, or an even bigger difference, would have occurred by chance less than 5% of the time if the null hypothesis were true.

in practice, this means that the study supports the theory that lesions to the thalamus decrease pain perception

significant result

This means you have support for the idea that the surgery did affect pain perception.

**Confidence Interval**

Basic formula for all confidence intervals:

$CI = \text{statistic} \pm (\text{critical value})(\text{standard error})$

for a difference of sample means

$CI = (X_1 - X_2) \pm t_{cv}s_{X_1 - X_2}$

We already have most of the terms (we get $t_{cv}$ from the Inverse $t$-distribution calculator, so

$CI_{95} = (0.875 - 1.3625) \pm (2.1448)(0.2015)$

$CI_{95} = (-0.9197, -0.0553)$

**Online Calculator**

The calculations are not complicated, but it is usually better to use a computer. You have to properly format the data.

**Online Calculator**

You need to understand how to pull out the information you want.
ASSUMPTIONS

The t-test that we use for hypothesis tests of means is based on three key assumptions:

1. The population distributions are normally distributed. Matters for small sample sizes.
2. Independent scores. For a two-sample t-test, the scores are uncorrelated between populations. (We deal with this case soon.)
3. Homogeneity of variance. For a two-sample t-test, the populations have the same variance (or standard deviation).

If these assumptions do not hold, then the t-distribution that we calculate is not an accurate description of the sampling distribution.

ROBUSTNESS?

Deviation from normal distributions for the populations does not matter very much, especially for large samples. If we run many tests, we keep the Type I error rate pretty close to what is intended by setting $\alpha$ (e.g., $\alpha = 0.05$)

Show in Robustness Simulation Demonstration in the textbook (12.4)
This is true for varying sample sizes.

HOMOGENEITY OF VARIANCE

to carry out hypothesis testing we need to calculate standard error

to get standard error we need to estimate (or know) the standard deviation of the population

since we sample two groups, we used a pooled estimate of $\sigma^2$

to get a pooled estimate we need to be certain that $\sigma_1^2 = \sigma_2^2$

we need consider what happens when homogeneity does not hold.
HOMOGENEITY OF VARIANCE

We are not actually going to do the hypothesis test for homogeneity of variance. It is messy and (a bit) confusing.

Just remember:

1. If the sample sizes are equal, then you are fine with the standard method.
2. If the sample sizes are unequal, then you might want to worry about homogeneity of variance. If \( s_1^2 \approx s_2^2 \), then you are probably also fine.

If you think you do not have homogeneity of variance, then you can run a revised version of the test. Some people (including your textbook) recommend this as the default approach.

CONCLUSIONS

comparing two means
independent samples
more flexible than one-sample case
many more experiments can be tested
same basic technique

NEXT TIME

Welch’s test
Power
Planning a replication study