Hypothesis Testing

Tests of hypotheses

Confidence interval: Form an interval (on the basis of data) of plausible values for a population parameter.

Test of hypothesis: Answer a yes or no question regarding a population parameter.

Examples:

→ Do the two strains have the same average response?
→ Is the concentration of substance X in the water supply above the safe limit?
→ Does the treatment have an effect?
Example

We have a quantitative assay for the concentration of antibodies against a certain virus in blood from a mouse.

We apply our assay to a set of ten mice before and after the injection of a vaccine. (This is called a “paired” experiment.)

Let $X_i$ denote the differences between the measurements (“after” minus “before”) for mouse $i$.

We imagine that the $X_i$ are independent and identically distributed $\text{Normal}(\mu, \sigma)$.

→ Does the vaccine have an effect? In other words: Is $\mu \neq 0$?

The data

![Graph showing before and after measurements with a line of best fit and a plot of differences.](image)
Hypothesis testing

We consider two hypotheses:

Null hypothesis, \( H_0: \mu = 0 \)  
Alternative hypothesis, \( H_a: \mu \neq 0 \)

Type I error: Reject \( H_0 \) when it is true (false positive)
Type II error: Fail to reject \( H_0 \) when it is false (false negative)

We set things up so that a Type I error is a worse error (and so that we are seeking to prove the alternative hypothesis). We want to control the rate (the significance level, \( \alpha \)) of such errors.

\[ \text{Test statistic: } T = \frac{(X - 0)}{(S/\sqrt{10})} \]

\[ \text{We reject } H_0 \text{ if } |T| > t^*, \text{ where } t^* \text{ is chosen so that} \]
\[ \Pr(\text{Reject } H_0 \mid H_0 \text{ is true}) = \Pr(|T| > t^* \mid \mu = 0) = \alpha. \]
\[ (\text{generally } \alpha = 5\%) \]

Example (continued)

Under \( H_0 \) (i.e., when \( \mu = 0 \)),

\[ T = \frac{(\bar{X} - 0)}{(S/\sqrt{10})} \sim t(df = 9) \]

We reject \( H_0 \) if \( |T| > 2.26 \).

As a result, if \( H_0 \) is true, there’s a 5% chance that you’ll reject it!

For the observed data:
\[ \bar{x} = 1.93, \; s = 2.24, \; n = 10 \]
\[ T = \frac{(1.93 - 0)}{(2.24/\sqrt{10})} = 2.72 \]
\[ \rightarrow \text{ Thus we reject } H_0. \]
The goal

→ We seek to prove the alternative hypothesis.

→ We are happy if we reject $H_0$.

→ In the case that we reject $H_0$, we might say:

*Either $H_0$ is false, or a rare event occurred.*

Another example

Question: is the concentration of substance X in the water supply above the safe level?

$X_1, X_2, \ldots, X_4 \sim \text{iid Normal}(\mu, \sigma)$.

→ We want to test $H_0: \mu \geq 6$ (unsafe) versus $H_a: \mu < 6$ (safe).

Test statistic: $T = \frac{\bar{X} - 6}{S/\sqrt{4}}$

If we wish to have the significance level $\alpha = 5\%$, the rejection region is $T < t^* = -2.35$. 

![t(df=3) distribution](t_distribution.png)
One-tailed vs two-tailed tests

If you are trying to prove that a treatment improves things, you want a one-tailed (or one-sided) test.

You’ll reject $H_0$ only if $T < t^\star$.

If you are just looking for a difference, use a two-tailed (or two-sided) test.

You’ll reject $H_0$ if $T < t^\star$ or $T > t^\star$.

P-values

P-value: $\rightarrow$ the smallest significance level ($\alpha$) for which you would fail to reject $H_0$ with the observed data.

$\rightarrow$ the probability, if $H_0$ was true, of receiving data as extreme as what was observed.

$X_1, \ldots, X_{10} \sim$ iid Normal($\mu, \sigma$), $H_0: \mu = 0; H_a: \mu \neq 0$.

$\bar{x} = 1.93; s = 2.24$

$T_{obs} = \frac{1.93 - 0}{2.24/\sqrt{10}} = 2.72$

P-value = $Pr(|T| > T_{obs}) = 2.4\%$.

$2*pt(-2.72, 9)$
Another example

\[ X_1, \ldots, X_4 \sim \text{Normal}(\mu, \sigma) \quad \text{H}_0: \mu \geq 6; \text{H}_a: \mu < 6. \]

\[ \bar{x} = 5.51; s = 0.43 \]

\[ T_{\text{obs}} = \frac{5.51 - 6}{0.43/\sqrt{4}} = -2.28 \]

P-value = \text{Pr}(T < T_{\text{obs}} \mid \mu = 6) = 5.4\%.

\[ p_t(-2.82, 3) \]

→ The P-value quantifies how likely it is to get data as extreme as the data observed, assuming the null hypothesis was true.

Recall: We want to prove the alternative hypothesis (i.e., reject \text{H}_0, receive a small P-value)

Hypothesis tests and confidence intervals

→ The 95\% confidence interval for \( \mu \) is the set of values, \( \mu_0 \), such that the null hypothesis \( \text{H}_0 : \mu = \mu_0 \) would not be rejected by a two-sided test with \( \alpha = 5\% \).

The 95\% CI for \( \mu \) is the set of plausible values of \( \mu \). If a value of \( \mu \) is plausible, then as a null hypothesis, it would not be rejected.

For example:

9.98 9.87 10.05 10.08 9.99 9.90 \hspace{1cm} \text{assumed to be iid Normal}(\mu, \sigma)

\[ \bar{x} = 9.98; \ s = 0.082; \ n = 6; \ qt (0.975, 5) = 2.57 \]

The 95\% CI for \( \mu \) is

\[ 9.98 \pm 2.57 \times 0.082 / \sqrt{6} = 9.98 \pm 0.086 = (9.89, 10.06) \]
Power

The power of a test = \( \Pr(\text{reject } H_0 \mid H_0 \text{ is false}) \).

The power depends on:
- The null hypothesis and test statistic
- The sample size
- The true value of \( \mu \)
- The true value of \( \sigma \)

Why “fail to reject”? 

If the data are insufficient to reject \( H_0 \), we say,

*The data are insufficient to reject \( H_0 \).*

We shouldn’t say, *We have proven \( H_0 \).*

\( \rightarrow \) We may only have low power to detect anything but extreme differences.

\( \rightarrow \) We control the rate of type I errors (“false positives”) at 5% (or whatever), but we may have little or no control over the rate of type II errors.
The effect of sample size

Let $X_1, \ldots, X_n$ be iid Normal($\mu$, $\sigma$).

We wish to test $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$.

Imagine $\mu = \mu_a$.

**n = 4**

**n = 16**

Testing the difference between two means

Strain A: $X_1, \ldots, X_n \sim$ iid Normal($\mu_A$, $\sigma_A$)

Strain B: $Y_1, \ldots, Y_m \sim$ iid Normal($\mu_B$, $\sigma_B$)

Test $H_0 : \mu_A = \mu_B$ vs $H_a : \mu_A \neq \mu_B$

Test statistic: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}}$

Reject $H_0$ if $|T| > t_{\alpha/2}$

If $H_0$ is true, then $T$ follows (approximately) a t distr'n with $k$ d.f.

$k$ according to the nasty formula from a previous lecture.
Strain A: n = 12, sample mean = 103.7, sample SD = 7.2
Strain B: n = 9, sample mean = 97.0, sample SD = 4.5

\[ \hat{SD}(\bar{X} - \bar{Y}) = \sqrt{\frac{7.2^2}{12} + \frac{4.5^2}{9}} = 2.57 \]

\[ T = \frac{(103.7 - 97.0)}{2.57} = 2.60. \]

k = \ldots = 18.48, so C = 2.10. Thus we reject H_0 at \( \alpha = 0.05 \).

**What to say**

When rejecting H_0:
- The difference is statistically significant.
- The observed difference can not reasonably be explained by chance variation.

When failing to reject H_0:
- There is insufficient evidence to conclude that \( \mu_A \neq \mu_B \).
- The difference is not statistically significant.
- The observed difference could reasonably be the result of chance variation.
What about a different significance level?

Recall $T = 2.60 \quad k = 18.48$

If $\alpha = 0.10, \quad C = 1.73 \implies \text{Reject } H_0$

If $\alpha = 0.05, \quad C = 2.10 \implies \text{Reject } H_0$

If $\alpha = 0.01, \quad C = 2.87 \implies \text{Fail to reject } H_0$

If $\alpha = 0.001, \quad C = 3.90 \implies \text{Fail to reject } H_0$

P-value: the smallest $\alpha$ for which you would still reject $H_0$ with the observed data.

With these data, $P = 2 \times (1 - pt(2.60, 18.48)) = 0.018$.

Another example

Suppose I measure the blood pressure of 6 mice on a low salt diet and 6 mice on a high salt diet. We wish to prove that the high salt diet causes an increase in blood pressure.

We imagine $X_1, \ldots, X_n \sim \text{iid Normal}(\mu_L, \sigma_L) \quad \text{low salt}$

$Y_1, \ldots, Y_m \sim \text{iid Normal}(\mu_H, \sigma_H) \quad \text{high salt}$

We want to test $H_0 : \mu_L = \mu_H$ versus $H_a : \mu_L < \mu_H$

$\implies \text{Are the data compatible with } H_0$?
A one-tailed test

Test statistic: \( T = \frac{\bar{X} - \bar{Y}}{\hat{SD}(\bar{X} - \bar{Y})} \)

Since we seek to prove that \( \mu_L \) is smaller than \( \mu_H \), only large negative values of the statistic are interesting.

Thus, our rejection region is \( T < C \) for some critical value \( C \).

We choose \( C \) so that \( \Pr( T < C \mid \mu_L = \mu_H ) = \alpha \).

The example

Low salt: \( n = 6; \) sample mean = 51.0, sample SD = 10.0
High salt: \( n = 6; \) sample mean = 69.1, sample SD = 15.1

\( \bar{x} - \bar{y} = -18.1 \quad \hat{SD}(\bar{X} - \bar{Y}) = 7.40 \quad T = \frac{-18.1}{7.40} = -2.44 \)

\( k = 8.69. \) If \( \alpha = 0.05 \), then \( C = -1.84. \)

Since \( T < C \), we reject \( H_0 \) and conclude that \( \mu_L < \mu_H. \)

Note: P-value = \( pt(-2.44, 8.69) = 0.019. \)
Always give a confidence interval!

P = 0.019
95% CI: (−34.9, −1.2)

P = 0.019
95% CI: (−13.6, −0.5)

Make a statistician happy: draw a picture of the data.

Example

Suppose I do some pre/post measurements.
I make some measurement on each of 5 mice before and after some treatment.

Question: Does the treatment have any effect?

<table>
<thead>
<tr>
<th>Mouse</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>18.6</td>
<td>14.3</td>
<td>21.4</td>
<td>19.3</td>
<td>24.0</td>
</tr>
<tr>
<td>After</td>
<td>17.8</td>
<td>24.1</td>
<td>31.9</td>
<td>28.6</td>
<td>40.0</td>
</tr>
</tbody>
</table>
Pre/post example

In this sort of pre/post measurement example, study the differences as a single sample.

Why? The pre/post measurements are likely associated, and as a result one can more precisely learn about the effect of the treatment.

<table>
<thead>
<tr>
<th>Mouse</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>18.6</td>
<td>14.3</td>
<td>21.4</td>
<td>19.3</td>
<td>24.0</td>
</tr>
<tr>
<td>After</td>
<td>17.8</td>
<td>24.1</td>
<td>31.9</td>
<td>28.6</td>
<td>40.0</td>
</tr>
<tr>
<td>Difference</td>
<td>−0.8</td>
<td>9.8</td>
<td>10.5</td>
<td>9.3</td>
<td>16.0</td>
</tr>
</tbody>
</table>

n = 5; mean difference = 8.96; SD difference = 6.08.
95% CI for underlying mean difference = . . . = (1.4, 16.5)
P-value for test of $\mu_{\text{before}} = \mu_{\text{after}}$: 0.03.

Summary

• Tests of hypotheses $\rightarrow$ answering yes/no questions regarding population parameters.

• There are two kinds of errors:
  ○ Type I: Reject $H_0$ when it is true.
  ○ Type II: Fail to reject $H_0$ when it is false.

• We seek to reject the null hypothesis.

• If we fail to reject $H_0$, we do not “accept $H_0$”.

• P-value $\rightarrow$ the probability, if $H_0$ is true, of obtaining data as extreme as was observed. $\Pr(\text{data} | \text{no effect})$ rather than $\Pr(\text{no effect} | \text{data})$.

• Power $\rightarrow$ the probability of rejecting $H_0$ when it is false.
Was the result important?

- **Statistically significant** is not the same as important.
- A difference is “statistically significant” if it cannot reasonably be ascribed to chance variation.
- With lots of data, small (and unimportant) differences can be statistically significant.
- With very little data, quite important differences will fail to be significant.
- Always look at the confidence interval as well as the P-value.

Does the difference prove the point?

- A test of significance **does not** check the design of the study.
- With observational studies or poorly controlled experiments, the proof of statistical significance may not prove what you want.
- **Example:** consider the tick/deer leg experiment. It may be that ticks are not attracted to deer-gland-substance but rather despise the scent of latex gloves and deer-gland-substance masks it.
- **Example:** In a study of gene expression, if cancer tissue samples were always processed first, while normal tissue samples were kept on ice, the observed differences might not have to do with normal/cancer as with iced/not iced.
- Don’t forget the science in the cloud of data and statistics.