Statistics: Unlocking the Power of Data

STAT 101
Dr. Kari Lock Morgan

Hypothesis Testing: Significance

SECTION 4.3, 4.5
• Significance level (4.3)
• Statistical conclusions (4.3)
• Type I and II errors (4.3)
• Statistical versus practical significance (4.5)
• Multiple testing (4.5)

Review of Last Class
• The randomization distribution shows what types of statistics would be observed, just by random chance, if the null hypothesis were true
• A p-value is the chance of getting a statistic as extreme as that observed, if H₀ is true
• A p-value can be calculated as the proportion of statistics in the randomization distribution as extreme as the observed sample statistic
• The smaller the p-value, the stronger the evidence against H₀

p-value and H₀
Which of the following p-values gives the strongest evidence against H₀?

a) 0.005
b) 0.1
c) 0.32
d) 0.56
e) 0.94

p-value and H₀
Which of the following p-values gives the strongest evidence against H₀?

a) 0.22
b) 0.45
c) 0.03
d) 0.8
e) 0.71

p-value and H₀
Two different studies obtain two different p-values. Study A obtained a p-value of 0.002 and Study B obtained a p-value of 0.2. Which study obtained stronger evidence against the null hypothesis?

a) Study A
b) Study B

Formal Decisions
• If the p-value is small:
  ○ REJECT H₀
  ○ the sample would be extreme if H₀ were true
  ○ the results are statistically significant
  ○ we have evidence for H₁

• If the p-value is not small:
  ○ DO NOT REJECT H₀
  ○ the sample would not be too extreme if H₀ were true
  ○ the results are not statistically significant
  ○ the test is inconclusive; either H₀ or H₁ may be true
Formal Decisions

A formal hypothesis test has only two possible conclusions:
1. The p-value is small: reject the null hypothesis in favor of the alternative
2. The p-value is not small: do not reject the null hypothesis

How small?

Significance Level

• The significance level, \( \alpha \), is the threshold below which the p-value is deemed small enough to reject the null hypothesis

\[
p\text{-value} < \alpha \implies \text{Reject } H_0 \\
p\text{-value} \geq \alpha \implies \text{Do not Reject } H_0
\]

Red Wine and Weight Loss

• Resveratrol, an ingredient in red wine and grapes, has been shown to promote weight loss in rodents, and has recently been investigated in primates (specifically, the Grey Mouse Lemur).
• A sample of lemurs had various measurements taken before and after receiving resveratrol supplementation for 4 weeks


Red Wine and Weight Loss

In the test to see if the mean resting metabolic rate is higher after treatment, the p-value is 0.013. Using \( \alpha = 0.05 \), is this difference statistically significant? (should we reject \( H_0 \): no difference?)

a) Yes
b) No

In the test to see if the mean body mass is lower after treatment, the p-value is 0.007. Using \( \alpha = 0.05 \), is this difference statistically significant? (should we reject \( H_0 \): no difference?)

a) Yes
b) No
Red Wine and Weight Loss

In the test to see if locomotor activity changes after treatment, the p-value is 0.980.
Using $\alpha = 0.05$, is this difference statistically significant? (should we reject $H_0$: no difference?)

a) Yes
b) No

Elephant Example

$H_0$: $X$ is an elephant
$H_a$: $X$ is not an elephant
Would you conclude, if you get the following data?
- $X$ walks on two legs
- $X$ has four legs

Never Accept $H_0$

"Do not reject $H_0$" is not the same as "accept $H_0$"!
• Lack of evidence against $H_0$ is NOT the same as evidence for $H_0$!

“For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning...”
-Sir R. A. Fisher

Statistical Conclusions

In a hypothesis test of $H_0$: $\mu = 10$ vs $H_a$: $\mu < 10$
the p-value is 0.002. With $\alpha = 0.05$, we conclude:

a) Reject $H_0$
b) Do not reject $H_0$
c) Reject $H_a$
d) Do not reject $H_a$

In a hypothesis test of $H_0$: $\mu = 10$ vs $H_a$: $\mu < 10$
the p-value is 0.002. With $\alpha = 0.01$, we conclude:

a) There is evidence that $\mu = 10$
b) There is evidence that $\mu < 10$
c) We have insufficient evidence to conclude anything
Statistical Conclusions

In a hypothesis test of

\[ H_0: \mu = 10 \text{ vs } H_a: \mu < 10 \]

the p-value is 0.21. With \( \alpha = 0.01 \), we conclude:

a) Reject \( H_0 \)

b) Do not reject \( H_0 \)

c) Reject \( H_a \)

d) Do not reject \( H_a \)

Multiple Sclerosis and Sunlight

• It is believed that sunlight offers some protection against multiple sclerosis, but the reason is unknown

• Researchers randomly assigned mice to one of:
  - Control (nothing)
  - Vitamin D Supplements
  - UV Light

• All mice were injected with proteins known to induce a mouse form of MS, and they observed which mice got MS


Multiple Sclerosis and Sunlight

• For each situation below, write down
  - Null and alternative hypotheses
  - Informal description of the strength of evidence against \( H_0 \)
  - Formal decision about \( H_0 \) using \( \alpha = 0.05 \)
  - Conclusion in the context of the question

• In testing whether UV light provides protection against MS (UV light vs control group), the p-value is 0.002.

• In testing whether Vitamin D provides protection against MS (Vitamin D vs control group), the p-value is 0.47.
Multiple Sclerosis and Sunlight

- In testing whether Vitamin D provides protection against MS (Vitamin D vs control group), the p-value is 0.47.

Red Wine and Weight Loss

- In the test to see if resveratrol is associated with food intake, the p-value is 0.035.
  - If resveratrol is not associated with food intake, a Type I Error would have been made.
- In the test to see if resveratrol is associated with locomotor activity, the p-value is 0.980.
  - If resveratrol is associated with locomotor activity, a Type II Error would have been made.

Errors

There are four possibilities:

<table>
<thead>
<tr>
<th>Truth</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₀ true</td>
<td>Reject H₀</td>
</tr>
<tr>
<td>H₀ false</td>
<td>Do not reject H₀</td>
</tr>
</tbody>
</table>

- A Type I Error is rejecting a true null (false positive)
- A Type II Error is not rejecting a false null (false negative)

Analogy to Law

A person isinnocent until proven guilty.

Evidence must be beyond the shadow of a doubt.

Types of mistakes in a verdict?

Convict an innocent
Release a guilty

Probability of Type I Error

Distribution of statistics, assuming H₀ true:

- 5% of statistics will be in the most extreme 5%
- 5% of statistics will give p-values less than 0.05
- 5% of statistics will lead to rejecting H₀ at α = 0.05
- If α = 0.05, there is a 5% chance of a Type I error

Probability of Type I Error

Distribution of statistics, assuming H₀ true:

- 1% of statistics will be in the most extreme 1%
- 1% of statistics will give p-values less than 0.01
- 1% of statistics will lead to rejecting H₀ at α = 0.01
- If α = 0.01, there is a 1% chance of a Type I error
Probability of Type I Error

- The probability of making a Type I error (rejecting a true null) is the significance level, $\alpha$

Probability of Type II Error

- How can we reduce the probability of making a Type II Error (not rejecting a false null)?
  
  Option 1:
  a) Decrease the significance level
  b) Increase the significance level
  
  Option 2:
  a) Decrease the sample size
  b) Increase the sample size

Probability of Errors

- The probability of making a Type I error (rejecting a true null) if the null is true is the significance level, $\alpha$
- The probability of making a Type II error (not rejecting a false null) if the alternative is true depends on the significance level and the sample size (among other things)
- $\alpha$ should be chosen depending how bad it is to make a Type I or Type II error

Choosing $\alpha$

- By default, usually $\alpha = 0.05$
- If a Type I error (rejecting a true null) is much worse than a Type II error, we may choose a smaller $\alpha$, like $\alpha = 0.01$
- If a Type II error (not rejecting a false null) is much worse than a Type I error, we may choose a larger $\alpha$, like $\alpha = 0.10$

Significance Level

Come up with a hypothesis testing situation in which you may want to...

- Use a smaller significance level, like $\alpha = 0.01$
- Use a larger significance level, like $\alpha = 0.10$

Statistical vs Practical Significance

- With small sample sizes, even large differences or effects may not be significant
- With large sample sizes, even a very small difference or effect can be significant
- A statistically significant result is not always practically significant, especially with large sample sizes
**Statistical vs Practical Significance**

- Example: Suppose a weight loss program recruits 10,000 people for a randomized experiment.
- A difference in average weight loss of only 0.5 lbs could be found to be statistically significant.
- Suppose the experiment lasted for a year. Is a loss of ½ a pound practically significant?

**Diet and Sex of Baby**

- Are certain foods in your diet associated with whether or not you conceive a boy or a girl?
- To study this, researchers asked women about their eating habits, including asking whether or not they ate 133 different foods regularly.
- A significant difference was found for breakfast cereal (mothers of boys eat more), prompting the headline “Breakfast Cereal Boosts Chances of Conceiving Boys”.


**Hypothesis Tests**

For each of the 133 foods studied, a hypothesis test was conducted for a difference between mothers who conceived boys and girls in the proportion who consume each food.

- If there are NO differences (all null hypotheses are true), about how many significant differences would be found using \( \alpha = 0.05 \)?
- How might you explain the significant difference for breakfast cereal?

**Multiple Testing**

When multiple hypothesis tests are conducted, the chance that at least one test incorrectly rejects a true null hypothesis increases with the number of tests.

If the null hypotheses are all true, \( \alpha \) of the tests will yield statistically significant results just by random chance.

---

**“Breakfast Cereal Boosts Chances of Conceiving Boys”**

I’m had identical twin boys a year ago, and I eat breakfast cereal every morning. Do you think this helped to boost my chances of having boys?

a) Yes
b) No
c) Impossible to tell
Multiple Comparisons

• Consider a topic that is being investigated by research teams all over the world

⇒ Using $\alpha = 0.05$, 5% of teams are going to find something significant, even if the null hypothesis is true

---

Multiple Comparisons

• Consider a research team/company doing many hypothesis tests

⇒ Using $\alpha = 0.05$, 5% of tests are going to be significant, even if the null hypotheses are all true

---

Multiple Comparisons

• This is a serious problem

• The most important thing is to be aware of this issue, and not to trust claims that are obviously one of many tests (unless they specifically mention an adjustment for multiple testing)

• There are ways to account for this (e.g. Bonferroni’s Correction), but these are beyond the scope of this class

---

Publication Bias

• publication bias refers to the fact that usually only the significant results get published

• The one study that turns out significant gets published, and no one knows about all the insignificant results

• This combined with the problem of multiple comparisons, can yield very misleading results

---

Jelly Beans Cause Acne!

[Cartoon: http://xkcd.com/882/]

http://xkcd.com/882/
Summary

• Results are statistically significant if the p-value is less than the significance level, $\alpha$.

• In making formal decisions, reject $H_0$ if the p-value is less than $\alpha$, otherwise do not reject $H_0$.

• Not rejecting $H_0$ is NOT the same as accepting $H_0$.

• Two types of errors: rejecting a true null (Type I) and not rejecting a false null (Type II).

•Statistical vs practical significance.

• Using $\alpha = 0.05$, 5% of all hypothesis tests will lead to rejecting the null, even if all nulls are true.

To Do

• Project 1 proposal due TODAY at 11:59pm
• Read Section 4.3, 4.5
• Do HW 4 (due Monday, 2/24)