Hypothesis Testing

Frank Burkholder

Objectives

- Motivation for Hypothesis Testing
- Be able to describe what a Hypothesis Test is
- Be able to list the steps needed to perform a Hypothesis Test
    - Define significance level, test statistic, p-value, type I & II errors
- Select the appropriate Hypothesis Test for what you’re doing
- Correct for issues associated with multiple tests in the same data set
- AB Testing
- Experimental vs Observational Studies
- When to use a Chi square test

Motivating example

An economist wants to determine whether the monthly energy cost for families has changed from the previous year, when the mean cost per month was $260. The economist randomly samples 25 families and records their energy costs for the current year and gets the following values:

mean: $330   std. deviation: $150

Has the monthly energy cost changed or not?

Motivating example

An economist wants to determine whether the monthly energy cost for families has changed from the previous year, when the mean cost per month was $260. The economist randomly samples 25 families and records their energy costs for the current year and gets the following values:

mean: $330 std. deviation: $150

Has the monthly energy cost changed or not?

The picture is muddied because we're looking at a sample rather than the entire population. Due to sampling error, it's possible that while our sample mean is $330, the population mean could still be $260. If we repeated the experiment, the second sample mean could be close to $260. A hypothesis test helps assess the likelihood of this possibility.

Minitab blog

What is a Hypothesis Test

A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

Real world examples of Hypothesis Tests

- Testing whether more men than women suffer from nightmares
- Establishing authorship of documents
- Evaluating the effect of the full moon on behavior
- Determining the range at which a bat can detect an insect by echo
- Deciding whether hospital carpeting results in more infections
- Selecting the best means to stop smoking
- Checking whether bumper stickers reflect car owner behavior
- Testing the claims of handwriting analysts

Wikipedia

Hypothesis Testing vs Estimation

Estimation

- The value of the parameter is unknown
- Goal is to find an estimate and confidence interval for the likely value

Hypothesis Testing

- The value of the parameter is stated
- Goal is to see if the parameter value is different than the stated value

Hypothesis Testing Steps

Overview

1. Formulate your two, mutually exclusive hypotheses.
    a. Null, Alternate
2. Choose a level of significance.
    a. alpha
3. Choose a statistical test and find the test statistic.
    a. t or Z, usually
4. Compute the probability of your results* assuming the null hypothesis is true.
    a. p-value
5. Compare p and alpha to draw a conclusion:
    a. p <= alpha, Reject Null in favor of Alternate
    b. p > alpha, Fail to reject Null

*Similar or *more extreme* results

Step 1 - Formulating your hypotheses

Null Hypothesis (H0)

- Typically a measure of the status quo (no effect)
- The null hypothesis is assumed to be true.

Alternative Hypothesis (Ha)

- Typically the effect that the researcher hopes to detect

Your hypotheses must be mutually exclusive.

The logical framework of hypothesis testing is proof by contradiction.

Step 1 - What would your hypotheses be?

- Testing whether more men than women suffer from nightmares
- Establishing authorship of documents
- Determining the range at which a bat can detect an insect by echo
- Deciding whether hospital carpeting results in more infections
- Establishing innocence or guilt in a court of law

Step 2 - Choosing your significance level

You can’t know things perfectly...

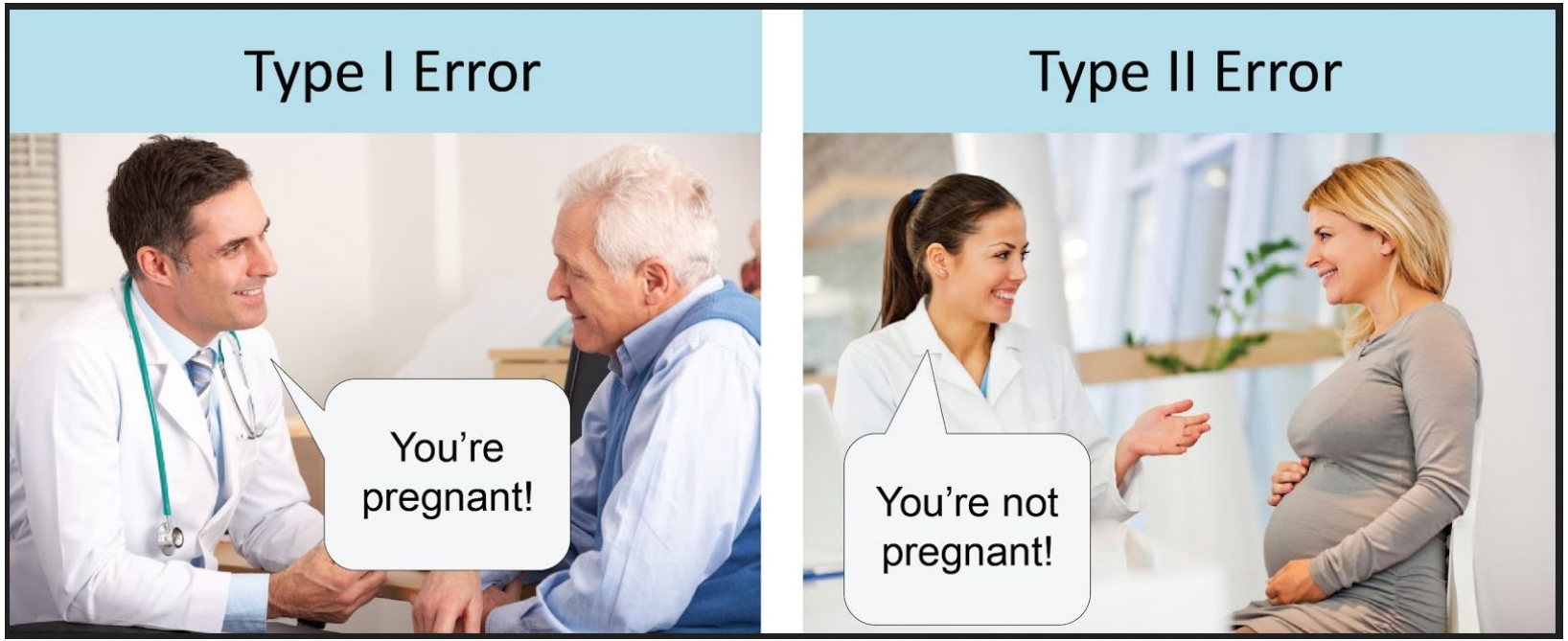
Significance level (alpha) is the probability of rejecting the null hypothesis when it is true.

It’s the chance of making a type I error.

Type I error - False Positive
Type II error - False Negative

Helping you remember Type I and Type II errors

Null hypothesis: **Not Pregnant**



Type I (alpha), False Positive

Type II (beta), False Negative

Step 2 - Add your desired significance level

- Testing whether more men than women suffer from nightmares
- Establishing authorship of documents
- Determining the range at which a bat can detect an insect by echo
- Deciding whether hospital carpeting results in more infections
- Establishing innocence or guilt in a court of law

Step 2 - Adjusting your significance level

When making multiple comparisons, need to adjust significance rates of individual tests (α_i) so that the overall experimental significance level remains the same (α_E)

The Bonferroni correction is straightforward and conservative:

$$\alpha_I = \frac{\alpha_E}{m} \text{ where } m \text{ is number of comparisons}$$

Look at False Discovery Rate for something more standard, less conservative.

Step 2 - Bonferroni correction, a bit more...

Say you have a set of hypotheses (e.g., 20) that you wish to test simultaneously, and a significance level of 0.05.

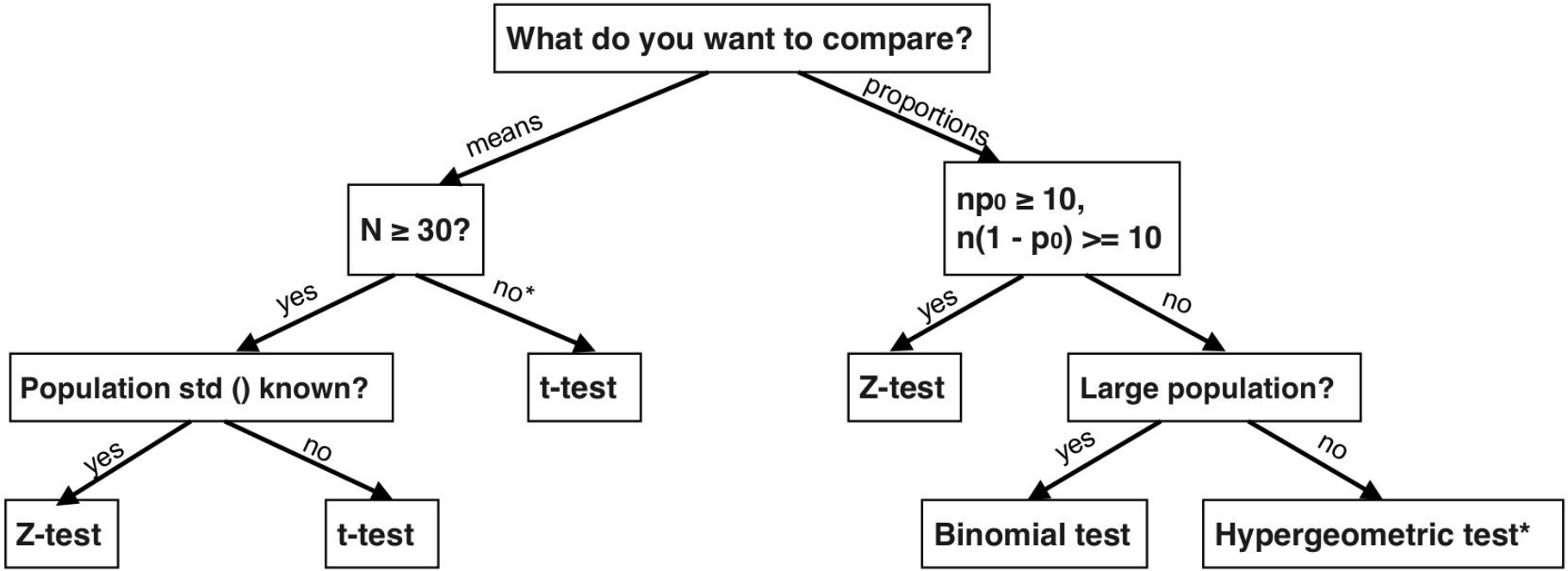
What's the probability of observing at least one significant result just due to chance?

$$
\begin{aligned}
P(\text{at least one significant result}) &= 1 - P(\text{no significant results}) \\
&= 1 - (1 - 0.05)^{20} \\
&\approx 0.64
\end{aligned}
$$

So, with 20 tests being considered, we have a 64% chance of observing at least one significant result, even if all of the tests are actually not significant.

source

Step 3 - Choose your test and test statistic



See the formulas in **hypo_formulas.pdf** and **stat_cheatsheet.pdf** provided with this lecture.

Step 3 - Choose your test and test statistic

Conceptually, what does the test statistic represent?

Step 3 - Choose your test and test statistic

Conceptually, what does the test statistic represent?

It’s the non-dimensionalized distance (in terms of standard deviations) between your two hypotheses.

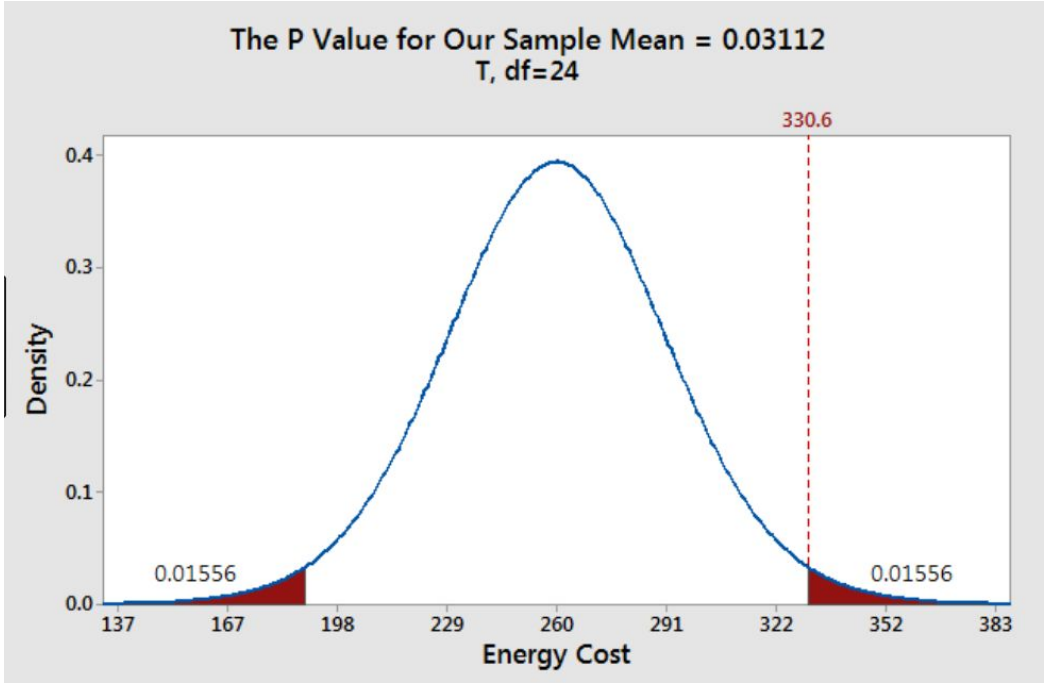
Large number (~2) of standard deviations means they are far apart, low number of standard deviations (~0.5) means they are close together...

Step 4 - Compute your p-value

p-value: The probability of your results (or more extreme) assuming the Null Hypothesis is true.

How:

- Using tables (Z or t)
- Using Python
  - Cumulative distribution function
  - SciPy Stats!



Minitab blog

Step 4 - Compute your p-value

Your p-value will also depend on what type of Hypothesis test you are performing:

Direction	H_0	H_A	P-value
2-sided Test	$=$	$\neq$	One half of P-value in each tail
Left-Tail	$\geq$	$<$	All of P-value in left tail
Right-Tail	$\leq$	$>$	All of P-value in right tail

Step 5 - Draw a conclusion

p <= alpha, Reject Null in favor of Alternate

p > alpha, Fail to reject Null

Do you ever prove the Null hypothesis?

Do you ever prove the Alternate hypothesis?

Hypothesis Testing Steps

Overview

1. Formulate your two, mutually exclusive hypotheses.
    a. Null, Alternate
2. Choose a level of significance.
    a. alpha
3. Choose a statistical test and find the test statistic.
    a. t or Z, usually
4. Compute the probability of your results* assuming the null hypothesis is true.
    a. p-value
5. Compare p and alpha to draw a conclusion:
    a. p <= alpha, Reject Null in favor of Alternate
    b. p > alpha, Fail to reject Null

*Similar or *more extreme* results

Hypothesis Testing Breakout

A principal claims that the students in his school are above average intelligence - the school has a mean score of 112.

Assume the mean population IQ is 100 with a standard deviation of 15.

Is the principal right?

Do it in Python and plot it!

AB-Testing

A/B testing (sometimes called split testing) is comparing two versions of a web page to see which one performs better. It's a Hypothesis Test.

Performance quantified using the Click-Through-Rate.

$$CTR = \frac{number\ of\ people\ that\ click\ on\ something}{number\ of\ people\ that\ view\ it}$$

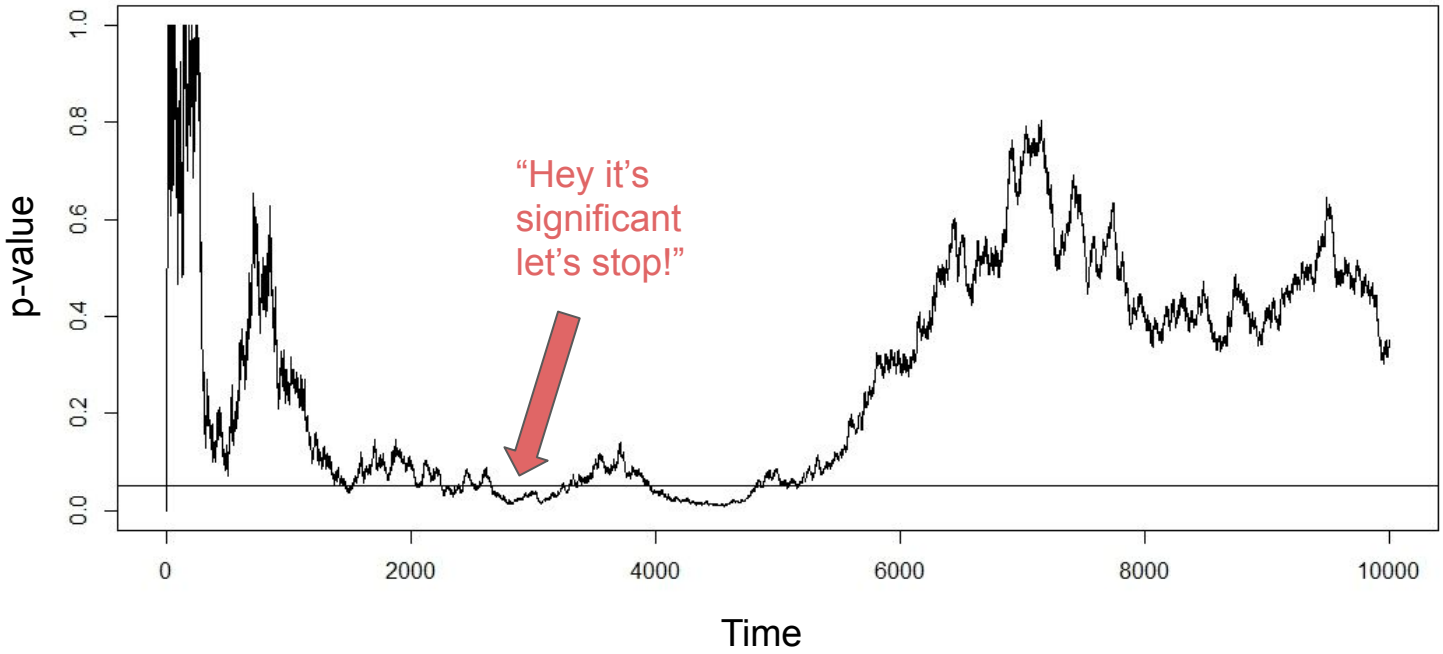
Why? A higher CTR associated with revenue -> more $!

AB-Testing - how to do it

1. Get a good baseline (CTR of Null Hypothesis)
2. Construct an alternative hypothesis
    a. Changing the size of the “checkout” button will increase the CTR
3. Decide for how long you will test (statistical power - later)
    a. You want your test to be able to reject a false null hypothesis (detect an effect if it exists)
4. Start the test, and while testing randomly direct users to one of the two web pages.
5. Don’t peak!
    a. But of course people do - and that gets them into trouble
6. Make your conclusion

AB-Testing

[How Not to Run an A/B Test]



Experimental Studies versus Observational Studies

Experiments:

- Randomly assign subjects to treatments
  - minimizes effect of confounding variables (more in a bit)
- Apply treatments to subjects
- CAN be used to determine causality

Observational:

- Subjects self-select into treatment groups
- Confounding variables often a problem
- CanNOT be used to establish causality (well, very hard: link to paper)

Confounding variables

An attribute (third variable) correlated with both the dependent and the independent variable that affects the association between them.

Example:
Studying relationship between birth order (1st child, 2nd child, etc.) and the presence of Down's Syndrome in the child.

Maternal age is a confounding variable:

1. Higher maternal age is directly associated with Down's Syndrome, regardless of birth order (a mother having her 1st vs 3rd child at age 50 confers the same risk)
2. Maternal age is directly associated with birth order (the 2nd child, except in the case of twins, is born when the mother is older than she was for the birth of the 1st child)
3. Maternal age is not a consequence of birth order (having a 2nd child does not change the mother's age)

Think of another example?

Chi-squared test

Hypothesis test where:

- the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true.
  - compare to *z* and *t* tests where the sampling distribution is assumed to follow the normal or *t* distributions.
- Use: Is there a significant difference between expected and observed frequencies in one or more categories?
  - Goodness-of-fit
    (Is a die fair based on 120 rolls?)
    - see **chi2_goodness_of_fit.ipynb** for example
  - Independence
    (Is the amount you smoke independent of your fitness level?)
    - see **chi2_independence_ex.py** for example

Objectives

- Motivation for Hypothesis Testing
- Be able to describe what a Hypothesis Test is
- Be able to list the steps needed to perform a Hypothesis Test
    - Define significance level, test statistic, p-value, type I & II errors
- Select the appropriate Hypothesis Test for what you’re doing
- Correct for issues associated with multiple tests in the same data set
- AB Testing
- Experimental vs Observational Studies
- When to use a Chi square test