Hypothesis Testing

Joe

galvanize

Introduction

Morning Session Objective

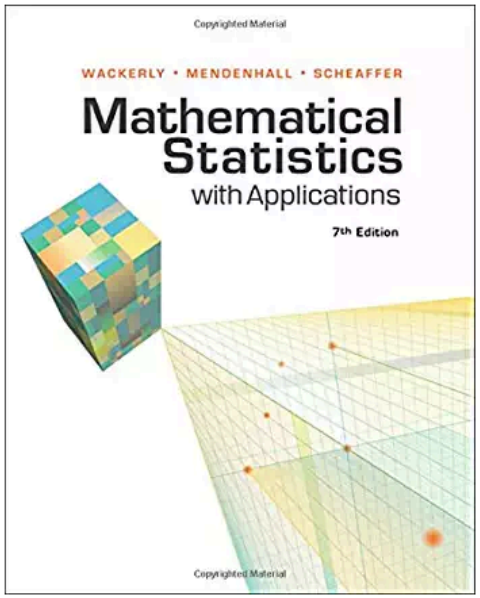
1. Describe the elements of a statistical test

2. Use appropriate terminology to describe accuracy of tests

3. Introduce statistical tests that use the CLT, and their variations

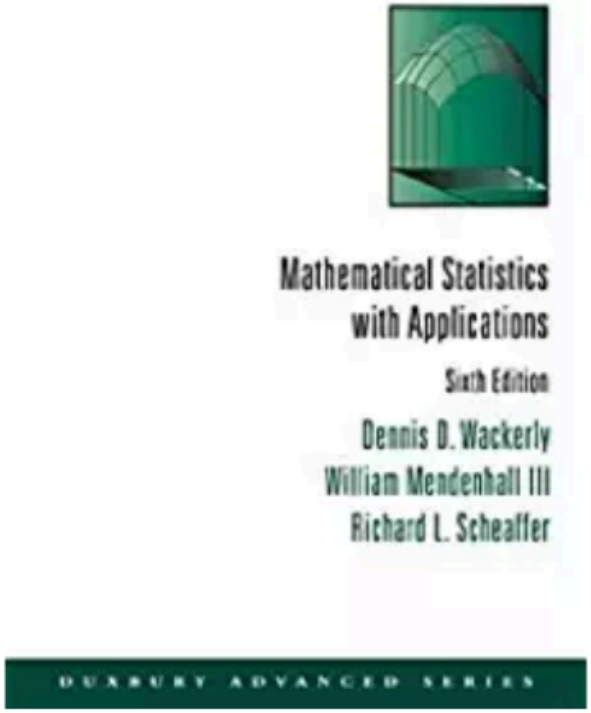
   I really liked Scott Schwartz's lecture on this subject (his Ph.D. is in stats… so yeah). I borrowed some of his derivations but his notebook is really worth a look through.

Resource



Mathematical Statistics with Applications 2008
by Dennis Wackerly and William Mendenhall

Hardcover
$40.71 to rent prime
$181.39 to buy prime
Get it by **Tomorrow, Oct 1**
FREE Shipping on eligible orders



Mathematical Statistics wit
May 30, 2001
by Dennis Wackerly and William Mende

Hardcover
$27.99 (50 used & new offers)

Paperback
$23.13 (13 used & new offers)

Other Formats: Hardcover
See newer edition of this book

Same book

galvanize

Why Perform Statistical Tests

Suppose you have a candidate who needs 50% of the vote to win.

You conduct a poll of 15 people and find 4 people prefer your candidate.

What, if anything, can you conclude?

Statistical Testing Basics

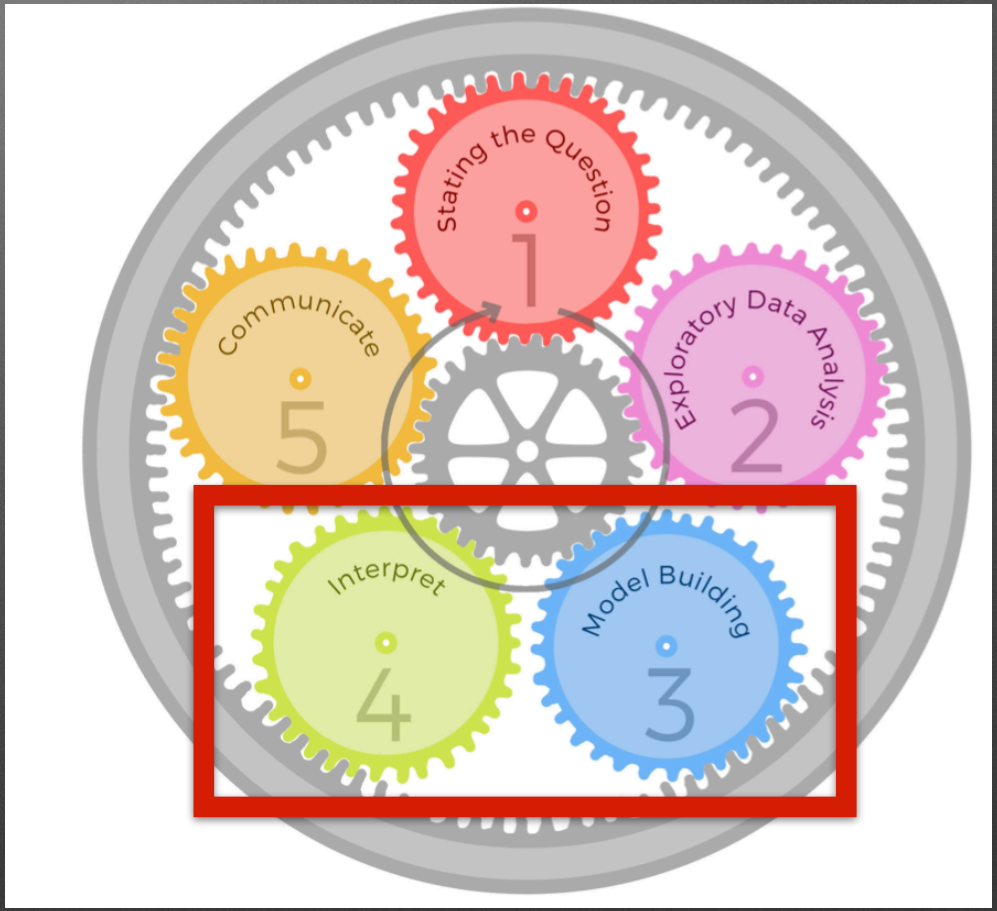
Statistical Tests in Data Science

Statistical tests allow us to:

1. Understand how much data we need to perform a test

2. Determine if a sample allows us to make conclusions about the population

3. Quantify the accuracy of any conclusion we may draw



Elements of a Statistical Test

1. A Null Hypothesis (H_0)
2. An Alternative Hypothesis (H_a)
3. A Test Statistic
4. A Rejection Region

What do these phrases mean to you?

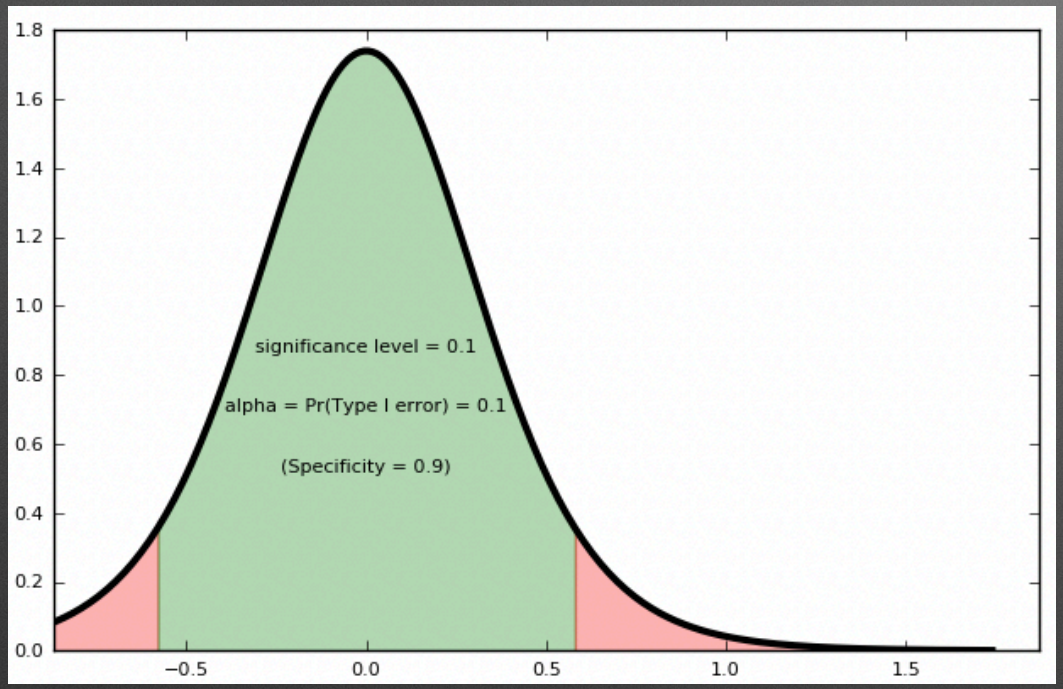
H_0 and H_a

Null Hypothesis (H_0)
The null hypothesis describes how the universe exists in isolation of the effect under study

Alternative Hypothesis (H_a)
A hypothesis that is possible iff the test rejects the null hypothesis

N.B. We do not prove the null hypothesis, we only fail to reject it.

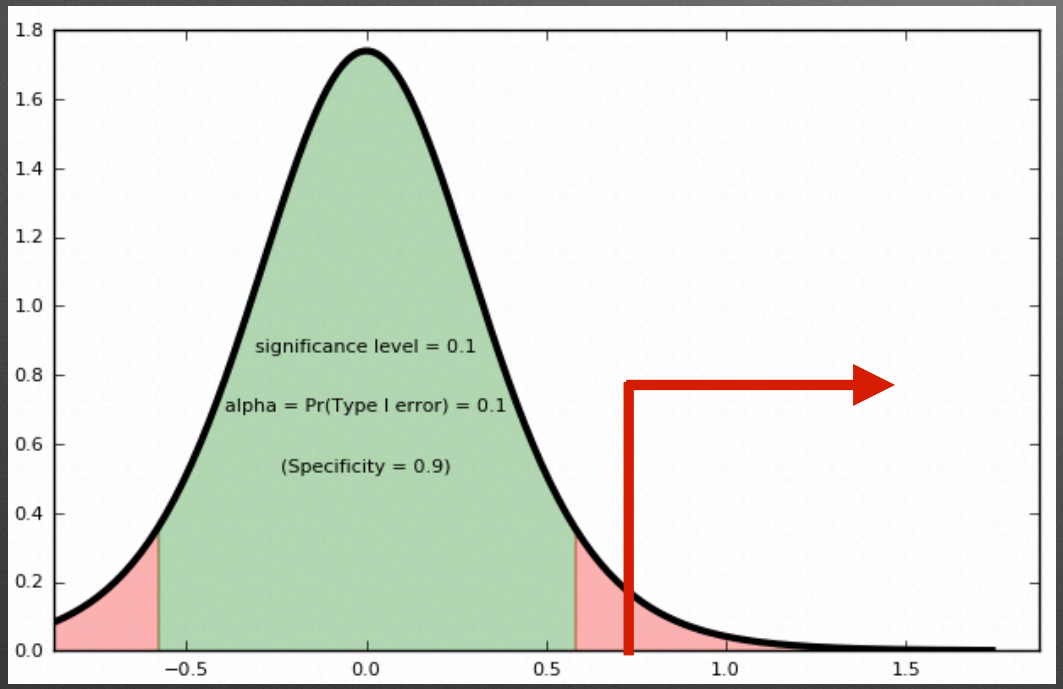
Test Statistic and the Rejection Region



Test Statistic - a function that we use to estimate the properties of the population based on a sample

Rejection Region (RR) - the values of the test statistic for which the null hypothesis is rejected

Test Statistic and the Rejection Region



Test Statistic - a function that we use to estimate the properties of the population based on a sample

Rejection Region (RR) - the values of the test statistic for which the null hypothesis is rejected

A sample will yield some measurement that we compare to our model. A p-value is the probability of producing as extreme an observation as we have, given the null hypothesis is true. Typically, significance levels are set to be 5%.

A Basic Statistical Test

From the Introduction

Suppose you have a candidate who needs 50% of the vote to win.

You conduct a poll of 15 people and find 4 people prefer your candidate.

Let's design our statistical test.

Building a Test

Suppose you have a candidate who needs 50% of the vote to win.

You conduct a poll of 15 people and find 4 people prefer your candidate.

1. Our Null Hypothesis (H_0)
2. Our Alternative Hypothesis (H_a)
3. Our Test Statistic
4. A Rejection Region

Building a Test

Suppose you have a candidate who needs 50% of the vote to win.

You conduct a poll of 15 people and find 4 people prefer your candidate.

1. Our Null Hypothesis (H_0)
   50% of people prefer our candidate
2. Our Alternative Hypothesis (H_a)
   < 50% of people prefer our candidate
3. Our Test Statistic
   Binomial probability based on p=.5
4. A Rejection Region
   $\alpha = .05$

Let's hop into the first N.B. to check this out.

CLT Tests

Recall CLT

- CLT states that a sample statistic, such as a mean, has a distribution that is approximately normal when the sample size is large.

- Here, we are able to model the distribution of say, likely voters, with the normal distribution, where H_0 is the candidates are neck in neck, and H_a is that our candidate has a lead.

T-Test

- The “Student’s T-Test” was developed by a Guinness employee named William Gosset
  - Called “Student” because Guinness employees were not permitted to publish their work
- The T-Test attempts to leverage the CLT in low statistic samples
- Low stats + CLT means a T distribution is correct, but it also means the applicability is limited..

Deriving the T-Test

Let $X_i \overset{i.i.d.}{\sim} f(\theta)$, for $i = 1, \cdots n$, and suppose we are interested in testing:

$$\begin{cases} H_0 : \mathrm{E}[X_i] = \mu_0 \\ H_a : \mathrm{E}[X_i] \neq \mu_0 \end{cases}$$

If X_i is distributed normally or the CLT applies* than if H_0 is true

$$\bar{X} - \mu_0 \overset{approx}{\sim} N\left(0, \frac{\mathrm{Var}[X]}{n}\right)$$

However, if we *do not know* Var[X] we would need to estimate it,

and can do so in an unbiased manner with

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

Unfortunately, if the CLT has not yet ``kicked in"
then *only if* $f(\theta) \sim N(\mu, \sigma^2)$ do we have that

$$\frac{\bar{X} - \mu_0}{\sqrt{s^2/n}} \sim t_{n-1}$$

But as $n \to \infty$,

$$t_n \longrightarrow N(0, 1)$$

Deriving the T-Test

Let $X_i \overset{i.i.d.}{\sim} f(\theta)$, for $i = 1, \cdots n$, and suppose we are interested in testing:

$$\begin{cases} H_0 : \mathrm{E}[X_i] = \mu_0 \\ H_a : \mathrm{E}[X_i] \neq \mu_0 \end{cases}$$

We will discuss 1 vs 2 tailed tests in greater detail tomorrow

If X_i is distributed normally or the CLT applies* than if H_0 is true

$$\bar{X} - \mu_0 \overset{approx}{\sim} N\left(0, \frac{\mathrm{Var}[X]}{n}\right)$$

However, if we *do not know* Var[X] we would need to estimate it,

and can do so in an unbiased manner with

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

Unfortunately, if the CLT has not yet ``kicked in''
then *only if* $f(\theta) \sim N(\mu, \sigma^2)$ do we have that

$$\frac{\bar{X} - \mu_0}{\sqrt{s^2/n}} \sim t_{n-1}$$

But as $n \to \infty$,

$$t_n \longrightarrow N(0, 1)$$

Let's hop into the next n.b. for a quick walkthrough

Connection to Confidence Intervals

Recall:

$$P(\bar{X} - 1.96 * \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 * \frac{\sigma}{\sqrt{n}}) = .95$$

iff

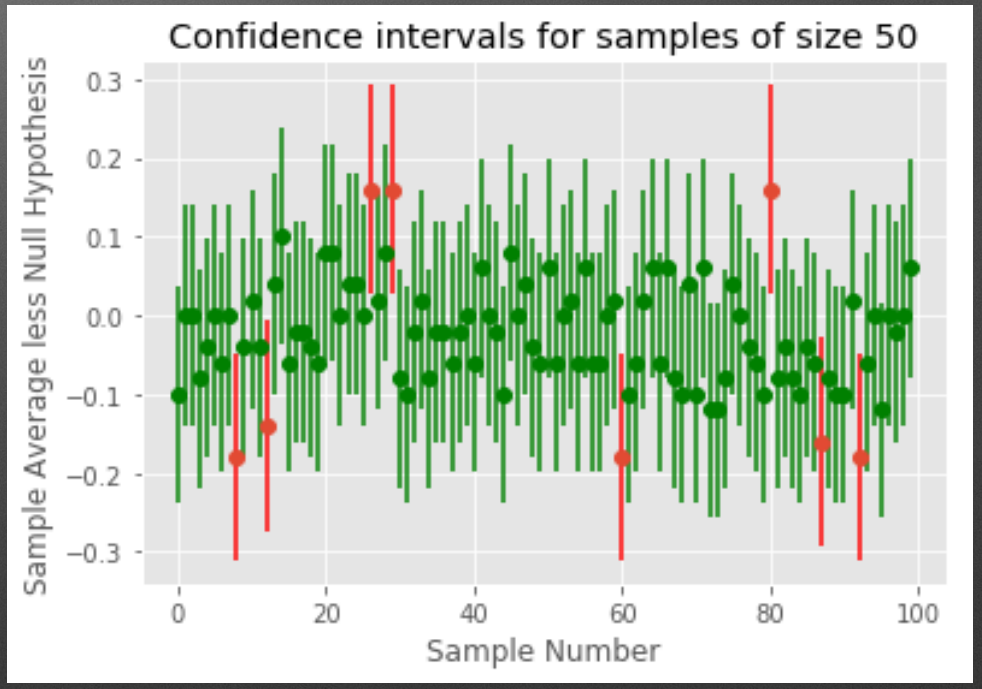
$$X_i \approx N(\mu, \sigma)$$

More generally,

$$P(\bar{X} - t_{n-1}^{\alpha/2} * \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + t_{n-1}^{\alpha/2} * \frac{\sigma}{\sqrt{n}}) = .95$$

The connection to hypothesis testing is that a 100%*(1-α) c.i.:

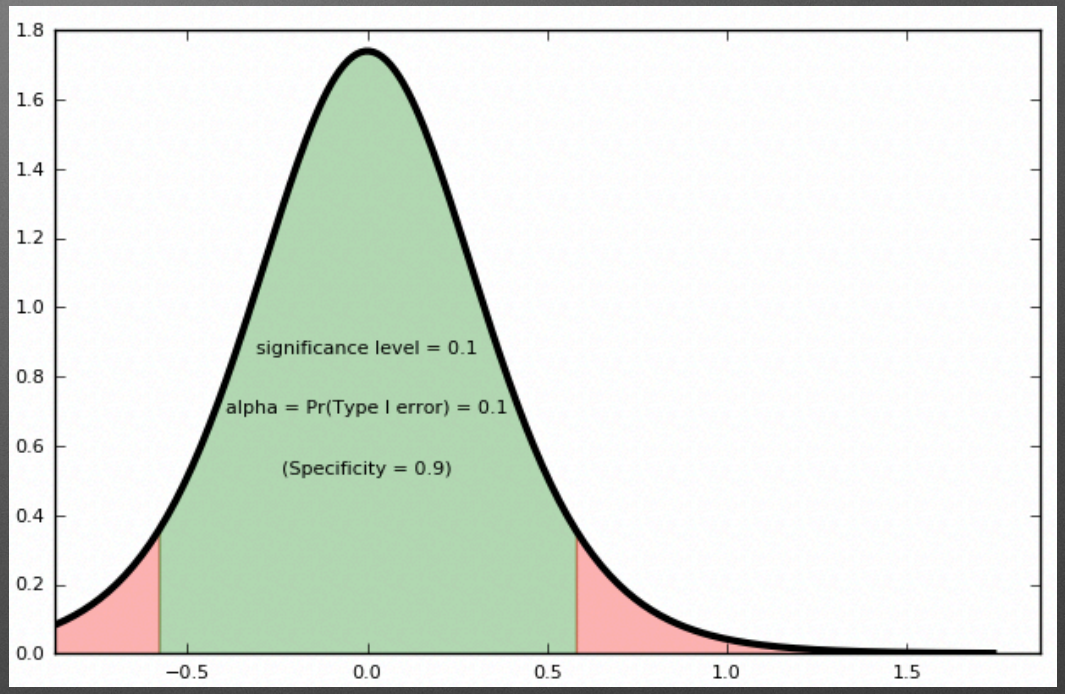
- that does not contain μ_0 rejects the null hypothesis in a two tailed test.
- that does contain μ_0 fails to reject the null hypothesis in a two tailed test.



galvanize

Z-Test

- A Z-test is a T-test in the limit where the sample is large enough that it can be safely approximated by the normal distribution
- Typically, you will calculate a Z-score (shown below) which is compared to a table value which will tell you
- Q: What is the Z score for the two tailed Z-test that achieves significance of .05?



$$Z = \frac{\bar{X} - \mathrm{E}[X]}{\sigma(X)/\sqrt{n}}$$

Two Sample T-Test

Suppose we have 2 samples X_i and Y_i, we desire to know if they have different means:

$H_0 : E[\bar{X}] = E[\bar{Y}]$

$H_a : E[\bar{X}]! = E[\bar{Y}]$

This test can be performed via the 2 distribution T-test. For the 'students' version of this test, our t statistic can be expressed as:

$$t = \frac{\mu_1 - \mu_2}{s_p^2 * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \text{ and } n_{dof} = n_1 + n_2 - 2$$

While this is the 'industry standard' way of proceeding, the students two sample has an implicit assumption that $Var(X) = Var(Y)$, and in cases where this assumption is not true, the test can lead to high type 1 & 2 errors. We prefer the unequal variance test which has the relation:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and

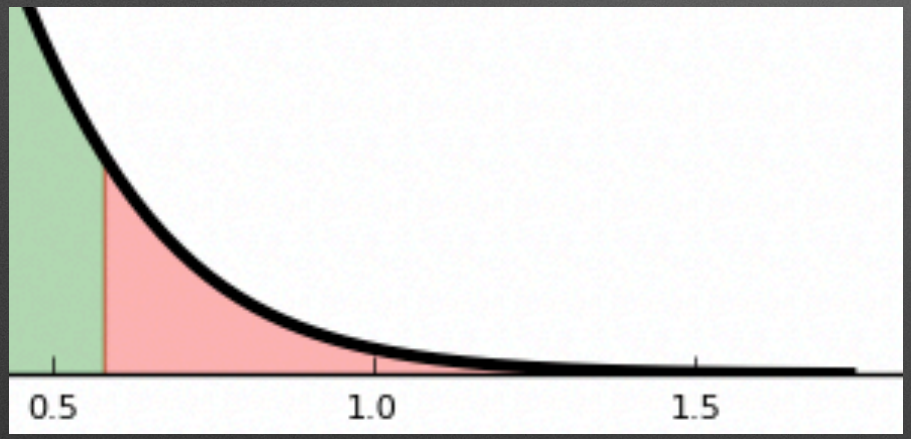
$$n_{dof} = \frac{(\frac{1}{n_1} + \frac{u}{n_2})^2}{\frac{1}{n_1^2(n_1-1)} + \frac{u^2}{n_2^2(n_2-1)}} \text{ where } u = \frac{s_2^2}{s_1^2}$$

You'll be using the two sample test in the assignments, so you'll get to see how to run these tests first hand

Error Analysis

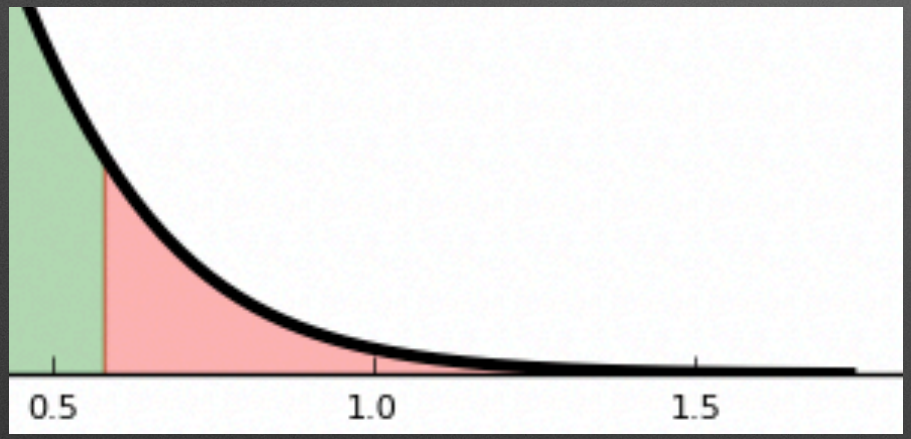
Defining RR

Q: What makes a good statistical test?



Q: How does that inform how we define RR?

Defining RR



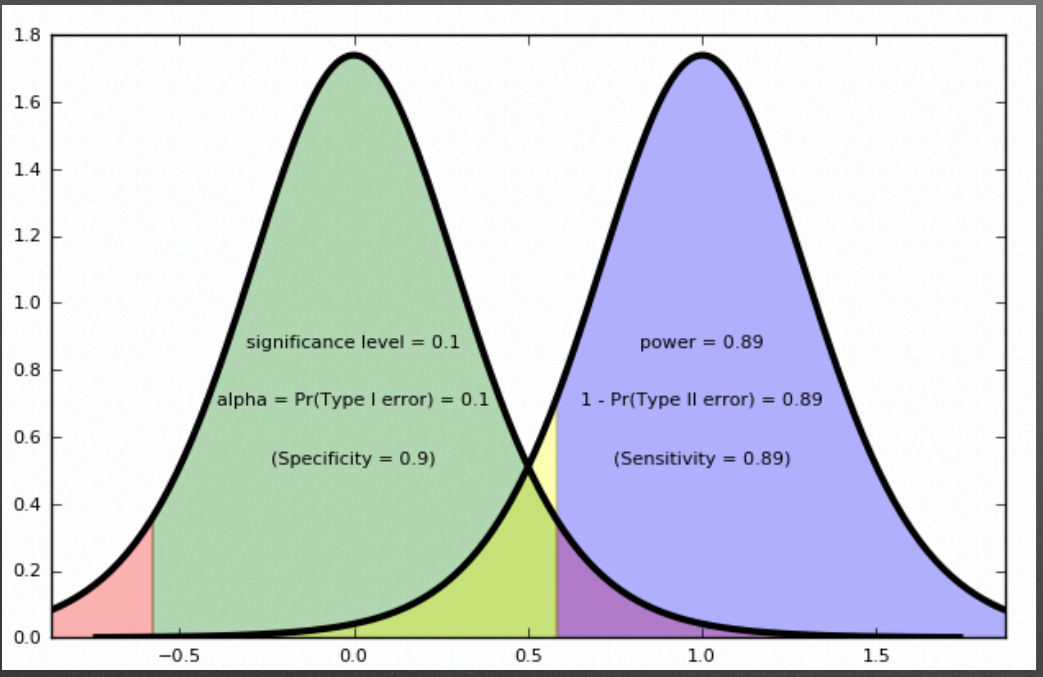
Q: What makes a good statistical test?

A: A good statistical test is one in which we only reject the null hypothesis if it is untrue

Q: How does that inform how we define RR?

A: We define the RR through a significance level α

Error Types



Type 1:

$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$

We call α our significance level

Type 2:

$\beta = P(\text{Fail to reject H0} \mid H_A \text{ is true})$

We call $(1-\beta)$ our test power

We'll discuss type 2 error in detail on Thursday morning

Questions

What are the differences between type 1 & type 2 errors?

What is α?

What is β?

What is the difference between α and a p-value?

P-Value

Quoting Scott Schwartz's lecture "P-value blunders for which I'll never forgive you, and which will haunt you for the rest of your natural life":

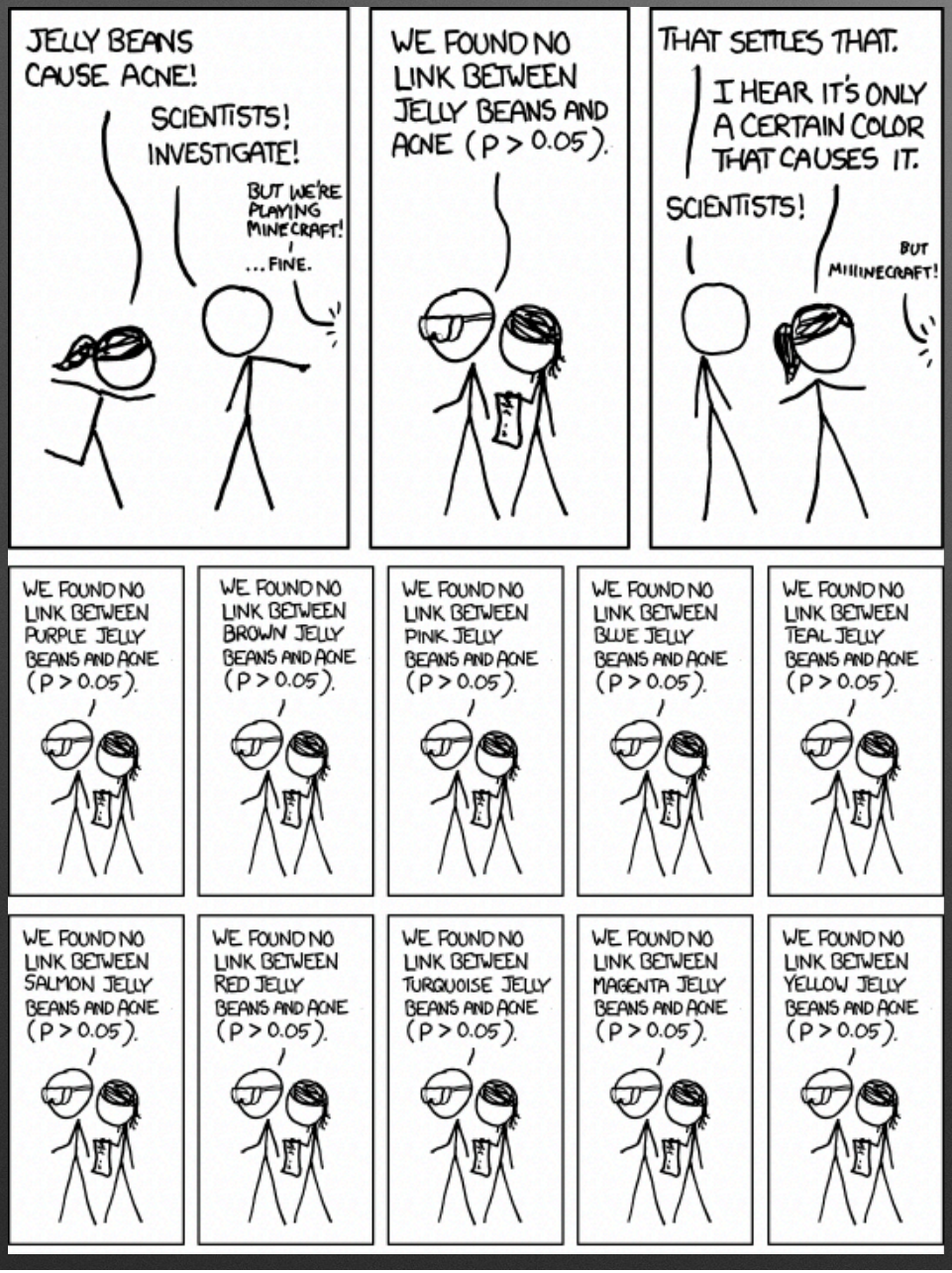
- A p-value *is not* the probability H_0 is False
  - H_0 is true or false

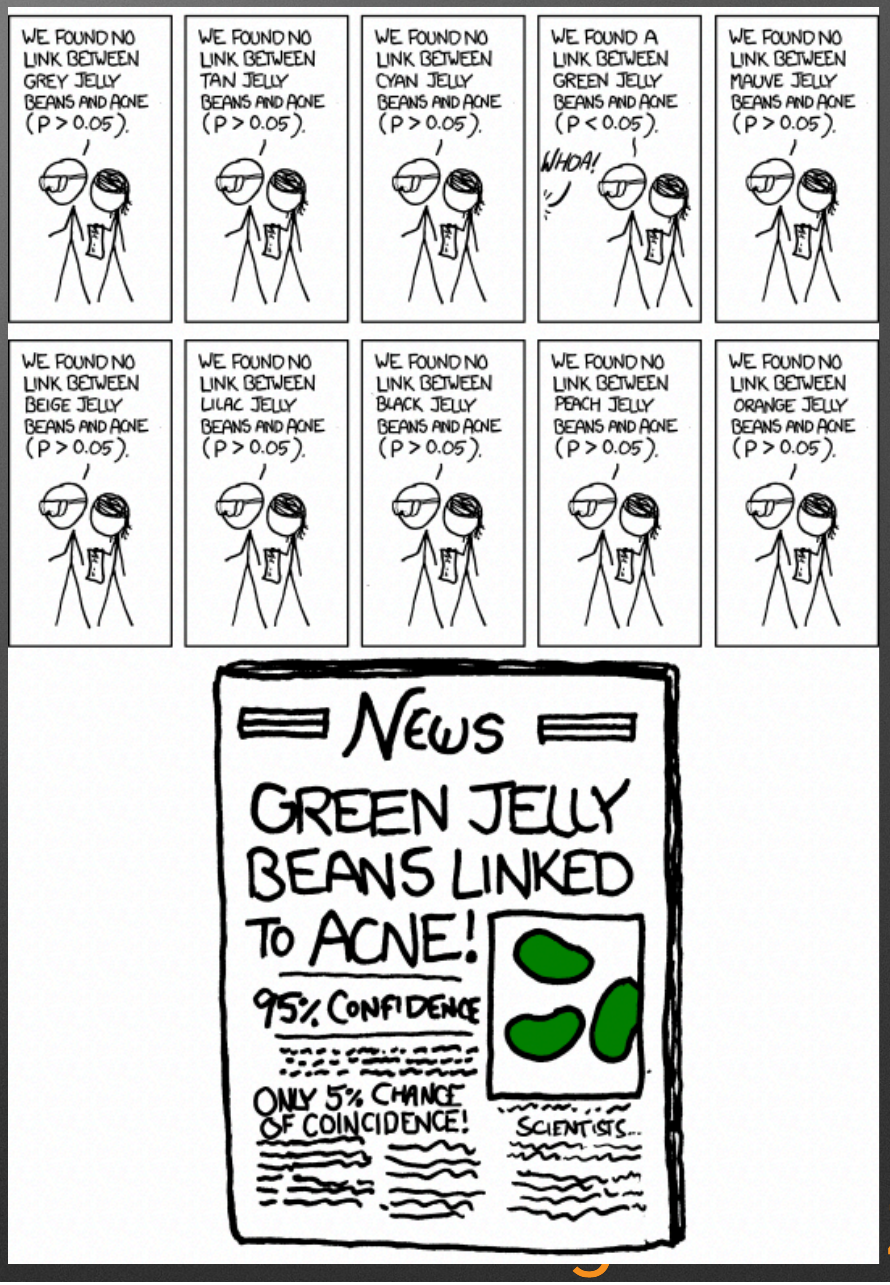
- A p-value *is not* the probability of incorrectly rejecting H_0
  - Significance level α is the probability of wrongly rejecting H_0

- A p-value *is :*
  P(seeing something as or more extreme than what you saw | H_0 is true)

Words have meanings! p-value is defined for a sample, α is defined for an experiment.

Multiple Testing





Multiple Testing

α is the probability that for a given test, we wrongly reject H_0

Thus, if we run multiple test, we are confident that we will randomly reject H_0 even though it is true.

To account for this factor, we enforce the Bonferroni correction which states that $\alpha' = \alpha/N$ give an α chance all tests are correct

Alternatively, one can use the False Discovery Rate (FDR), which for a set of tests is the proportion q of the tests called incorrectly

Morning Session Objective

1. Describe the elements of a statistical test
2. Use appropriate terminology to describe accuracy of tests
3. Introduce statistical tests that use the CLT, and their variations

End of Morning's Lectures

Afternoon Session Objective

1. Use alternative statistical tests to solve a variety of alternate statistical problems

2. Outline the process of A/B testing and how to conduct such tests

Other Statistical Tests

Scenario 2

Suppose the initial example was data for the state of Texas. Now I want to determine if my candidate polls differently in different states.

Example Problem - Fisher's Exact

PRESIDENTIAL PREFERENCE	TEXAS	CALIFORNIA	ROW TOTALS
DEMOCRAT	4	15	19
REPUBLICAN	11	5	16
COLUMN TOTALS	15	20	35

More Generally…

CLASSIFICATION	GROUP 1	GROUP 2	ROW TOTALS
CLASS 1	A	B	A+B
CLASS 2	C	D	C+D
COLUMN TOTALS	A+C	B+D	N = A + B + C +C

Fisher's Exact Test

Fisher's Exact Test states that for the above scenario, we can use combinatorics to describe the probabilities of seeing those distributions from random chance.

The distribution that describes this probability is the hypergeometric distribution

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

Fisher's Exact Test (cont)

- This hypergeometric probability only describes the probability of getting the exact observed distribution.

- In order to conduct the test completely, we need to test our observed data, and the more extreme cases
  - The alternative cases need to preserve row/column totals
  - For our example, we would need to calculate all 5 hypergeometric probabilities….thankfully computers exist

4	15	19
11	5	16
15	20	35

3	16	19
12	4	16
15	20	35

2	17	19
13	3	16
15	20	35

1	18	19
14	2	16
15	20	35

0	19	19
15	1	16
15	20	35

```
>>> import scipy.stats as stats
>>> oddsratio, pvalue = stats.fisher_exact([[4,15],[11, 5]], alternative='less')
>>> pvalue
0.005785816769034808 1
```

Scenario 3

Suppose we have speaking points that our candidate recites during speeches, and polling data for people who were undecided on the candidate before the lecture, and formed an opinion on them afterwards. Do talking points matter when swaying undecideds?

Candidate Speaking Points

POST SPEECH OPINION	MILITARY	ECONOMY	IMMIGRATION
PREFERS CANDIDATE	27	15	30
DOES NOT PREFER CANDIDATE	18	15	8

Pearson's χ^2 Test

- Pearson's χ^2 test (typically just called a χ^2 test), is a test for independence on two sets of data based on some categorization.
- H_0 is that the the distributions are independent of the categories
- H_a is that they are not
- There are other χ^2 tests which will be covered later on.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(O_i/N - p_i)^2}{p_i}$$

where

χ^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.

O_i = the number of observations of type *i*.

N = total number of observations

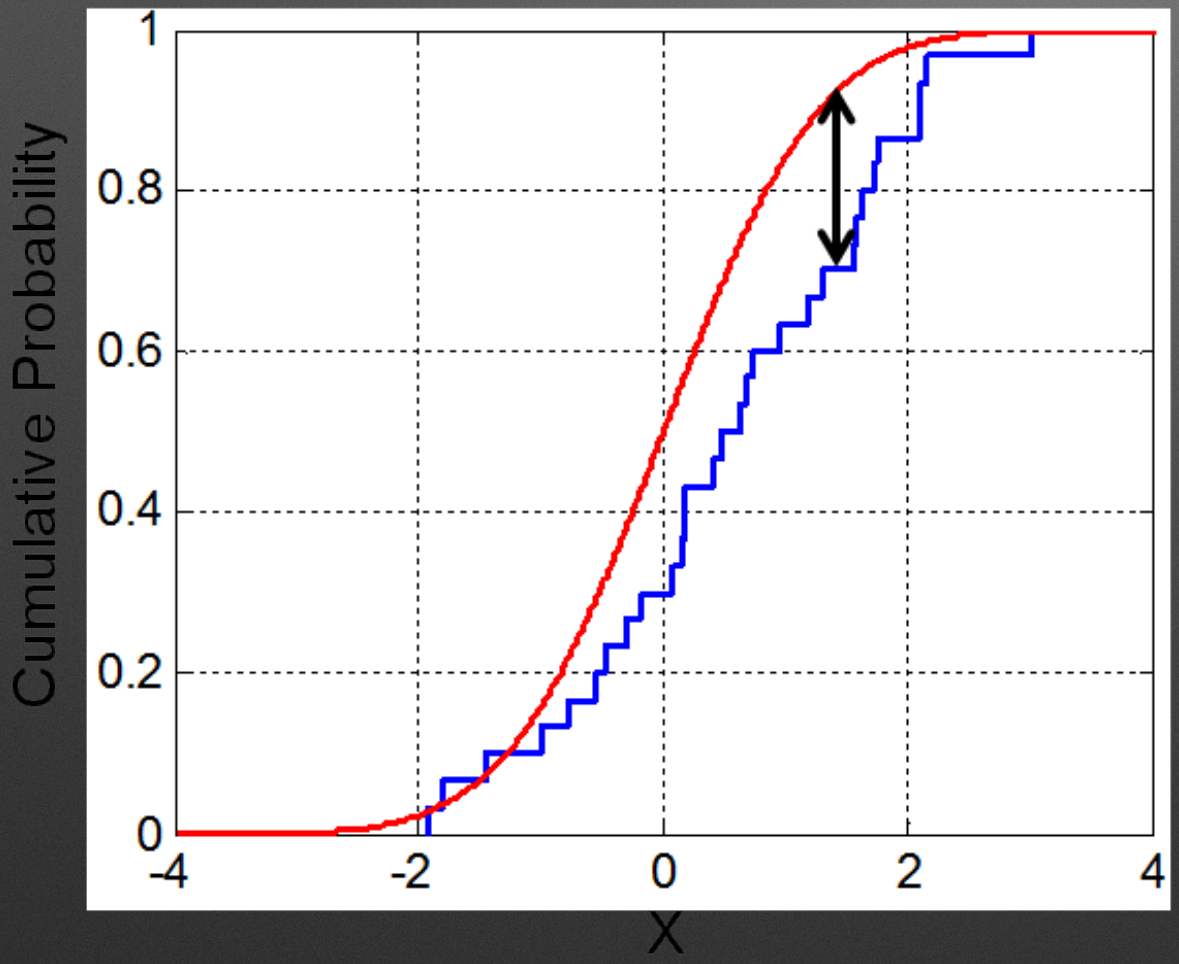
$E_i = Np_i$ = the expected (theoretical) frequency of type *i*, asserted by the null hypothesis that the fraction of type *i* in the population is p_i

n = the number of cells in the table.

Scenario 4

Suppose we have data that you are unsure of the distribution, but have a guess. Is there a function that can test to see if a distribution is a possible match without feeding model parameters?

Kolmogorov-Smirnov

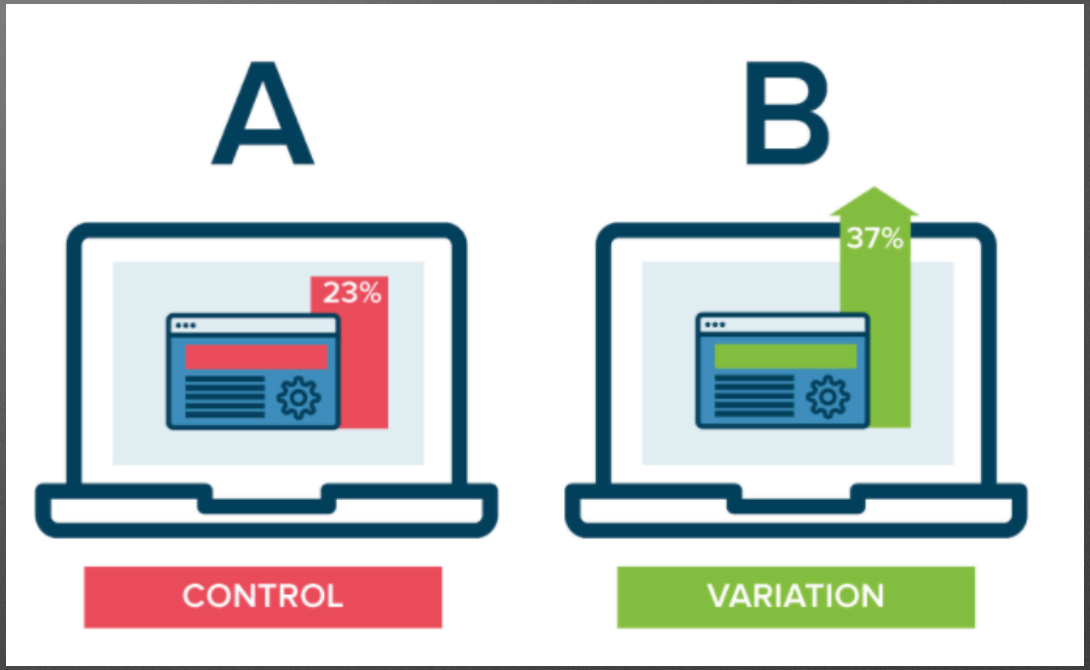


- Kolmogovor-Smirnov Tests (KS Tests) are non-parametric tests that order data, and note differences between the data and the cumulative density function.
- KS Tests typically optimize the CDF so parameters are not included in the fit.

A/B Testing

A/B Testing

- A/B testing is a common task is website design and marketing campaigns



- Changes to layouts, button placement, and other factors can all be test

- Process
  1. Create a page that is variation on the default version of the page
  2. Display the pages in a way that isolates the changes you are hoping to test
  3. Collect data for the performance of A & B, and run a hypothesis test.

Q: What are the elements of the hypothesis test?

Sensitivity of A/B Testing

- There is value to stable formats for websites, as such typically a minimum threshold can be calculated to determine when a change should be made
- Tomorrow, we'll be covering power calculations, which will help us uncover other important properties of a hypothesis test.

Afternoon Session Objective

1. Use alternative statistical tests to solve a variety of alternate statistical problems

2. Outline the process of A/B testing and how to conduct such tests