Statistical Hypothesis Testing

Schwartz

September 23, 2017

A Brief History of Statistics (Part I)

Significance testing is largely the product of Karl Pearson (1857–1936), William Sealy Gosset (1876–1937), and Ronald Fisher (1890–1962), although evidence of its use dates back to Laplace (1749–1827) in the 1770's. Pearson created the notion of a *p-value* and (Pearson's) *chi-squared test* and founded the world's first statistics department at University College London in 1911. Gosset developed and penned the *t-distribution* and *t-test* under the pseudonym Student due to the objections of his employer – the original Guinness Brewery in Dublin, Ireland – regarding publication of internal practices. And Fisher created *analysis of variance* and popularized the notions of *null hypothesis* and *significance test*. In addition to being regarded as the father of modern statistical science and experimental design, Fisher also made significant contributions to agricultural biology and genetics. Indeed, Richard Dawkins named him "the greatest biologist since Darwin".

Hypothesis testing was developed by Jerzy Neyman (1894 – 1981) and Egon Pearson (1895–1980, son of Karl Pearson). Building on these ideas, Neyman later introduced *confidence intervals* into the statistics landscape. At the time of the publication of their work on hypothesis testing in 1933, Neyman and Pearson (along with Fisher) were faculty members at the University College London in the department of statistics (founded by the older Pearson). While Fisher as a result of his agricultural background emphasized rigorous experimental design and methods to extract a result from few samples assuming Gaussian distributions, Neyman (who teamed with the younger Pearson) emphasized mathematical rigor and methods to obtain more results from many samples and a wider range of distributions.

Initially a *Bayesian*, Fisher but sought to provide a more "objective" approach to inference. The significance testing he developed did not use the notion of an alternative hypothesis – only a null hypothesis – and hence did not involve the notion of *Type II error*. Fisher's interpretation of p-values was informal: p-values were only meant to provide guidance for potential future experiments. Neyman and Pearson on the other hand formalized hypothesis testing with *Type I/II errors* and developed a procedure to choose between competing hypotheses. They considered their formulation to be an improved and more objective generalization of significance testing as it provided a decision making tool to determine researcher behavior without requiring any inductive inference on the part of the researcher.

A Brief History of Statistics (Part II)

Fisher and Neyman/Pearson clashed bitterly, and often – they all shared the same building at the University College London they had ample opportunity to cross paths (and swords – although only Fisher was ever knighted – and not until many years later – and Neyman was, after all, Polish, not English). They disagreed about the proper role of models in statistical inference. Fisher thought the Neyman/Pearson approach was not applicable to scientific research because (1) initial assumptions about the null hypothesis are often discovered to be questionable as unexpected sources of error appear over the course of the experiment and (2) rigid reject/accept decisions based on models formulated before data is collected are incompatible with the real-world scenario faced by scientists and attempts to apply such formulations to scientific research would lead to mass confusion (as it has).

In 1938 Neyman left University College London and moved to the University of California, Berkeley. This put much of the planetary diameter between both his partnership with Pearson and his dispute with Fisher. A further respite in the debate was provided by World War II. Nonetheless, the disagreement between Fisher and Neyman only terminated (unresolved after 27 years) with Fisher's death in 1962. Neyman wrote a well-regarded eulogy of Fisher upon his death. And some of Neyman's later publications reported p-values and significance levels.

Afterword:

In an apparent effort to provide a "non-controversial" theory (as well as likely from confusion and misunderstanding of the topic, *per se*) the modern version of hypothesis testing used today is an inconsistent hybrid of the "Fisher versus Neyman/Pearson" formulations developed in the early 20th century. Rather than comparing two directly competing realistic hypotheses, one of the hypotheses is made to be a "no effect null hypothesis" so (despite great conceptual differences and caveats) p-values can be interpreted from both the Fisher and the Neyman/Pearson perspectives. Neyman and Pearson provided the stronger terminology, the more rigorous mathematics and the more consistent philosophy, but the hypothesis testing used today has more similarities with Fisher's method than theirs.

Outline

- Know what a **null hypothesis** is
  - Know what an α-**significance level** is
    - Know what a **two-tailed versus one-tailed test** is
    - Know how this relates to **confidence intervals**
  - Know what a **p-value** is (don't you *dare* mess this up *EVAR*)
- Know what an **alternative hypothesis** is
  - Know what **power** is
- Know what Type I and Type II errors are
- And know yourself a little *Bonferroni* and *FDR*

- Be able to navigate in the hypothesis testing universe:
  - z/t-test with common or unique variances, or paired samples
  - Pearson's χ^2-test, and other χ^2-tests
  - Fisher's exact test
  - The other "usual suspects" in terms of non-parametric tests
  - Kolmogorov-Smirnov (K-S) test
  - F-test
  - And be able to correctly apply them where appropriate

Hypothesis Testing Concepts I: **The Scientific Method**

0. What is your *question*?

Hypothesis Testing Concepts I: **The Scientific Method**

0. What is your *question*?
1. What is your *population* (i.e., statistical distribution)?

Hypothesis Testing Concepts I: **The Scientific Method**

0. What is your *question*?
1. What is your *population* (i.e., statistical distribution)?
2. Specify a *null hypothesis* H_0

Null hypothesis H_0

A "no effect" / "boring baseline" statement to "disprove"

E.g., proposing no difference between two population means

Hypothesis Testing Concepts I: **The Scientific Method**

0. What is your *question*?
1. What is your *population* (i.e., statistical distribution)?
2. Specify a *null hypothesis* H_0

   Null hypothesis H_0

   A "no effect" / "boring baseline" statement to "disprove"

   E.g., proposing no difference between two population means

3. Choose a *test statistic* that is dependent on H_0

Hypothesis Testing Concepts I: **The Scientific Method**

0. What is your *question*?
1. What is your *population* (i.e., statistical distribution)?
2. Specify a *null hypothesis* H_0

   Null hypothesis H_0
   A "no effect" / "boring baseline" statement to "disprove"
   E.g., proposing no difference between two population means

3. Choose a *test statistic* that is dependent on H_0
4. Determine the statistic's *theoretical distribution* if H_0 is true

Hypothesis Testing Concepts I: **The Scientific Method**

0. What is your *question*?
1. What is your *population* (i.e., statistical distribution)?
2. Specify a *null hypothesis* H_0

   Null hypothesis H_0

   A "no effect" / "boring baseline" statement to "disprove"
   E.g., proposing no difference between two population means

3. Choose a *test statistic* that is dependent on H_0
4. Determine the statistic's *theoretical distribution* if H_0 is true
5. Specify a H_0 *rejection region*, i.e., values of the test statistic which cast great doubt (i.e., α level) on the accuracy of H_0

   Significance level α

$$\alpha = \Pr(\text{rejecting } H_0 | H_0 \text{ is true}) = \Pr(\text{"Type I error"})$$

Hypothesis Testing Concepts I: **The Scientific Method**

0. What is your *question*?
1. What is your *population* (i.e., statistical distribution)?
2. Specify a *null hypothesis* H_0

Null hypothesis H_0

A "no effect" / "boring baseline" statement to "disprove"

E.g., proposing no difference between two population means

3. Choose a *test statistic* that is dependent on H_0
4. Determine the statistic's *theoretical distribution* if H_0 is true
5. Specify a H_0 *rejection region*, i.e., values of the test statistic which cast great doubt (i.e., α level) on the accuracy of H_0

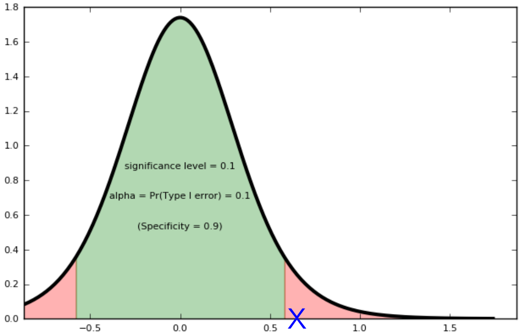
Significance level α

$$\alpha = \Pr(\text{rejecting } H_0 | H_0 \text{ is true}) = \Pr(\text{"Type I error"})$$

6. Collect data and *compute the test statistic*

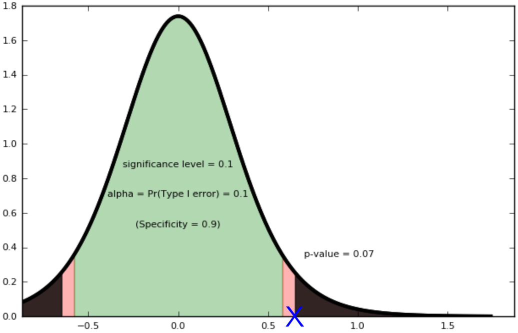
Hypothesis Testing Concepts II: **REJECTING** H_0

7. Reject H_0 if the test statistic looks "strange". The statistic is a sample from the null distribution conditional on H_0 being true, so if it doesn't look right then H_0 seems wrong...



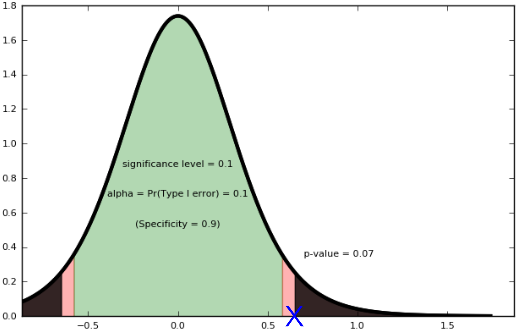
Hypothesis Testing Concepts II: **REJECTING H_0**

7. Reject H_0 if the test statistic looks "strange". The statistic is a sample from the null distribution conditional on H_0 being true, so if it doesn't look right then H_0 seems wrong...



Hypothesis Testing Concepts II: **REJECTING** H_0

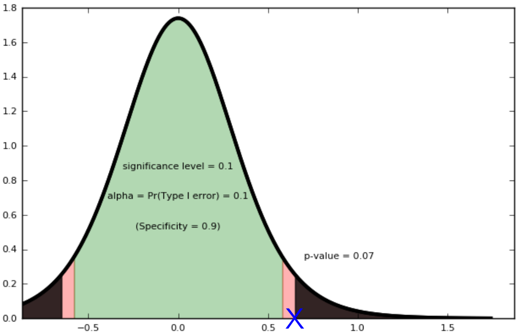
7. Reject H_0 if the test statistic looks "strange". The statistic is a sample from the null distribution conditional on H_0 being true, so if it doesn't look right then H_0 seems wrong...



$$\frac{\text{p-value}}{\Pr(\text{seeing something as or more extreme than what you saw} | H_0 \text{ true})}$$

Hypothesis Testing Concepts II: **REJECTING** H_0

7. Reject H_0 if the test statistic looks "strange". The statistic is a sample from the null distribution conditional on H_0 being true, so if it doesn't look right then H_0 seems wrong...



p-value

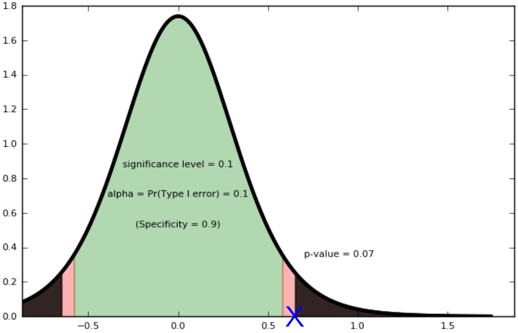
Pr(seeing something as or more extreme than what you saw$|H_0$ true)

One-tailed VS. two-tailed tests

How probability α defining a Type I error is allotted over H_0

Hypothesis Testing Concepts II: **REJECTING** H_0

7. Reject H_0 if the test statistic looks "strange". The statistic is a sample from the null distribution conditional on H_0 being true, so if it doesn't look right then H_0 seems wrong...



Oh, and –
We *NEVER* *"accept"* H_0

p-value

Pr(seeing something as or more extreme than what you saw$|H_0$ true)

One-tailed VS. *two-tailed tests*

How probability α defining a Type I error is allotted over H_0

P-value *blunders for which I'll never forgive you*

and which will *haunt* you for the *rest* of your *natural life*

- X A p-value ***is not*** the probability H_0 is False
- ✓ H_0 is True, or it is not – there is no "sometimes/probability"

P-value *blunders for which I'll never forgive you*

and which will *haunt* you for the *rest* of your *natural life*

- X A p-value *is not* the probability H_0 is False
- ✓ H_0 is True, or it is not – there is no "sometimes/probability"

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

P-value *blunders for which I'll never forgive you*
and which will *haunt* you for the *rest* of your *natural life*

- X A p-value *is not* the probability H_0 is False
- ✓ H_0 is True, or it is not – there is no "sometimes/probability"

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

- X A p-value *is not anything else except*
  Pr(**seeing something as or more extreme than what you saw** $|H_0$ is true)

P-value *blunders for which I'll never forgive you*

and which will *haunt* you for the *rest* of your *natural life*

- X A p-value *is not* the probability H_0 is False
- ✓ H_0 is True, or it is not – there is no "sometimes/probability"

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

- X A p-value *is not anything else except*
  Pr(**seeing something as or more extreme than what you saw** $|H_0$ is true)

- ✓ A p-value is, *at all times, ever only and EXACTLY ONLY*
  Pr(**seeing something as or more extreme than what you saw** $|H_0$ is true)

The *Pivot* (i.e., 95% Confidence Intervals)

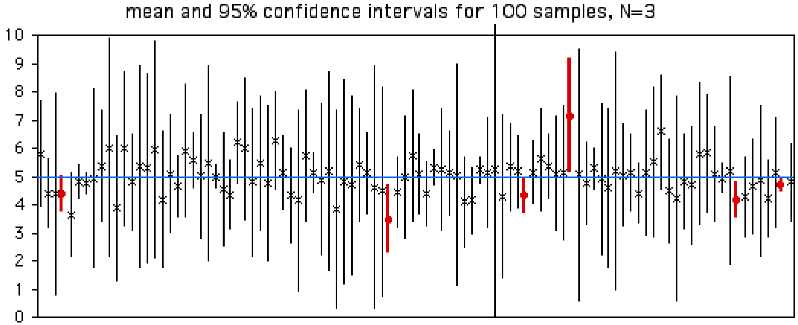
If H_0 is true, then

$$\Pr\left(-t_{n-1}^{\alpha/2} < \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} < t_{n-1}^{\alpha/2}\right) = \Pr\left(-\bar{x} - t_{n-1}^{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} < -\mu_0 < -\bar{x} + t_{n-1}^{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

$$= \Pr\left(\bar{x} + t_{n-1}^{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} > \mu_0 > \bar{x} - t_{n-1}^{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

$$= \Pr\left(\bar{x} - t_{n-1}^{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} < \mu_0 < \bar{x} + t_{n-1}^{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

"captures" μ_0 in 95% of hypothetically repeated experiments



Confidence Intervals and p-values *equivalence*

A $100(1-\alpha)\%$ confidence interval *does not contain* μ_0

Confidence Intervals and p-values *equivalence*

A $100(1-\alpha)\%$ confidence interval *does not contain* μ_0

$$\Longleftrightarrow$$

A two-sided test *rejects* H_0 at the α-significance level

Confidence Intervals and p-values *equivalence*

A $100(1-\alpha)\%$ confidence interval *does not contain* μ_0

$$\Longleftrightarrow$$

A two-sided test *rejects* H_0 at the α-significance level

A $100(1-\alpha)\%$ confidence interval *contains* μ_0

Confidence Intervals and p-values *equivalence*

A $100(1-\alpha)\%$ confidence interval *does not contain* μ_0

$$\Longleftrightarrow$$

A two-sided test *rejects* H_0 at the α-significance level

A $100(1-\alpha)\%$ confidence interval *contains* μ_0

$$\Longleftrightarrow$$

A two-sided test *fails to reject* H_0 at the α-significance level

Confidence Intervals and p-values *equivalence*

If H_0 is true, then

$$\alpha = \Pr_{\bar{x}}\left(\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}\right)$$

The observed p-value under H_0 is

$$p = \Pr_Z\left(Z > \left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right|\right)$$

- If $p < \alpha$ then $\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}$

  $\implies \mu_0 < \bar{x} - Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$ or $\mu_0 > \bar{x} + Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$

Confidence Intervals and p-values *equivalence*

If H_0 is true, then

$$\alpha = \Pr_{\bar{x}}\left(\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}\right)$$

The observed p-value under H_0 is

$$p = \Pr_Z\left(Z > \left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right|\right)$$

- If $p < \alpha$ then $\left|\frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > Z_{\alpha/2}$

  $\Longrightarrow \mu_0 < \bar{x} - Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$ or $\mu_0 > \bar{x} + Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$

  $\Longrightarrow \mu_0 \notin \left(\bar{x} - Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$

i.e., the $100(1-\alpha)\%$ confidence interval *does not* contain μ_0

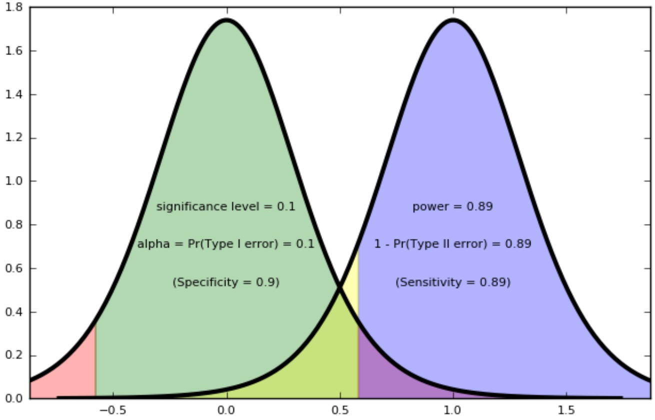
Hypothesis Testing Concepts III

- Alternative hypothesis (H_A)
  An "actual" hypothesized truth about a distribution on data, distinct from H_0's, often used to characterize a tests power
  E.g., assuming group A's mean is twice that of group B

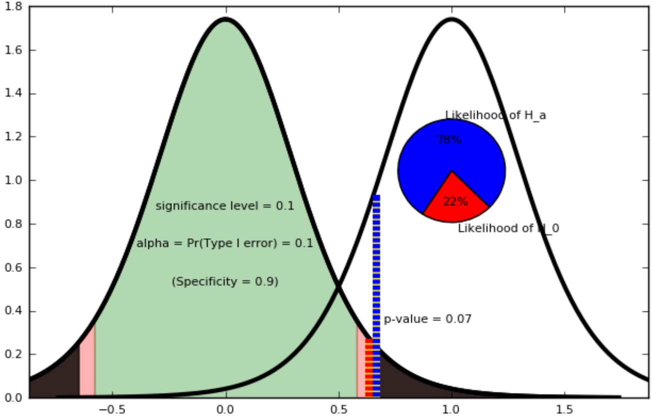
- Test *power* β

$$\begin{aligned}\beta =&\Pr(\text{Fail to rejecting } H_0 | H_A \text{ is true})\\ =&\Pr(\text{Type II error})\end{aligned}$$

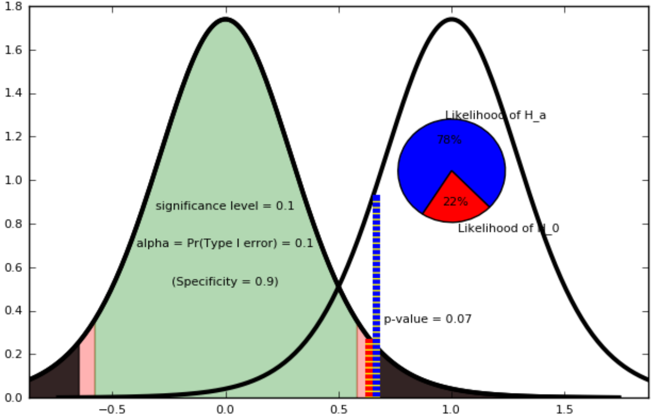
Hypothesis Testing Concepts III



Hypothesis Testing Concepts III



Hypothesis Testing Concepts III



P-values are not related to relative likelihoods. Compared to all alternatives, p-values ~ 0.05 are very strong evidence for H_0.

If some extreme H_A is true then you'd never see anything like 0.05, but if H_0 is true then you would...

Practice

- Yankees fan proportion versus Mets fan proportion

Practice

- Yankees fan proportion versus Mets fan proportion

- Cat weights versus Dog weights

More Practice

		
Trial 1		
Trial 2		
Trial 3		
Trial 4		
$\vdots$		
Trial $n_A + n_B$		
Total		

More Practice

		
Trial 1	$\emptyset$	
Trial 2		
Trial 3		
Trial 4		
$\vdots$		
Trial $n_A + n_B$		
Total		

More Practice

		
Trial 1	$\emptyset$	
Trial 2		$\emptyset$
Trial 3		
Trial 4		
$\vdots$		
Trial $n_A + n_B$		
Total		

More Practice

		
Trial 1	$\emptyset$	
Trial 2		$\emptyset$
Trial 3		$\checkmark$
Trial 4		
$\vdots$		
Trial $n_A + n_B$		
Total		

More Practice

		
Trial 1	$\emptyset$	
Trial 2		$\emptyset$
Trial 3		$\checkmark$
Trial 4	$\emptyset$	
$\vdots$		
Trial $n_A + n_B$		
Total		

More Practice

		
Trial 1	$\emptyset$	
Trial 2		$\emptyset$
Trial 3		$\checkmark$
Trial 4	$\emptyset$	
$\vdots$		
Trial $n_A + n_B$		
Total	$\hat{p}_A = \frac{\sum X_A^{(i)}}{n_A}$	$\hat{p}_B = \frac{\sum X_B^{(j)}}{n_B}$

More Practice: A/B testing

$$\hat{p}_A = \frac{\sum X_A^{(i)}}{n_A} \qquad \hat{p}_B = \frac{\sum X_B^{(j)}}{n_B}$$

$$X_A^{(i)} \sim \text{Bern}\left(\theta_A\right) = \text{Binomial}\left(\theta_A, N_A = 1\right)$$
$$X_B^{(j)} \sim \text{Bern}\left(\theta_B\right) = \text{Binomial}\left(\theta_B, N_B = 1\right)$$

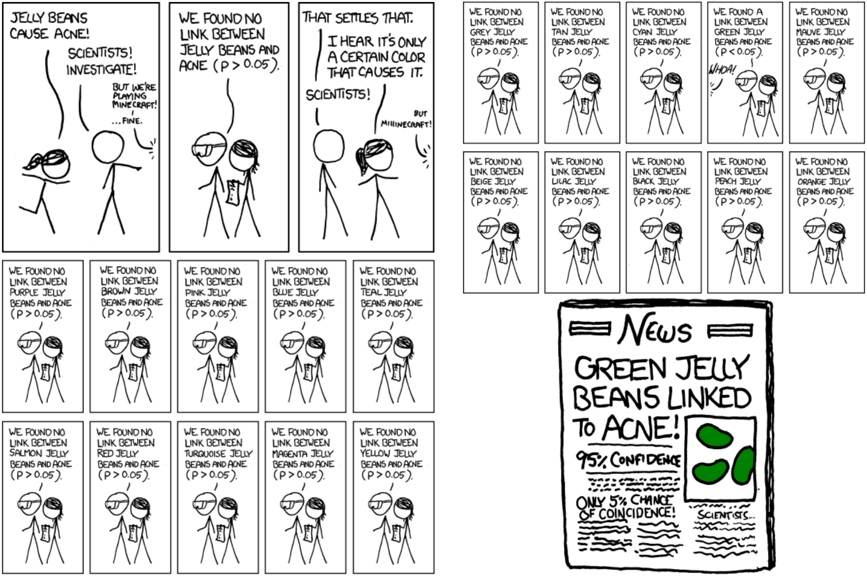
IF $\theta_A = \theta_B$ $[H_0]$

THEN $Var\left(X_A^{(i)}\right) = Var\left(X_B^{(j)}\right) = \ ?$

SO $\hat{p}_A - \hat{p}_B \sim \ ?$ [By CLT]

AND what is a good estimator of $\theta = \theta_A = \theta_B$?

Multiple Testing



Multiple Testing

- Each time we do a hypothesis test [what?]

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong

  So if we do N tests, and H_0 is true for all of them

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong

  So if we do N tests, and H_0 is true for all of them

  we still expect to wrongly reject H_0 about $\alpha \times N$ times!

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong

  So if we do N tests, and H_0 is true for all of them

  we still expect to wrongly reject H_0 about $\alpha \times N$ times!
- Testing at $\alpha' = \alpha / N$ gives an α chance all tests are right

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong

  So if we do N tests, and H_0 is true for all of them

  we still expect to wrongly reject H_0 about $\alpha \times N$ times!
- Testing at $\alpha' = \alpha / N$ gives an α chance all tests are right
- This is called *Bonferroni correction*

Multiple Testing

- Each time we do a hypothesis test [what?]

  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong

  So if we do N tests, and H_0 is true for all of them

  we still expect to wrongly reject H_0 about $\alpha \times N$ times!
- Testing at $\alpha' = \alpha/N$ gives an α chance all tests are right
- This is called *Bonferroni correction*

  and it guarantees a α *familly-wise error rate*

Multiple Testing

- Each time we do a hypothesis test [what?]
  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong
  So if we do N tests, and H_0 is true for all of them
  we still expect to wrongly reject H_0 about $\alpha \times N$ times!
- Testing at $\alpha' = \alpha / N$ gives an α chance all tests are right
- This is called *Bonferroni correction*
  and it guarantees a α *familly-wise error rate*

- Bonferroni correction is really quite stringent...

Multiple Testing

- Each time we do a hypothesis test [what?]
  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong
  So if we do N tests, and H_0 is true for all of them
  we still expect to wrongly reject H_0 about $\alpha \times N$ times!
- Testing at $\alpha' = \alpha / N$ gives an α chance all tests are right
- This is called *Bonferroni correction*
  and it guarantees a α *familly-wise error rate*

- Bonferroni correction is really quite stringent...

- An alternative is the *False Discovery Rate (FDR) q*

Multiple Testing

- Each time we do a hypothesis test [what?]
  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong
  So if we do N tests, and H_0 is true for all of them
  we still expect to wrongly reject H_0 about $\alpha \times N$ times!
- Testing at $\alpha' = \alpha/N$ gives an α chance all tests are right
- This is called *Bonferroni correction*
  and it guarantees a α *familly-wise error rate*

- Bonferroni correction is really quite stringent...

- An alternative is the *False Discovery Rate (FDR) q*
  which for a set of tests (e.g., tests significant at the α-level)

Multiple Testing

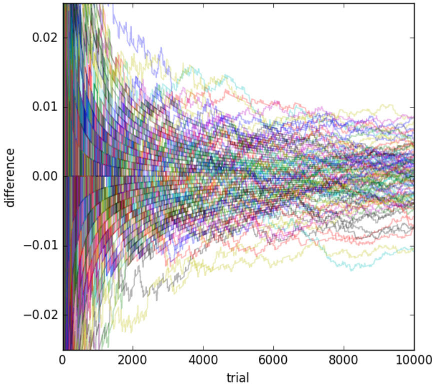
- Each time we do a hypothesis test [what?]
  There's a chance we are wrong about our decision
- If H_0 is true, an α chance of being wrong
  So if we do N tests, and H_0 is true for all of them
  we still expect to wrongly reject H_0 about $\alpha \times N$ times!
- Testing at $\alpha' = \alpha / N$ gives an α chance all tests are right
- This is called *Bonferroni correction*
  and it guarantees a α *familly-wise error rate*

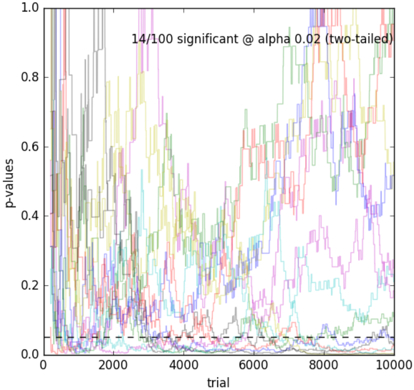
- Bonferroni correction is really quite stringent...

- An alternative is the *False Discovery Rate (FDR) q*
  which for a set of tests (e.g., tests significant at the α-level)
  is the proportion q of the tests called incorrectly ("FDR")

Multiple A/B Testing

- There is no difference in conversion rates in these simulations





- Continuous (multiple) testing does not achieve α-significance