CSCI 443: LECTURE 15 HYPOTHESIS TESTING: DEEPER AND DANGERS

Professor David Harrison

OFFICE HOURS

Tuesday	4:00-5:00 PM
Wednesday	12:30-2:30 PM

HOMEWORK 4

Due Tonight at 11:00pm.

Handout Homework 5 next week.

HOMEWORK 4 PROBLEM 2.4

Problem 2.4 Skewness of a distribution is defined as

$$\cancel{E[(X-\mu)^3] = \int_{-\infty}^{\infty} (x-\mu)^3 f(x)\, dx}$$

Revised

$$\frac{E[(X-\mu)^3]}{\sigma^3} = \int_{-\infty}^{\infty} \frac{(x-\mu)^3}{\sigma^3} f(x)\, dx$$

HOMEWORK 4 PROBLEM 4

Sometimes misleading result.

Depends on random seed.

Binomial distribution shows skew when p is near 0 or near 1. Problem 4 aimed to show this, but only required 1000 binomial samples (binomial trials). For p=0.2 and p=0.8, the effect is more consistent with 10000 binomial trials or more.

HOMEWORK 4
PROBLEM 4.4, 4.5, 4.6

Problem 4.4 Using your function implemented for Problem 2, compute the sample skewness of the samples in Problem ~~3.1~~ 4.1.

Problem 4.5 In the same way, compute the sample skewness of the samples in Problem ~~3.2~~ 4.2 .

Problem 4.6 In the same way, compute the sample skewness of the samples in Problem ~~3.3~~ 4.3.

HOMEWORK 4 PROBLEM 4.7

Problem 4.7 (Original wording) For a binomial distribution with $n = 5$ and $p = 0.2$, simulate drawing 1000 sample sets each of size 5. Plot the sampling distribution of the sample proportion (i.e., the percentage of outcomes with successes). On the same plot place the PDF of a Gaussian random variable $N(p, \sigma/\sqrt{n})$. What does the Gaussian PDF represent? Is the sampling distribution skewed or symmetric? How does it compare to the original distribution?

(Revised wording) *For a binomial distribution with* $n = 5$ *and* $p = 0.2$, *simulate drawing 1000 samples of* $X \sim Bin(n, p)$ *and computing the sample proportion . The distribution of the sample proportion is the sampling distribution of* p. *Plot this sampling distribution. On the same plot place the PDF of a Gaussian random variable* $N(p, \sqrt{p(1-p)/n}\,)$. *How does the Gaussian PDF relate to computing confidence intervals? Is the sampling distribution skewed or symmetric? How does the Gaussian distribution compare to the distribution of the sample proportion?*

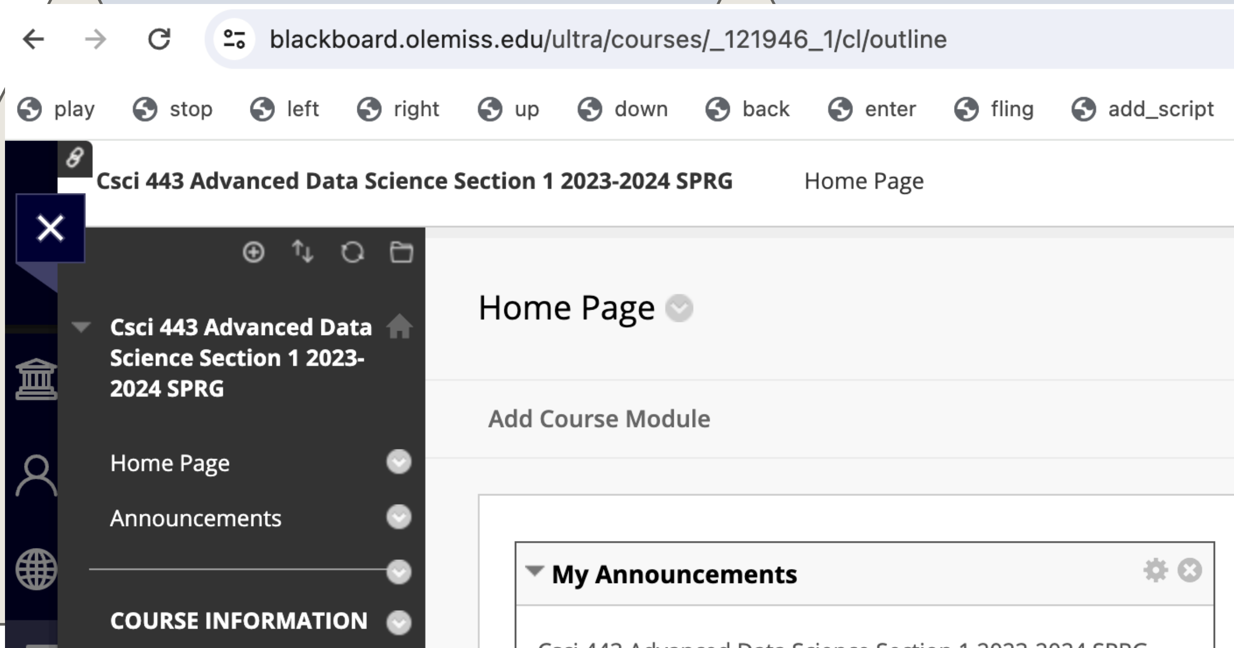
BLACKBOARD & GITHUB

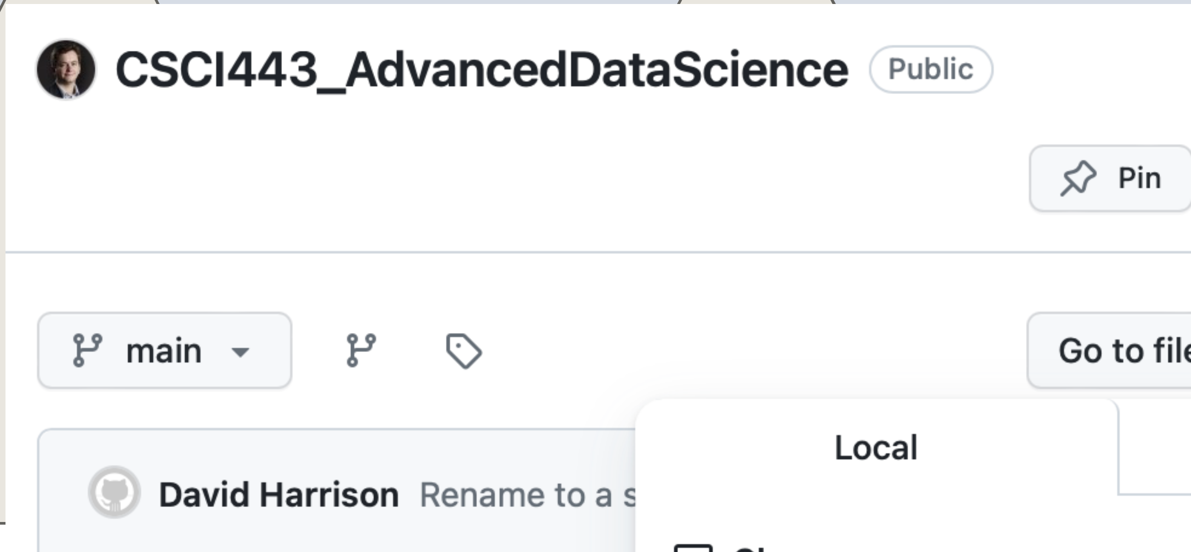
Slides up through lecture 13 on blackboard.

Lecture slides and examples committed to GitHub also up through lecture 13.

The project is at

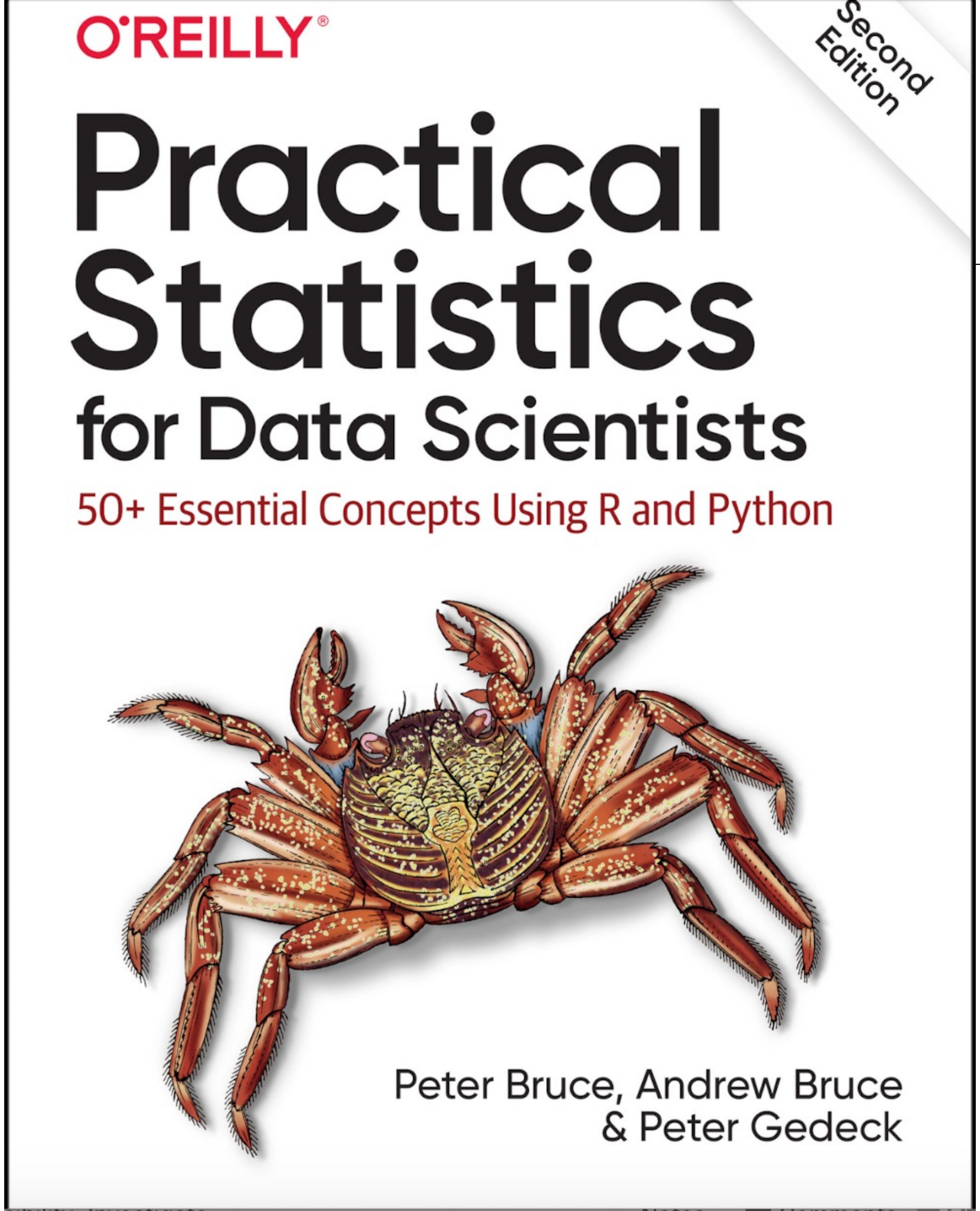
https://github.com/dosirrah/CSCI443_AdvancedDataScience





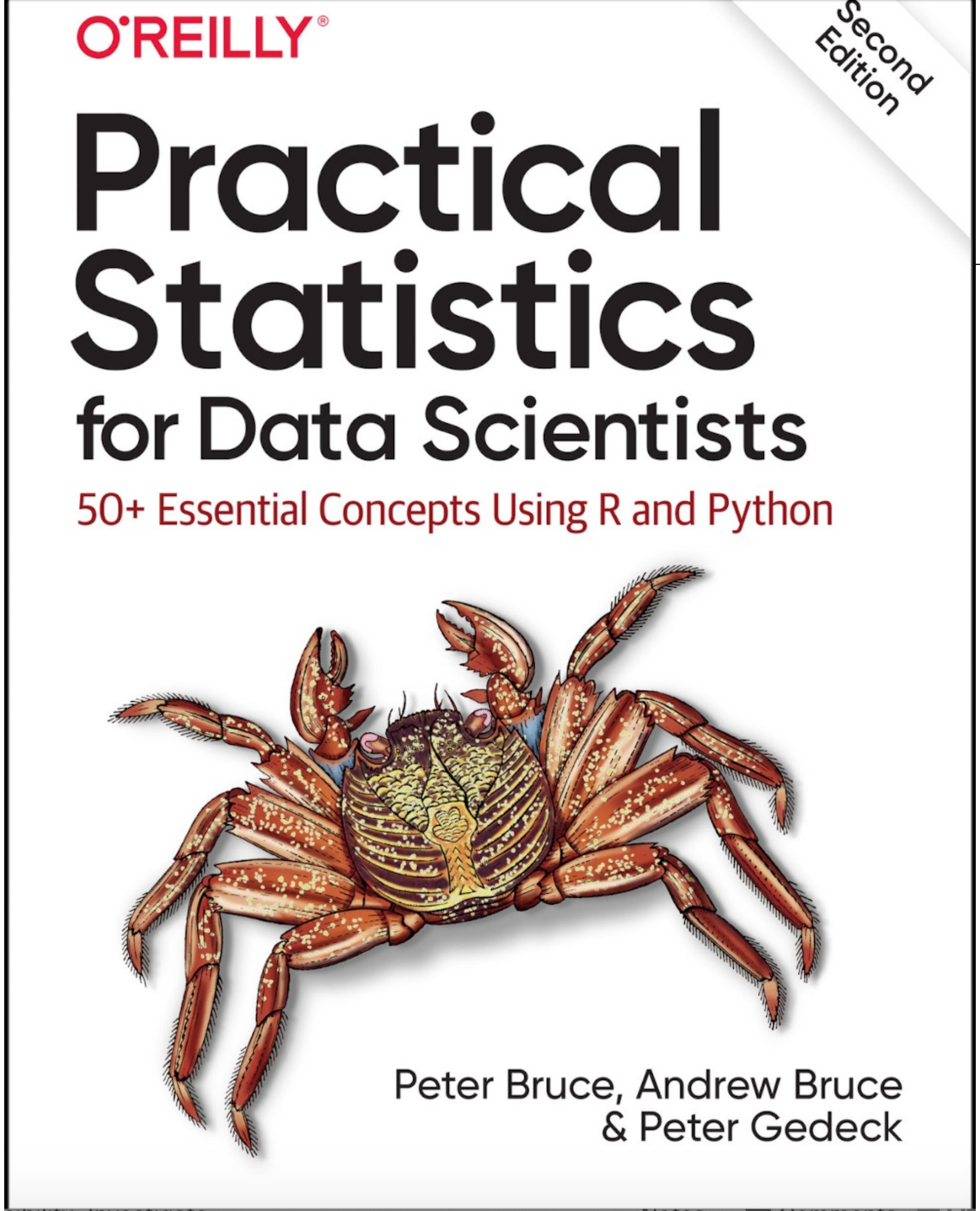
READ ABOUT

- chapter 3: experiments, hypothesis testing
  - [...]
  - Statistical Significance
  - P-values



THINGS I WANT TO COVER TODAY

- Little review
- Fix b0tched analysis
- Dig a little into the Ad Comparison example.
- Failures of hypothesis testing.
- Holdouts
- Cross-validation
- Alternate mechanisms for avoiding Type I (false positive) Errors.



FROM PREVIOUS LECTURE:
STATISTICAL INFERENCE PIPELINE

•

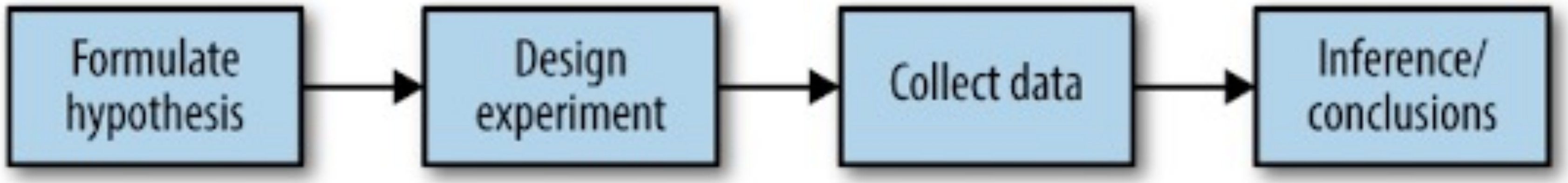


Figure 3-1. The classical statistical inference pipeline

FROM PREVIOUS LECTURE: A/B TESTING

- Experiment with two groups.
- For example,
  - exposed vs. not exposed.
  - treated vs. not treated
  - Between two headlines, which produces more clicks.
  - Between two prices, which produces more profits.
- Ensure all factors are the same, except for the 1 factor being varied.

FROM PREVIOUS LECTURE: OBSERVATIONAL VS. CONTROLLED STUDIES

Observational studies show **correlation**. May be used to establish hypotheses for controlled studies.

A well-designed, randomized, double-blind, placebo-controlled study is taken to show **causality**.

FROM PREVIOUS LECTURE: EVEN A CONTROLLED STUDY CAN BE WRONG!

Given A/B test with

- A=treated

vs

- B=placebo control.

Randomized, double-blind, placebo-controlled study shows A does better than B.

Two major reasons for this outcome:

1. The treatment is better.
2. Chance.

FROM PREVIOUS LECTURE: STATISTICAL HYPOTHESIS TESTING

"A statistical hypothesis test is [..] analysis of an A/B test [to] assess whether random chance is a reasonable explanation for the observed difference between groups A and B."

--Bruce, Peter, et al. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*

FROM PREVIOUS LECTURE: NULL HYPOTHESIS TESTING

Null hypothesis

The hypothesis that chance is to blame.

H_0

Alternative hypothesis

Counterpoint to the null (what you hope to prove).

H_1

Null hypothesis = “proposes there is no significant difference or effect”

PREVIOUS LECTURE: AD COMPARISON

Testing two ads A and B. Which generates more click-throughs?

Using click-through as a proxy variable for revenue.

Two-sided tests:

H_0: there is no significant difference in click-through rates.

H_1: there is a significant difference in click-through rates.

Choose α=5%.

A:

B:

BERNOULLI RANDOM VARIABLES

A Bernoulli random variable takes a value of 1 with probability p and 0 with probability (1-p).

Ex: flipping a coin.

H=1 has p=1/2, T=0 has p=1-1/2=1/2

Ex: clicking through an ad.

1 = clicked has probability p.

0 = didn’t click has probability 1-p.

$$P(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

BINOMIAL RANDOM VARIABLES.

A binomial is the sum of series of Bernoulli trials.

Ex coin flips: H, T, H, H, H, T, …

1 + 0 + 1 + 1 + 1 + 0+…

Ex click throughs: clicked, didn't, clicked

1 + 0 + 1 + …

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$P(X = k)$ is the probability of getting exactly k successes out of n trials,

$$\text{Mean of } X = \mu = np$$

$$\sigma = \sqrt{n \cdot p \cdot (1-p)}$$

OF CLICK THROUGHS IS BINOMIAL.

The number of click-throughs is a binomial $X_A \sim Bin(n_A, p_A)$

For all binomial random variables

$$\sigma = \sqrt{n \cdot p \cdot (1-p)} \quad \mathrm{Var}(X) = np(1-p)$$

The click through rate is an estimate of the sample proportion. $\hat{p} = X/n$

PREVIOUS LECTURE: AD COMPARISON

A:

B:

1000 views of each.

0.1% click through rate for A.

0.5% click through rate for B.

MODIFIED FROM PREVIOUS LECTURE: AD COMPARISON

A:

B:

$\hat{p}_A$ = click through rate for A
= Bin(n_A, p_A) / n_A

$\hat{p}_B$ = click through rate for B
Bin(n_B, p_B) / n_B

PREVIOUS LECTURE: AD COMPARISON

See lecture notes about sample proportions and how sample proportion is the sampling distribution for p.

A:

B:

VARIANCE OF THE SAMPLE PROPORTION

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{\mathrm{Var}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

See lecture notes for details.

PREVIOUS LECTURE: AD COMPARISON

We want the probability of an outcome for the difference between the two click through rates.

D = difference in clickthrough rates.

$$Z = \frac{D - \mu_D}{SE}$$

According to our null hypothesis, the clickthrough rates are not significantly different.

$$\mu_D = 0 \qquad Z = \frac{D - 0}{SE} = \frac{D}{SE}$$

A:

B:

FROM PREVIOUS LECTURE: AD COMPARISON

$$Z = \frac{D - 0}{SE} = \frac{D}{SE}$$

$$D = \hat{p}_A - \hat{p}_B$$

$$Z = \frac{\hat{p}_A - \hat{p}_B}{SE}$$

A:

B:

HOW DO WE GET THE STANDARD ERROR OF D?

Assuming people seeing A and people seeing B don't coordinate, we can assume the click throughs for A and B are independent.

$$\mathrm{Var}(\hat{p}_A - \hat{p}_B) = \mathrm{Var}(\hat{p}_A) + \mathrm{Var}(\hat{p}_B)$$

See lecture notes for proof.

FIND STANDARD ERROR OF D

Assuming people seeing A and people seeing B don't coordinate, we can assume the click throughs for A and B are independent.

$$\mathrm{Var}(\hat{p}_A - \hat{p}_B) = \mathrm{Var}(\hat{p}_A) + \mathrm{Var}(\hat{p}_B)$$

$$SE_D = \sqrt{\frac{\hat{p}_A \cdot (1 - \hat{p}_A)}{n_A} + \frac{\hat{p}_B \cdot (1 - \hat{p}_B)}{n_B}}$$

See lecture notes for proof.

NOW COMPUTE THE Z VALUE

Z-value for the hypothesis test:

$$Z = \frac{\hat{p}_A - \hat{p}_B}{SE_D}$$

$$SE_D = \sqrt{\frac{\hat{p}_A \cdot (1 - \hat{p}_A)}{n_A} + \frac{\hat{p}_B \cdot (1 - \hat{p}_B)}{n_B}}$$

$$SE_D = \sqrt{\frac{0.001 \cdot (1 - 0.001)}{1000} + \frac{0.005 \cdot (1 - 0.005)}{1000}}$$

$$SE_D = 0.00244\ldots$$

NOW COMPUTE THE Z VALUE

Z-value for the hypothesis test:

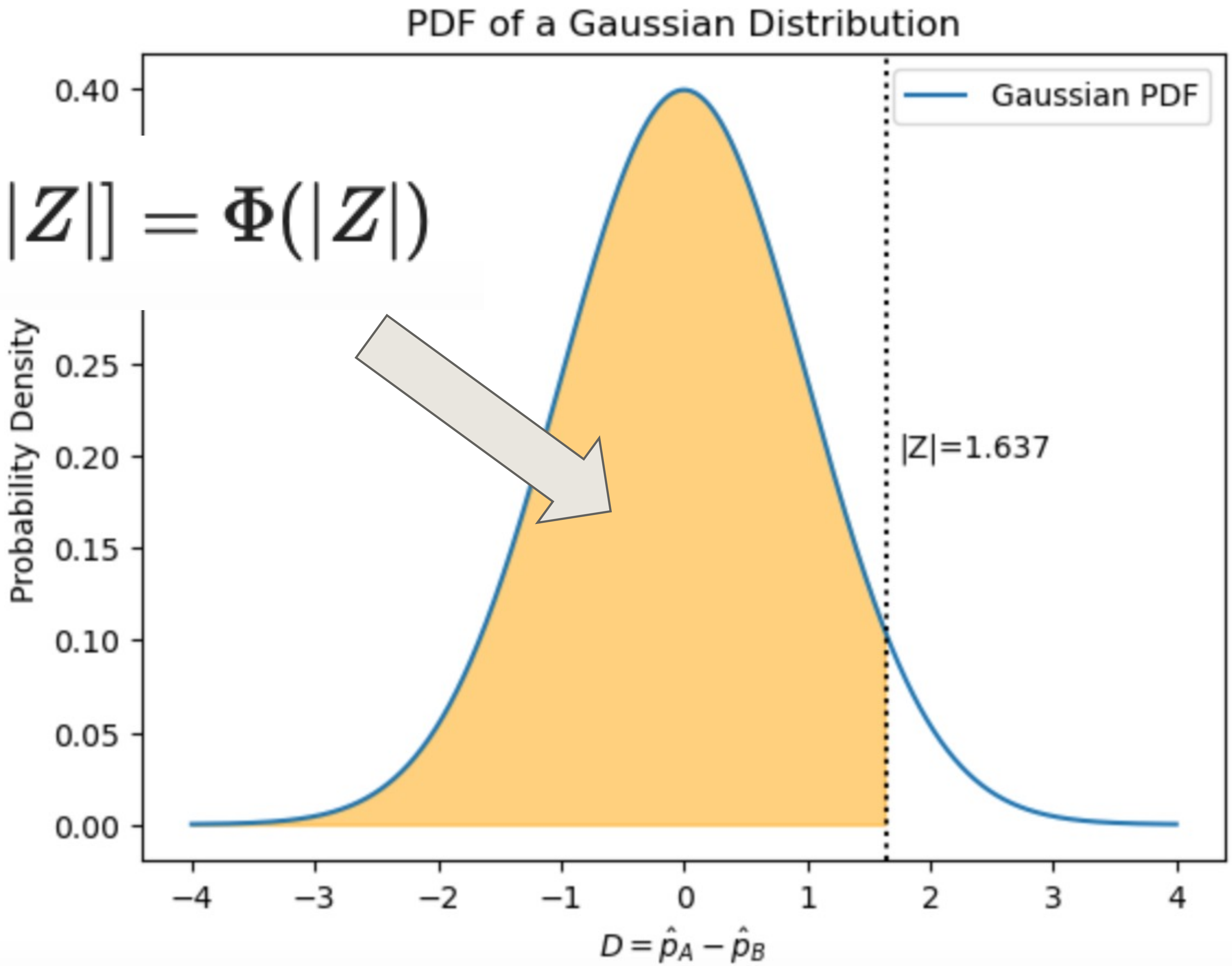
$$Z = \frac{\hat{p}_A - \hat{p}_B}{SE} \qquad Z = \frac{0.001 - 0.005}{0.00244}$$

$$Z \approx -1.637$$

To rule out the null hypothesis, we find the probability of encountering a value equal or more extreme than the Z value for our sample.

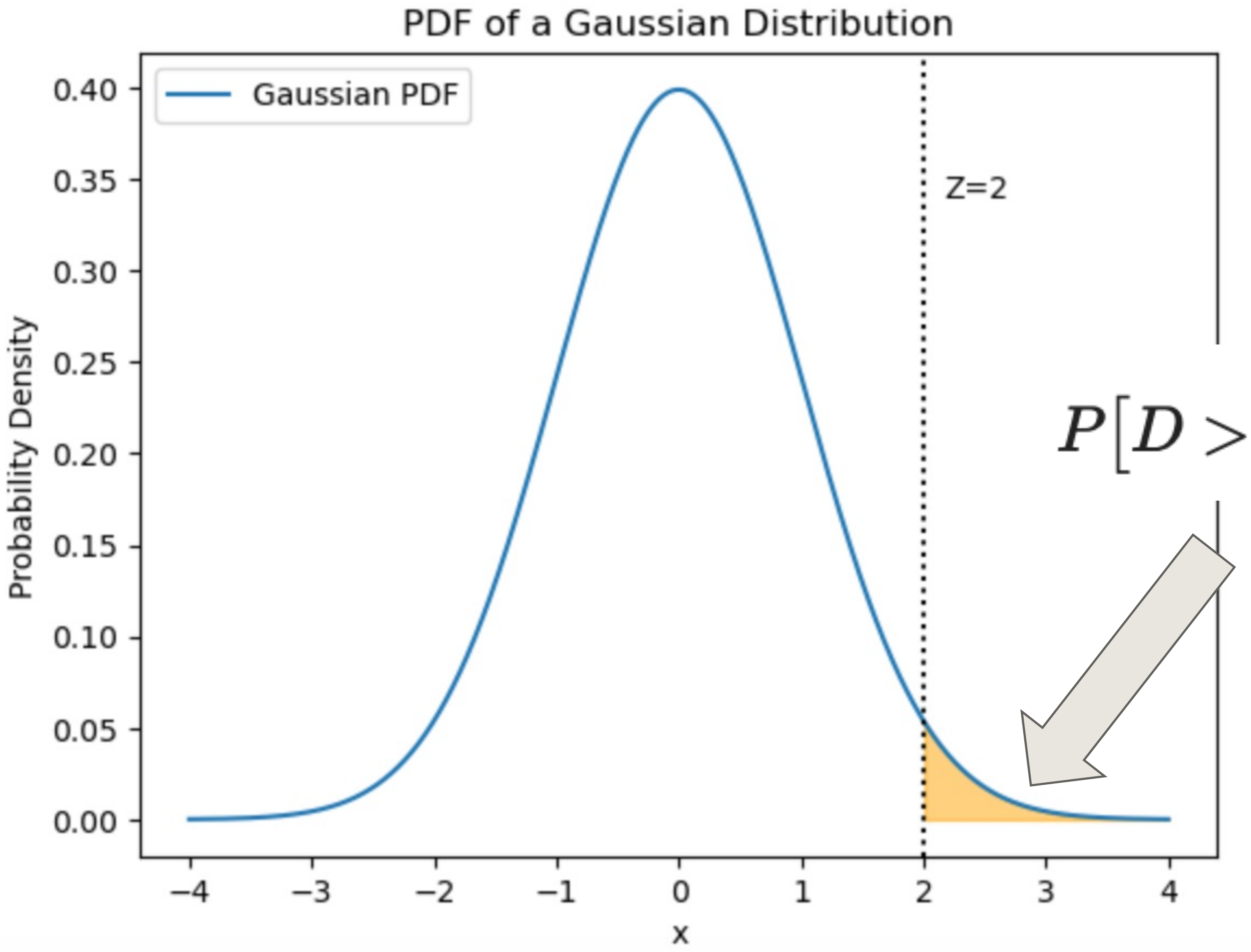
We are using a 2-sided test, meaning we care if the two values are different in either direction. P[D<-|Z| or D >|Z|]

P[D < |Z|]



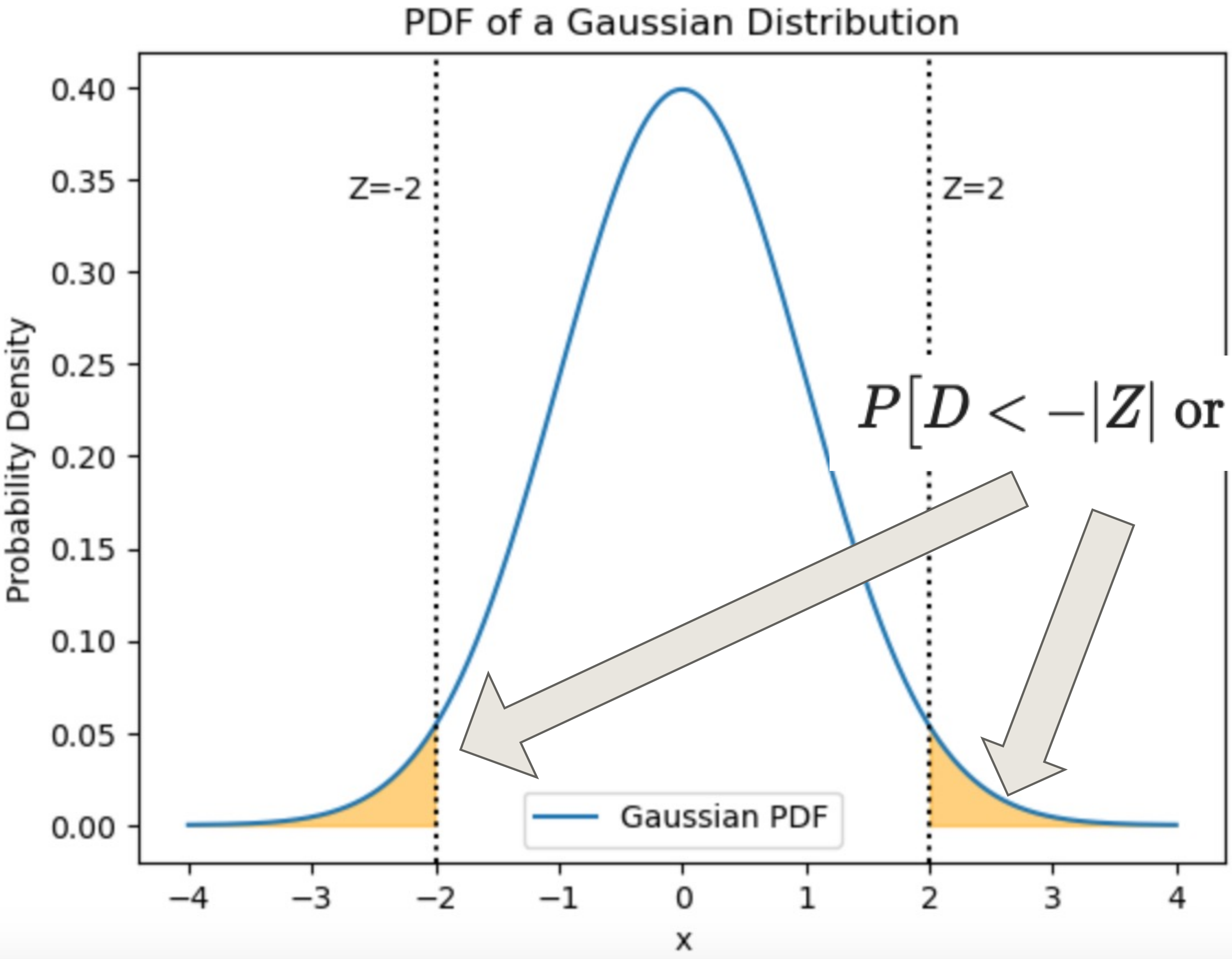
$$P[D \leq |Z|] = \Phi(|Z|)$$

P[D>|Z|]



$$P\left[D > |Z|\right] = 1 - \Phi\left(|Z|\right)$$

NOW COMPUTE THE Z VALUE



$$P\left[D < -|Z| \text{ or } D > |Z|\right] = 2\left(1 - \Phi\left(|Z|\right)\right)$$

NOW COMPUTE THE Z VALUE

Z-value for the hypothesis test:

$$Z \approx -1.637$$

$$\text{p-value} = P\left[D < -|Z| \text{ or } D > |Z|\right] = 2\left(1 - \Phi\left(|Z|\right)\right)$$

$$\text{p-value} = 2\left(1 - \Phi\left(1.637\right)\right)$$

$$\text{p-value} \approx 0.102$$

$$\alpha = 5\% = 0.05$$

$$\boxed{0.102 >> 0.05}$$

NOT SIGNIFICANTLY DIFFERENT

DANGERS OF NULL HYPOTHESIS TESTING

The job of a data scientist performing an observational study is NOT to find the truth.

The job of a data scientist performing an observational study is to find indicia (hallmarks, anomalies,…)

EXAMPLE: REBELTON MAYORAL ELECTION

Todd Reb (the original "Hot Todd"), ne're-do-well great-grandson of the beloved Colonel Reb has a surprising victory in the race for Mayor of Rebelton. Data scientists descend on Rebelton to find out whether the election outcome was affected by fraud.

Daytona Truthy uncovers an anomaly.

The number of people over 90 who voted went up 3x from the previous 10 elections.

This is a 4σ event.

$$p\text{-value} = P[X < -Z \text{ or } X > Z] = 2(1 - \Phi(4)) \approx 0.0063\%$$

This is lower than any typical α significance level.

Ms. Truthy declares fraud!

EXAMPLE: TRUTHY FINDS MORE!

Daytona Truthy finds six more anomalies that each meet a confidence level of α=5%.

What is the problem?

EXAMPLE: TRUTHY FINDS MORE!

Daytona Truthy's results may be perfectly valid…

"If you torture the data long enough, it will confess"

- Ronald Coasce, British economist, Nobel laureate (at least according to Irving John Good in a 1971 lecture at a meeting of the Institute of Mathematical Statistics)

If you test for a hundred anomalies. α=0.05.

Assuming none of the effects tested for are real. Probability of correct result is 95%. The probability that 100 do NOT generate at least one false positive 0.95^{100}= 0.59%.

Your job as a data scientist is to hand off the positive results to investigators.

THANK YOU

David Harrison

Harrison@cs.olemiss.edu