CLASSIFYING BIOLOGICAL SEX

Drew Fustin

PhD, Physics

Lead Data Scientist

SPOT HERO

Course Instructor

METIS®

drewfustin@gmail.com | 1.10.2017 | Metis | Intro to Data Science Open House

CLASSIFYING BIOLOGICAL SEX

No, not that.

METIS

INTRODUCTION TO DATA SCIENCE

JAN 24 to MAR 2
Tuesdays & Thursdays
6:30pm to 9:30pm
thisismetis.com/introduction-to-data-science

METIS

INTRODUCTION TO DATA SCIENCE

Schedule:

Tuesday	Thursday
1/24: Introduction to Python and Version Control	1/26: Introduction to Linear Algebra and Statistics in Python
1/31: Pandas and Exploratory Data Analysis	2/2: Pandas and Data Visualization
2/7: Supervised Machine Learning and Basic Model Evaluation	2/9: Unsupervised Machine Learning and Basic Model Evaluation
2/14: Advanced Supervised Learning	2/16: Data Modeling: Feature Engineering and Basic Cross Validation
2/21: Data Modeling: Regularization, Feature Decomposition	2/23: Advanced Unsupervised Learning
2/28: Advanced Model Evaluation and Pipelines	3/2: Project Presentations and Course Wrap-up

METIS

INTRODUCTION TO DATA SCIENCE

- Complete exercises 1-7, 13, 18-21, 27-35, 38, 39 of Learn Python The Hard Way (learnpythonthehardway.org/book)
- Watch the linear algebra review videos from Andrew Ng's excellent Coursera ML course. They are labeled III. Linear Algebra Review (Week 1). (class.coursera.org/ml-005/lecture/preview)
- Complete the exercises in chapters 2 and 3 of OpenIntro Statistics. (openintro.org/stat/textbook.php)

THE PROBLEM

Different sexes behave differently.
Knowing sex of user could help in predictions.
We don't have user sex defined in our data.
We *do* have first names.
Can we determine sex given first names only?

THE PROBLEM

Hey look, here comes Beyoncé.
Yup, she's female.
Can I do that automatically?

THE QUESTION

p = probability a person with a given name is male
(1 – p = probability this same person is female)
question: what is my 95% confidence interval on p?

THE QUESTION

p = probability a person with a given name is male
(1 – p = probability this same person is female)
question: what is my 95% confidence interval on p?

There is an exact answer for p,
but I don't know exactly what it is.
Have to sample data to model it.

AN EXAMPLE

Suppose, in reality, $p = 0.75$.
That is, of all our users with a given name Yxelyfyx,
75% of them are male and 25% are female.
This answer is unknown. I'm trying to determine it
by making observations on a sample of our users.

AN EXAMPLE

Before I know *anything* about Yxelyfyx's, what's my guess?

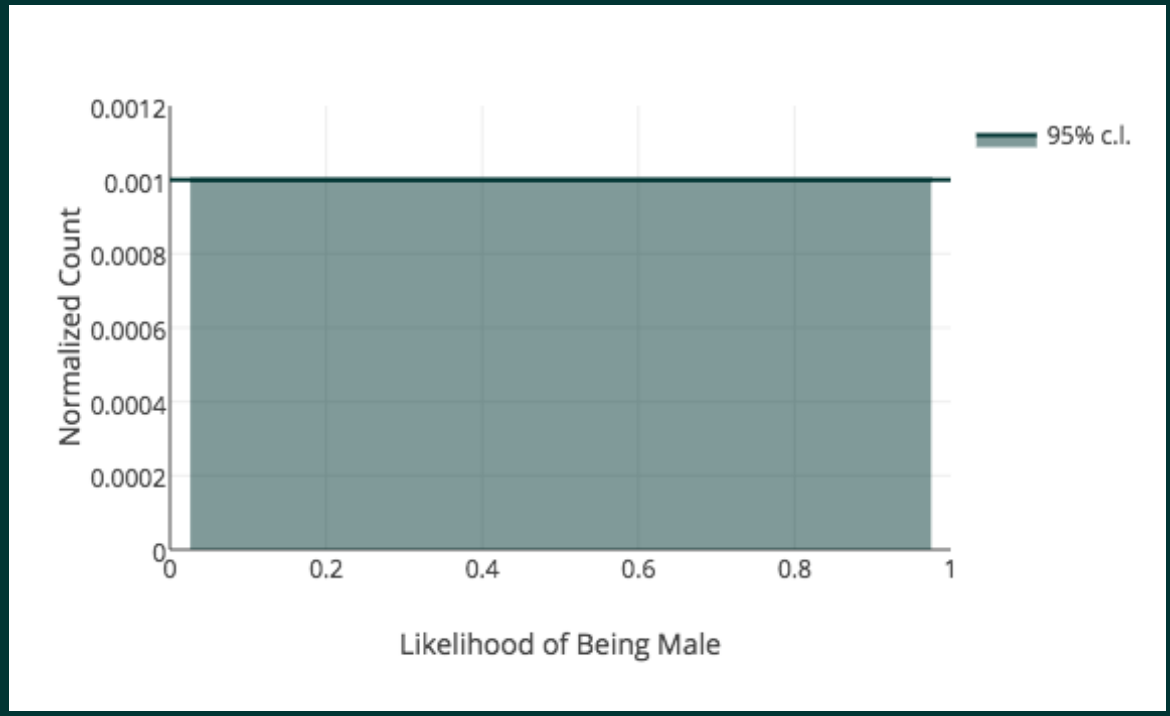
AN EXAMPLE

Before I know *anything* about Yxelyfyx's,
what's my guess?
Sure, p = 0.5 is a good guess for the expectation value,
but what about the 95% confidence interval?

AN EXAMPLE

I've made no observations yet.

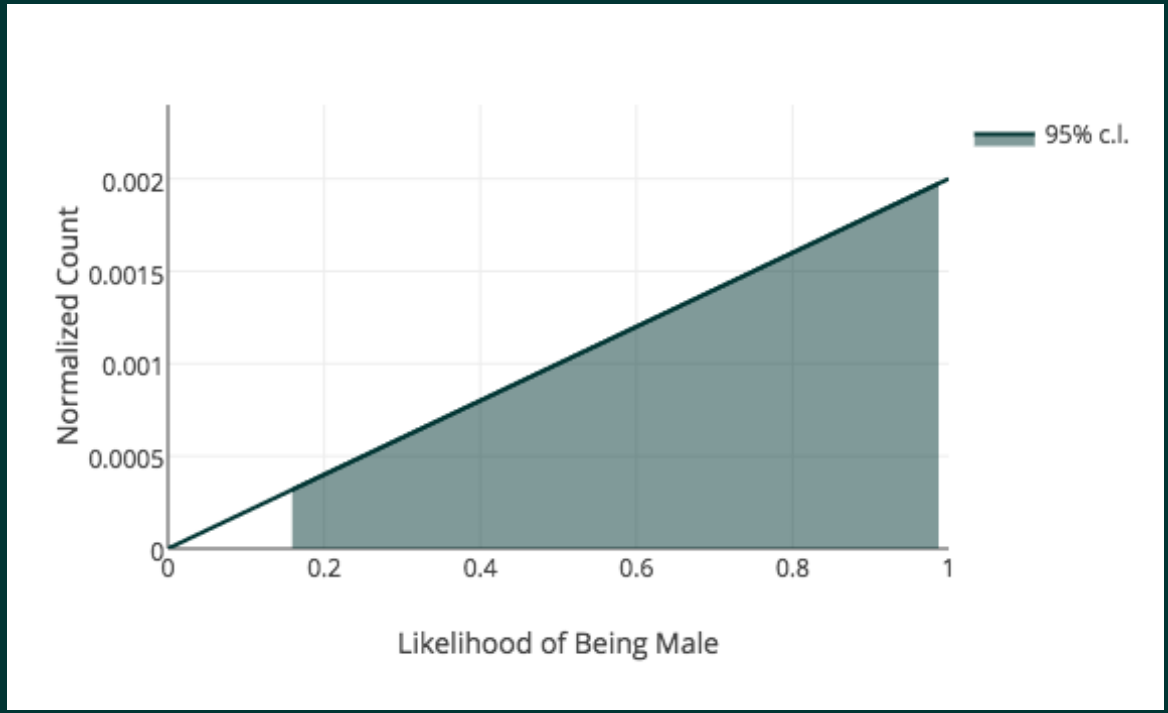
All likelihoods are equally possible.
I'm 95% confident it's between [0.025, 0.975].



AN EXAMPLE

Observe one person. They're male.

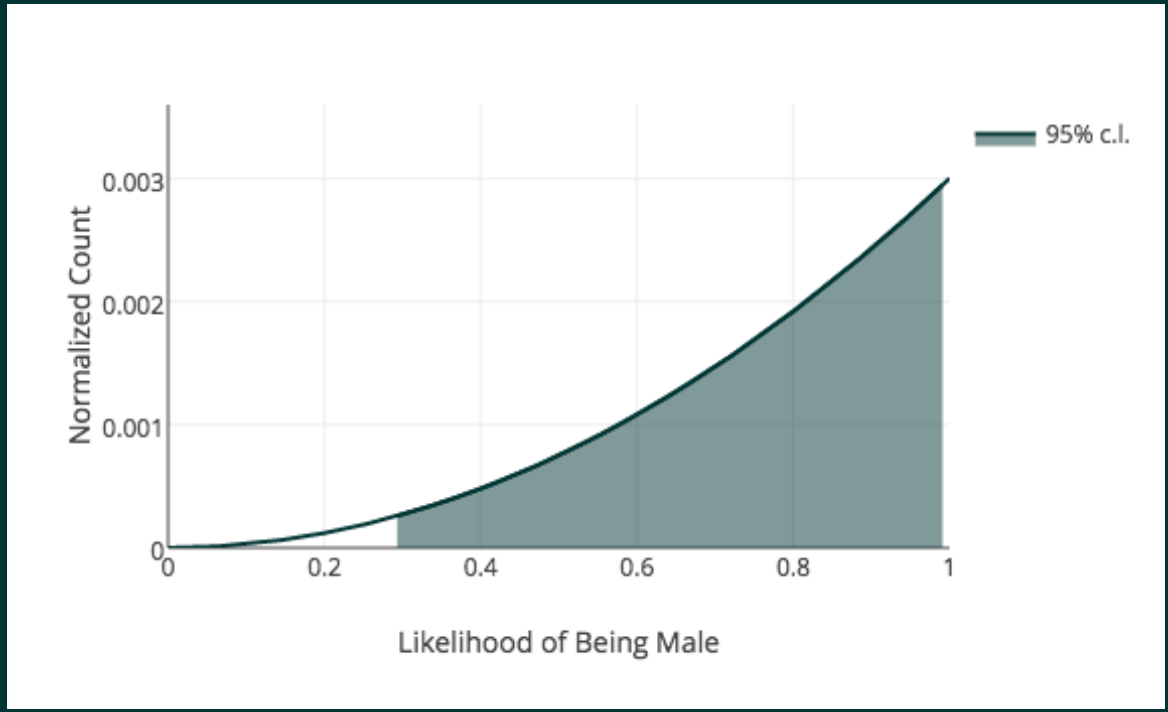
Now, 95% confidence limits are [0.159, 0.987].



AN EXAMPLE

Another man comes along. Now, observed [M, M].

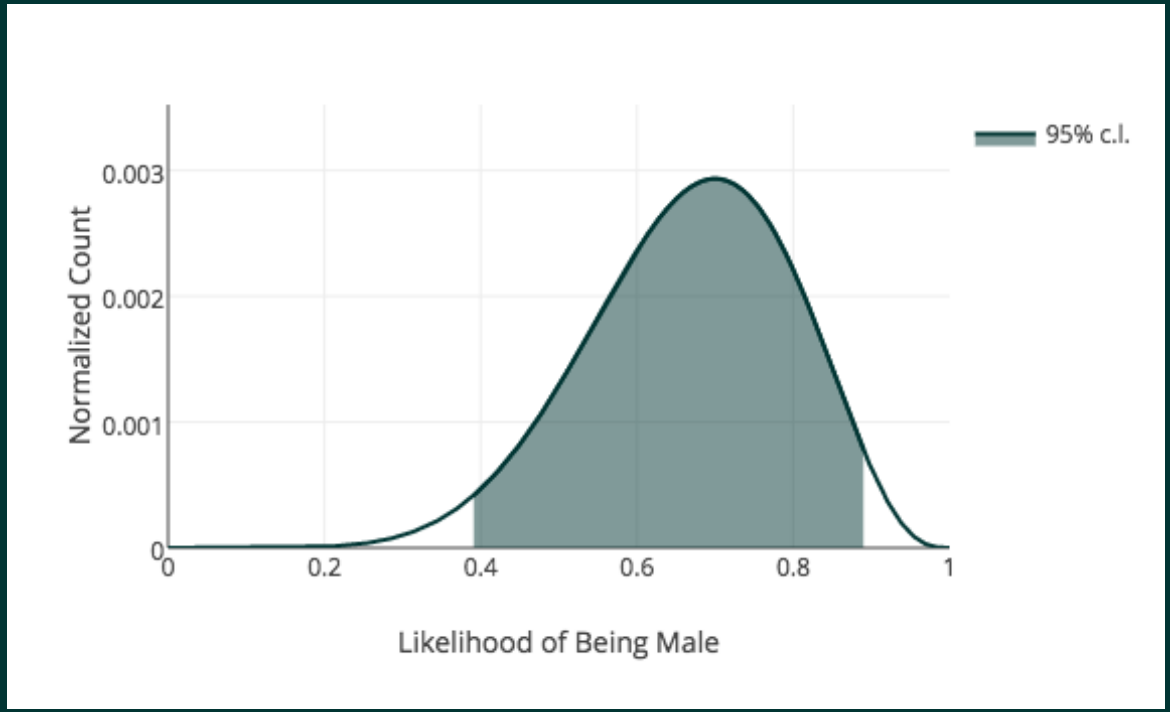
Now, 95% confidence limits are [0.293, 0.991].



AN EXAMPLE

Observe [M, M, M, F, F, M, M, F, M, M].

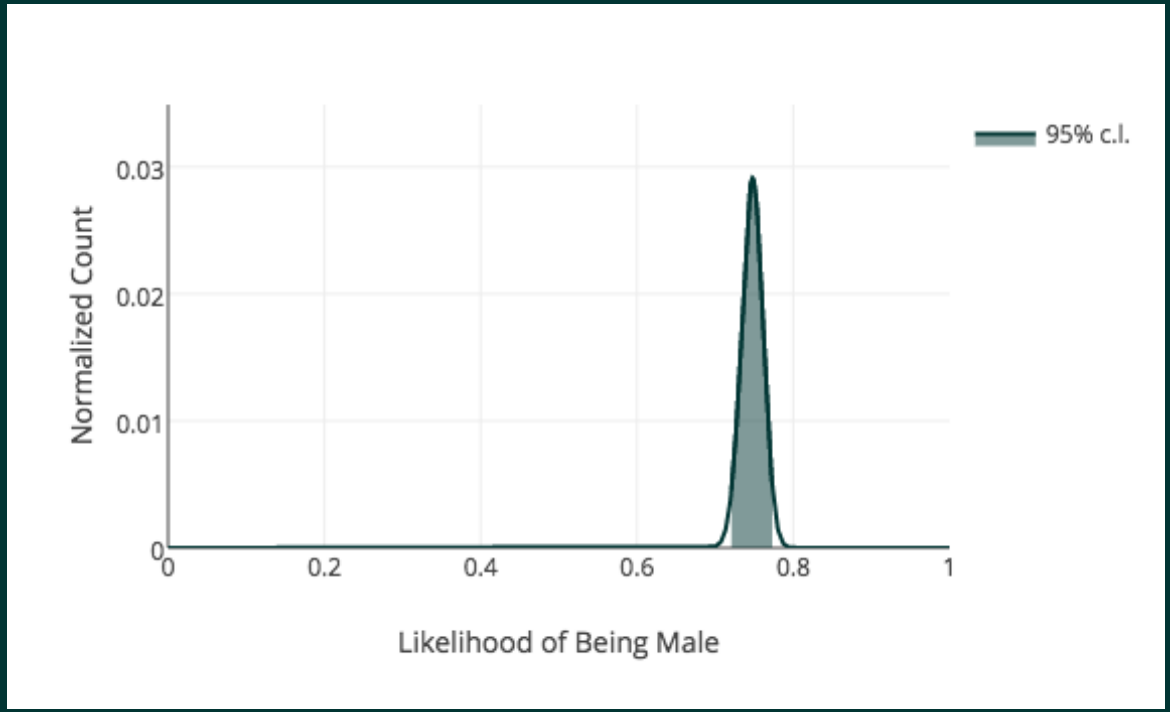
Now, 95% confidence limits are [0.391, 0.980].



AN EXAMPLE

Observe M = 748, F = 252.

Now, 95% confidence limits are [0.721, 0.773].



YOUR TASK

Finance wants you to develop a database of first names and the 95% confidence interval that a user with that first name is male.

YOUR TASK

Finance wants you to develop a database of first names and the 95% confidence interval that a user with that first name is male.

TO THE NOTEBOOK!