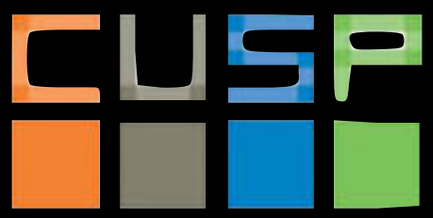
Urban Informatics

Fall 2015

dr. federica bianco fb55@nyu.edu

 @fedhere 

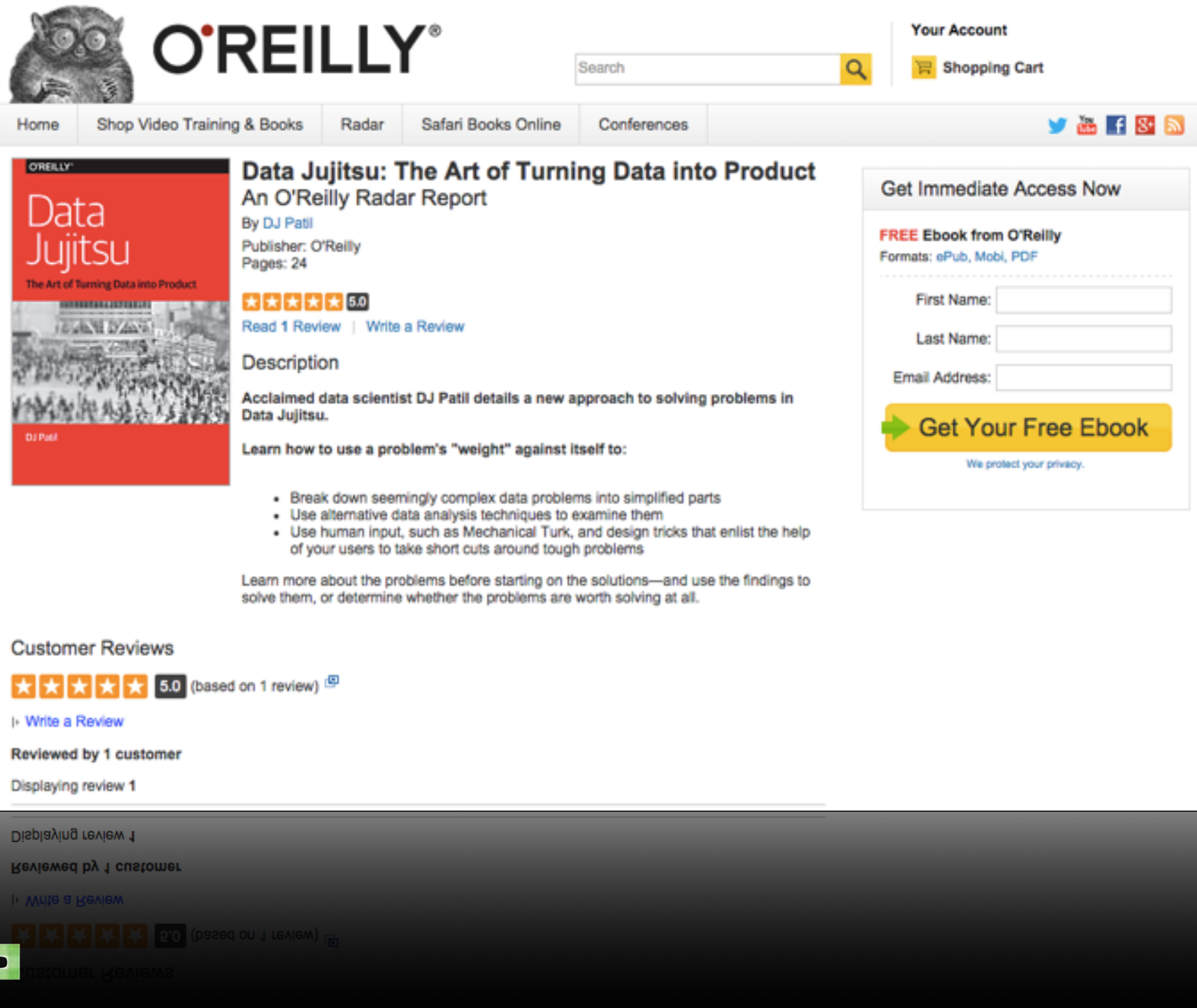


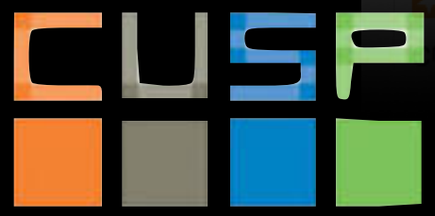
Summary:

- **Epistemological concepts:**
  falsifiability, law of parsimony,
- **Good scientific practice:**
  reproducibility of research

Summary:

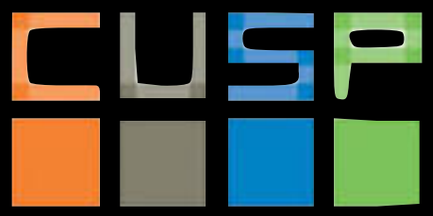
- **Epistemological concepts:**
  falsifiability, law of parsimony,
- **Good scientific practice:**
  reproducibility of research
- **Gathering parsing data, API:**
  data munging or wrangling, data jujitsu


O'REILLY®
Search
Your Account
Shopping Cart
Home
Shop Video Training & Books
Radar
Safari Books Online
Conferences
O'REILLY
Data Jujitsu
The Art of Turning Data into Product
DJ Patil
Data Jujitsu: The Art of Turning Data into Product
An O'Reilly Radar Report
By DJ Patil
Publisher: O'Reilly
Pages: 24
5.0
Read 1 Review | Write a Review
Description
Acclaimed data scientist DJ Patil details a new approach to solving problems in Data Jujitsu.
Learn how to use a problem's "weight" against itself to:
Break down seemingly complex data problems into simplified parts
Use alternative data analysis techniques to examine them
Use human input, such as Mechanical Turk, and design tricks that enlist the help of your users to take short cuts around tough problems
Learn more about the problems before starting on the solutions—and use the findings to solve them, or determine whether the problems are worth solving at all.
Customer Reviews
5.0 (based on 1 review)
Write a Review
Reviewed by 1 customer
Displaying review 1
Get Immediate Access Now
FREE Ebook from O'Reilly
Formats: ePub, Mobi, PDF
First Name:
Last Name:
Email Address:
Get Your Free Ebook
We protect your privacy.



[…]data products are unique in that they are often extremely difficult, and seemingly intractable for small teams with limited funds. Yet, they get solved every day.

How? Are the people who solve them superhuman data scientists who can come up with better ideas in five minutes than most people can in a lifetime? Are they magicians of applied math who can cobble together millions of lines of code for high-performance machine learning in a few hours? No. Many of them are incredibly smart, but meeting big problems head-on usually isn't the winning approach. There's a method to solving data problems that avoids the big, heavyweight solution, and instead, concentrates building something quickly and iterating. Smart data scientists don't just solve big, hard problems; they also have an instinct for making big problems small.

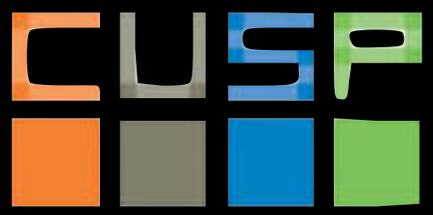


Data Definitions

- **Data:** observations that have been collected
- **Population:** the complete body of subjects we want to infer about
- **Sample:** the subset of the population we actually studied

- **Census:** collection of data from the entire population

- **Parameter:** numerical value describing an attribute of the *population*
- **Statistics:** numerical value describing an attribute of the *sample*



Types of Data:

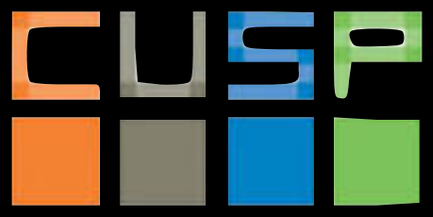
- **Continuous:** distance to the closest park
- **Ordinal:** survey response Good/Fair/Poor
- **Categorical:** gender, race

Data may also be:

- **Censored:** age>90
- **Missing:** "Prefer not to answer" (NA / NaN)

Numerical data may also be:

- **Ordinal:** ranks of Colleges
- **Interval:** F temperature - interval size preserved
- **Ratio:** Car speed - 0 is naturally defined



- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)
- statistics
  - models (regression)
  - prediction
  - validation (simulations)
- interpretation
- presentation
  - visualization
  - write a paper!

FORMULATING HYPOTHESIS

- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)
- statistics
  - models (regression)
  - prediction
  - validation (simulations)
- interpretation
- presentation
  - visualization
  - write a paper!

- IDEA
- dataset

FORMULATING HYPOTHESIS

  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)

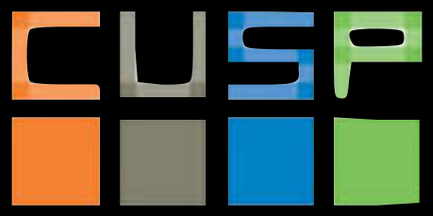
- statistics
  - models (regression)
  - prediction
  - validation (simulations)

HYPOTHESIS TESTING

- interpretation
- presentation
  - visualization
  - write a paper!

CUSP

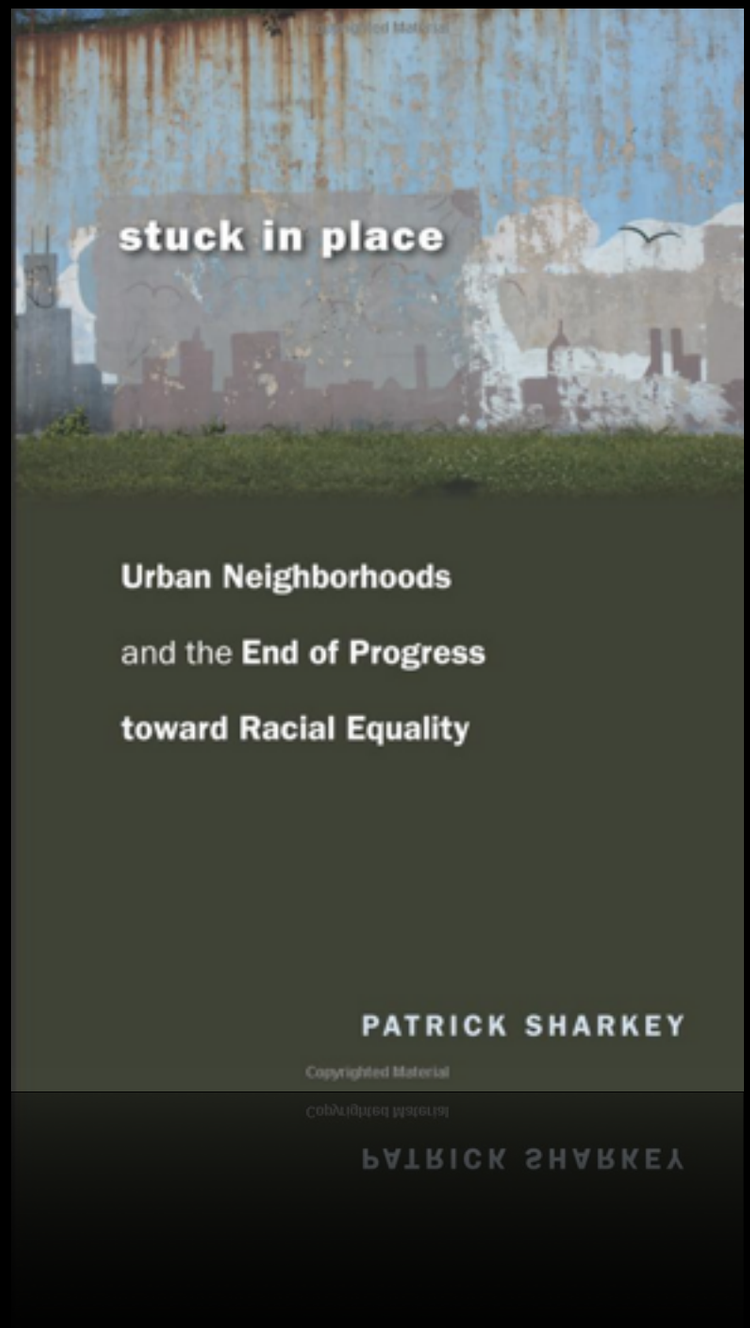
from idea to Null Hypothesis



III: Introduction to statistics

- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)
- statistics
  - models (regression)
  - prediction
  - validation (simulations)
- interpretation
- presentation
  - visualization
  - write a paper!

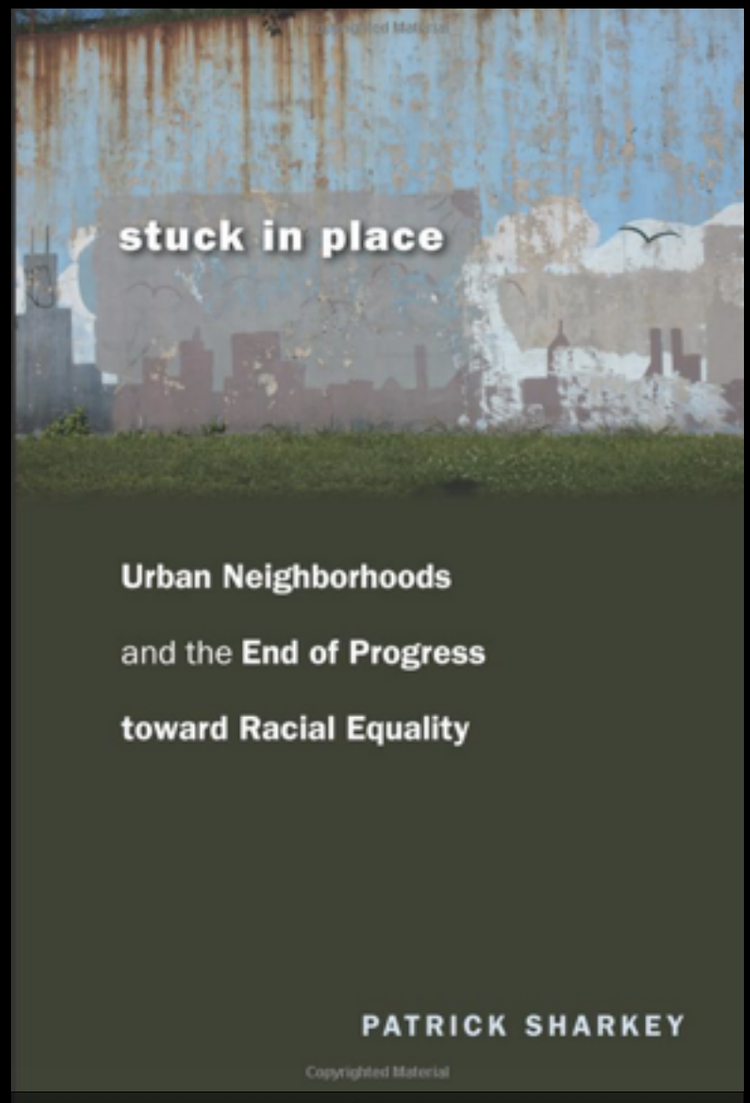
CUSP seminar
Friday 9/18



III: Introduction to statistics

- IDEA
- dataset
  1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis
     - define ideal data
     - figure out best data available
     - figure out if you can get new data
     - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)
- statistics
  - models (regression)
  - prediction
  - validation (simulations)
- interpretation
- presentation
  - visualization
  - write a paper!

CUSP seminar Friday 9/18



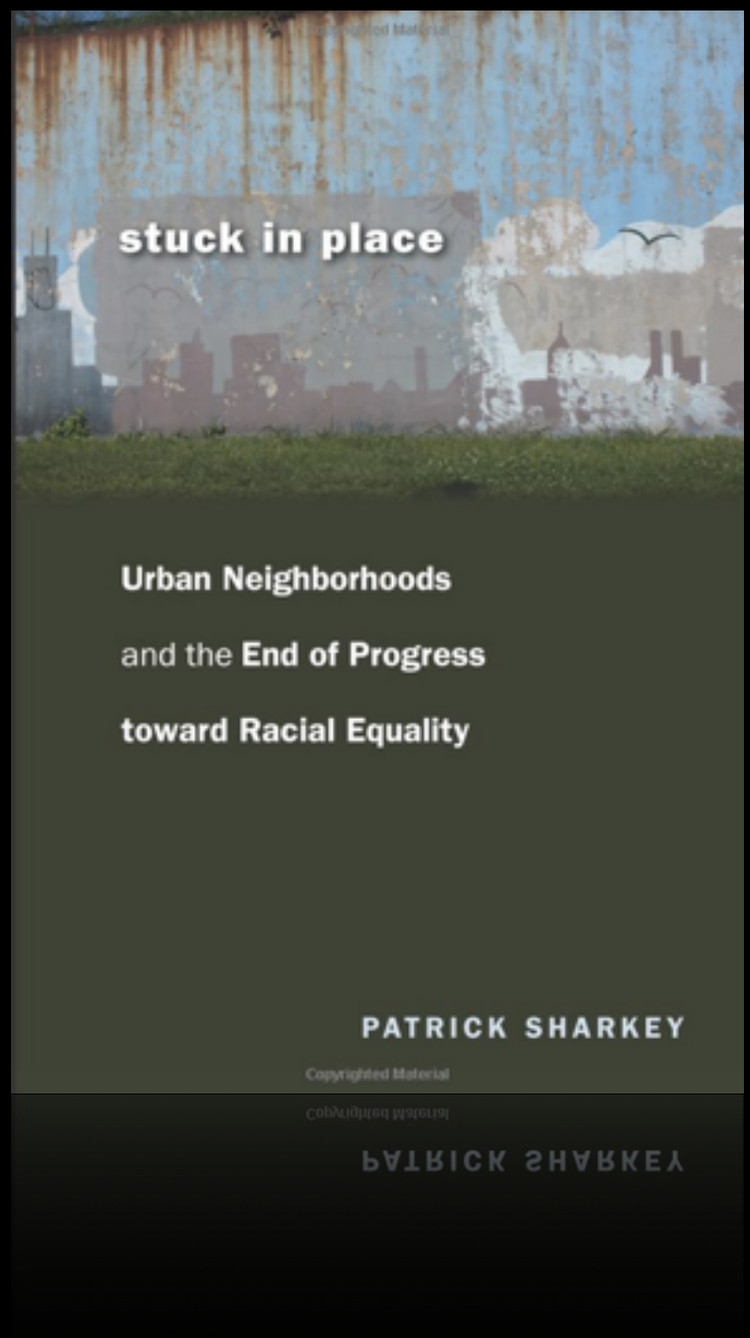
III: Introduction to statistics

- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
  - data handling
    - joining databases
    - formatting data
  - exploratory data analysis
    - machine learning (clustering? dimensionality reduction?)
  - statistics
    - models (regression)
    - prediction
    - validation (simulations)
  - interpretation
  - presentation
    - visualization
    - write a paper!

1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis

e.g.:

QUESTION: does proximity to violence affect children's development ?

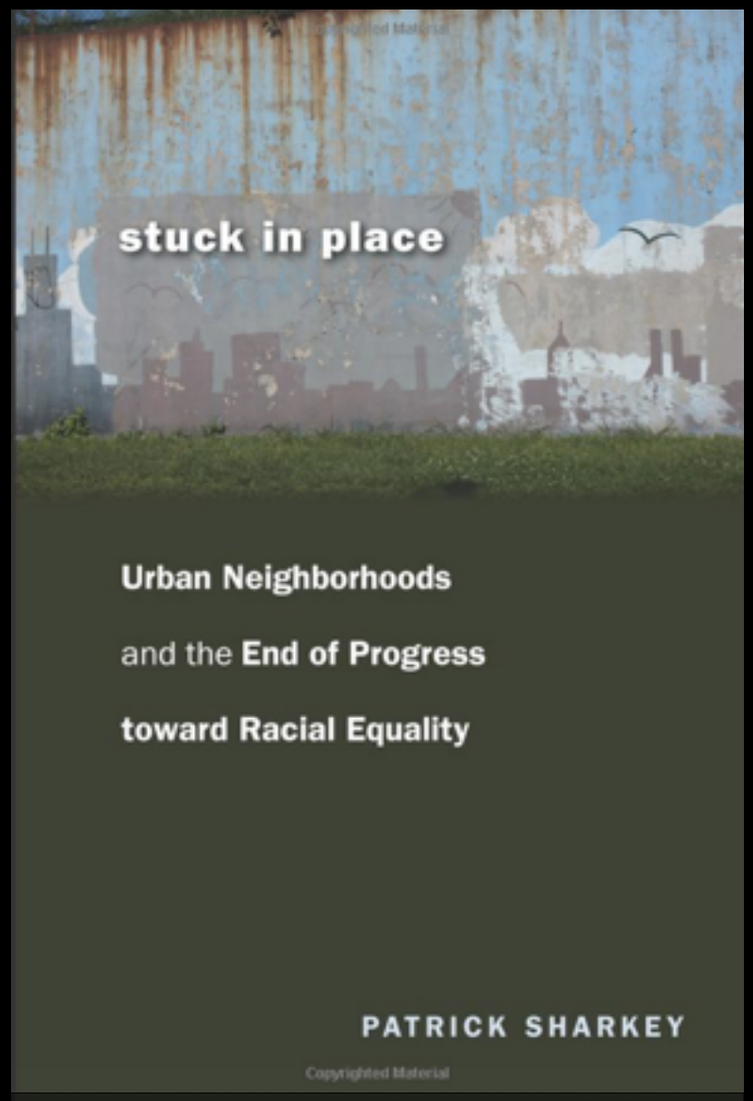


- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)
- statistics
  - models (regression)
  - prediction
  - validation (simulations)
- interpretation
- presentation
  - visualization
  - write a paper!

1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis

QUESTION: does proximity to violence affect children's development?

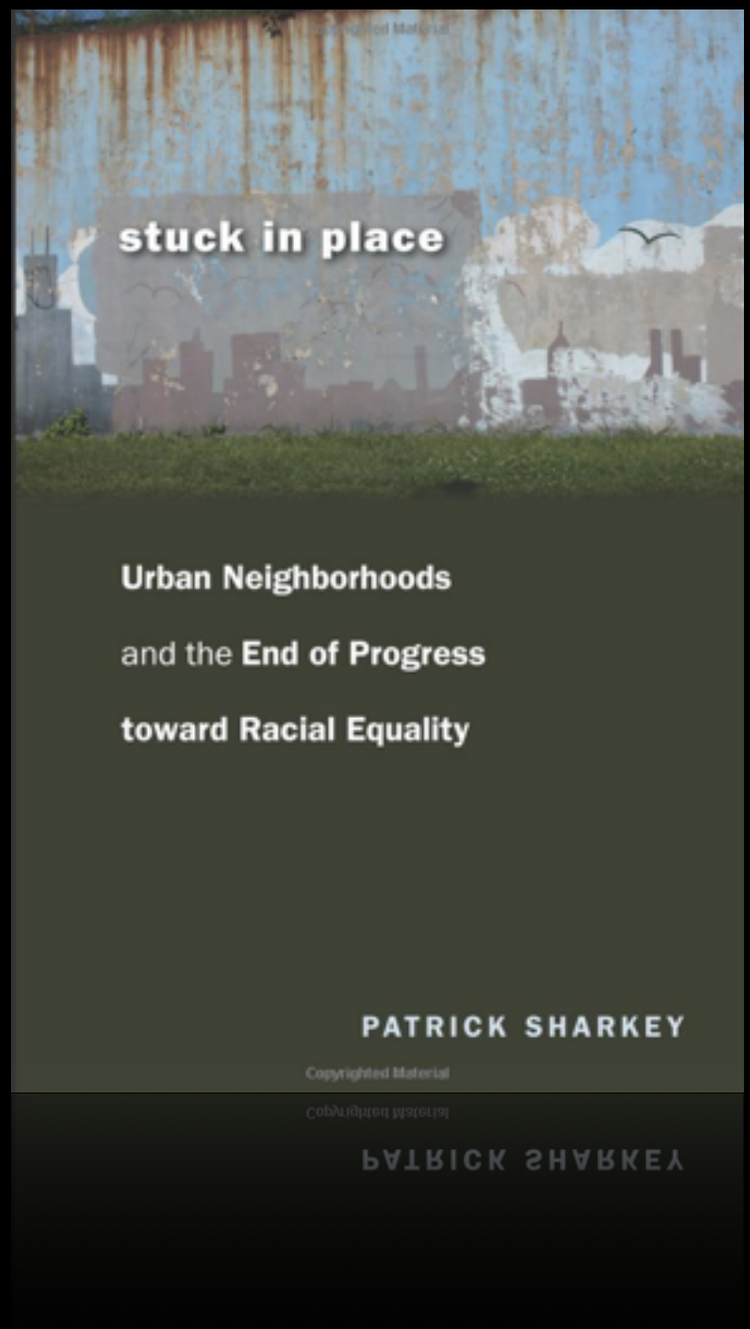
HYPOTHESIS: the reading test score of children who live near the site of a violent crime is lower after the crime occurred



1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis

HYPOTHESIS: the reading test score of children who live near the site of a violent crime is lower after the crime occurred

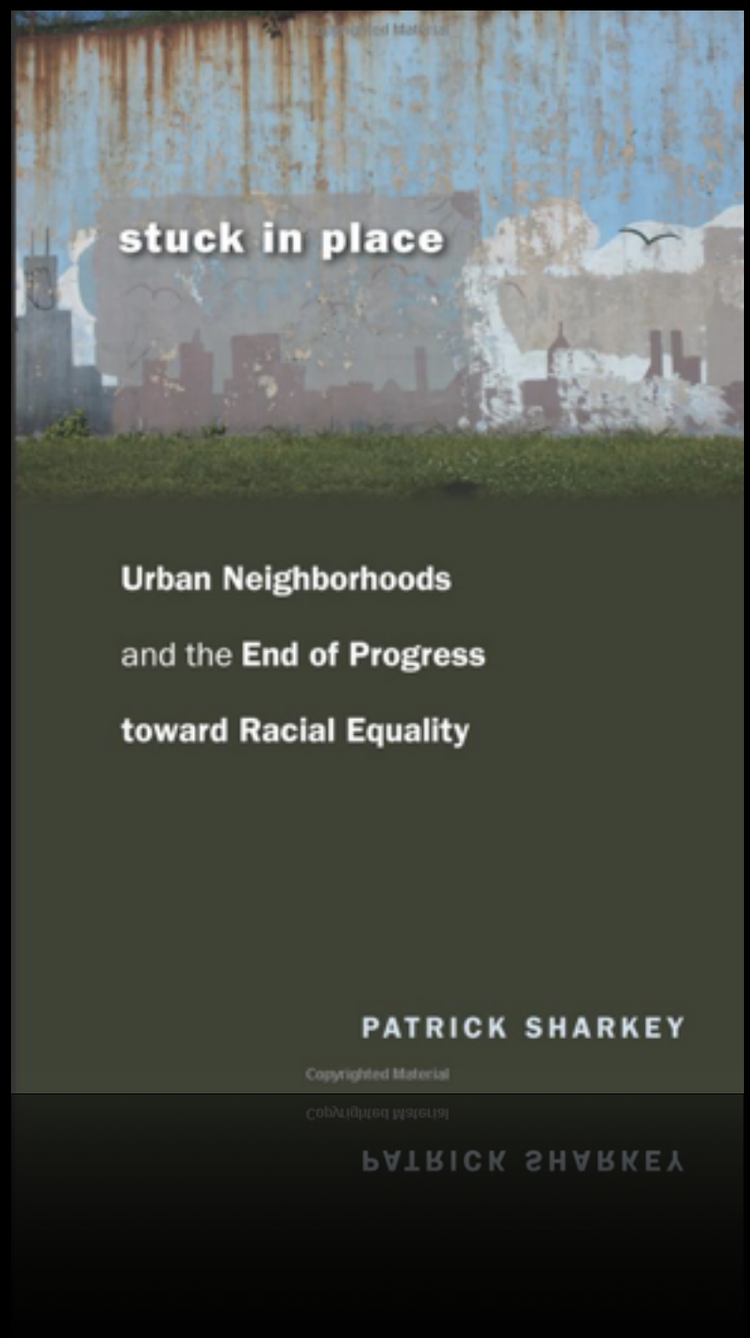
***TESTABLE* HYPOTHESIS**: the *average* test score of children who live *within a block of* the site of a violent crime is *significantly* lower in the days following the crime



- IDEA
- 1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis

***NULL* HYPOTHESIS**: the average reading test score of children who live within a block of the site of a violent crime is *the same or higher* than the average score for the control group in the days following the crime, *significance level p=0.05*

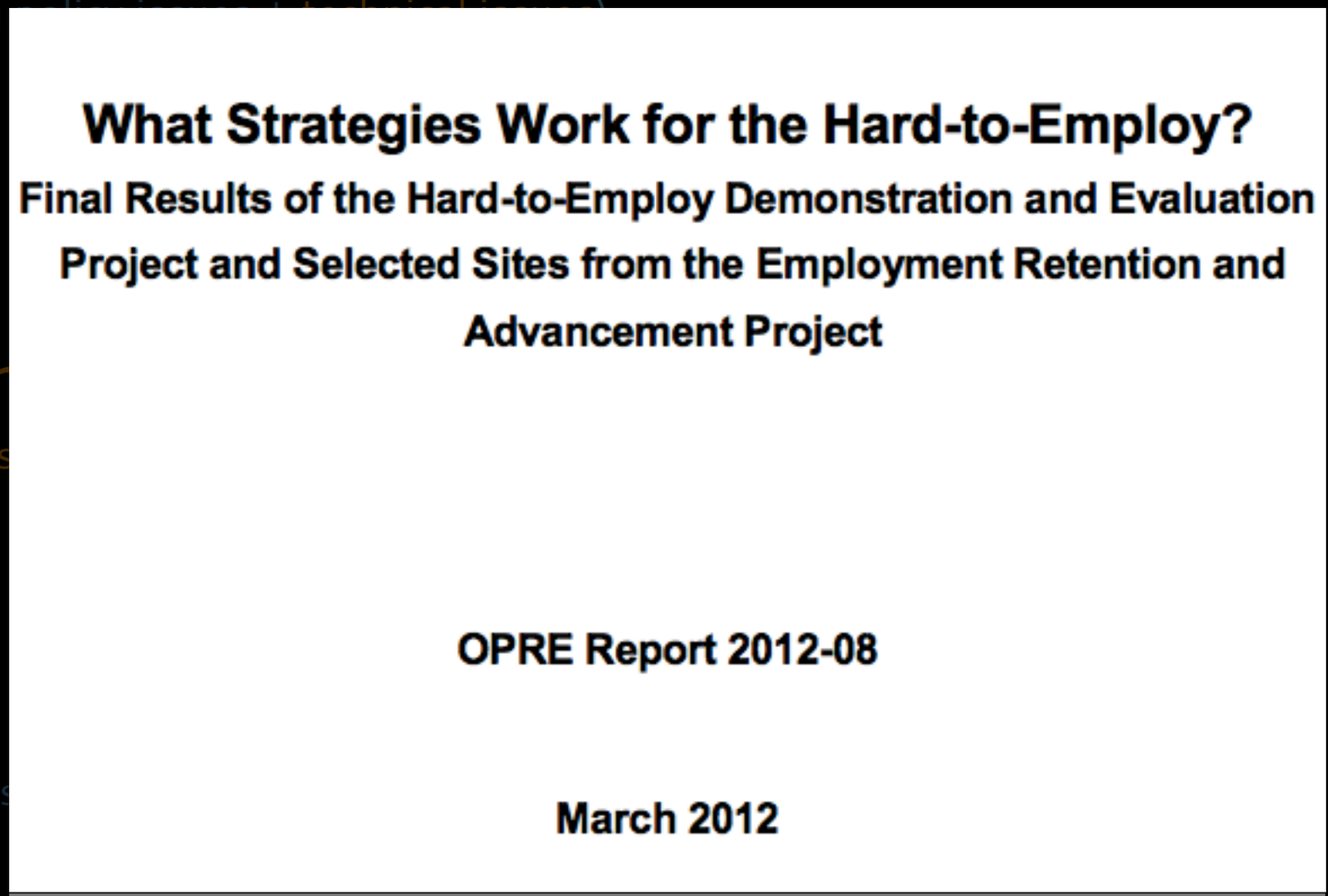
***ALTERNATIVE* HYPOTHESIS**: the *average* test score of children who live *within a block of* the site of a violent crime is *significantly* lower in the days following the crime



- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including
- data handling
  - joining databases
  - formatting data
- exploratory data an
  - machine learning (clus
- statistics
  - models (regression)
  - prediction
  - validation (simulations
- interpretation
- presentation
  - visualization
  - write a paper!

IDEA: does the NYC Post-Prison Employment Programs increase employment?

1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis



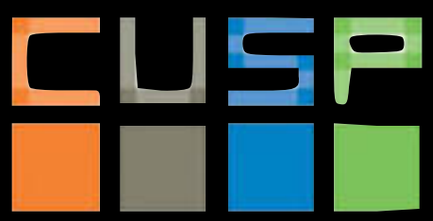
What Strategies Work for the Hard-to-Employ?

Final Results of the Hard-to-Employ Demonstration and Evaluation Project and Selected Sites from the Employment Retention and Advancement Project

OPRE Report 2012-08

March 2012

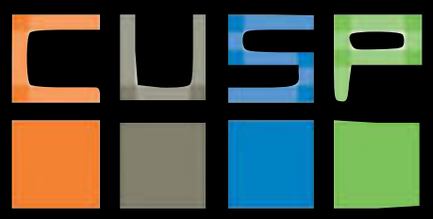
http://cosmo.nyu.edu/~fb55/UI_CUSP_2015/submit.html



- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)
- statistics
  - models (regression)
  - prediction
  - validation (simulations)
- interpretation
- presentation
  - visualization
  - write a paper!

QUESTION: does the NYC Post-Prison Employment Programs increase employment?

HYPOTHESIS: the number of former prisoners employed 3 years after release is higher for candidates who participated in the program

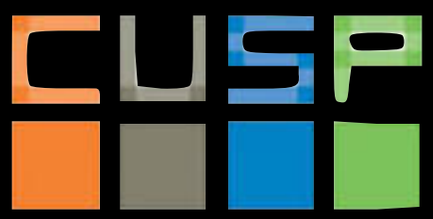


- IDEA
- dataset
  - define ideal data
  - figure out best data available
  - figure out if you can get new data
  - obtain data (including policy issues + technical issues)
- data handling
  - joining databases
  - formatting data
- exploratory data analysis
  - machine learning (clustering? dimensionality reduction?)
- statistics
  - inference (regression)
  - prediction
  - validation (simulations)
- interpretation
- presentation
  - visualization
  - write a paper!

1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis

HYPOTHESIS: the number of former prisoners employed 3 years after release is higher for candidates who participated in the program

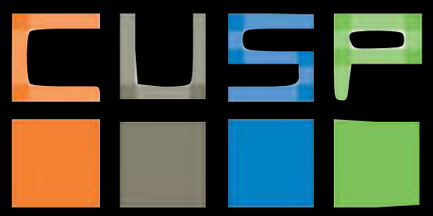
***TESTABLE* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *significantly* higher for candidates who participated in the program



- IDEA
  1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis

***NULL* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group, *significance level p=0.05*

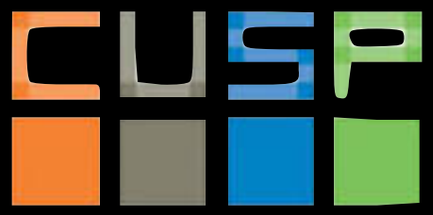
***ALTERNATIVE* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *significantly* higher for candidates who participated



1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis
2. choose and mangle datasets
3. choose suitable statistical tests
4. assess if your result rejects the null hypothesis

1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis
2. choose and mangle datasets
3. choose suitable statistical tests
4. assess if your result rejects the null hypothesis

Hypothesis testing: Statistical Analysis



III: Introduction to statistics

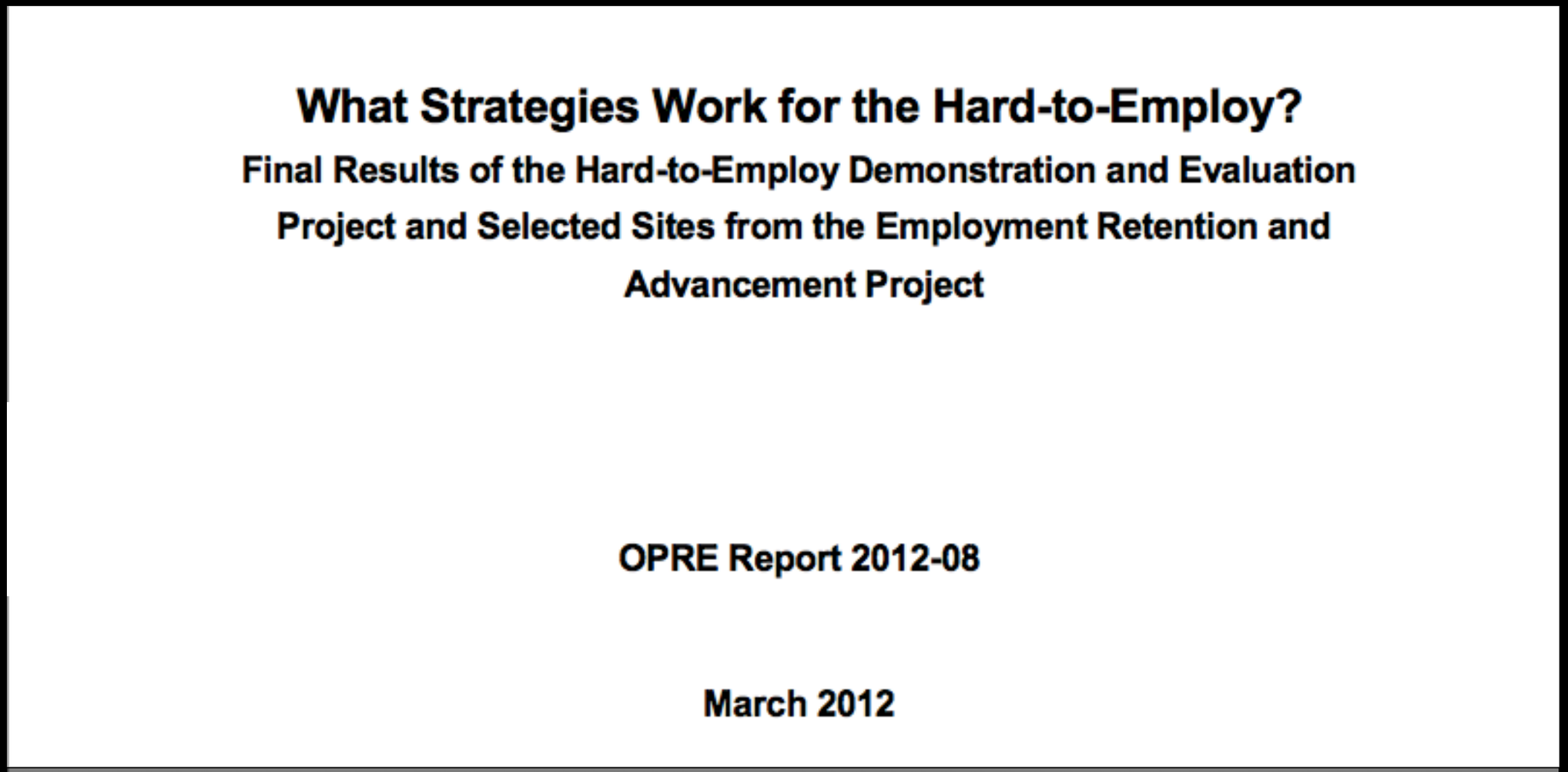
hypothesis testing

null hypothesis: no relationship between two measured phenomena, or no difference among groups
if you have a test control sample: test sample and control sample are the same - no effect

falsify the null hypothesis: do you see an effect?
do you see a difference b/w samples?

***NULL* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,

significance level p=0.05



What Strategies Work for the Hard-to-Employ?

Final Results of the Hard-to-Employ Demonstration and Evaluation Project and Selected Sites from the Employment Retention and Advancement Project

OPRE Report 2012-08

March 2012

http://www.mdrc.org/sites/default/files/What%20Strategies%20Work%20for%20the%20Hard%20FR.pdf

***NULL* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,

significance level p=0.05

The Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project

Table 2.1

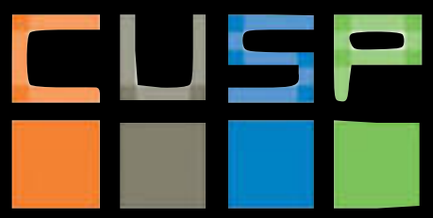
Summary of Impacts, New York City Center for Employment Opportunities

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Employment (Years 1-3) (%)				
Ever employed	83.8	70.4	13.4 ***	0.000
Ever employed in a CEO transitional job[a]	70.1	3.5	66.6 ***	0.000
Ever employed in an unsubsidized job	63.7	69.0	-5.3 *	0.078
Postprogram unsubsidized employment (Years 2-3)				
Ever employed in an unsubsidized job (%)	53.3	52.1	1.2	0.713
Employed in an unsubsidized job, average per quarter (%)	28.2	27.2	1.1	0.618
Employed for six or more consecutive quarters (%)	14.7	11.9	2.8	0.195
Total UI-covered earnings[b] ($)	10,435	9,846	589	0.658
Sample size (total = 973)[c]	564	409		

SOURCES: MDRC earnings calculations from the National Directory of New Hires (NDNH) database and employment calculations from the unemployment insurance (UI) wage records from New York State, MDRC calculations using data from the New York State Division of Criminal Justice Services (DCJS) and the New York City Department of Correction (DOC).

NOTES: Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.
The p-value indicates the likelihood that the difference between the program and control groups arose by chance.

http://www.mdrc.org/sites/default/files/What%20Strategies%20Work%20for%20the%20Hard%20FR.pdf



***NULL* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,

significance level p=0.05

The Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project

Table 2.1

Summary of Impacts, New York City Center for Employment Opportunities

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Employment (Years 1-3) (%)	P_1	P_0		
Ever employed	83.8	70.4	13.4 ***	0.000
Ever employed in a CEO transitional job[a]	70.1	3.5	66.6 ***	0.000
Ever employed in an unsubsidized job	63.7	69.0	-5.3 *	0.078
Postprogram unsubsidized employment (Years 2-3)				
Ever employed in an unsubsidized job (%)	53.3	52.1	1.2	0.713
Employed in an unsubsidized job, average per quarter (%)	28.2	27.2	1.1	0.618
Employed for six or more consecutive quarters (%)	14.7	11.9	2.8	0.195
Total UI-covered earnings[b] ($)	10,435	9,846	589	0.658
Sample size (total = 973)[c]	564	409		

SOURCES: MDRC earnings calculations from the National Directory of New Hires (NDNH) database and employment calculations from the unemployment insurance (UI) wage records from New York State, MDRC calculations using data from the New York State Division of Criminal Justice Services (DCJS) and the New York City Department of Correction (DOC).

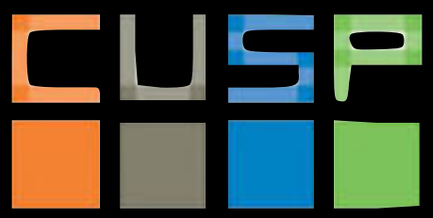
NOTES: Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.
The p-value indicates the likelihood that the difference between the program and control groups arose by chance.

H_0: $P_0 - P_1 > 0$

H_a: $P_0 - P_1 <= 0$

$\alpha=0.05$

http://www.mdrc.org/sites/default/files/What%20Strategies%20Work%20for%20the%20Hard%20FR.pdf



hypothesis testing

null hypothesis: no relationship between two measured phenomena, or no difference among groups
if you have a test control sample: test sample and control sample are the same - no effect

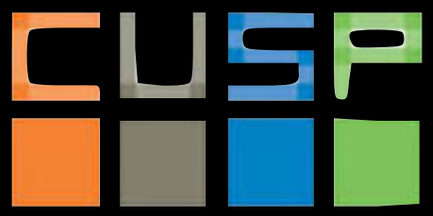
falsify the null hypothesis: do you see an effect?
do you see a difference b/w samples?

hypothesis testing

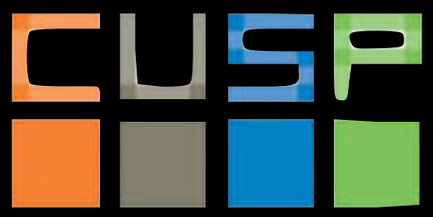
what is the probability that we would have gotten the same result out of just noise?

null hypothesis: no relationship between two measured phenomena, or no difference among groups
if you have a test control sample: test sample and control sample are the same - no effect

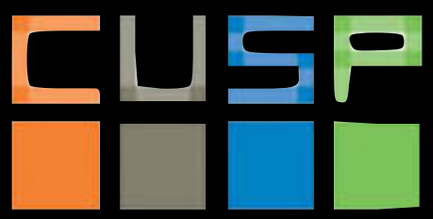
falsify the null hypothesis: do you see an effect?
do you see a difference b/w samples?
is this effect larger than the just by chance?



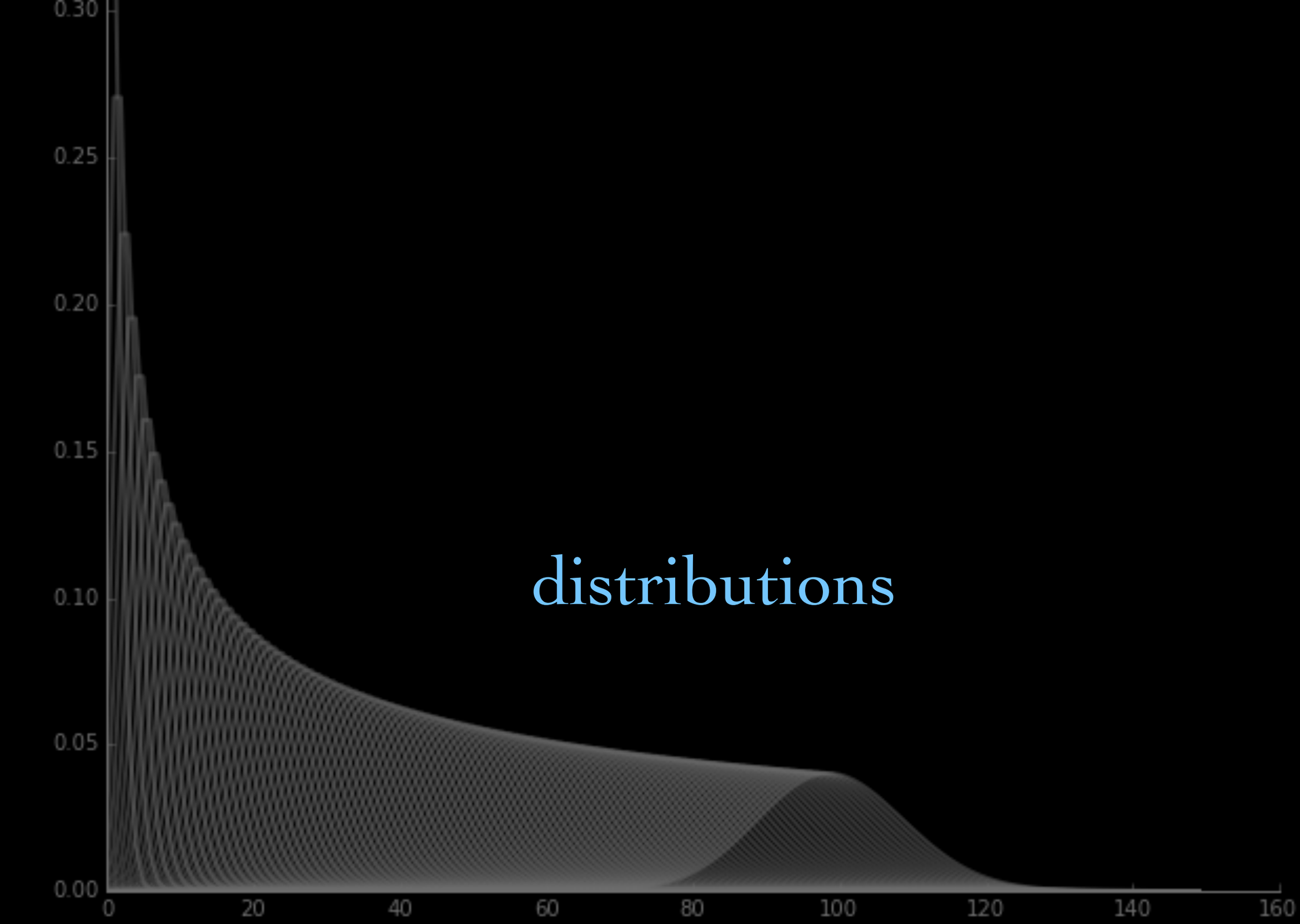
	H_0 is True	H_0 is False
H_0 is falsified	Type I error False Positive	True Positive
H_0 is not falsified	True Negative	Type II error False negative



	H_0 is True	H_0 is False
H_0 is falsified	Type I error False Positive important message gets spammed	True Positive
H_0 is not falsified	True Negative	Type II error False negative Spam in your Inbox



distributions



III: Introduction to statistics

treycausey.com/data_science_interviews.html

Apps Bookmarks gmail pedometer http://www.weather.c New folder SESN in literature nyu travel ccpp ccpp event CCPPshare papers Other Bookmarks

trey causey

Twitter

GitHub

LinkedIn

the spread

What it's like to be on the data science job market

September 20, 2015

Sooner or later you're going to find yourself looking for a data science job. Maybe it's your first one or maybe you're changing jobs. Even if you're fully confident in your skills, have no impostor syndrome, and have tons of inside leads at great companies, it's a tremendously stressful experience. The process of looking for a new job is often one that occurs secretly and confidentially and then is so exhausting that discussing the process is the last thing you want to do. I hope to change that.

I recently went through this myself and thought I'd record my thoughts on the process while they're still fresh. I interviewed a lot. Some went well, some didn't go well at all. The reason for this was sometimes me, sometimes them, often both. Sometimes I didn't get selected for an on-site interview. Other times I withdrew from the process after seeing that it wouldn't be a good fit for me. I took notes throughout, though, and here they are.

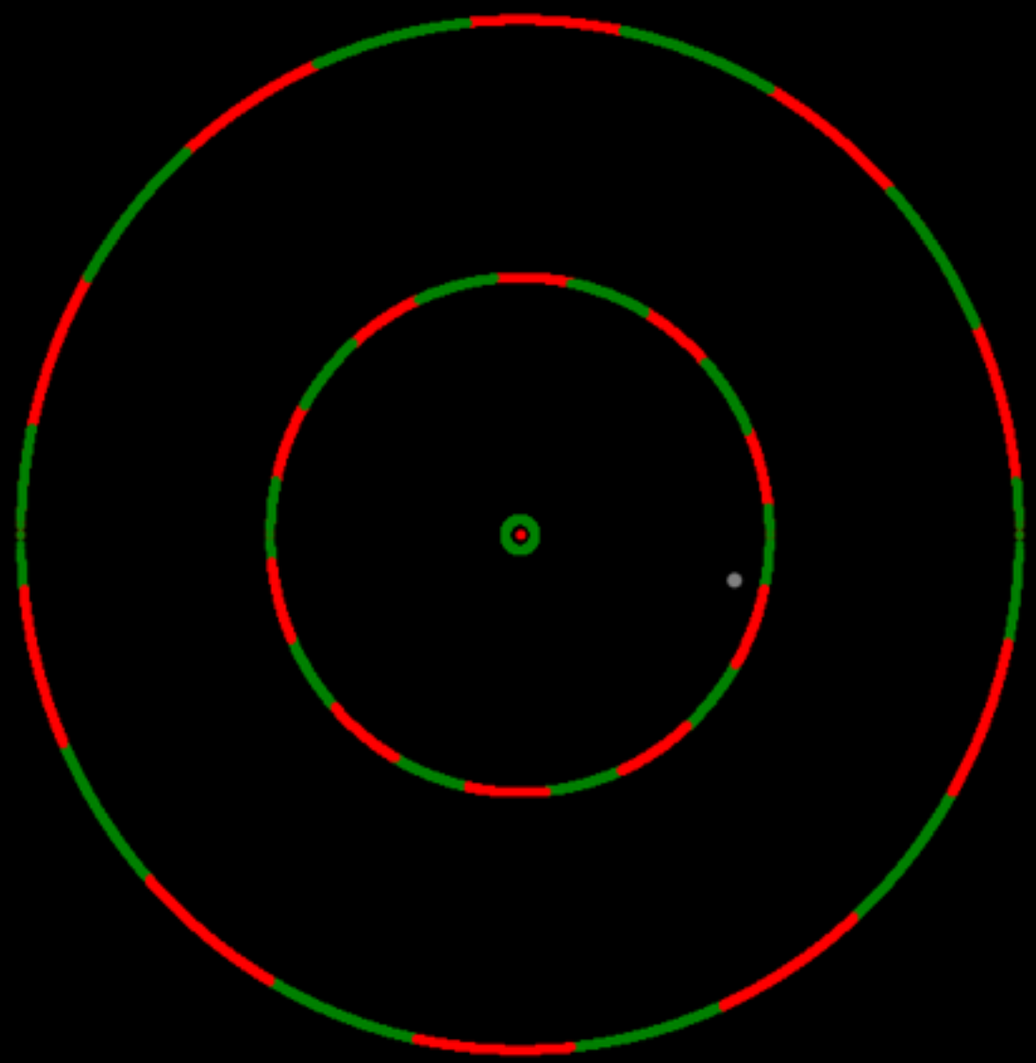
Warning: What follows are my personal thoughts, extrapolating from a small sample, and generalization from anecdotes. Precisely the kind of thing that data scientists hate! But, despite the frequent misquotation, the plural of anecdote is data, so this discussion should start somewhere.

Interview questions: What to expect

Prepare to be asked terrible questions. Some of the worst questions I've been asked:

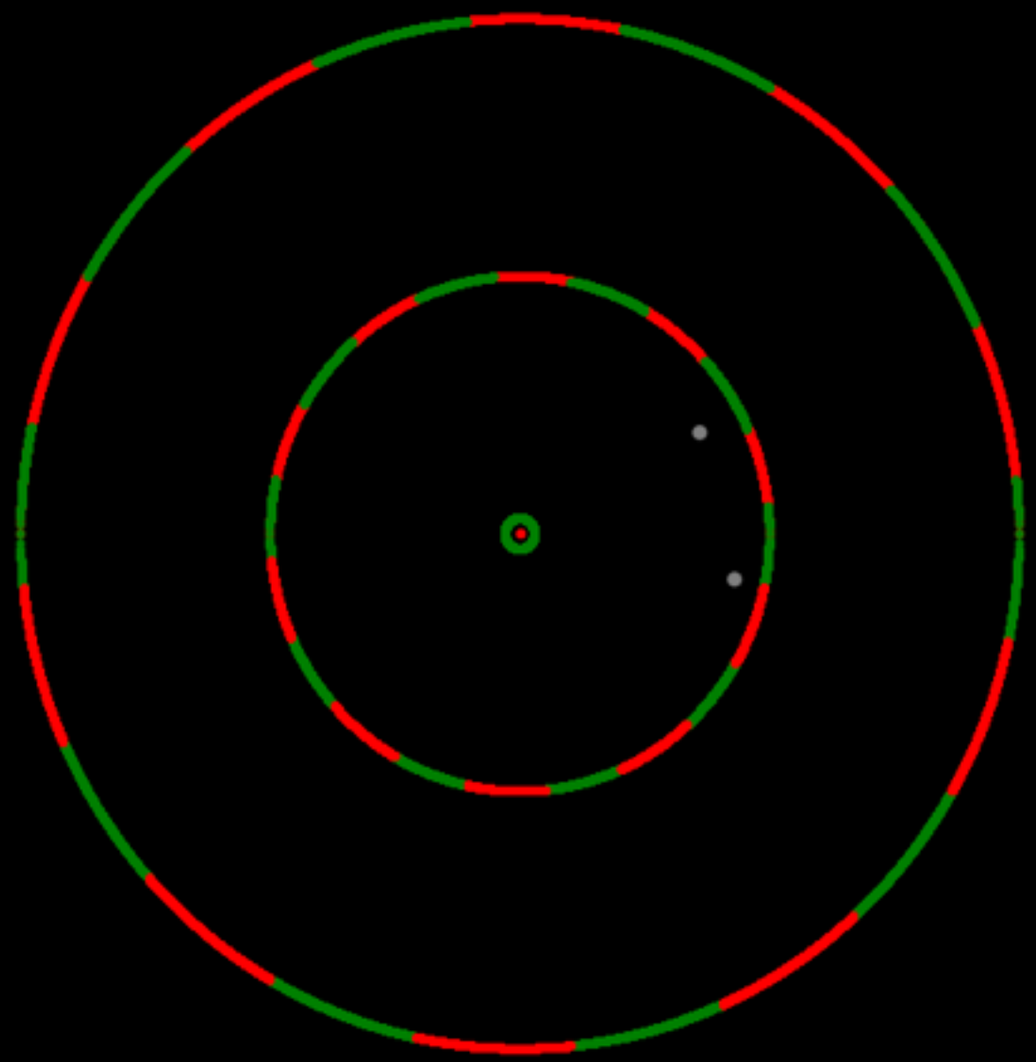
- I have a random number generator. What number does it produce and why?
- Anything involving dice or urns
- Can you tell me *certain piece of trivia* about *parameter rho* from *distribution delta*? (Names changed to protect the innocent PDFs)
- Here's a problem it took my team six months to solve. Please solve it for me on the whiteboard.

CUSP

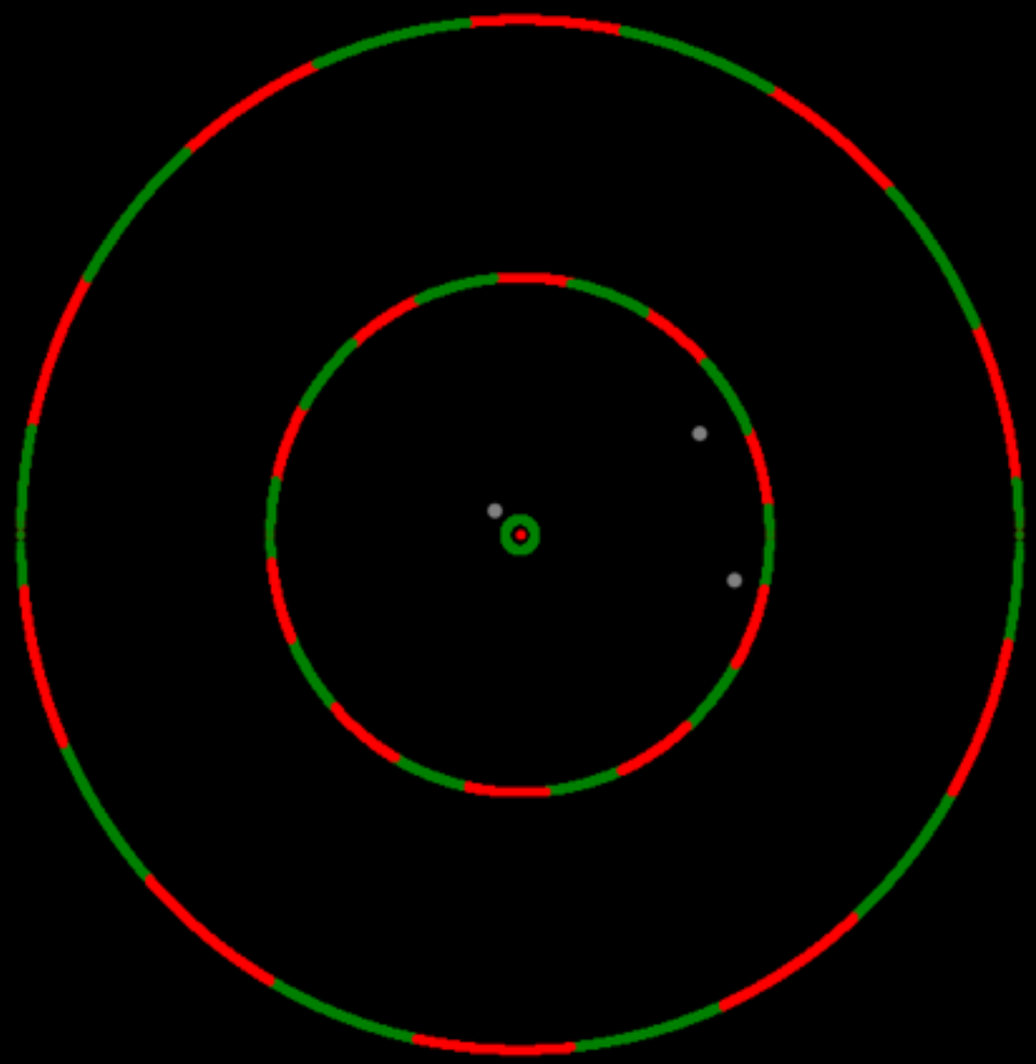


CUSP

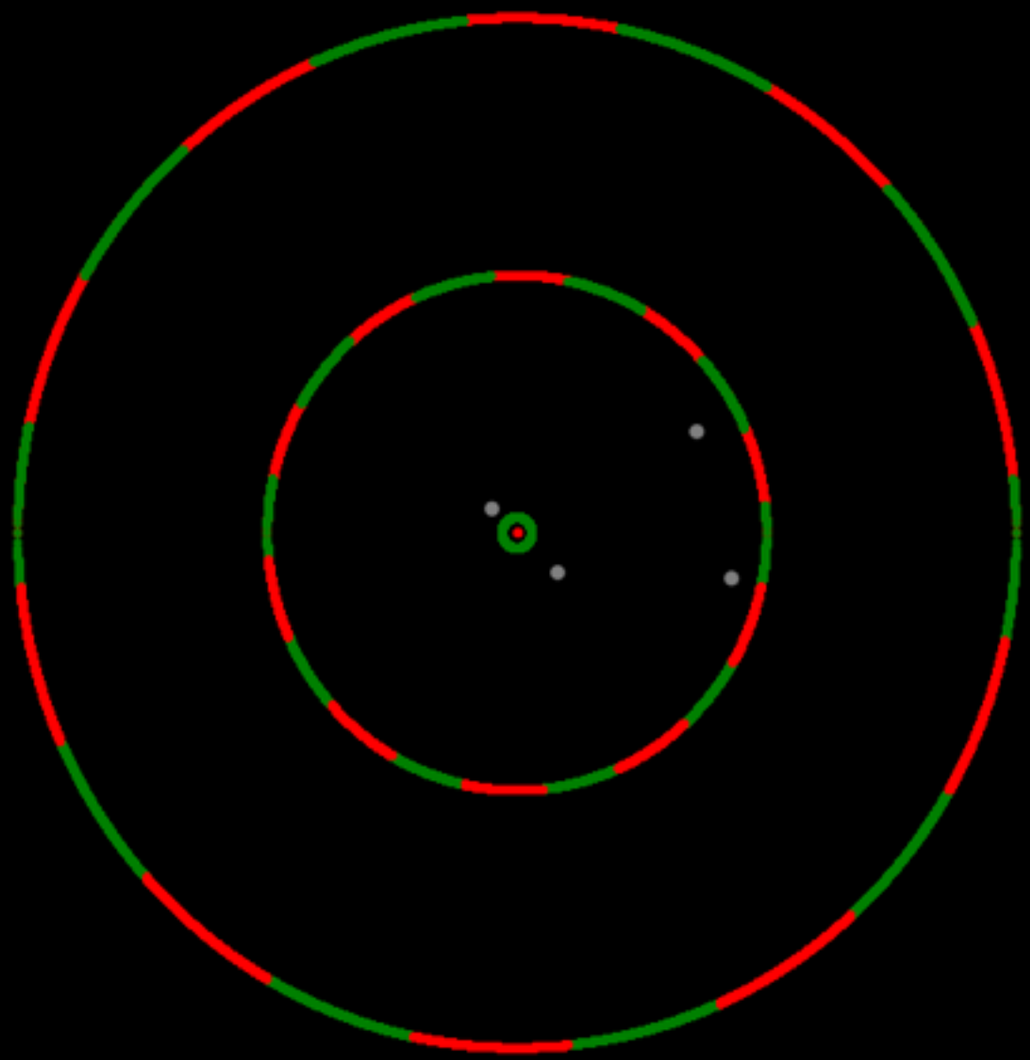
III: Introduction to statistics



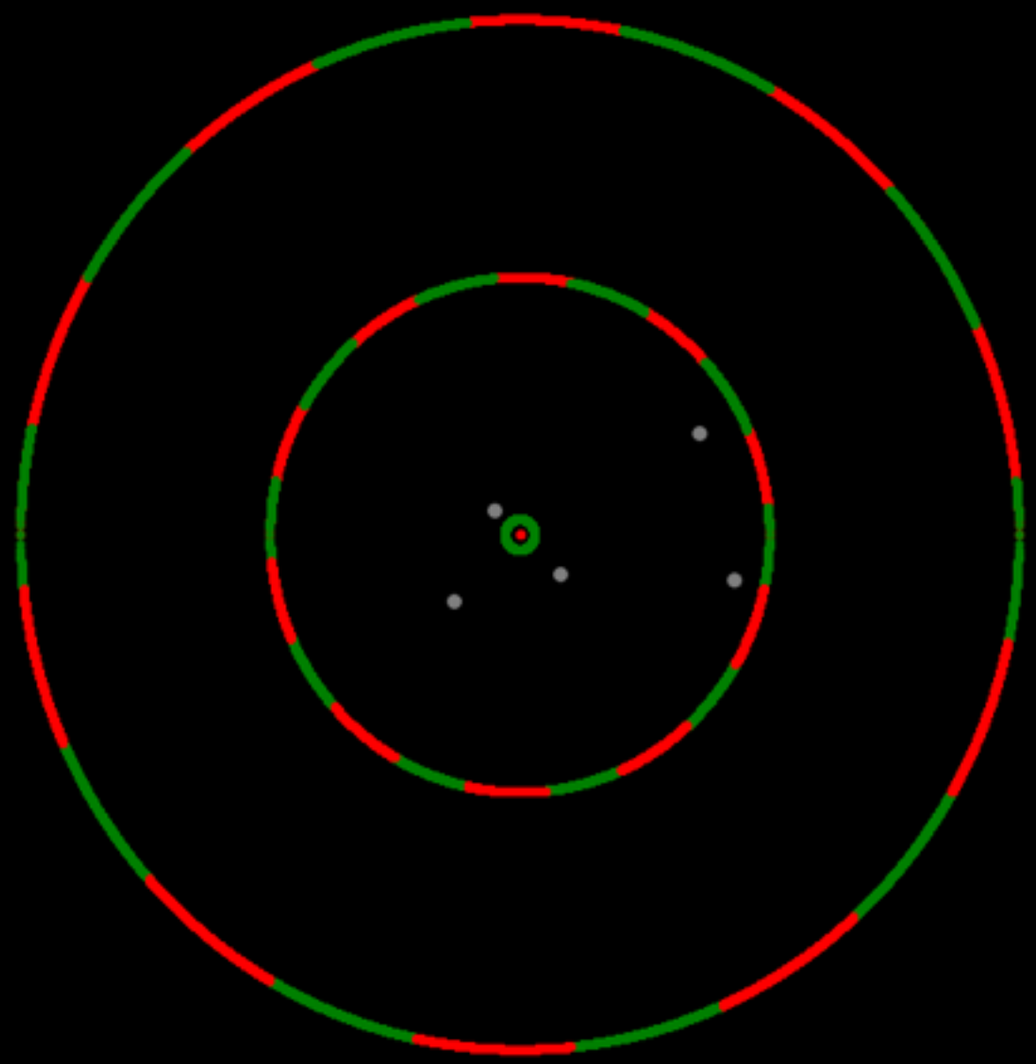
III: Introduction to statistics



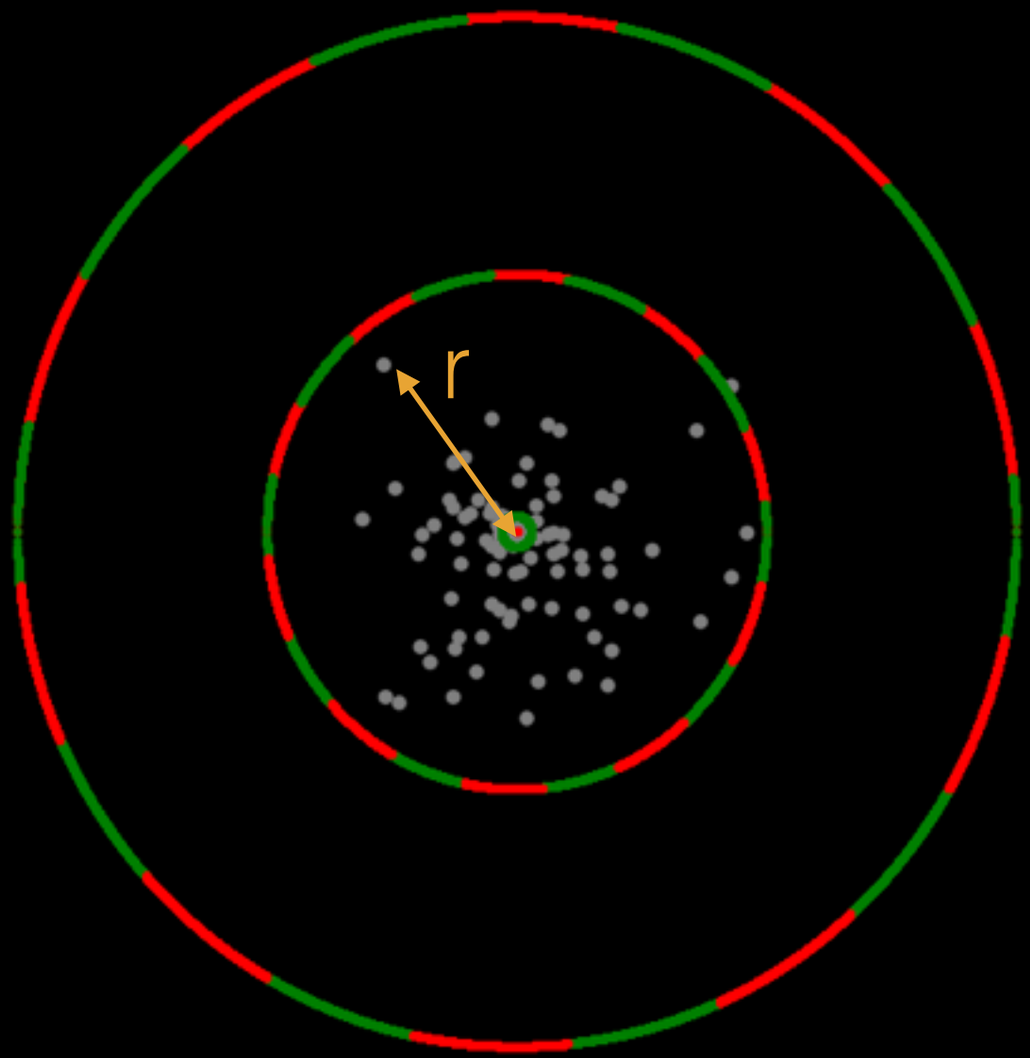
III: Introduction to statistics



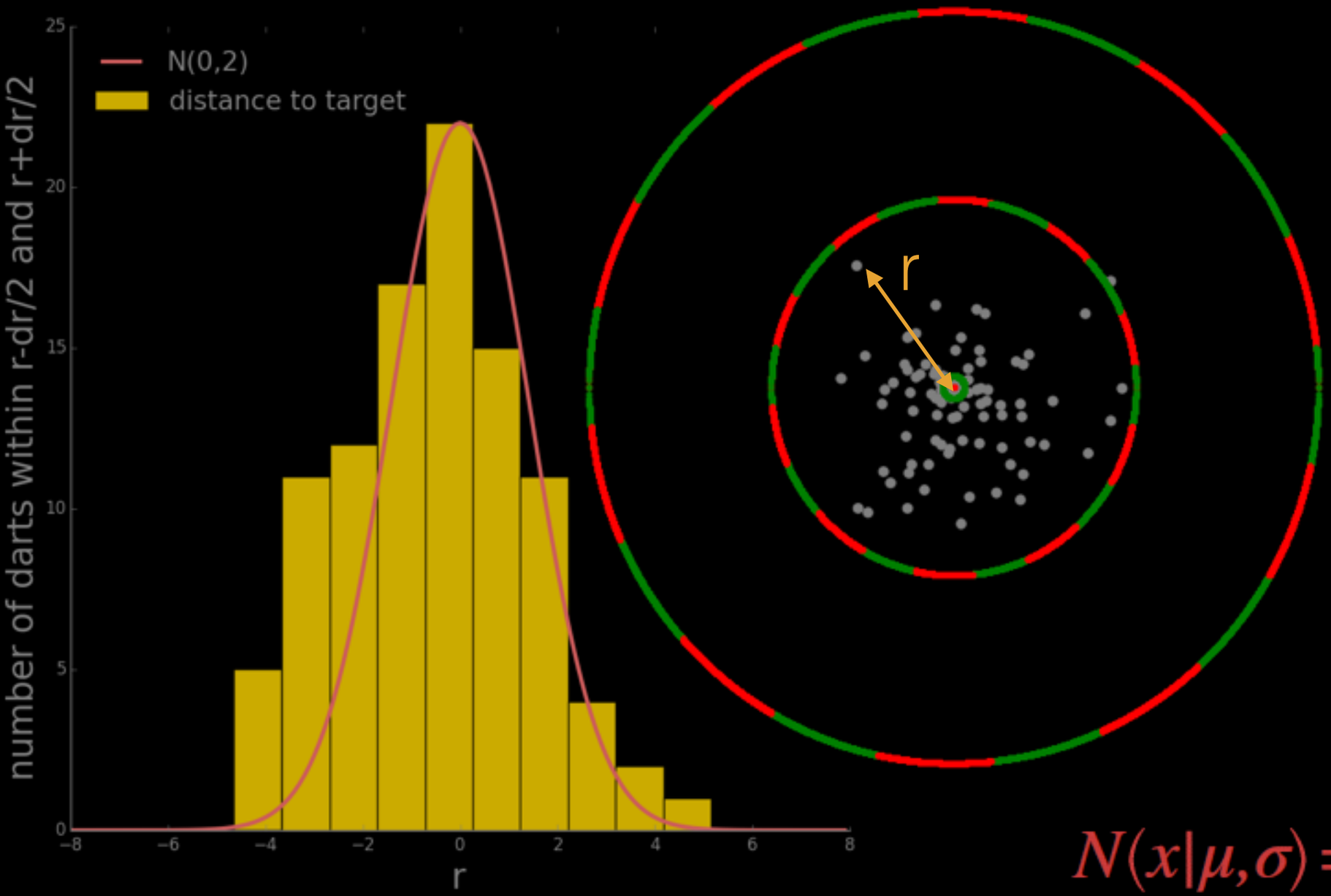
III: Introduction to statistics



III: Introduction to statistics



III: Introduction to statistics



$$N(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

III: Introduction to statistics

Random sampling (numpy.random)

Simple random data

rand(d0, d1, ..., dn)	Random values in a given shape.
randn(d0, d1, ..., dn)	Return a sample (or samples) from the "standard normal" distribution.
randint(low[, high, size])	Return random integers from *low* (inclusive) to *high* (exclusive).
random_integers(low[, high, size])	Return random integers between *low* and *high*, inclusive.
random_sample([size])	Return random floats in the half-open interval [0.0, 1.0).
random([size])	Return random floats in the half-open interval [0.0, 1.0).
ranf([size])	Return random floats in the half-open interval [0.0, 1.0).
sample([size])	Return random floats in the half-open interval [0.0, 1.0).
choice(a[, size, replace, p])	Generates a random sample from a given 1-D array
bytes(length)	Return random bytes.

Permutations

Save the figure

shuffle(x)	Modify a sequence in-place by shuffling its contents.
permutation(x)	Randomly permute a sequence, or return a permuted range.

Distributions

beta(a, b[, size])	The Beta distribution over `[0, 1]`.
binomial(n, p[, size])	Draw samples from a binomial distribution.
chisquare(df[, size])	Draw samples from a chi-square distribution.
dirichlet(alpha[, size])	Draw samples from the Dirichlet distribution.
exponential([scale, size])	Exponential distribution.

Table Of Contents

- Random sampling (`numpy.random`)
  - Simple random data
  - Permutations
  - Distributions
  - Random generator

Previous topic

numpy.RankWarning

Next topic

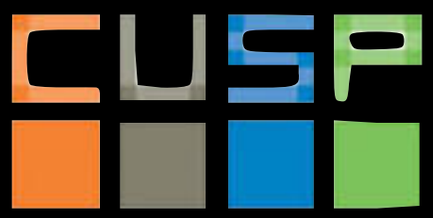
numpy.random.rand

distributions: moments

a distribution's moments summarize its properties:

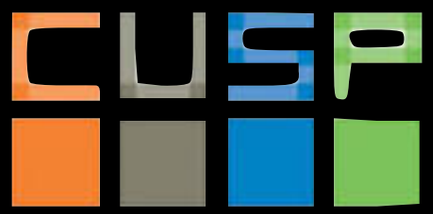
$$m_n = \int_{-\infty}^{\infty} (x-c)^n f(x)\,dx.$$

central tendency: mean (n=1), median, mode
spread: standard deviation/variance (n=2), quartiles range
symmetry: skewness (n=3)
CUSPiness: kurtosis (n=4)
…

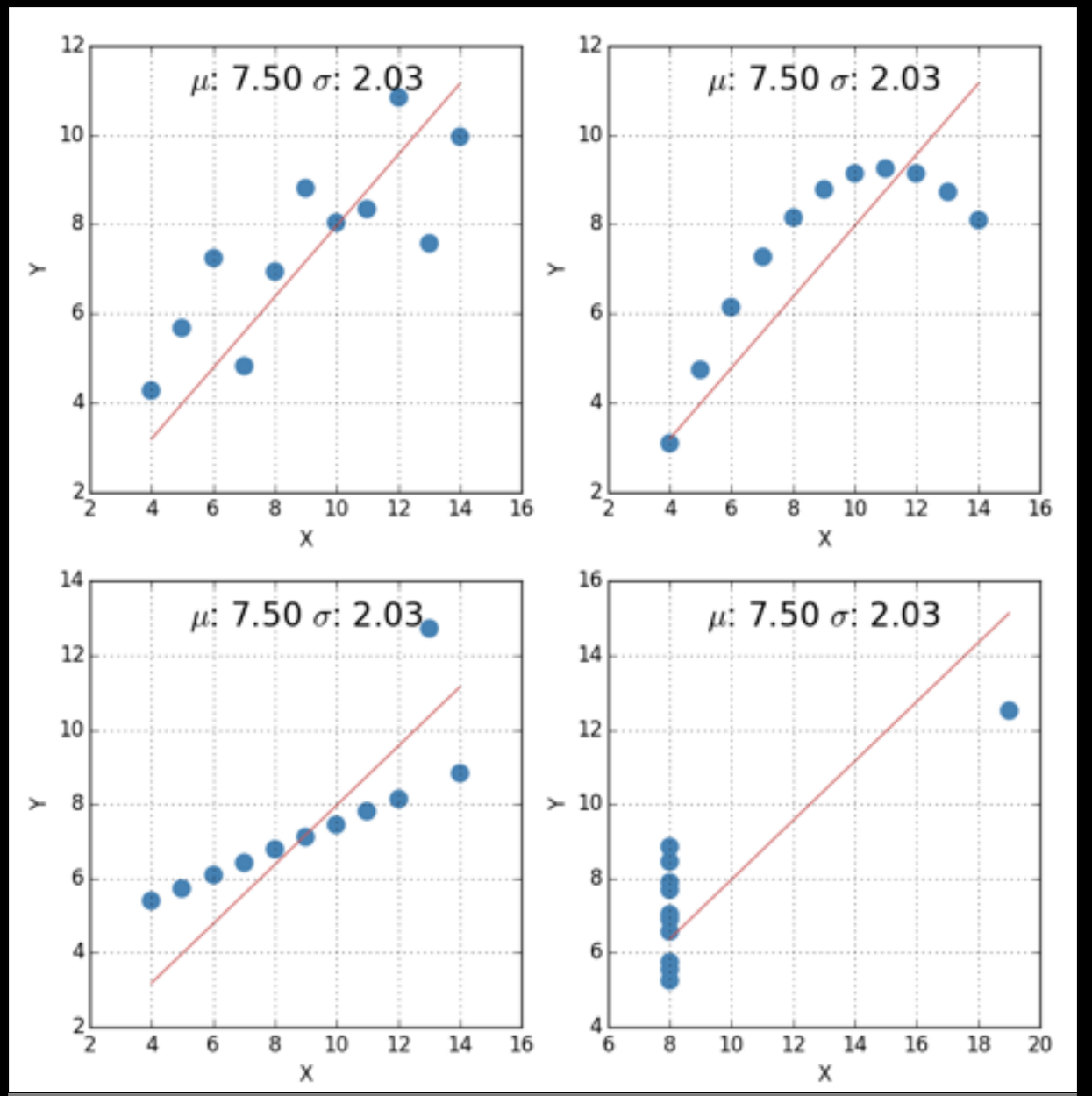


basic distributions and their moments

https://github.com/fedhere/UInotebooks/blob/master/poisson%20vs%20gaussian.ipynb



distributions: moments



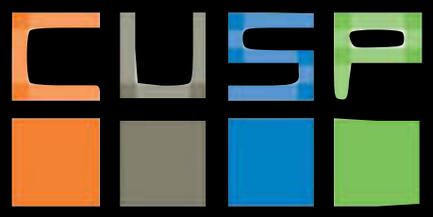
https://github.com/fedhere/UInotebooks/blob/master/Anscombe's%20Quartet.ipynb

distributions: Central Limit Theorem

Laplace (1700s)
but also: Poisson, Bessel, Dirichlet, Cauchy, Ellis

Let $X_1 \ldots X_N$ be an N-elements sample from a population whose distribution has
mean **μ** and standard deviation **σ**
In the limit of N-> infinity
the sample mean m approaches a Normal (Gaussian) distribution with mean μ and standard deviation σ
regardless of the distribution of X

$$\bar{x} \sim N\left(\mu, \sigma/\sqrt{N}\right)$$



distributions, moments, and Central Limit Theorem

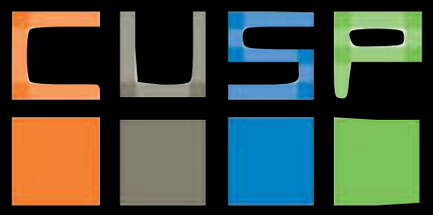
HOMEWORK 1 :

1. GENERATE 100 samples of different sizes N (N>10 & N<2000) from each of 6 different distributions (500 samples in total), all with the same *population* mean. Include a Normal, a Poisson, a Binomial, a Chi-Squared distribution, and 2 more of your choice.
2. For each sample plot the sample mean against the sample size N (if you want you can do it with the sample standard deviation as well). Can you describe the behavior you see in the plots?
3. PLOT the distributions of all sample means (together for all distributions). Mandatory: as a histogram,
   optional: in any other way you think is convincing
4. optional: FIT a gaussian to the distribution of means

e.g. how to fit function to data in numpy:

http://glowingpython.blogspot.com/2012/07/distribution-fitting-with-scipy.html

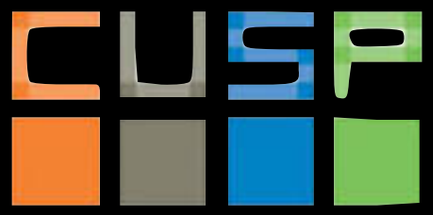
http://stackoverflow.com/questions/7805552/fitting-a-histogram-with-python



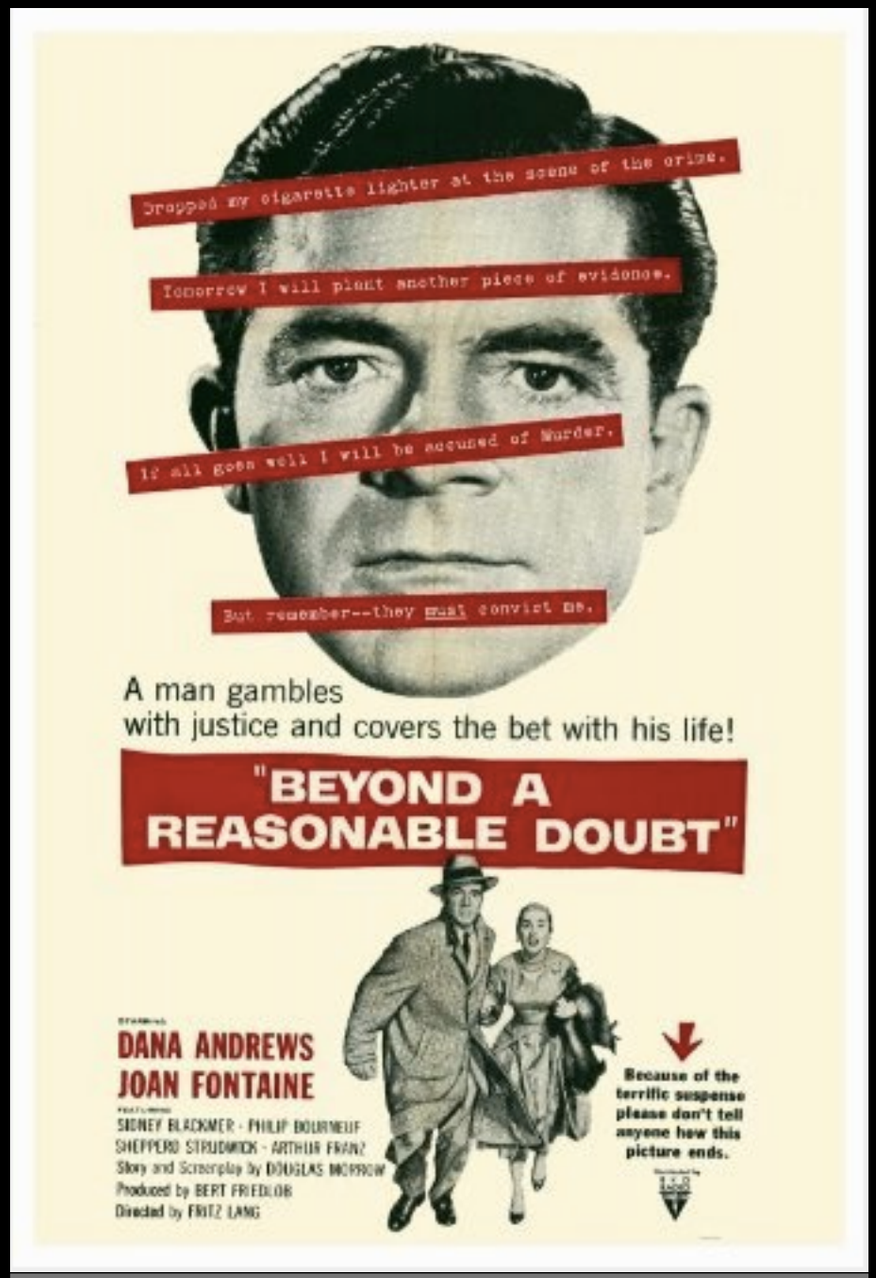
distributions, moments, and Central Limit Theorem

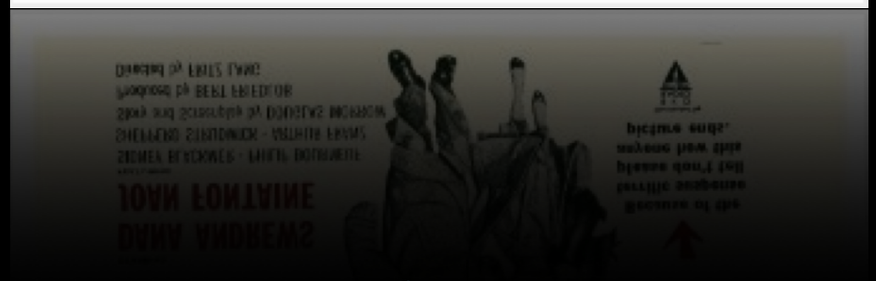
HOMEWORK 1 :

- create a new Github directory called HW3 inside of your PUI2015_<your name> repo.
- upload an ipython notebook, with the rendered plots.
- include a README.md which describes what you are doing, and, if appropriate, how to run the notebook (input variables? global variables that need to be setup?).
- 75% of the grade will be based on the rendered version of the plot, 25% will be awarded if the TA can download and run the notebook. If you include any package that was not in the standard Anaconda distribution state that in your README.md, so that the TA can download them.



rejecting the Null beyond any reasonable doubt



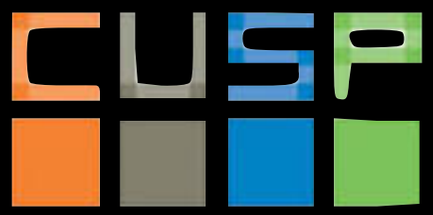


CUSP

III: Introduction to statistics

Rejecting the Null Hypothesis

is the probability of getting a result at least as extreme as the one you observed lower than the significance level you established?

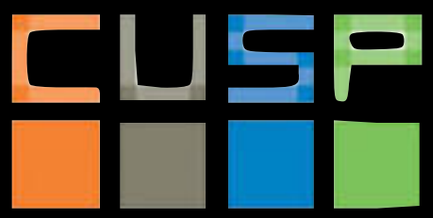


Rejecting the Null Hypothesis

is the probability of getting a result at least as extreme as the one you observed lower than the significance level you established?

- decide what the significance threshold is: typically 5%

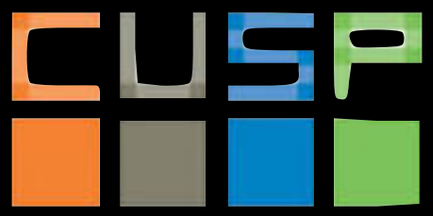
  $\alpha=0.05$



p-value

given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

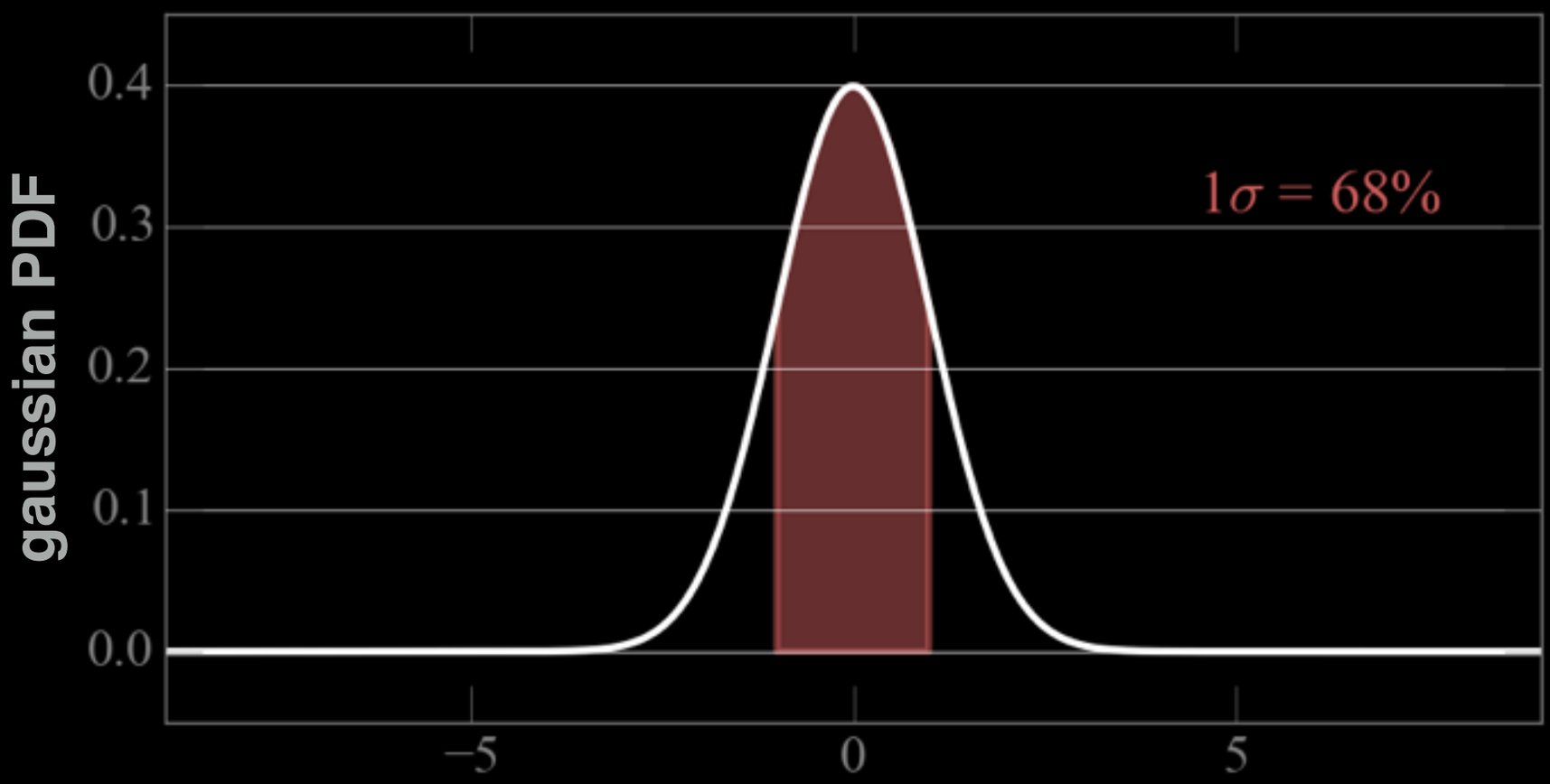
- decide what the significance threshold is: typically 5%
  α=0.05



p-value

given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

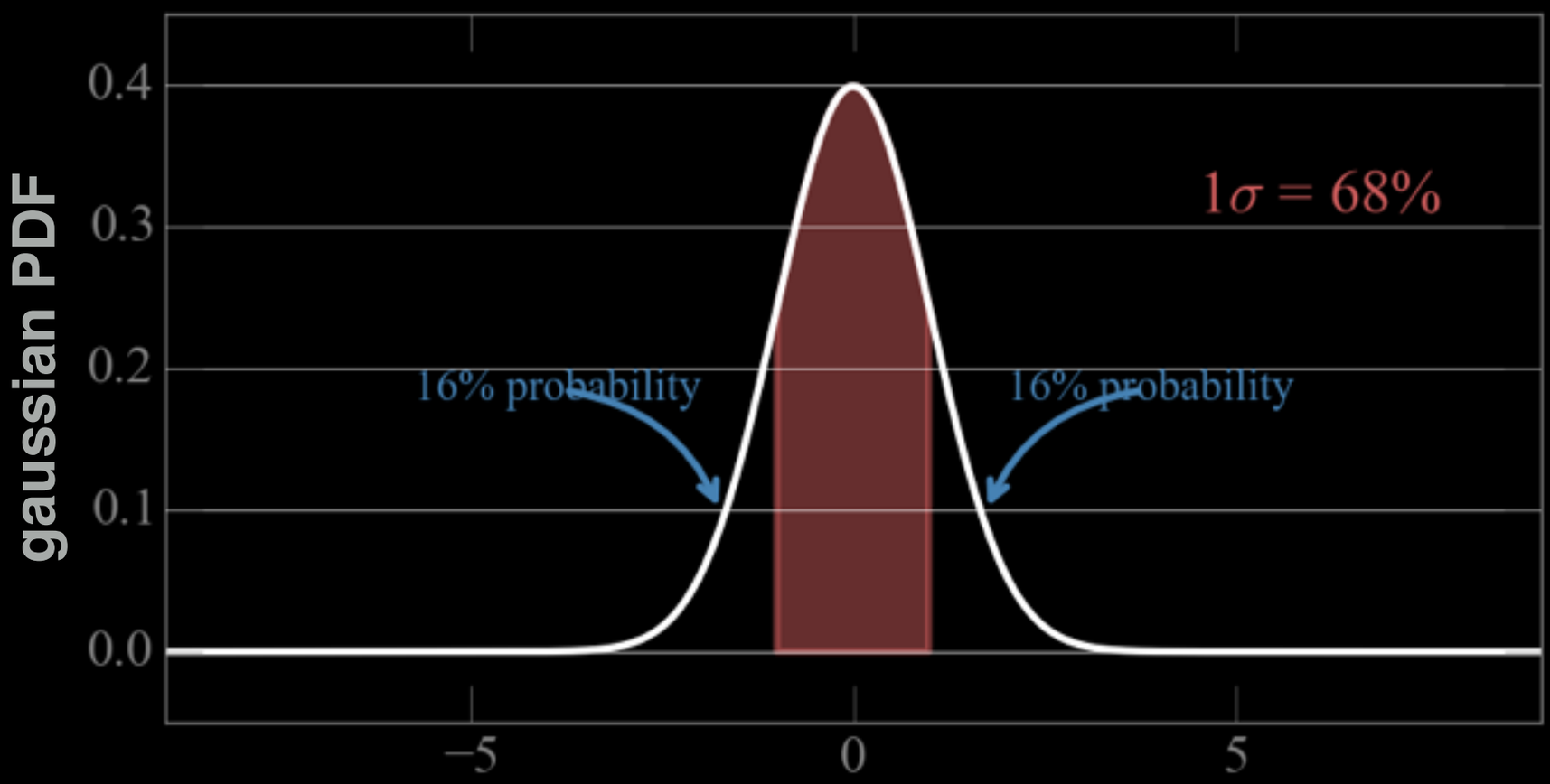
- decide what the significance threshold is: typically 5%
  α=0.05
- find the probability *p* your measurement for your test H_a



p-value

given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

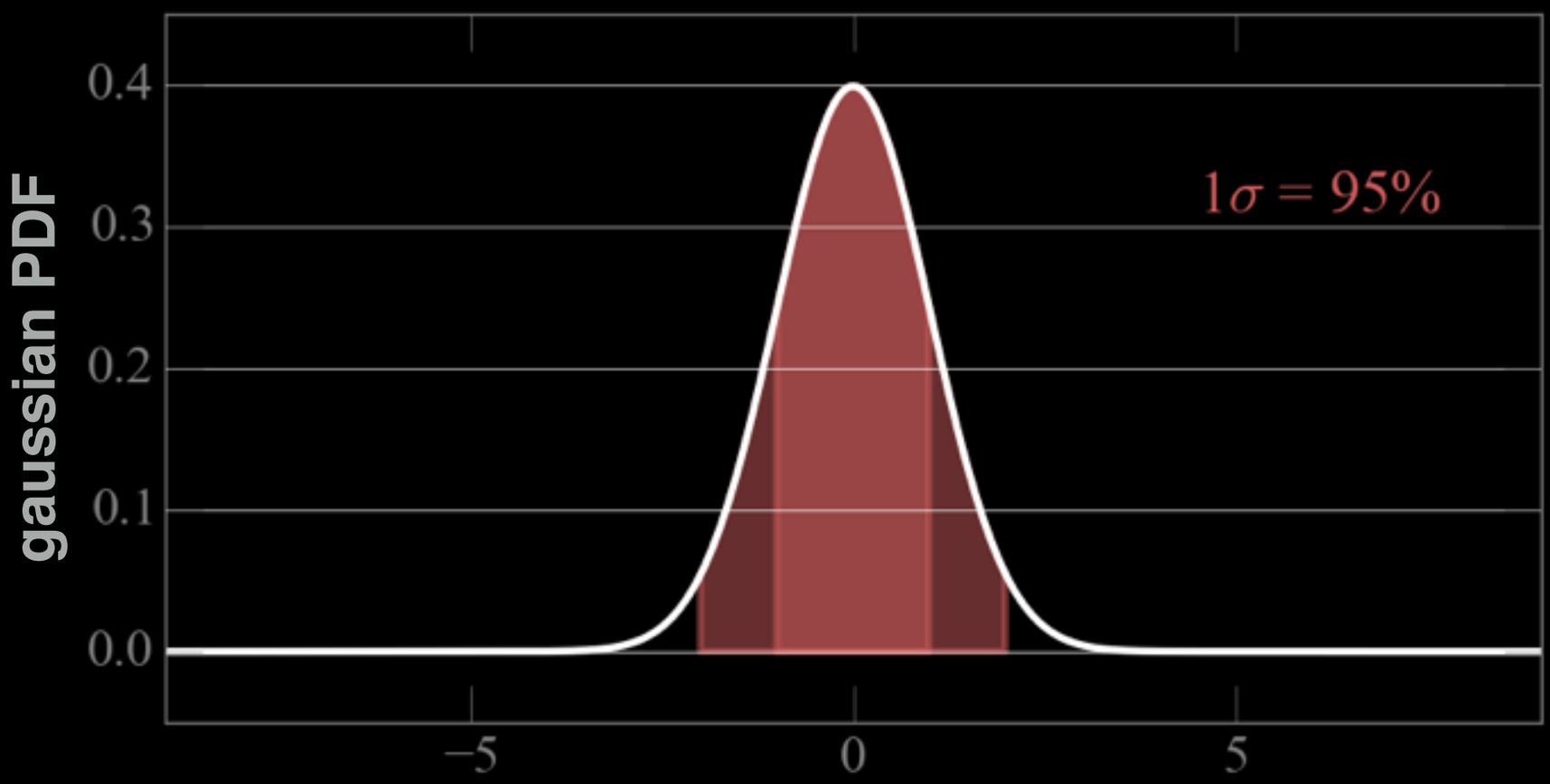
- decide what the significance threshold is: typically 5%
  α=0.05
- find the probability *p* your measurement for your test H_a



p-value

given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

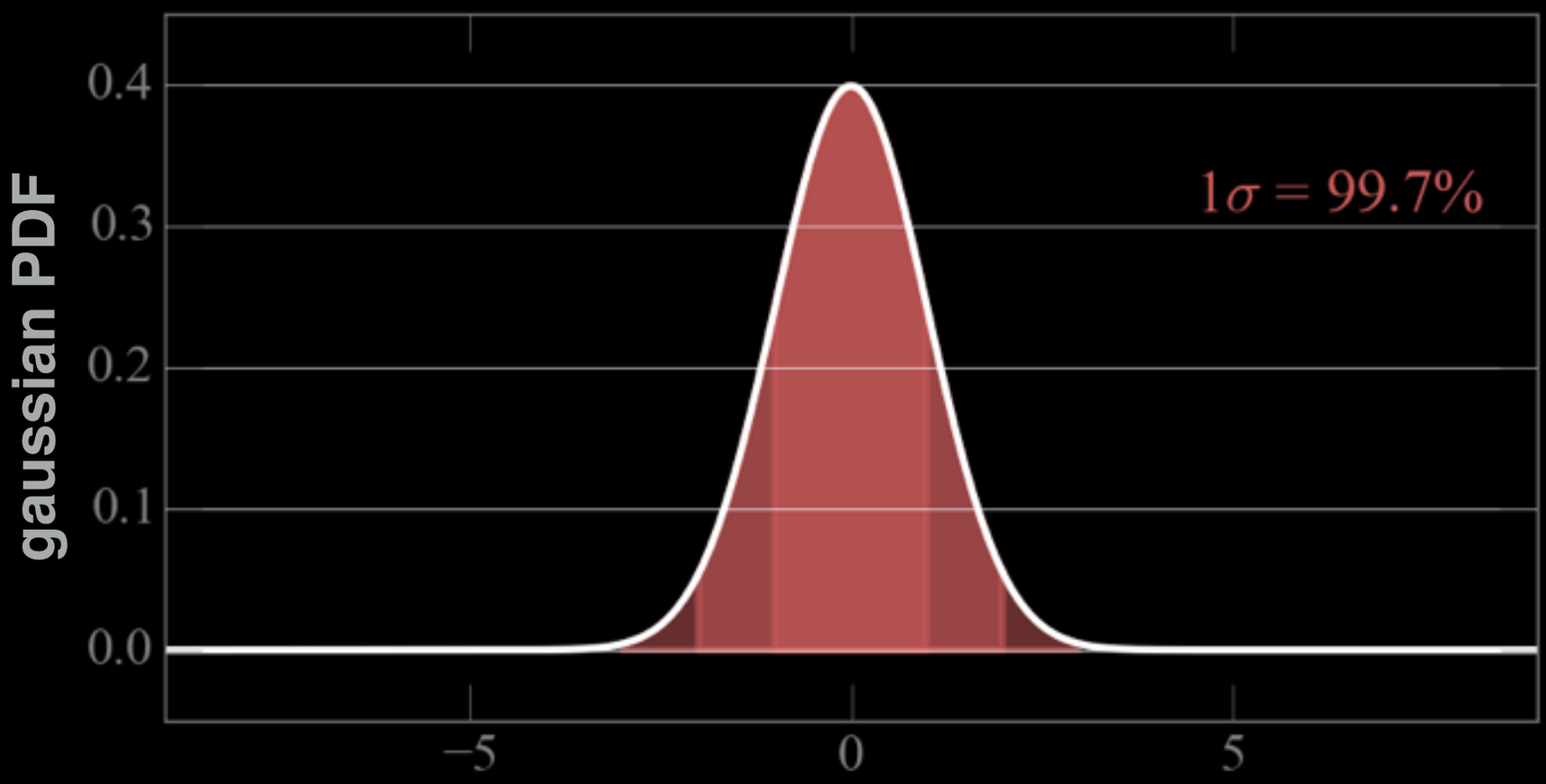
- decide what the significance threshold is: typically 5%
  α=0.05
- find the probability *p* your measurement for your test H_a



p-value

given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

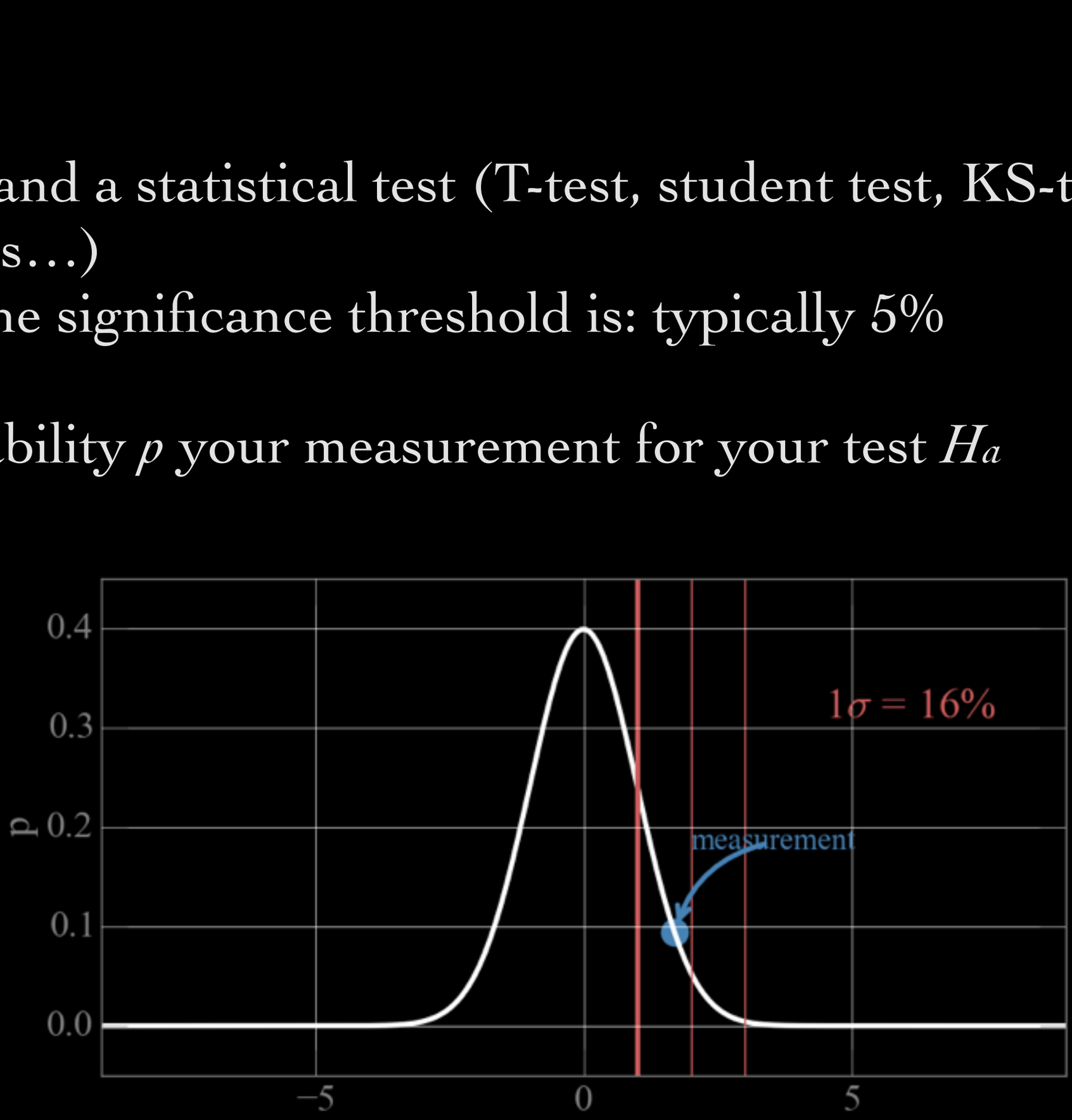
- decide what the significance threshold is: typically 5%
  α=0.05
- find the probability *p* your measurement for your test H_a



p-value

given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

- decide what the significance threshold is: typically 5%
  α=0.05
- find the probability *p* your measurement for your test H_a

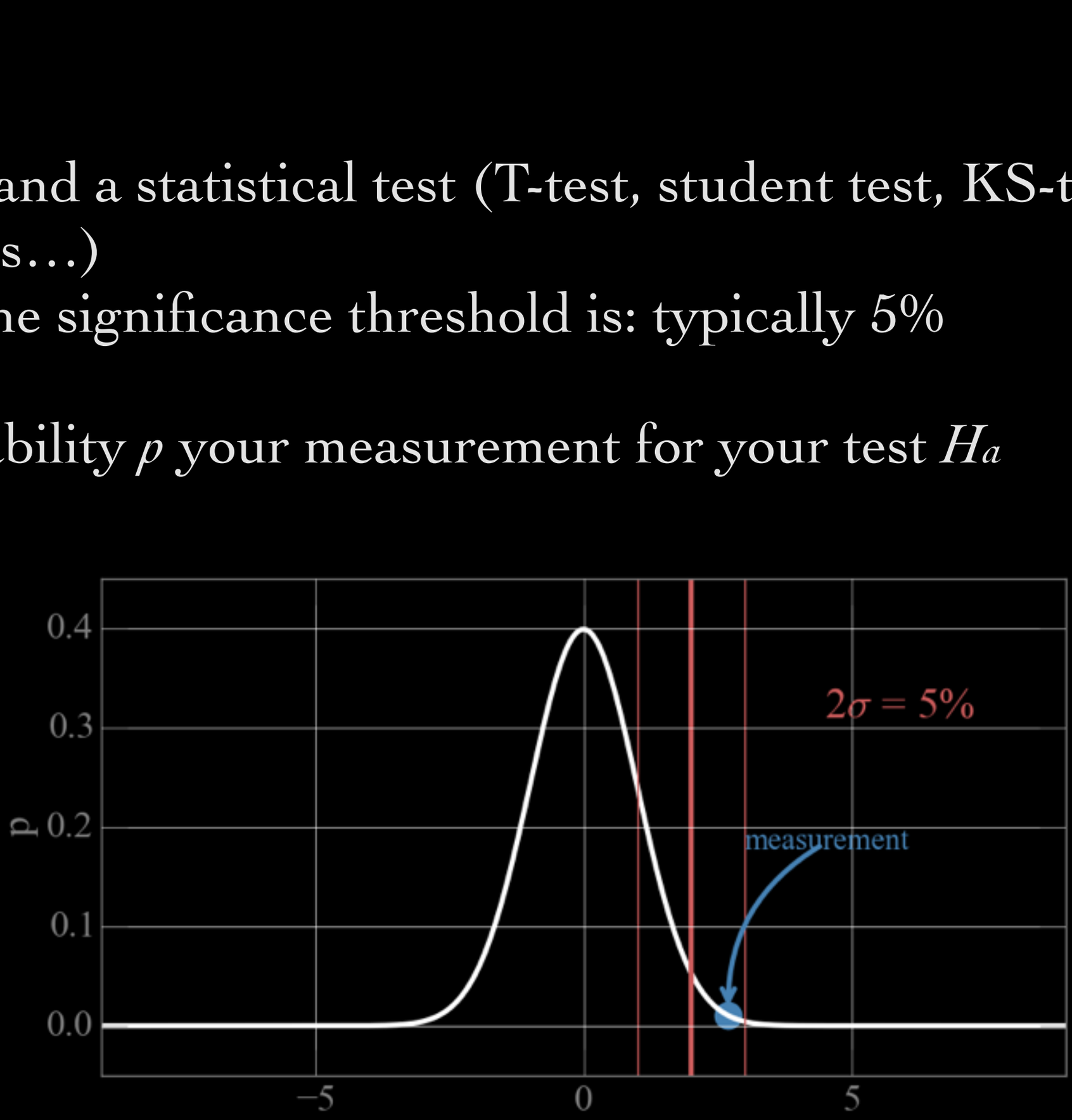


measurable effect

p-value

given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

- decide what the significance threshold is: typically 5%
  α=0.05
- find the probability *p* your measurement for your test H_a

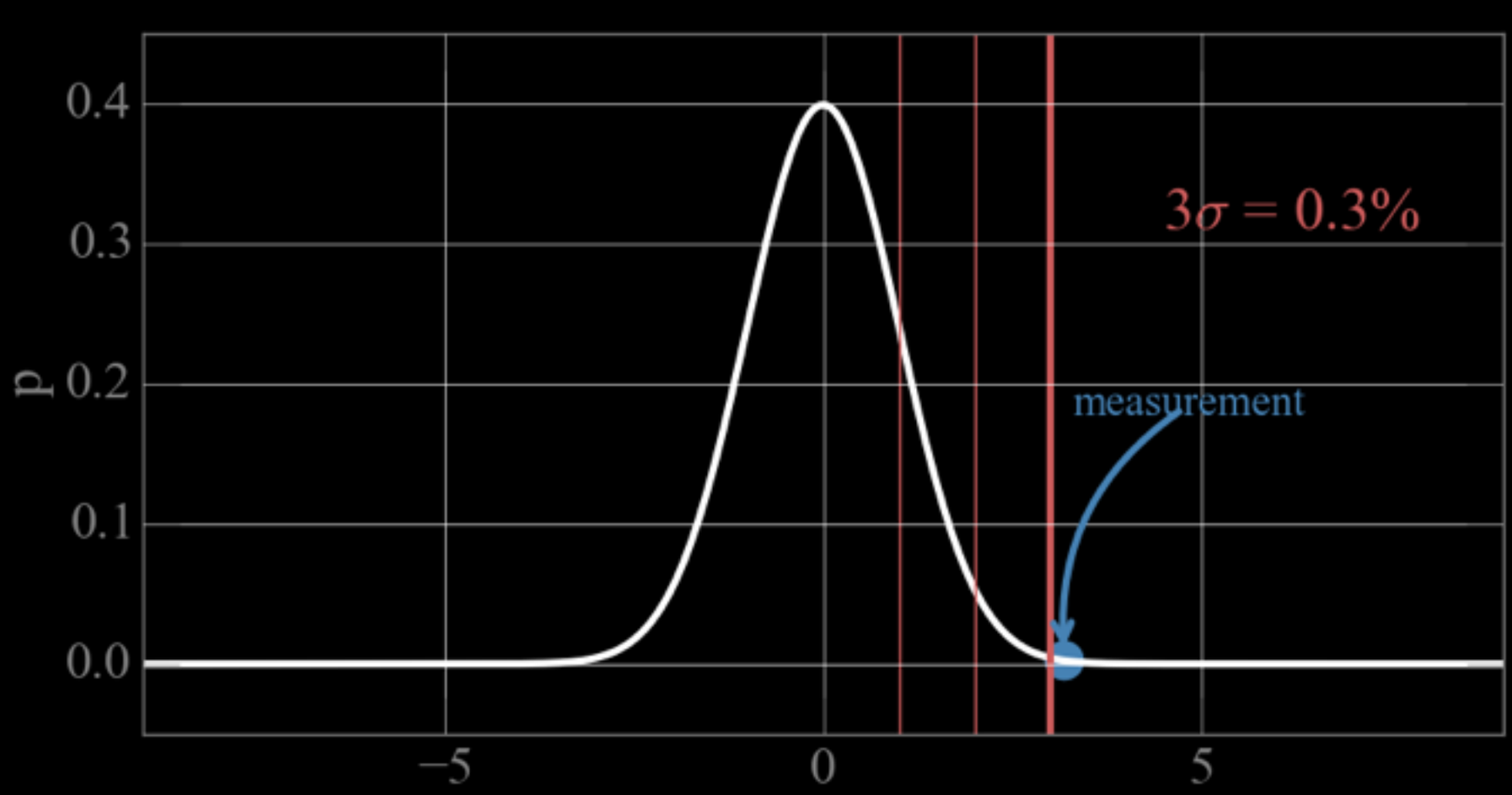


measurable effect

p-value

given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

- decide what the significance threshold is: typically 5%
  α=0.05
- find the probability *p* your measurement for your test H_a

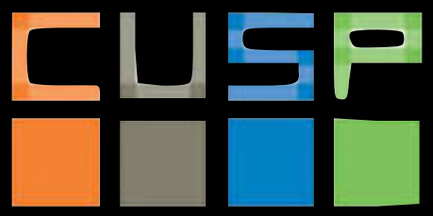


measurable effect

p-value

given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

- decide what the significance threshold is: typically 5%
  α=0.05
- find the probability *p* your measurement under H_a : $p(\mathrm{m} \mid H_a)$
- if $p(H_a)$-$p(H_0)$ > α the null hypothesis H_0 is falsified at the
  1-α confidence level

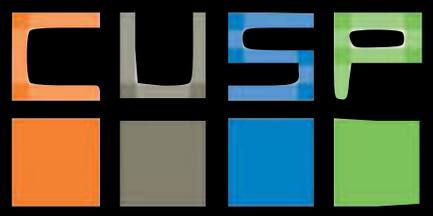


p-value

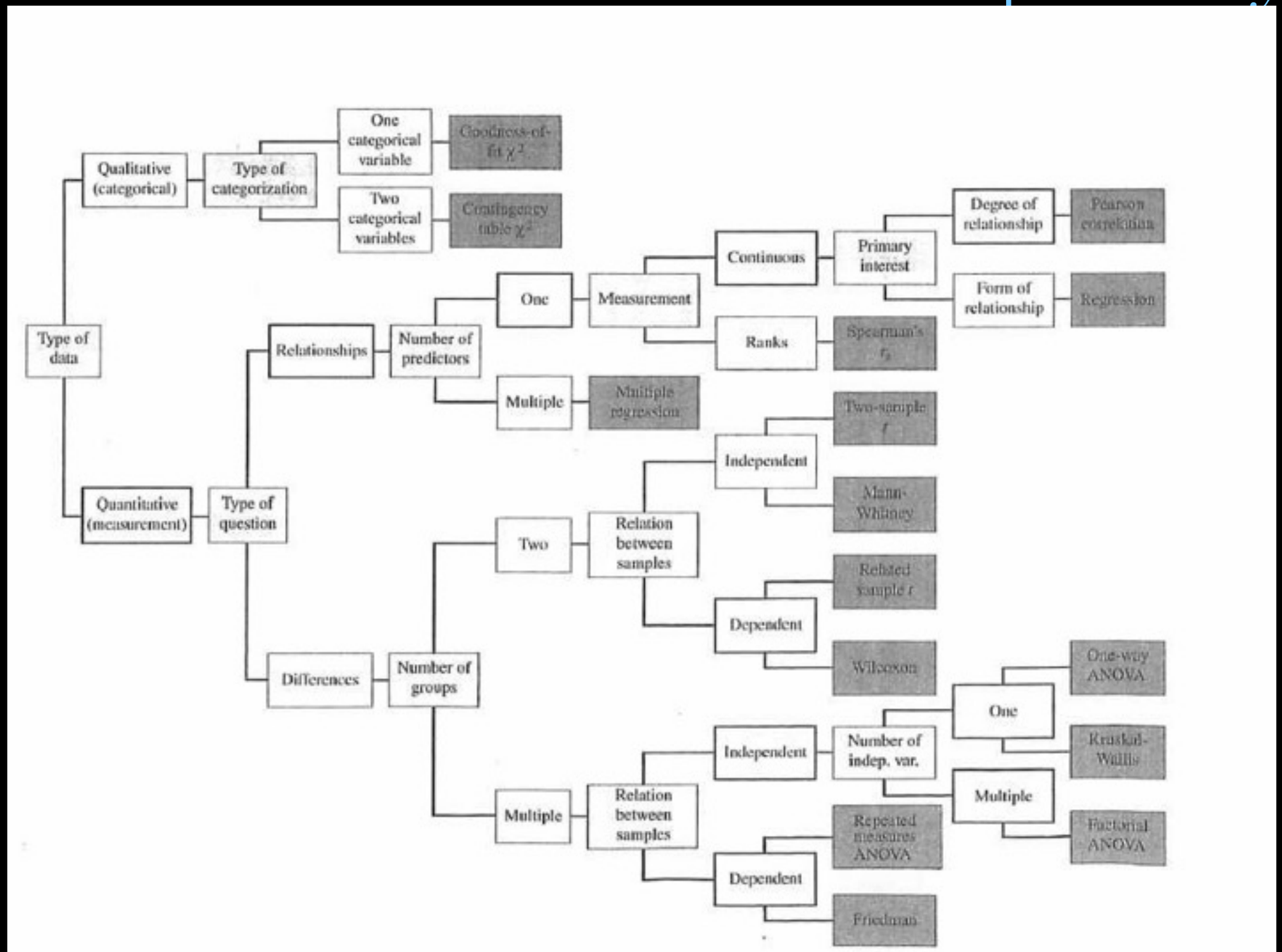
given a sample, and a statistical test (T-test, student test, KS-test, bayesian analysis…)

- decide what the significance threshold is: typically 5%
  α=0.05
- find the probability *p* your measurement under H_a : $p(\mathrm{m} \mid H_a)$
- if $p(H_a)$-$p(H_0)$ > α the null hypothesis H_0 is falsified at the
  1-α confidence level

The P-value is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true.



which is the correct statistical test?? it depends on your data!

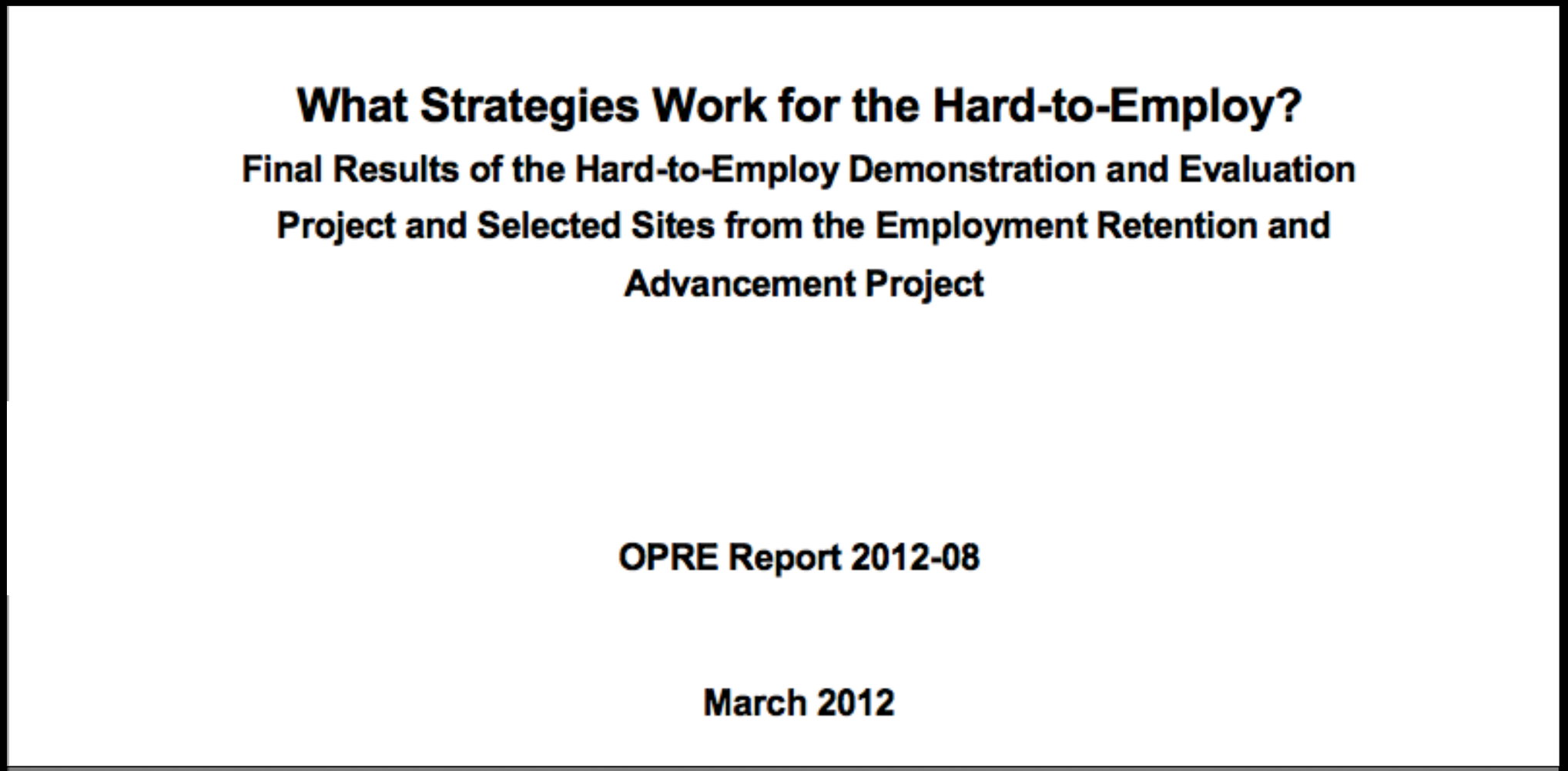


Chapter 5 of Statistics in a Nutshell
(photocopies in the library)

III: Introduction to statistics

***NULL* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,

significance level p=0.05



What Strategies Work for the Hard-to-Employ?

Final Results of the Hard-to-Employ Demonstration and Evaluation Project and Selected Sites from the Employment Retention and Advancement Project

OPRE Report 2012-08

March 2012

http://www.mdrc.org/sites/default/files/What%20Strategies%20Work%20for%20the%20Hard%20FR.pdf

***NULL* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,

significance level p=0.05

The Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project

Table 2.1

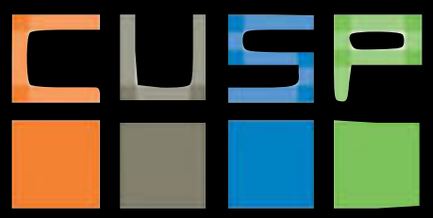
Summary of Impacts, New York City Center for Employment Opportunities

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Employment (Years 1-3) (%)				
Ever employed	83.8	70.4	13.4 ***	0.000
Ever employed in a CEO transitional job[a]	70.1	3.5	66.6 ***	0.000
Ever employed in an unsubsidized job	63.7	69.0	-5.3 *	0.078
Postprogram unsubsidized employment (Years 2-3)				
Ever employed in an unsubsidized job (%)	53.3	52.1	1.2	0.713
Employed in an unsubsidized job, average per quarter (%)	28.2	27.2	1.1	0.618
Employed for six or more consecutive quarters (%)	14.7	11.9	2.8	0.195
Total UI-covered earnings[b] ($)	10,435	9,846	589	0.658
Sample size (total = 973)[c]	564	409		

SOURCES: MDRC earnings calculations from the National Directory of New Hires (NDNH) database and employment calculations from the unemployment insurance (UI) wage records from New York State, MDRC calculations using data from the New York State Division of Criminal Justice Services (DCJS) and the New York City Department of Correction (DOC).

NOTES: Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.
The p-value indicates the likelihood that the difference between the program and control groups arose by chance.

http://www.mdrc.org/sites/default/files/What%20Strategies%20Work%20for%20the%20Hard%20FR.pdf



***NULL* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,

significance level p=0.05

The Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project

Table 2.1

Summary of Impacts, New York City Center for Employment Opportunities

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Employment (Years 1-3) (%)	P_1	P_0		
Ever employed	83.8	70.4	13.4 ***	0.000
Ever employed in a CEO transitional job[a]	70.1	3.5	66.6 ***	0.000
Ever employed in an unsubsidized job	63.7	69.0	-5.3 *	0.078
Postprogram unsubsidized employment (Years 2-3)				
Ever employed in an unsubsidized job (%)	53.3	52.1	1.2	0.713
Employed in an unsubsidized job, average per quarter (%)	28.2	27.2	1.1	0.618
Employed for six or more consecutive quarters (%)	14.7	11.9	2.8	0.195
Total UI-covered earnings[b] ($)	10,435	9,846	589	0.658
Sample size (total = 973)[c]	564	409		

SOURCES: MDRC earnings calculations from the National Directory of New Hires (NDNH) database and employment calculations from the unemployment insurance (UI) wage records from New York State, MDRC calculations using data from the New York State Division of Criminal Justice Services (DCJS) and the New York City Department of Correction (DOC).

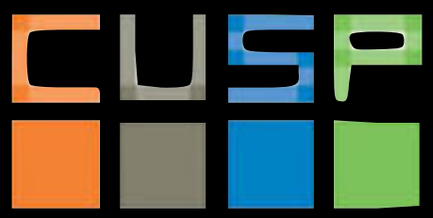
NOTES: Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.
The p-value indicates the likelihood that the difference between the program and control groups arose by chance.

H_0: $P_0 - P_1 > 0$

H_a: $P_0 - P_1 <= 0$

$\alpha=0.05$

http://www.mdrc.org/sites/default/files/What%20Strategies%20Work%20for%20the%20Hard%20FR.pdf



***NULL* HYPOTHESIS**: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,

significance level p=0.05

The Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project

Table 2.1

Summary of Impacts, New York City Center for Employment Opportunities

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Employment (Years 1-3) (%)	P_1	P_0		
Ever employed	83.8	70.4	13.4 ***	0.000
Ever employed in a CEO transitional job[a]	70.1	3.5	66.6 ***	0.000
Ever employed in an unsubsidized job	63.7	69.0	-5.3 *	0.078
Postprogram unsubsidized employment (Years 2-3)				
Ever employed in an unsubsidized job (%)	53.3	52.1	1.2	0.713
Employed in an unsubsidized job, average per quarter (%)	28.2	27.2	1.1	0.618
Employed for six or more consecutive quarters (%)	14.7	11.9	2.8	0.195
Total UI-covered earnings[b] ($)	10,435	9,846	589	0.658
Sample size (total = 973)[c]	564	409		

H_0: $P_0 - P_1 > 0$

Ha: $P_0 - P_1 <= 0$

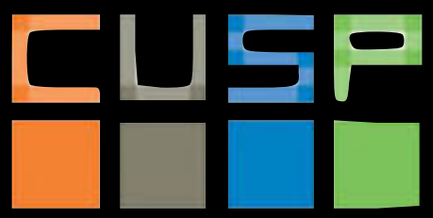
$\alpha=0.05$

https://github.com/fedhere/PUI2015_fbianco/blob/master/HW3/effectivenes%20of%20NYC%20Post-Prison%20Employment%20Programs.ipynb

HOMEWORK:

Reading: introduction to Data Jujitsu (DJ Patil)
http://www.oreilly.com/data/free/data-jujitsu.csp
Sections:

- Introduction
- No data vomit



distributions, moments, and Central Limit Theorem

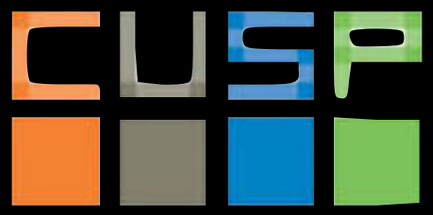
assignment 1 :

1. GENERATE 100 samples of different sizes N (N>10 & N<2000) from each of 6 different distributions (500 samples in total), all with the same *population* mean. Include a Normal, a Poisson, a Binomial, a Chi-Squared distribution, and 2 more of your choice.
2. For each sample plot the sample mean against the sample size N (if you want you can do it with the sample standard deviation as well). Can you describe the behavior you see in the plots?
3. PLOT the distributions of all sample means (together for all distributions). Mandatory: as a histogram,
   optional: in any other way you think is convincing
4. optional: FIT a gaussian to the distribution of means

e.g. how to fit function to data in numpy:

http://glowingpython.blogspot.com/2012/07/distribution-fitting-with-scipy.html

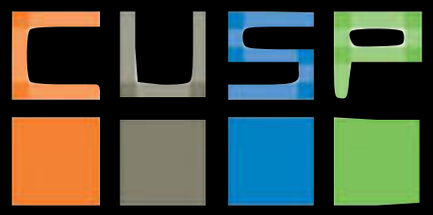
http://stackoverflow.com/questions/7805552/fitting-a-histogram-with-python



distributions, moments, and Central Limit Theorem

assignment 1 :

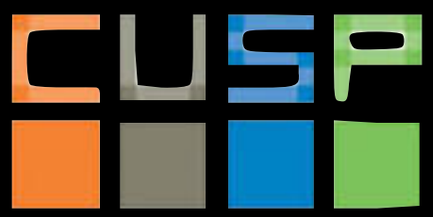
- create a new Github directory called HW3 inside of your PUI2015_<your name> repo.
- upload an ipython notebook, with the rendered plots.
- include a README.md which describes what you are doing, and, if appropriate, how to run the notebook (input variables? global variables that need to be setup?).
- 75% of the grade will be based on the rendered version of the plot, 25% will be awarded if the TA can download and run the notebook. If you include any package that was not in the standard Anaconda distribution state that in your README.md, so that the TA can download them.



distributions, moments, and Central Limit Theorem

assignment 2 : Z-test and chi sq test

- Fill in missing cells in (your own copy of) https://github.com/fedhere/PUI2015_fbianco/blob/master/HW3/effectivenes%20of%20NYC%20Post-Prison%20Employment%20Programs.ipynb
- turn in the python notebook in the HW3 directory (see assignment 1)



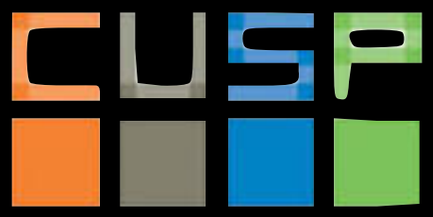
HYPOTHESIS TESTING HOMEWORK:

assignment 3 (best to work in groups! but 5 ppl max):

you will work on CitiBikes data to assess a proportion or a mean problem. I prepared an example here:
https://github.com/fedhere/UInotebooks/blob/master/citibikes_1950s.ipynb

you can test any breakdown, by age (older vs younger then), by gender, tourists vs locals…

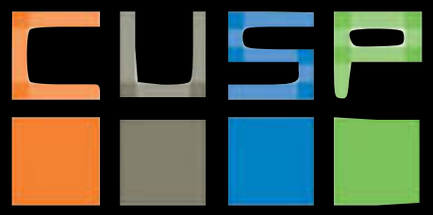
- describe your idea
- state your Null and alternative hypothesis
- choose a confidence level
- mangle your data
- choose a statistical test. Use z-score if the sample is small, while the chi square statistics if the sample is better if the sample is large.
- assess whether you can reject the Null Hypothesis



HYPOTHESIS TESTING HOMEWORK:

assignment 3 (best to work in groups! but 5 ppl max):
you will work on CitiBikes data to assess a proportion or a mean problem. I prepared an example here:
https://github.com/fedhere/UInotebooks/blob/master/citibikes_1950s.ipynb
you can test any breakdown, by age (older vs younger then), by gender, tourists vs locals…

- create a new Github directory called citibikes inside of your PUI2015_<your name> repo. We will work from your first notebook incrementally.
- make sure you include a README.md in your repo that describes the project and states your specific contribution.



MUST KNOWS:

- Null Hypothesis
- Normal (Gaussian)
- Poisson, Chi-Squared, t distribution
- Moments of a distribution
- *p*-value
- *z*-score
- chi-squared test

Resources:

Sarah Boslaugh, Dr. Paul Andrew Watters, 2008 (in the CUSP library)
Statistics in a Nutshell (Chapters 3,4,5)
https://books.google.com/books/about/Statistics_in_a_Nutshell.html?id=ZnhgO65Pyl4C

David M. Lane et al. (free online)
Introduction to Statistics (Chapter I, XI, XII)
http://onlinestatbook.com/Online_Statistics_Education.epub
http://onlinestatbook.com/2/index.html

Max Mether
The history of the central limit theorem
http://salserver.org.aalto.fi/vanhat_sivut/Opinnot/Mat-2.4108/pdf-files/emet03.pdf

William Chen & Joe Blitzstein
Probability Cheatsheet v2.0
http://alturl.com/b22bs

Various authors
Latex Wikibook
https://en.wikibooks.org/wiki/LaTeX

Buteler et al. 2012
What Strategies Work for the Hard-to-Employ?
http://www.mdrc.org/sites/default/files/What%20Strategies%20Work%20for%20the%20Hard%20FR.pdf