

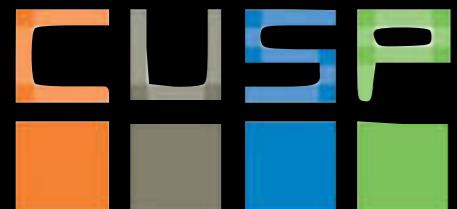
Urban Informatics

Fall 2015

dr. federica bianco fb55@nyu.edu

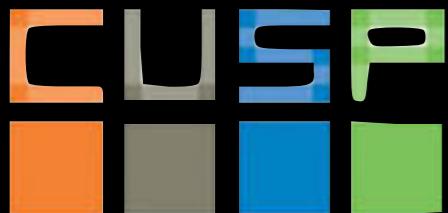


@fedhere



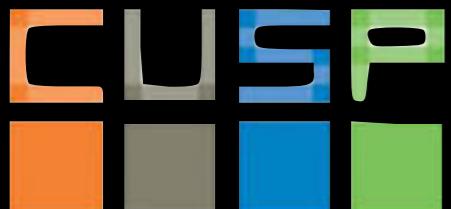
Summary:

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



Summary:

- **Epistemological concepts:**
falsifiability, law of parsimony,
- **Good scientific practice:**
reproducibility of research
- **Gathering parsing data, API:**
data munging or wrangling, data jujitsu
- **Formulating and testing a scientific Hypothesis**
Basic statistics: distributions and their moments
Hypothesis testing: p -value, statistical significance



- 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 FORMULATING HYPOTHESIS

• 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 FORMULATING HYPOTHESIS

• 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)

HYPOTHESIS TESTING

• interpretation

• presentation

- visualization
- write a paper!



FORMULATING HYPOTHESIS

idea → *Null Hypothesis*

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

CUSP seminar
Friday 9/18/16



Urban Neighborhoods
and the **End of Progress**
toward Racial Equality

PATRICK SHARKEY

Copyrighted Material

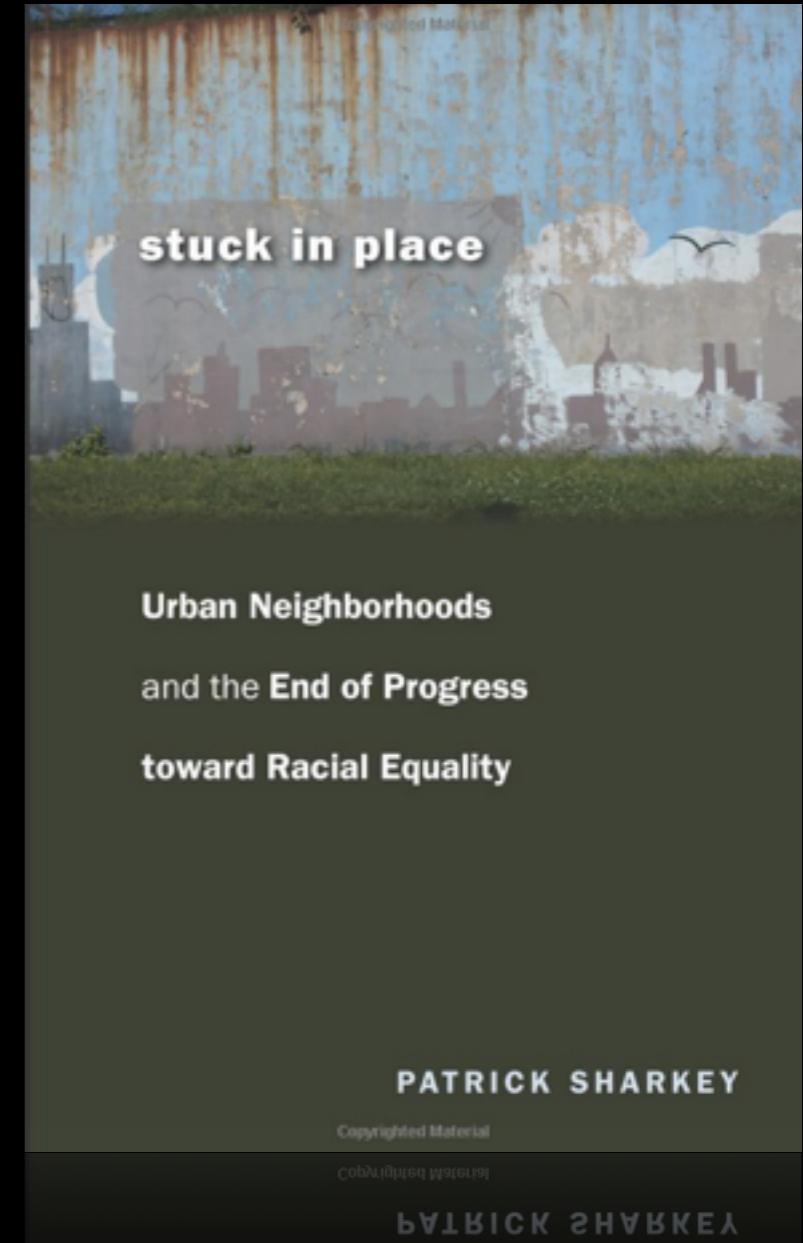
ISBN 978-0393249590

PATRICK SHARKEY

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

e.g.:

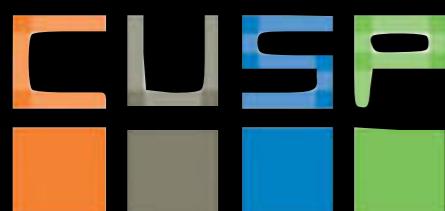
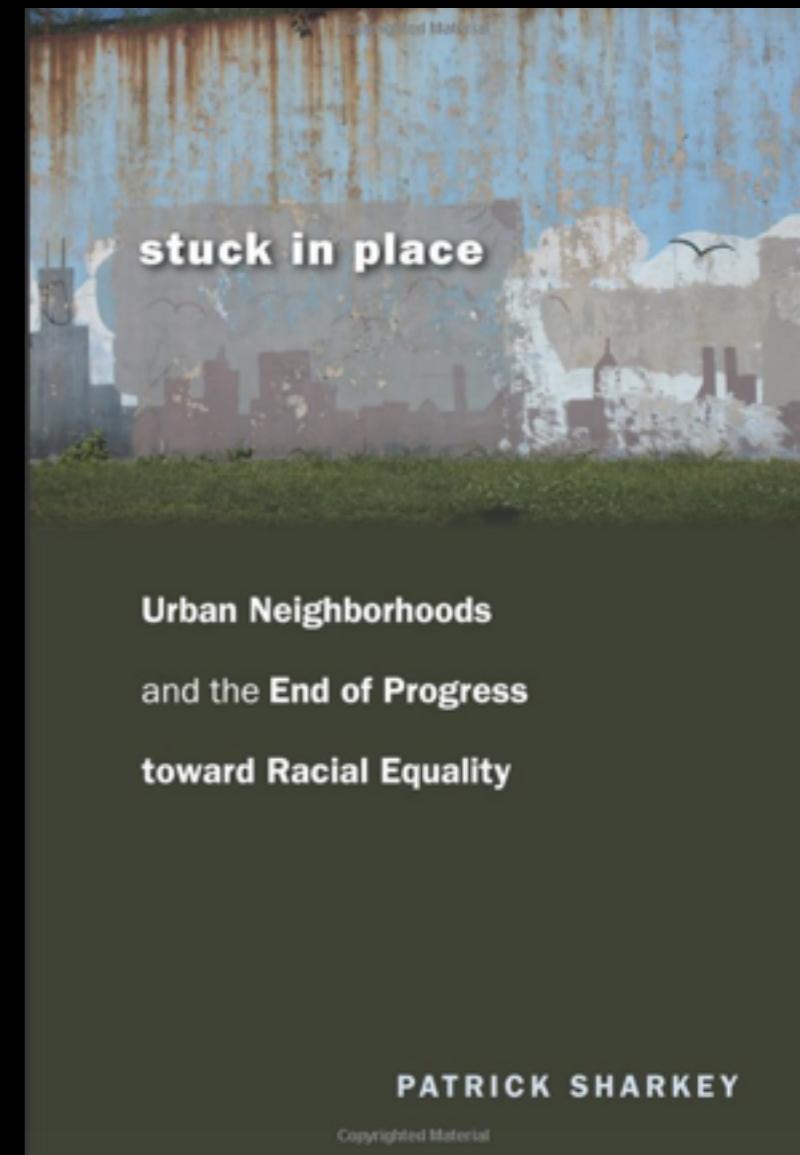
QUESTION: does proximity to violence affect children's development ?



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

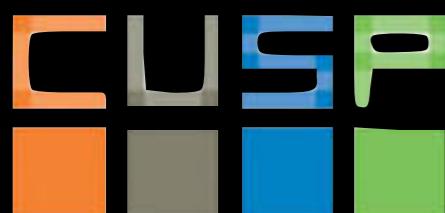
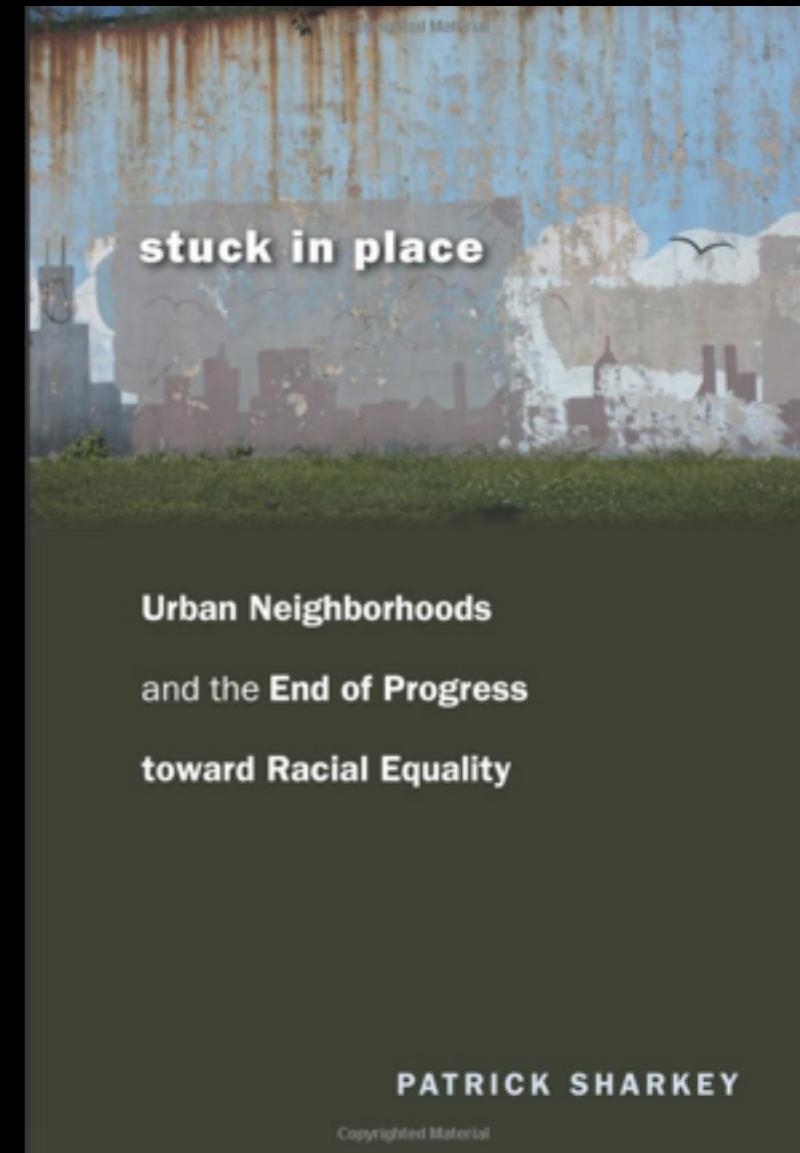


1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis

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

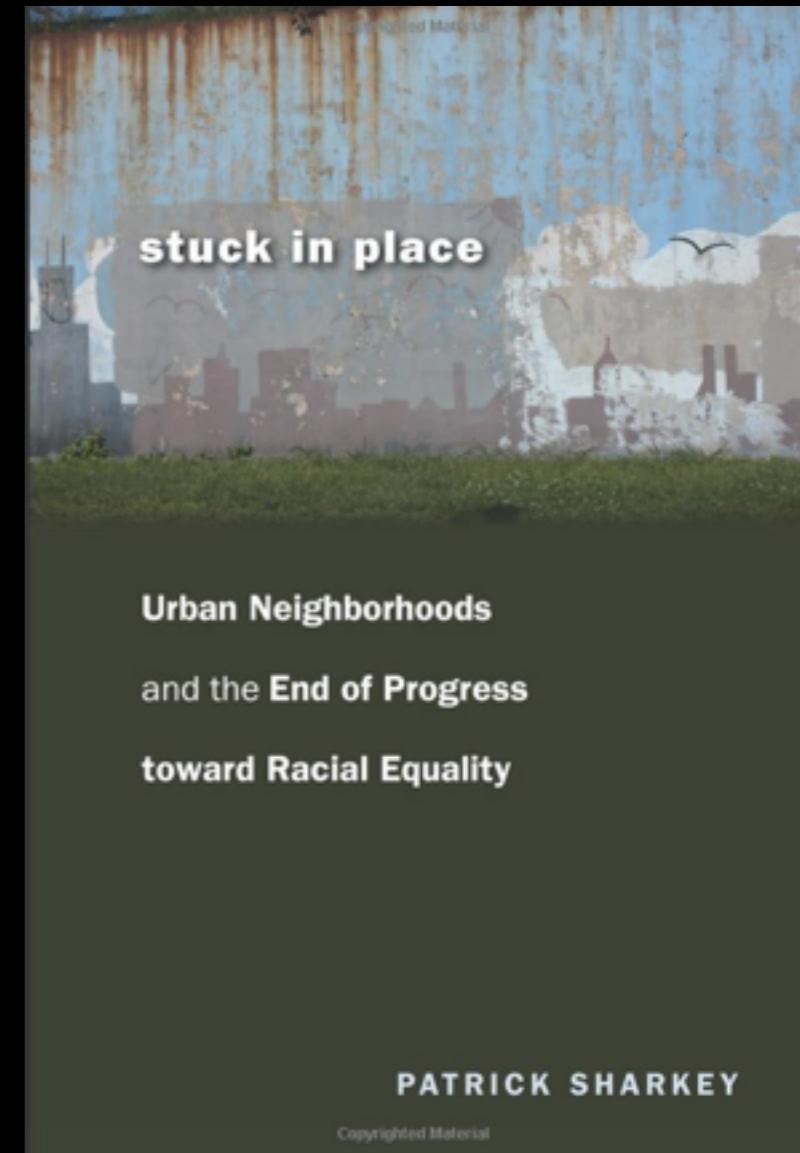
TESTABLE HYPOTHESIS: the average test score of children who live within a block of the site of a violent crime is significantly lower in the days following the crime



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



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

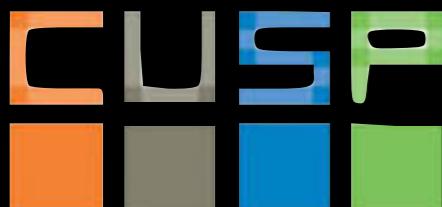
IDEA: does the NYC Post-Prison Employment Programs increase employment?

**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

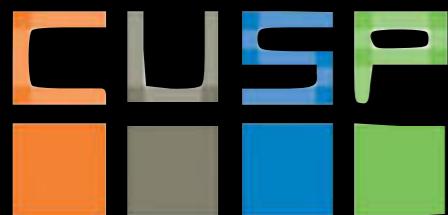
<http://cosmo.nyu.edu/~fb55/PUI2016/NullHypothesisSubmit.html>



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

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

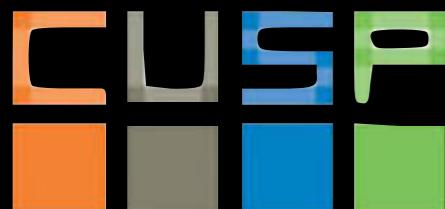


1. develop a hypothesis that can be tested mathematically & state the *Null hypothesis* and alternative hypothesis

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

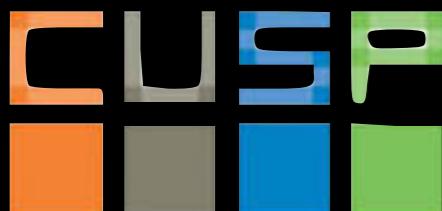
TESTABLE HYPOTHESIS: the % of former prisoners employed 3 years after release is *significantly* higher for candidates who participated in the program



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 $p=0.05$*

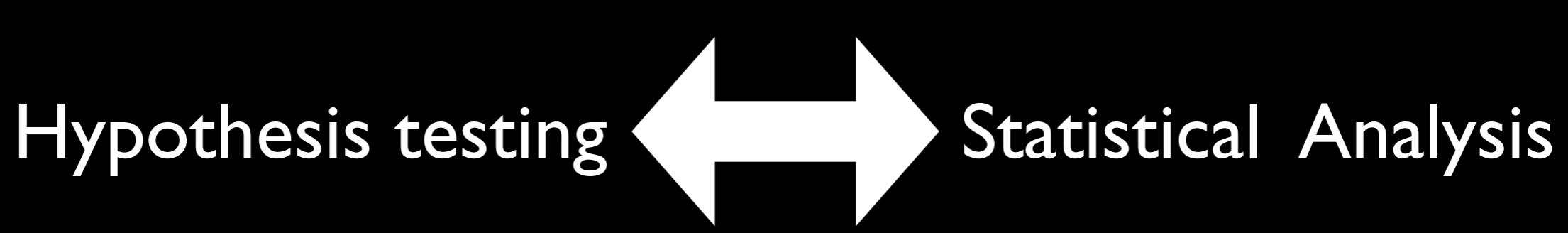
ALTERNATIVE HYPOTHESIS: the % of former prisoners employed 3 years after release is *significantly higher* for candidates who participated



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



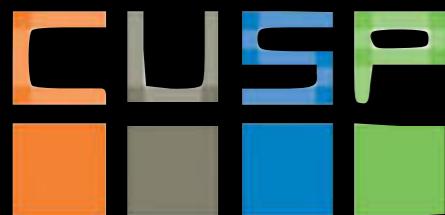
HYPOTHESIS TESTING



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?



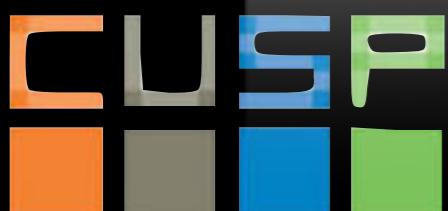
NULL HYPOTHESIS: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,
significance level p=0.05

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

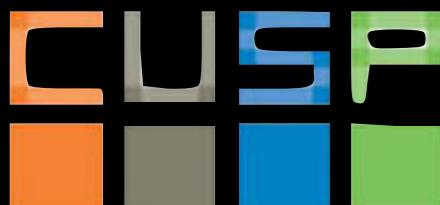
Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Employment (Years 1-3) (%)				
Ever employed	83.8	70.4	13.4 ***	0.000
Ever employed in a CEO transitional job ^a	70.1	3.5	66.6 ***	0.000
Ever employed in an unsubsidized job	63.7	69.0	-5.3 *	0.078
Postprogram unsubsidized employment (Years 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		

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

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.



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) (%)	P₁	P₀		
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		

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

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).

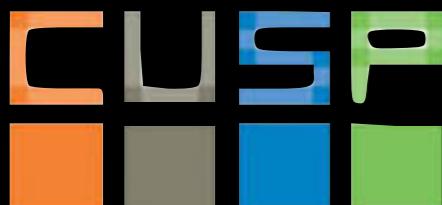
NOTES: Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

The p-value indicates the likelihood that the difference between the program and control groups arose by chance.

$$H_0: P_0 - P_1 \geq 0$$

$$H_a: P_0 - P_1 < 0$$

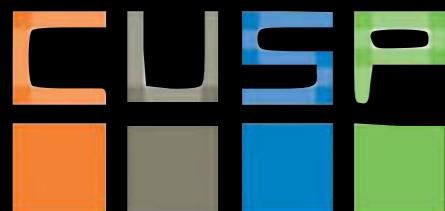
$$\alpha=0.05$$



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?

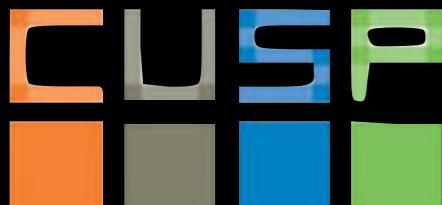


hypothesis testing

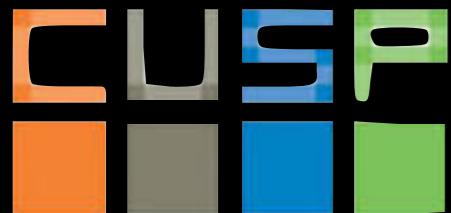
what is the probability that we would have gotten the same result out of just noise?

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?

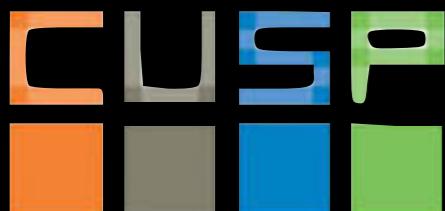


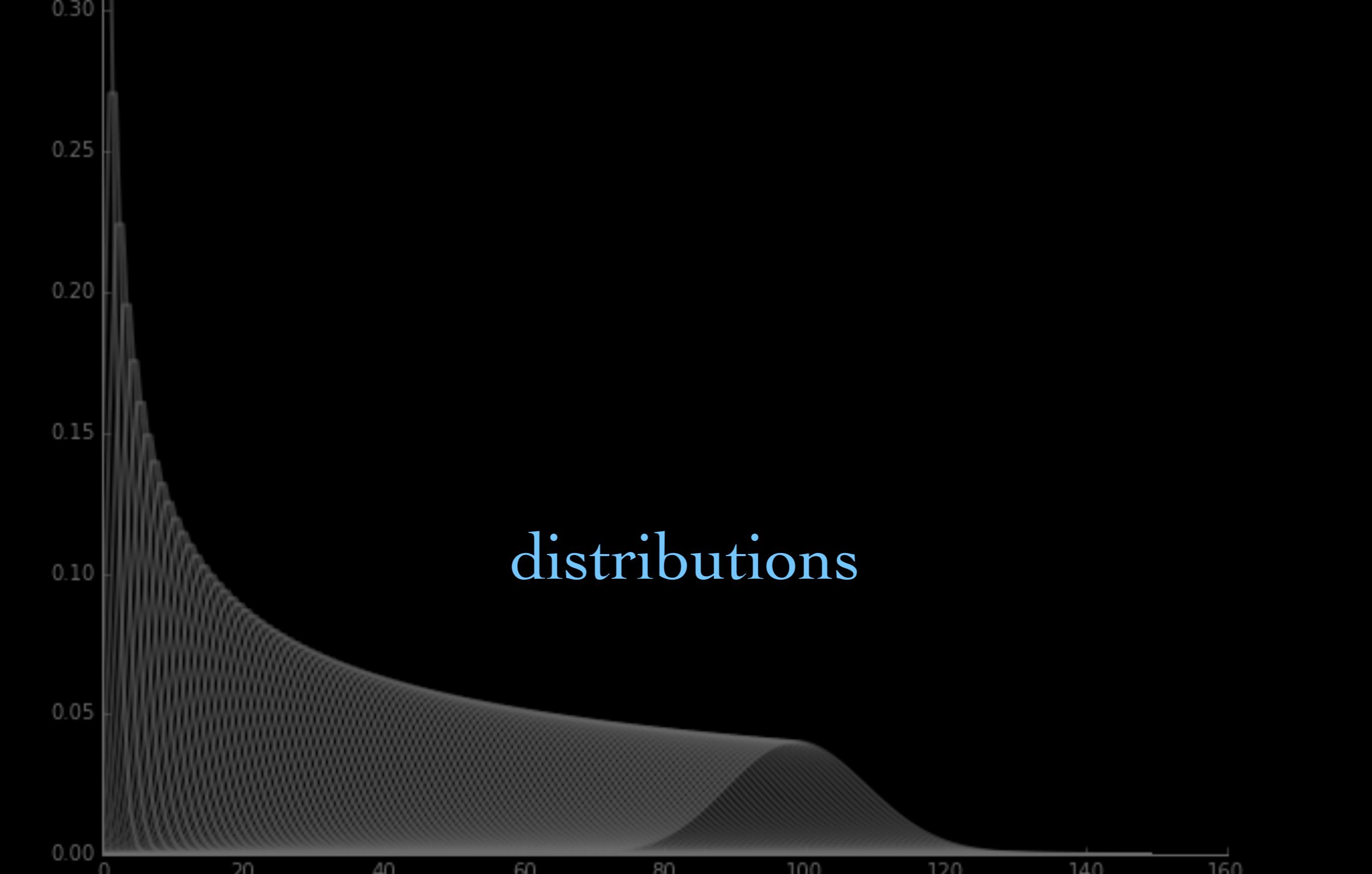
is this effect larger than the just by chance?



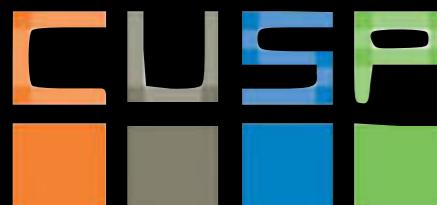
<i>reality</i> → <i>result of analysis</i> ↓	H_0 is True	H_0 is False
H_0 is falsified	Type I error False Positive	True Positive
H_0 is not falsified	True Negative	Type II error False negative

	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





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.

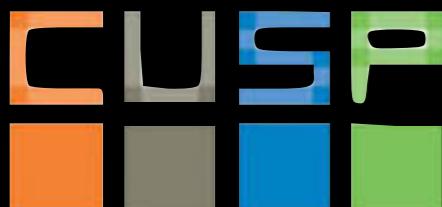
I recently went through this myself and thought I'd record my thoughts on the process while they're still fresh. I interviewed a lot. Some went well, some didn't go well at all. The reason for this was sometimes me, sometimes them, often both. Sometimes I didn't get selected for an on-site interview. 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.

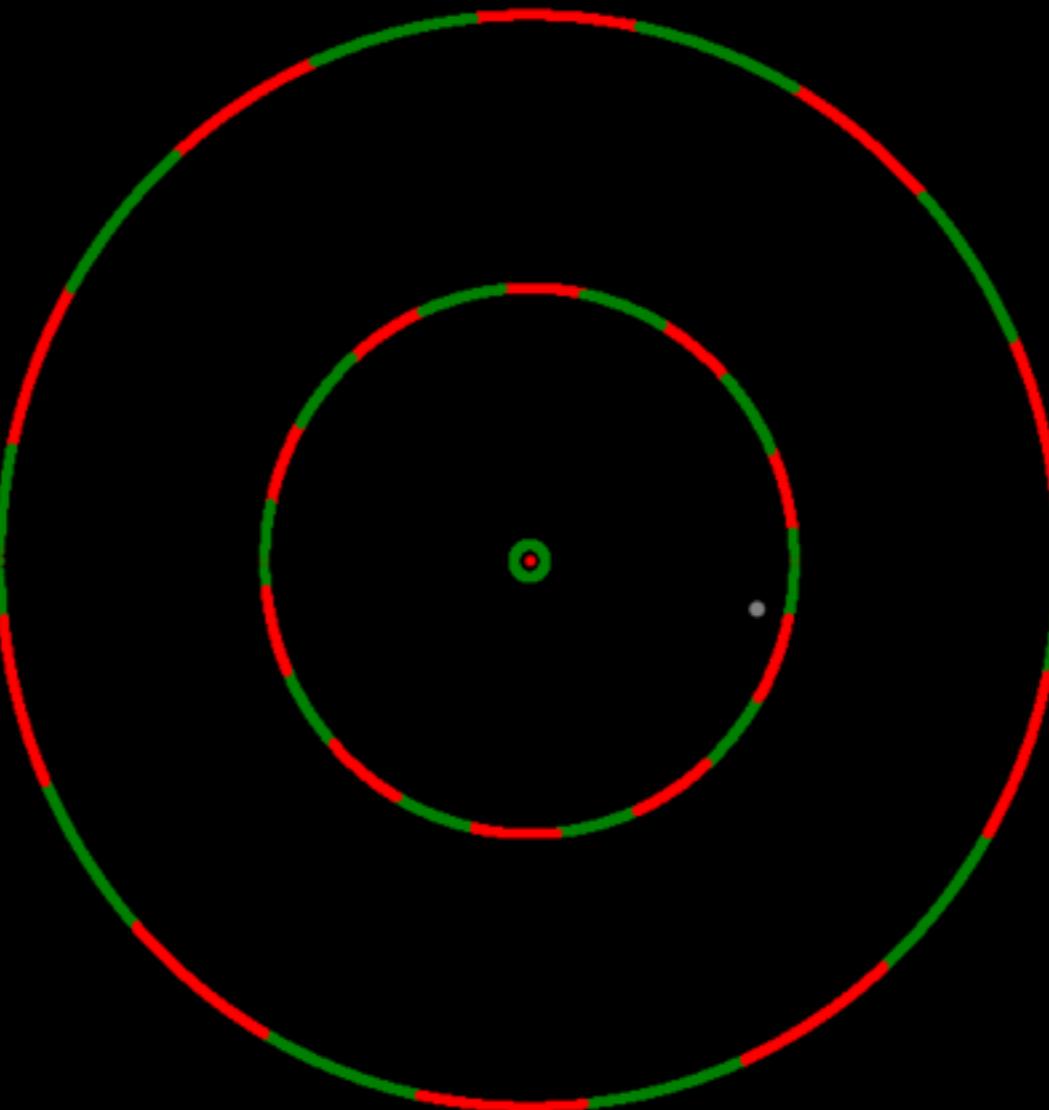
Warning: What follows are my personal thoughts, extrapolating from a small sample, and generalization from anecdotes. Precisely the kind of thing that data scientists hate! But, despite the frequent misquotation, the plural of anecdote is data, so this discussion should start somewhere.

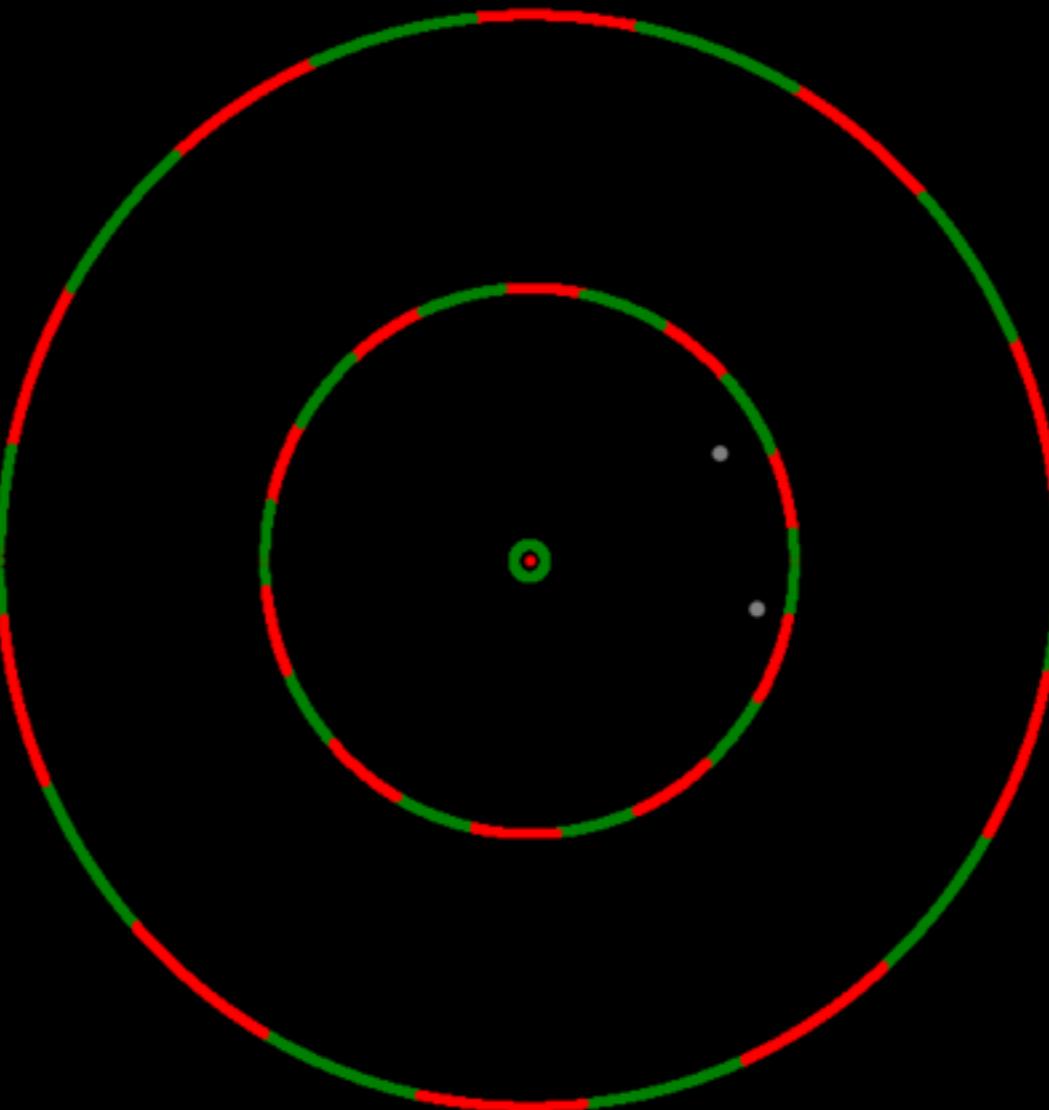
Interview questions: What to expect

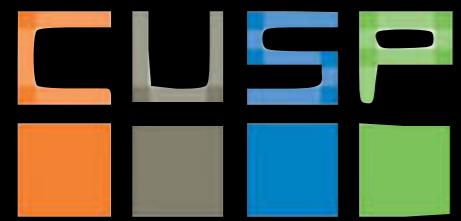
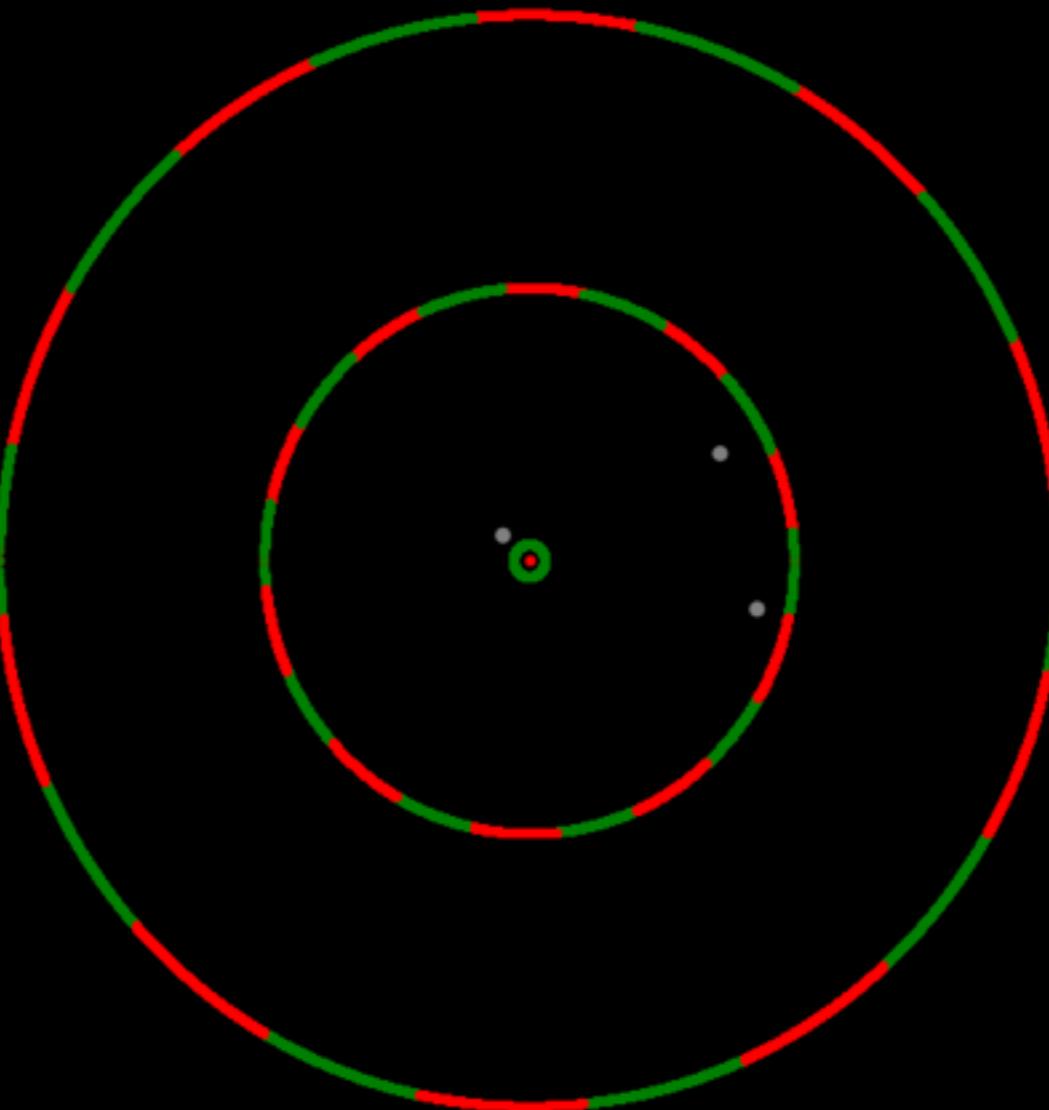
Prepare to be asked terrible questions. Some of the worst questions I've been asked:

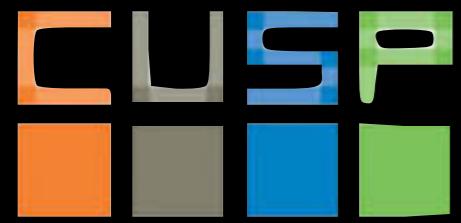
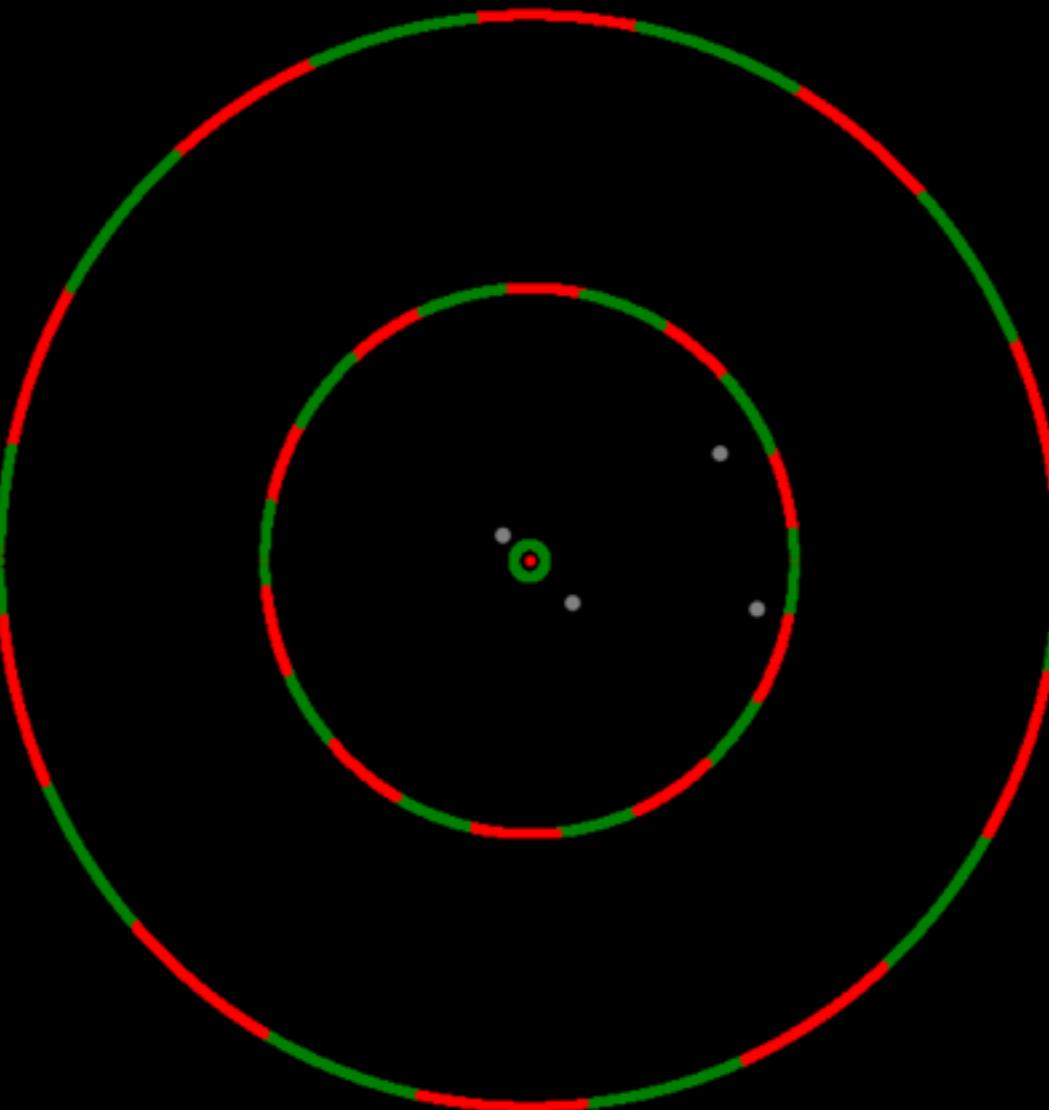
- I have a random number generator. What number does it produce and why?
- Anything involving dice or urns
- Can you tell me *certain piece of trivia* about *parameter rho* from *distribution delta*? (Names changed to protect the innocent PDFs)
- Here's a problem it took my team six months to solve.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

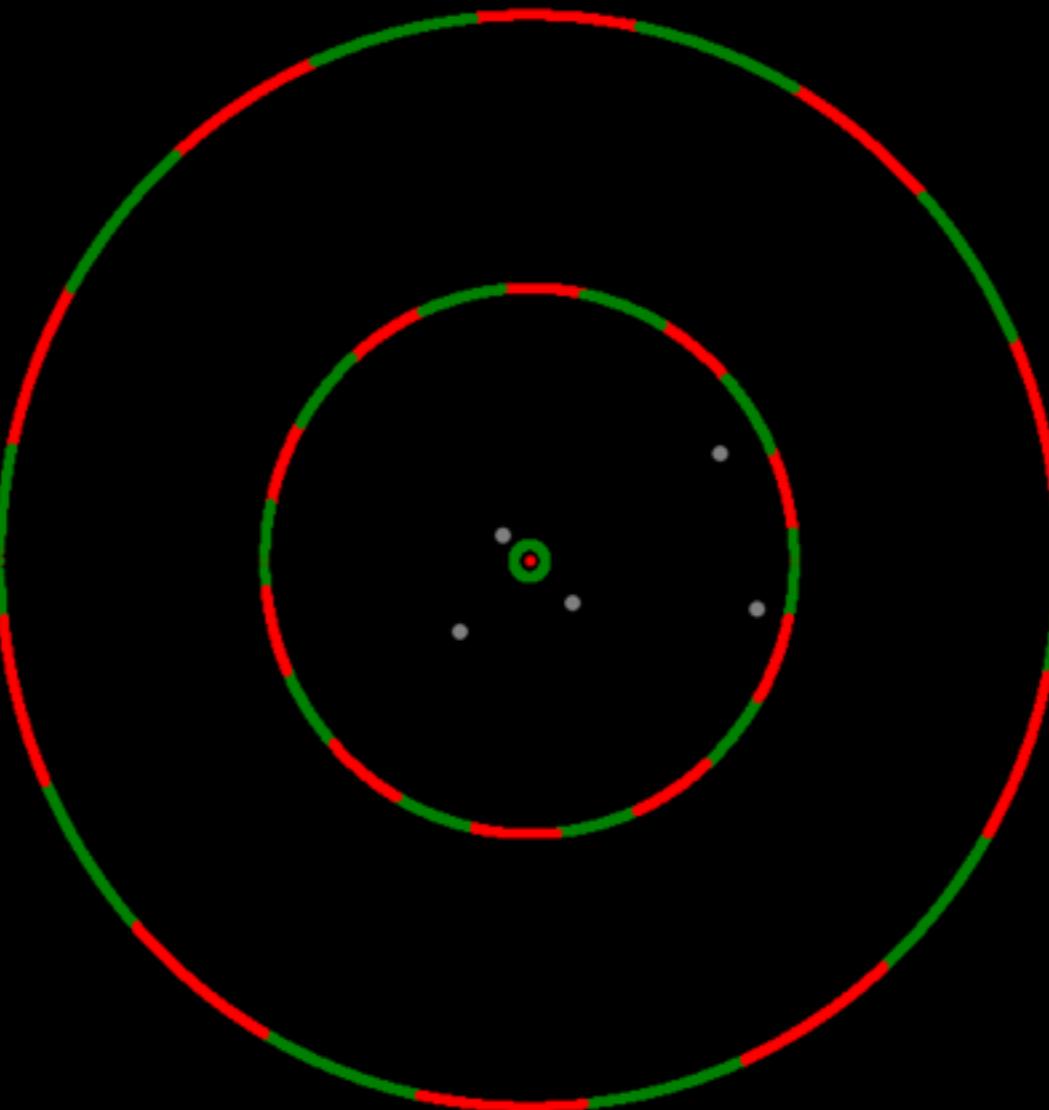


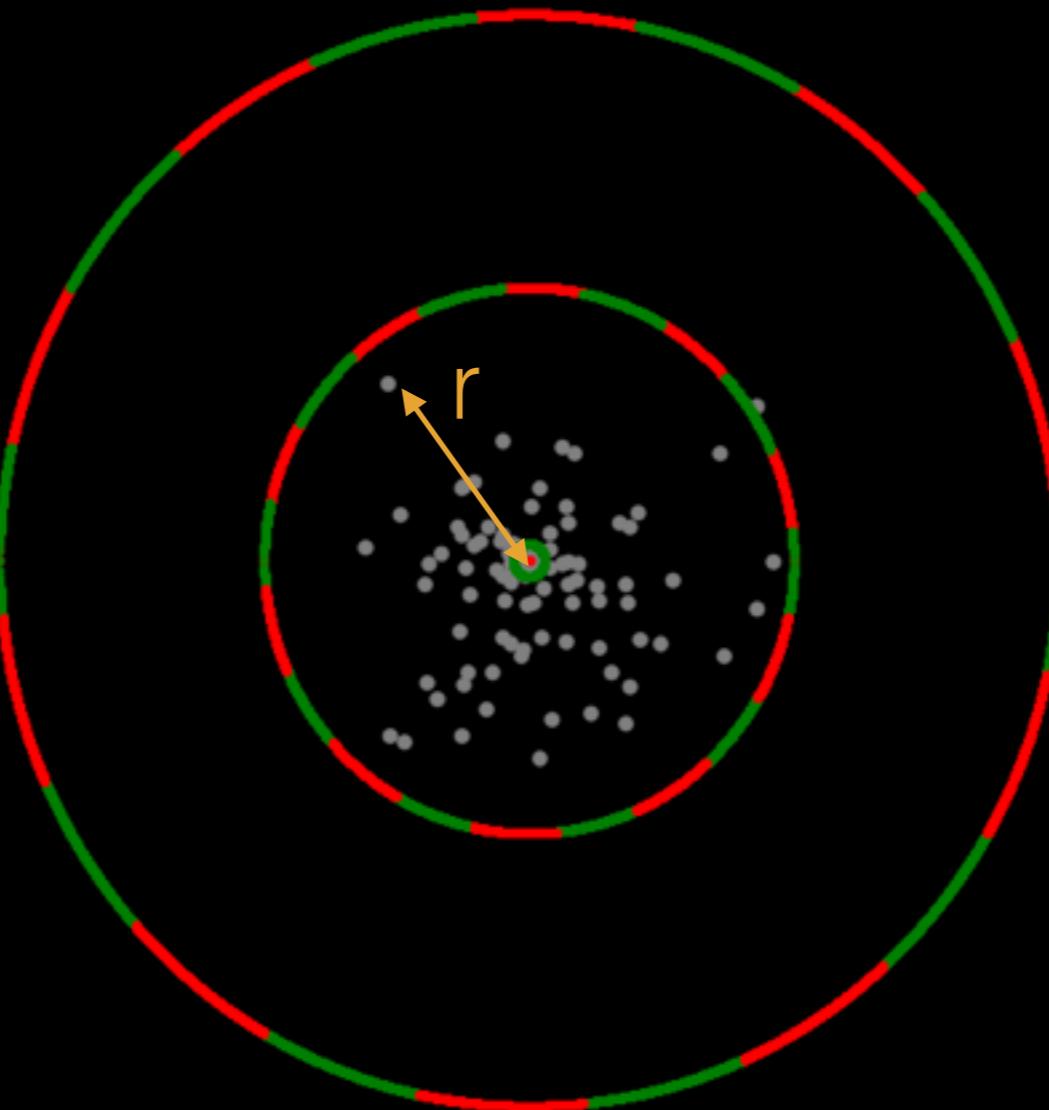


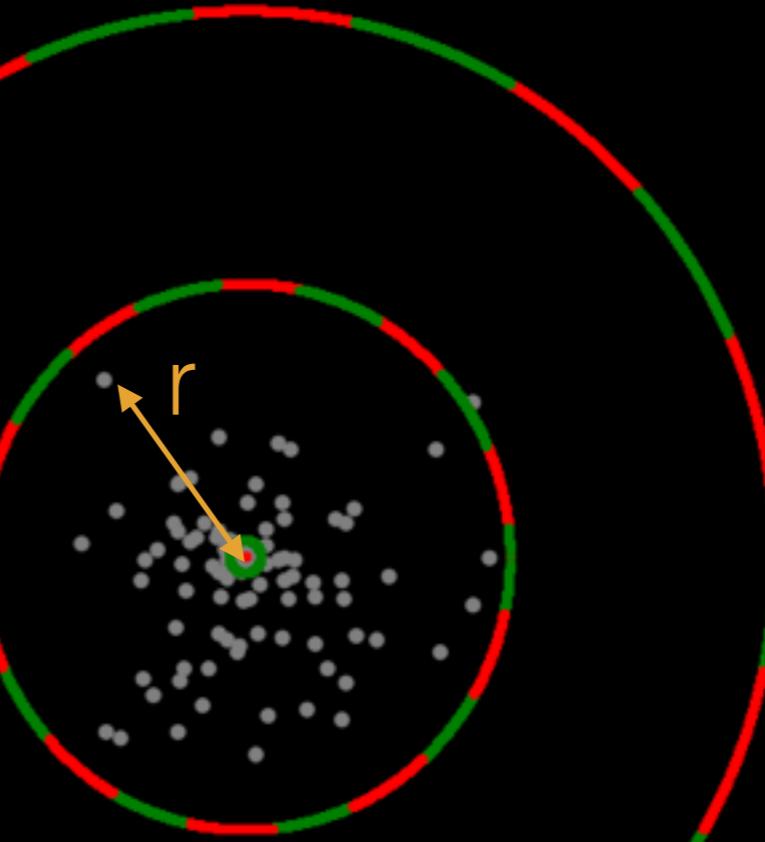
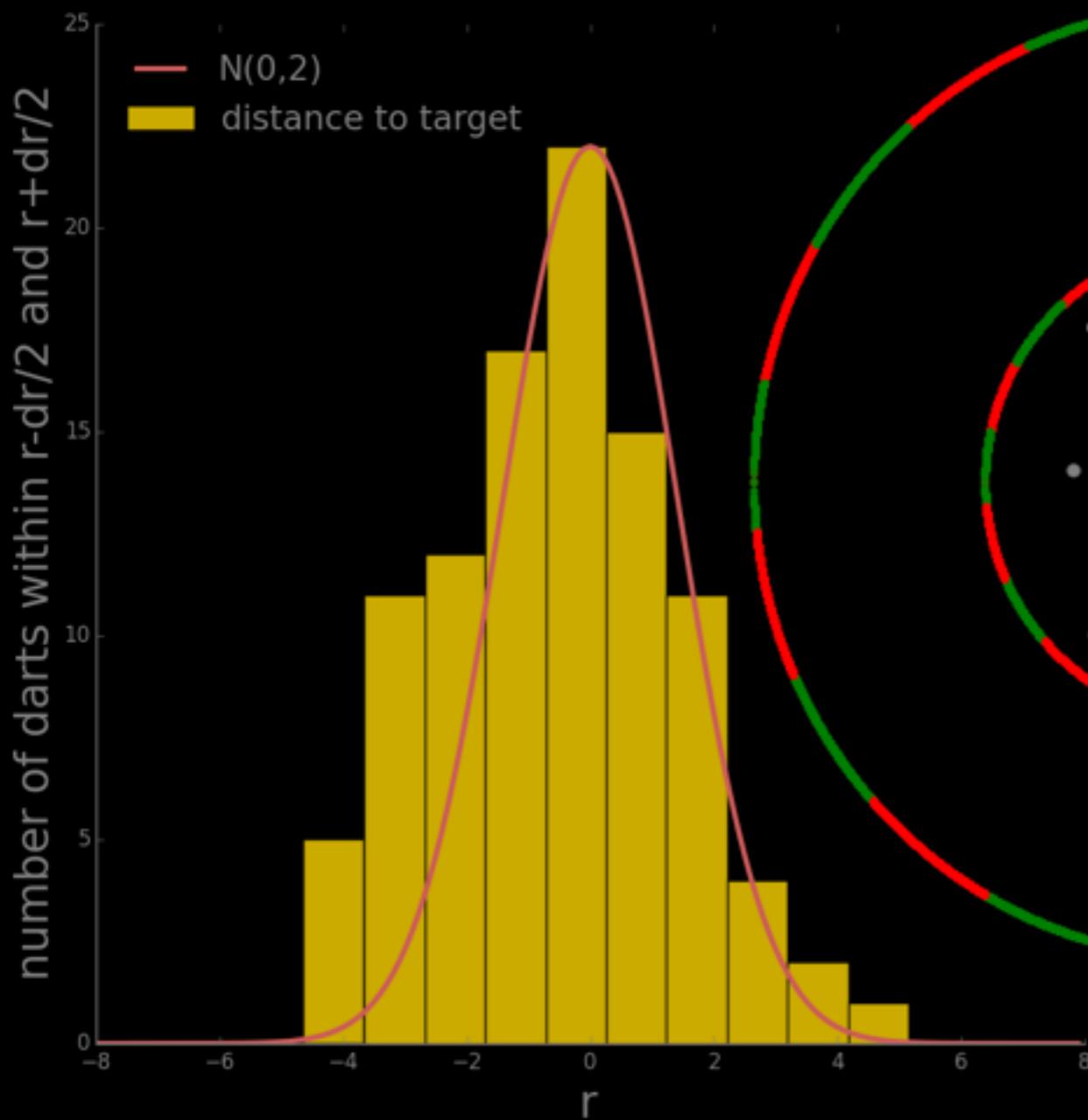




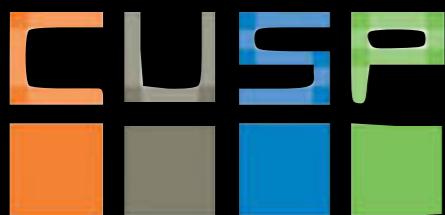








$$N(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Random sampling (numpy.random)

Simple random data

`rand(d0, d1, ..., dn)`

`randn(d0, d1, ..., dn)`

`randint(low[, high, size])`

`random_integers(low[, high, size])`

`random_sample([size])`

`random([size])`

`ranf([size])`

`sample([size])`

`choice(a[, size, replace, p])`

`bytes(length)`

Random values in a given shape.

Return a sample (or samples) from the “standard normal” distribution.

Return random integers from *low* (inclusive) to *high* (exclusive).

Return random integers between *low* and *high*, inclusive.

Return random floats in the half-open interval [0.0, 1.0).

Return random floats in the half-open interval [0.0, 1.0).

Return random floats in the half-open interval [0.0, 1.0).

Return random floats in the half-open interval [0.0, 1.0).

Generates a random sample from a given 1-D array

Return random bytes.

Table Of Contents

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

[Previous topic](#)

[numpy.RankWarning](#)

[Next topic](#)

[numpy.random.rand](#)

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.

`exponential([scale, size])`

Exponential distribution.



<http://docs.scipy.org/doc/numpy/reference/routines.random.html>

III: Introduction to statistics

distribution moments

a distribution's moments summarize its properties:

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

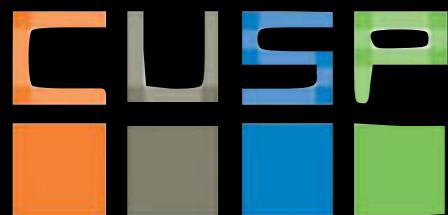
central tendency: mean ($n=1$), median, mode

spread: standard deviation/variance ($n=2$), quartiles range

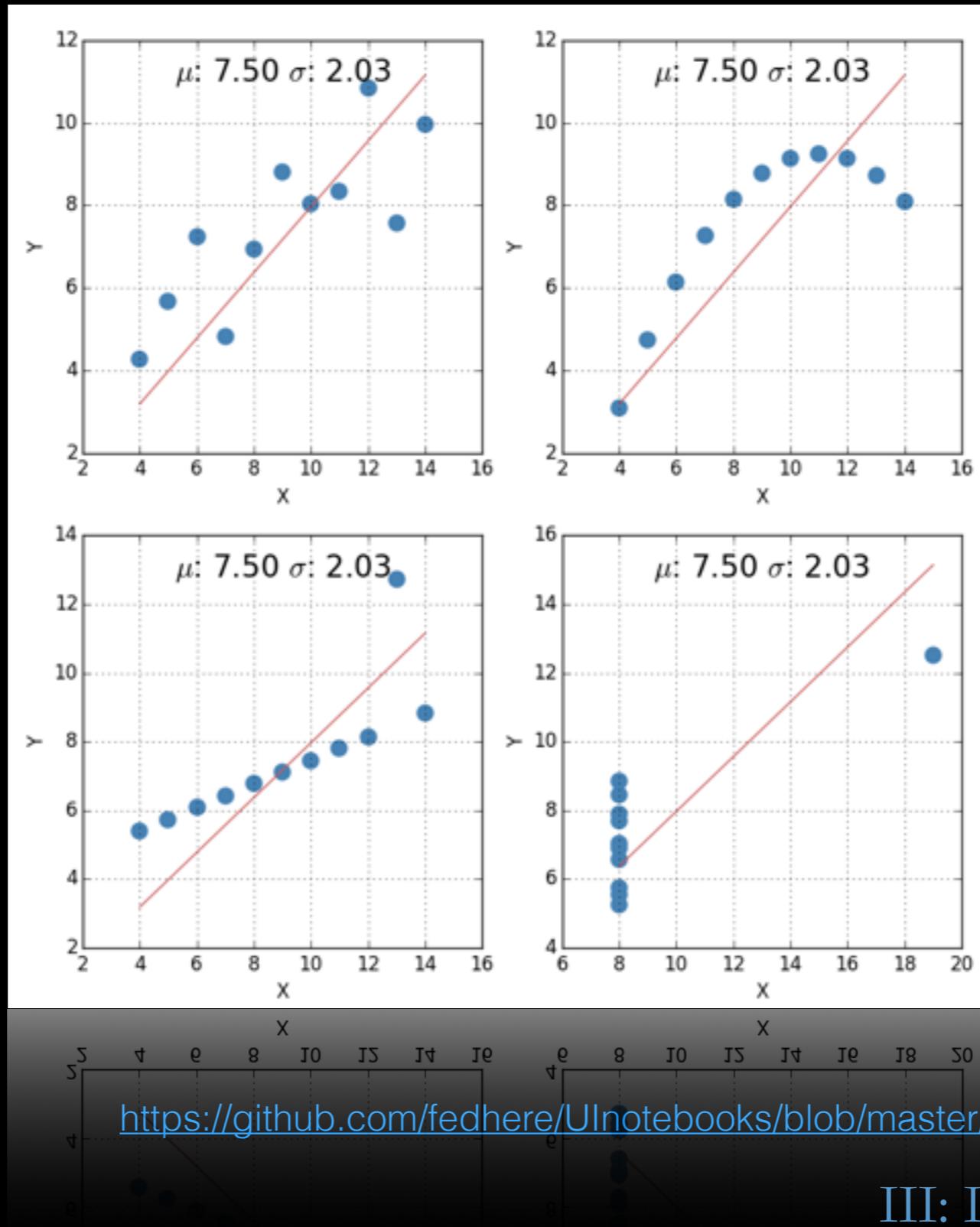
symmetry: skewness ($n=3$)

CUSPiness: kurtosis ($n=4$)

...



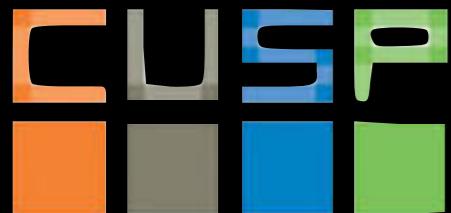
distribution moments



basic distributions and their moments

jupyter

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



distributions: Central Limit Theorem

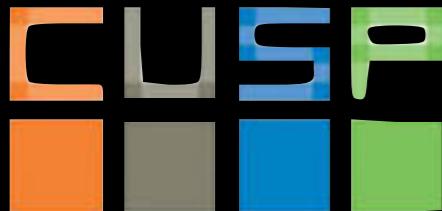
Laplace (1700s)

but also: Poisson, Bessel, Dirichlet, Cauchy, Ellis

Let $X_1 \dots X_N$ be an N-elements sample from a population whose distribution has mean μ and standard deviation σ

In the limit of $N \rightarrow \infty$
the sample mean m approaches a Normal (Gaussian) distribution with mean μ and standard deviation σ
regardless of the distribution of X

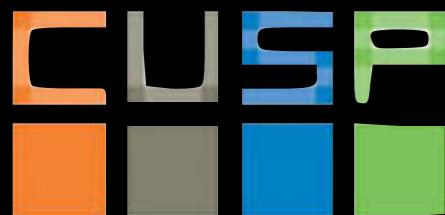
$$\bar{x} \sim N(\mu, \sigma/\sqrt{N})$$



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:

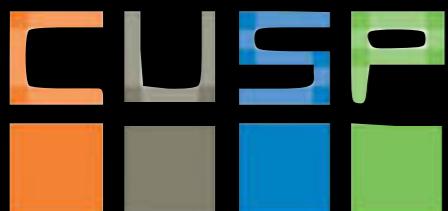


<http://glowingpython.blogspot.com/2012/07/distribution-fitting-with-scipy.html>
<http://stackoverflow.com/questions/7805552/fitting-a-histogram-with-python>

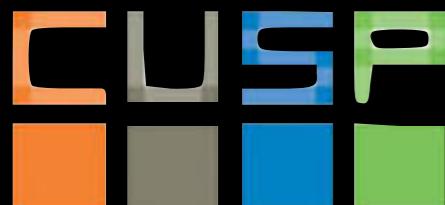
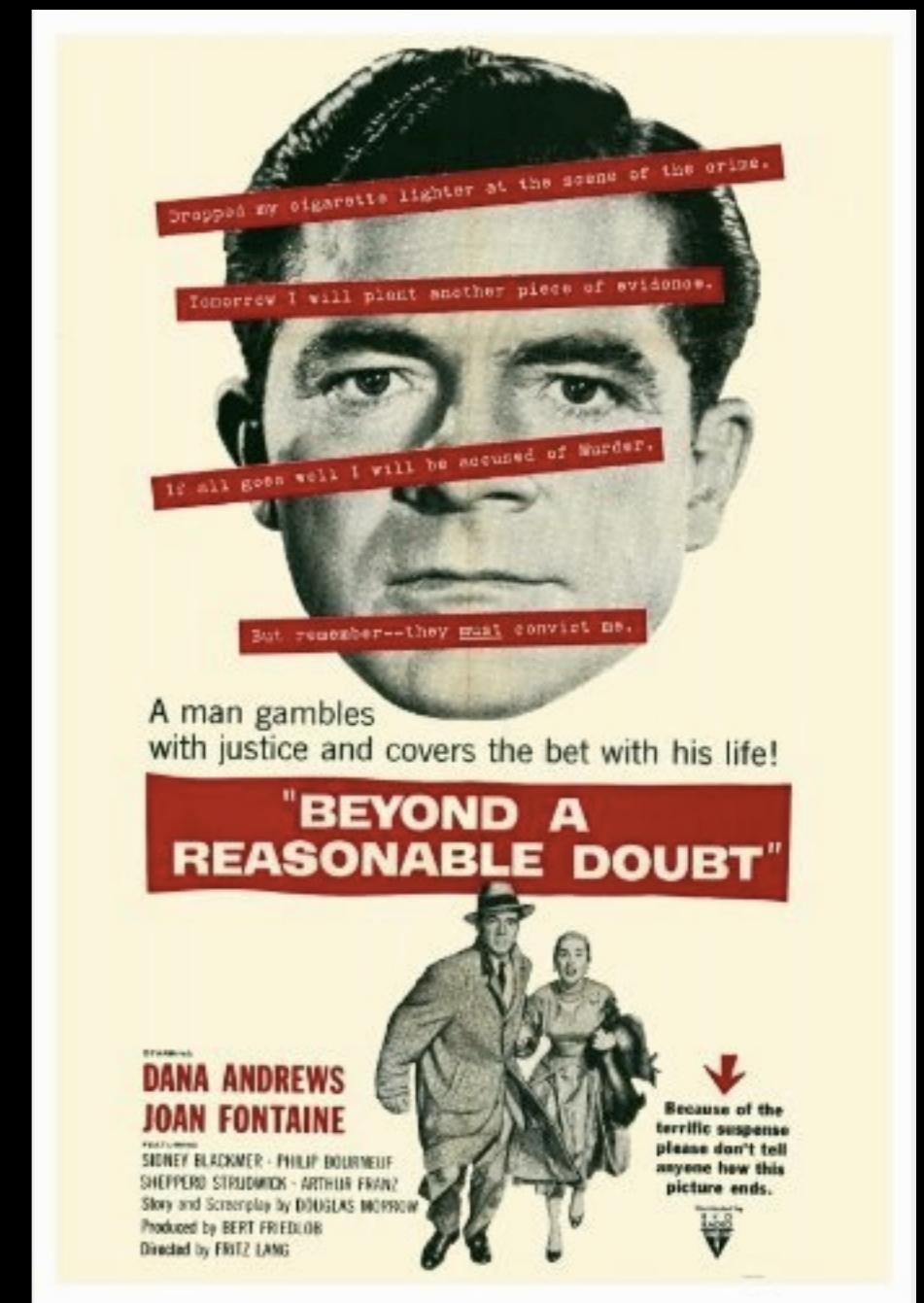
distributions, moments, and Central Limit Theorem

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.



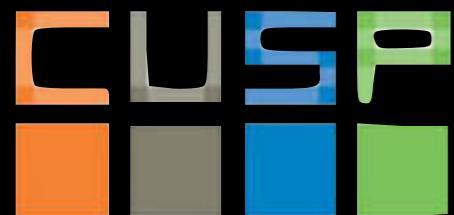
rejecting the Null
beyond any reasonable doubt



III: Introduction to statistics

Rejecting the Null Hypothesis: what is the p value?

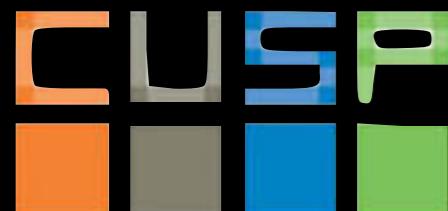
is the probability of getting a result at least as extreme as the one you observed just by chance lower than the significance level you established?



Rejecting the Null Hypothesis: what is the p value?

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

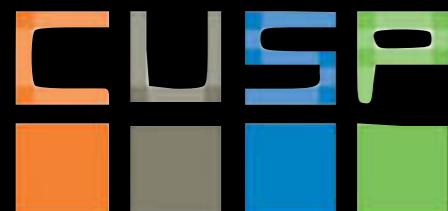
- decide what the significance threshold is: typically 5%
 $\alpha=0.05$



Rejecting the Null Hypothesis: what is the *p* value?

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

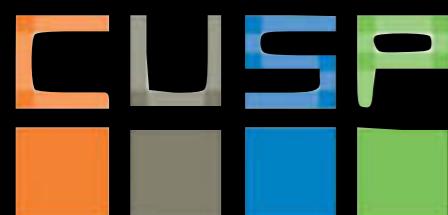
- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)



Rejecting the Null Hypothesis: what is the *p* value?

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

- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)
- find the probability p of your measurement for your test H_a in absence of effect

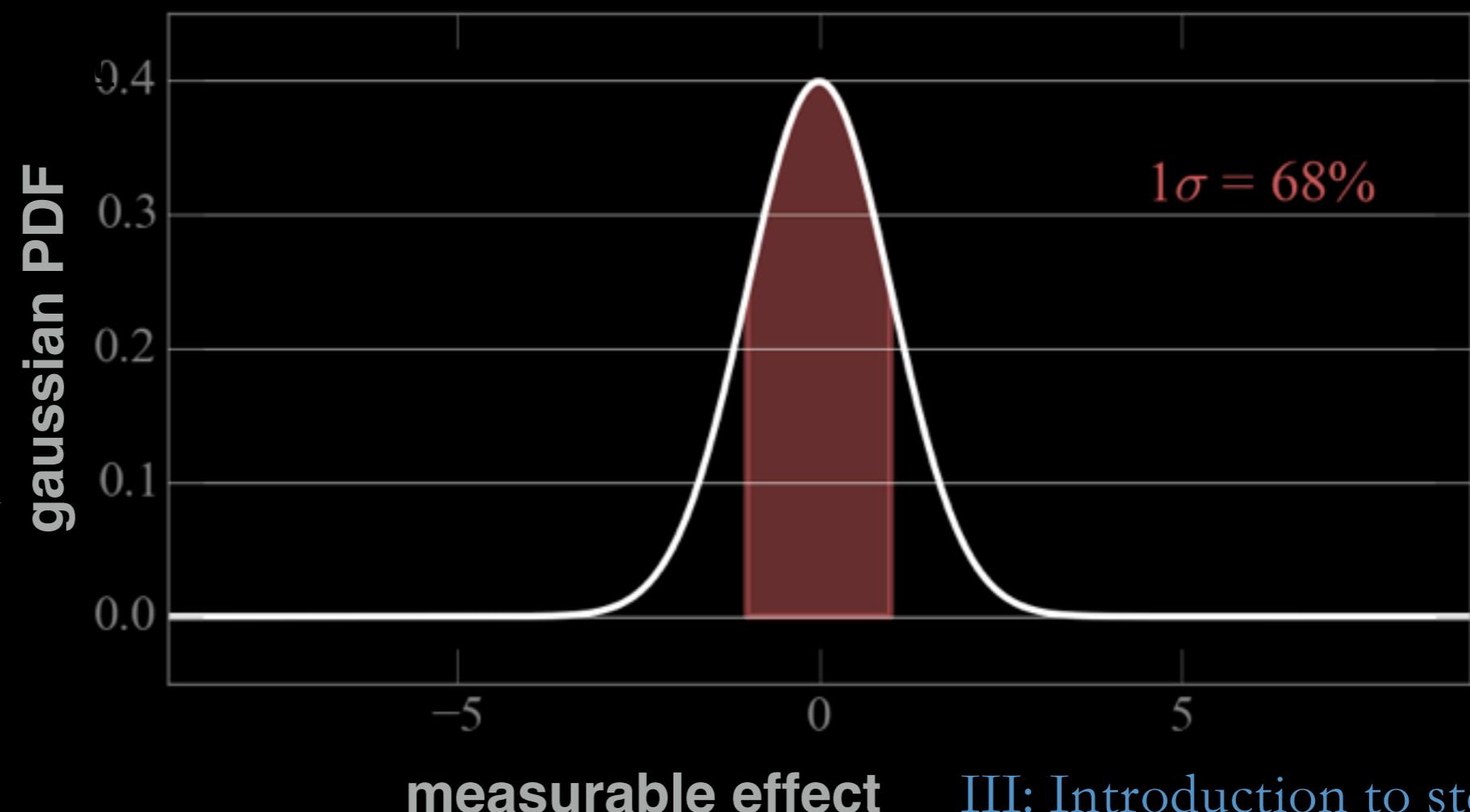


Rejecting the Null Hypothesis: what is the *p* value?

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

All statistical tests
rely on assuming
the possible results
of your
measurements (or
some quantities
derived from it)
follow a certain
distribution e.g.
Gaussian, Poisson
....

T-test, Z-test, bayesian analysis...



Rejecting the Null Hypothesis: what is the *p* value?

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

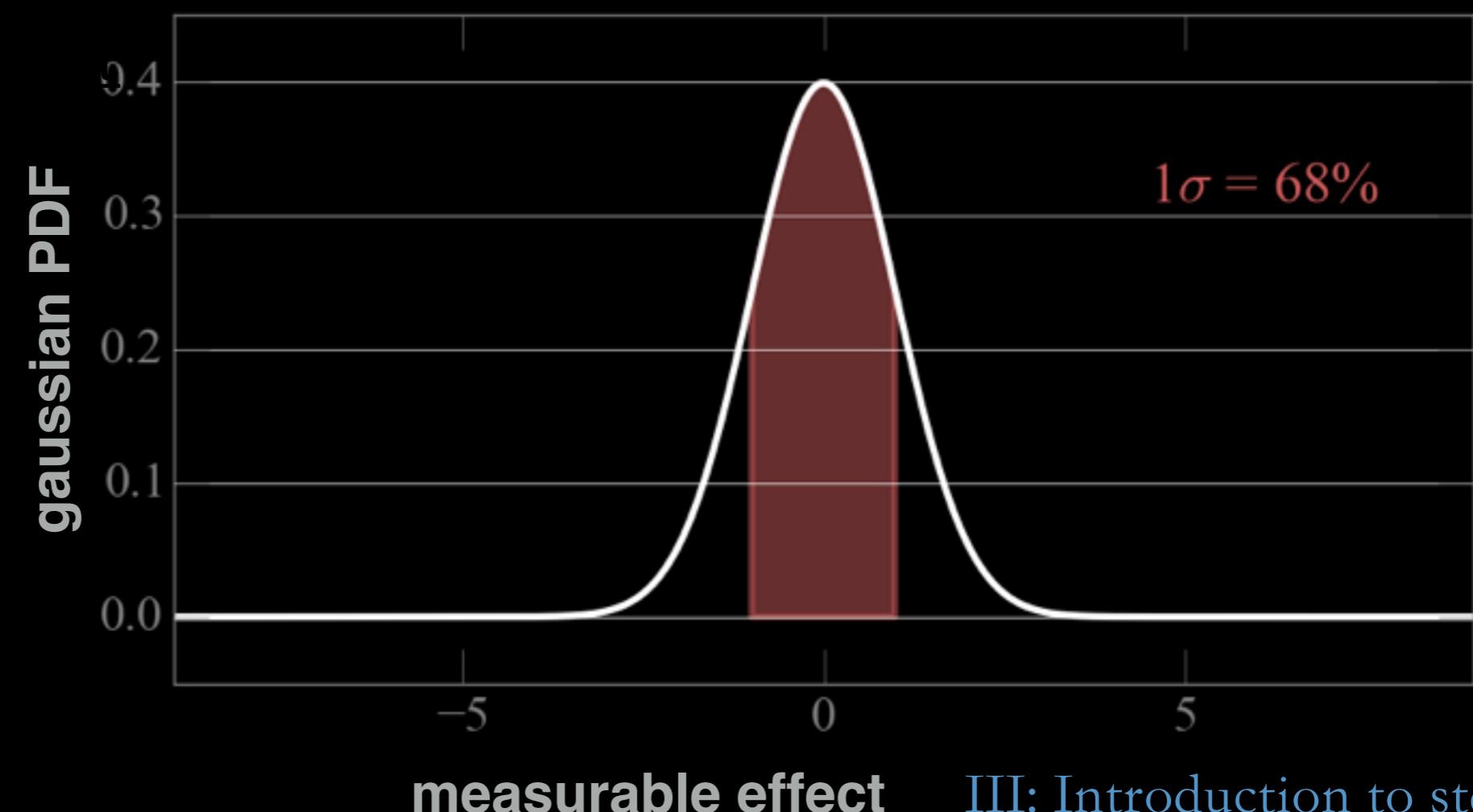
So you can answer the question:

“I find my measurement fall where the probability of the appropriate distribution is

$p_{\text{meas.}}$

Is $p_{\text{meas}} < p\text{-value?}$

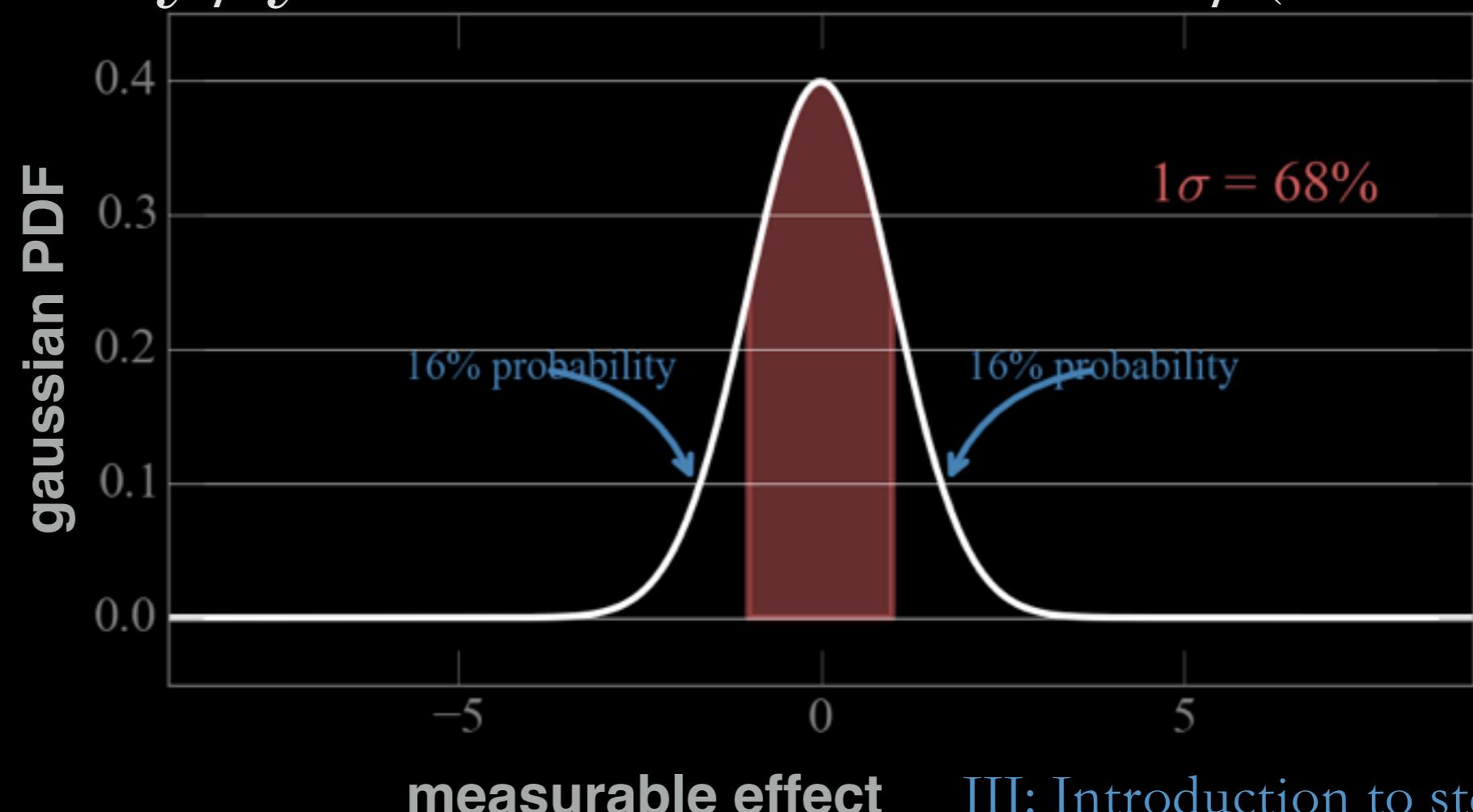
T-test, Z-test, bayesian analysis...



Rejecting the Null Hypothesis: what is the *p* value?

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

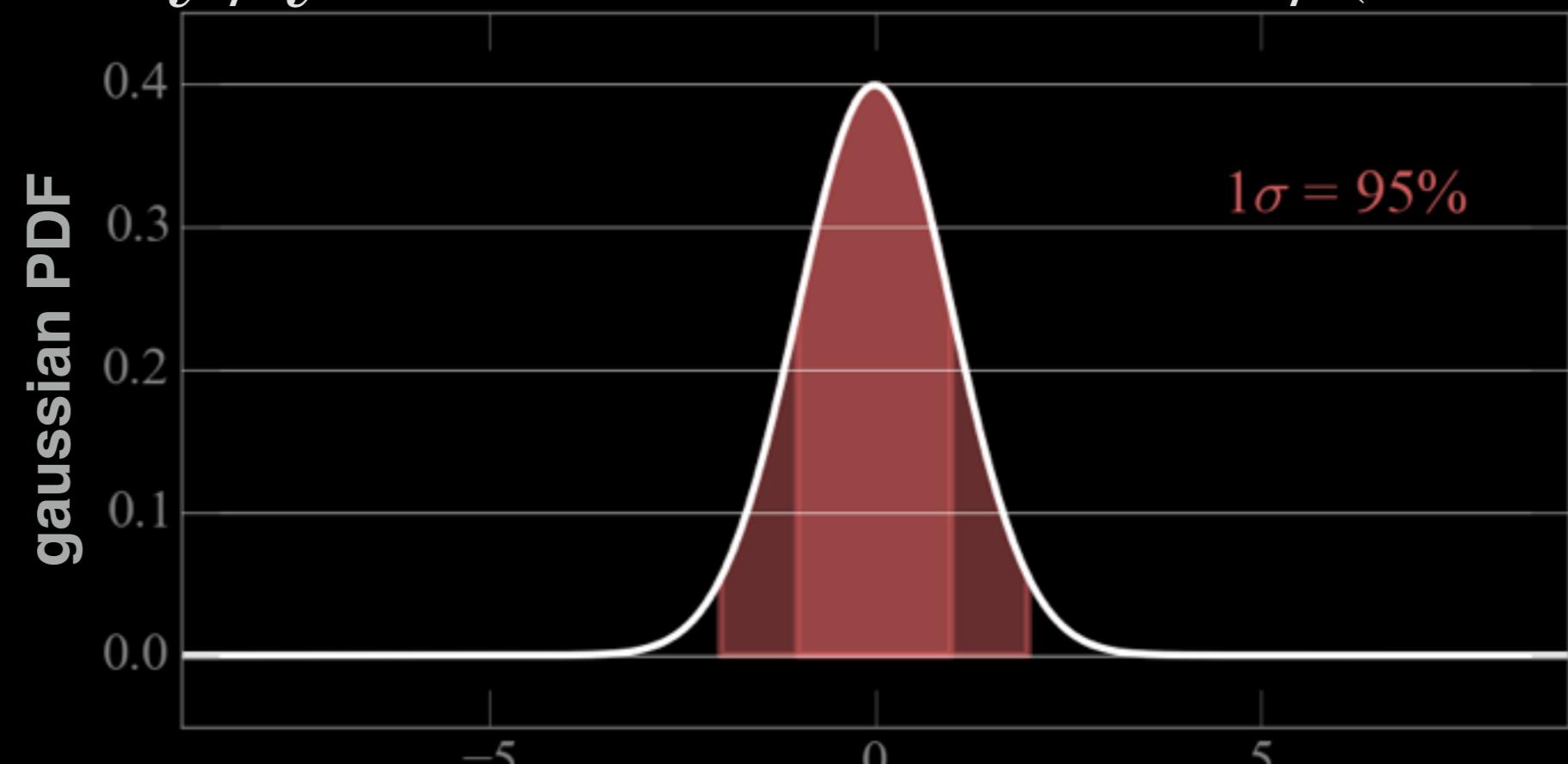
- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)
- find the probability p your measurement under $H_a: p(m | H_a)$



Rejecting the Null Hypothesis: what is the *p* value?

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

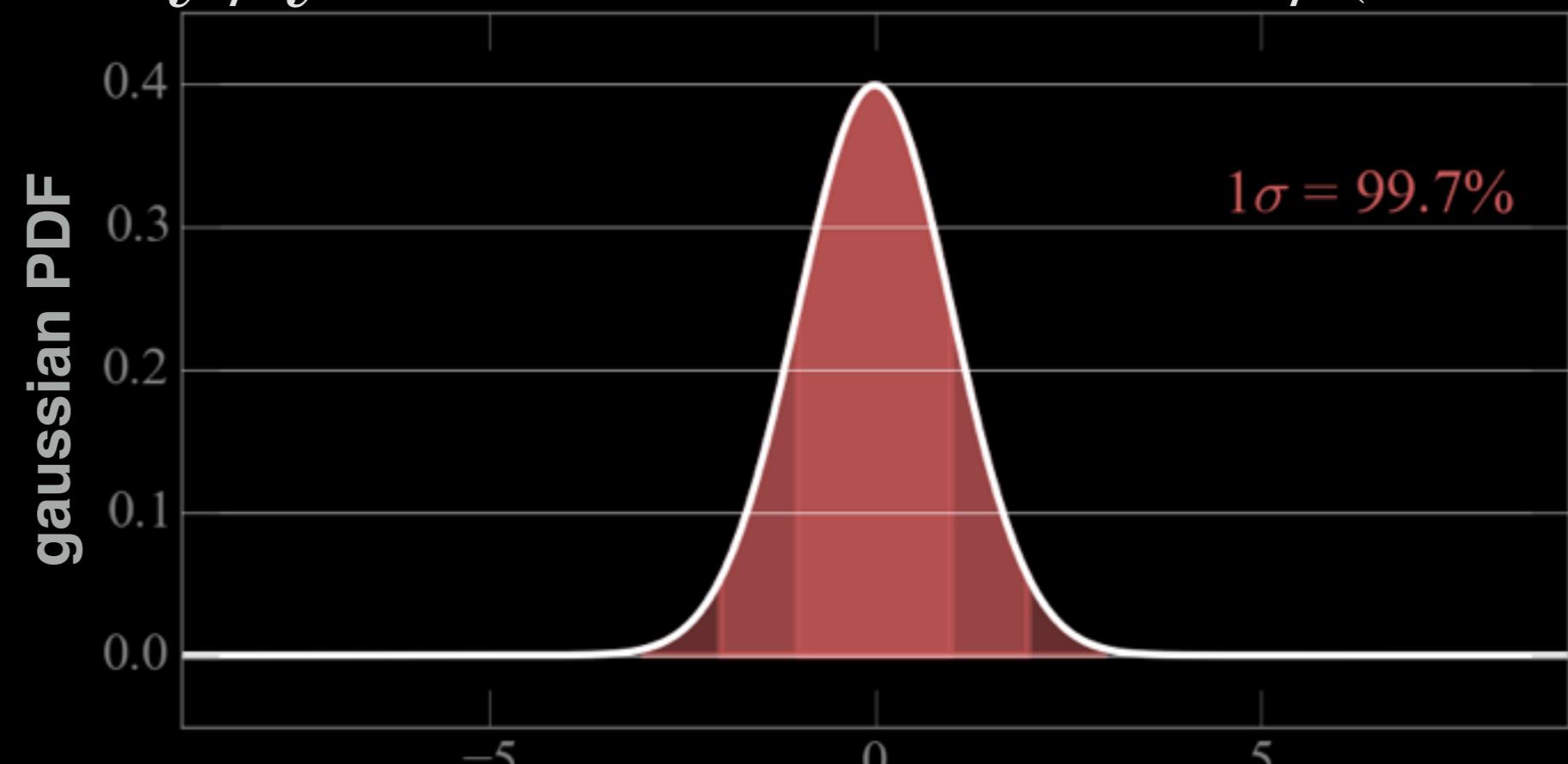
- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)
- find the probability p your measurement under $H_a: p(m | H_a)$



Rejecting the Null Hypothesis: what is the *p* value?

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

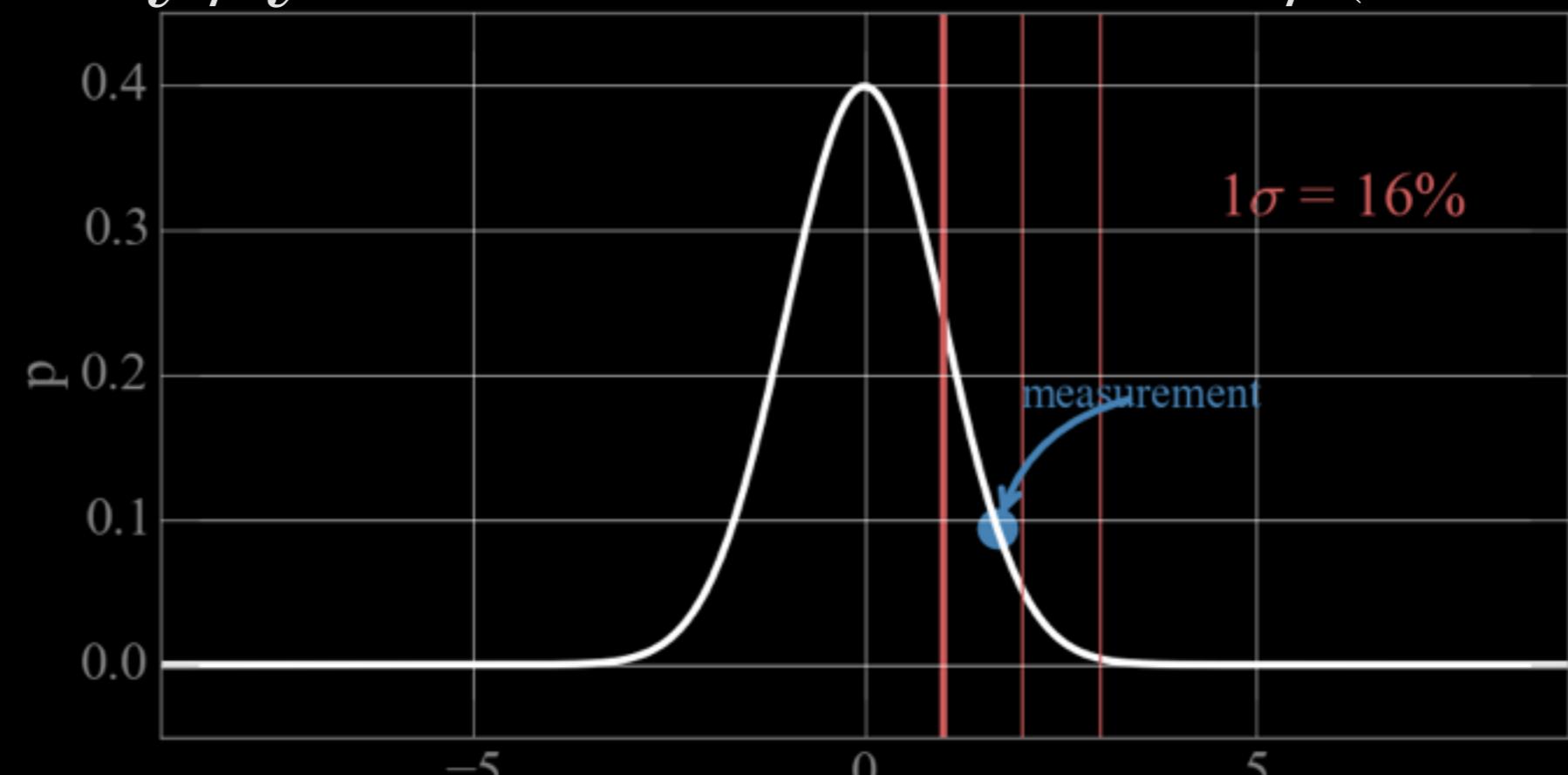
- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)
- find the probability p your measurement under $H_a: p(m | H_a)$



Rejecting the Null Hypothesis: what is the *p* value?

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

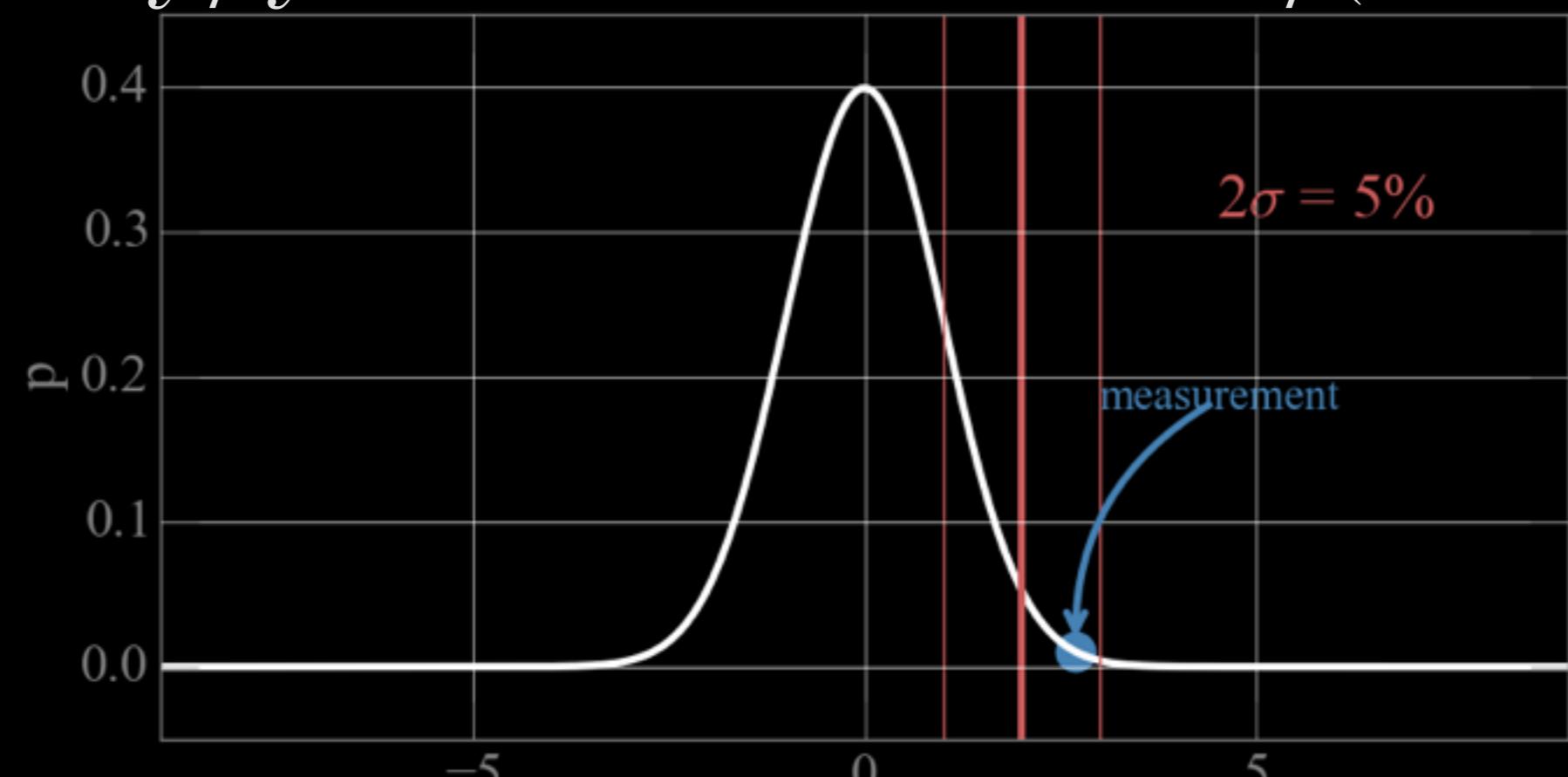
- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)
- find the probability p your measurement under $H_a: p(m | H_a)$



Rejecting the Null Hypothesis: what is the *p* value?

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

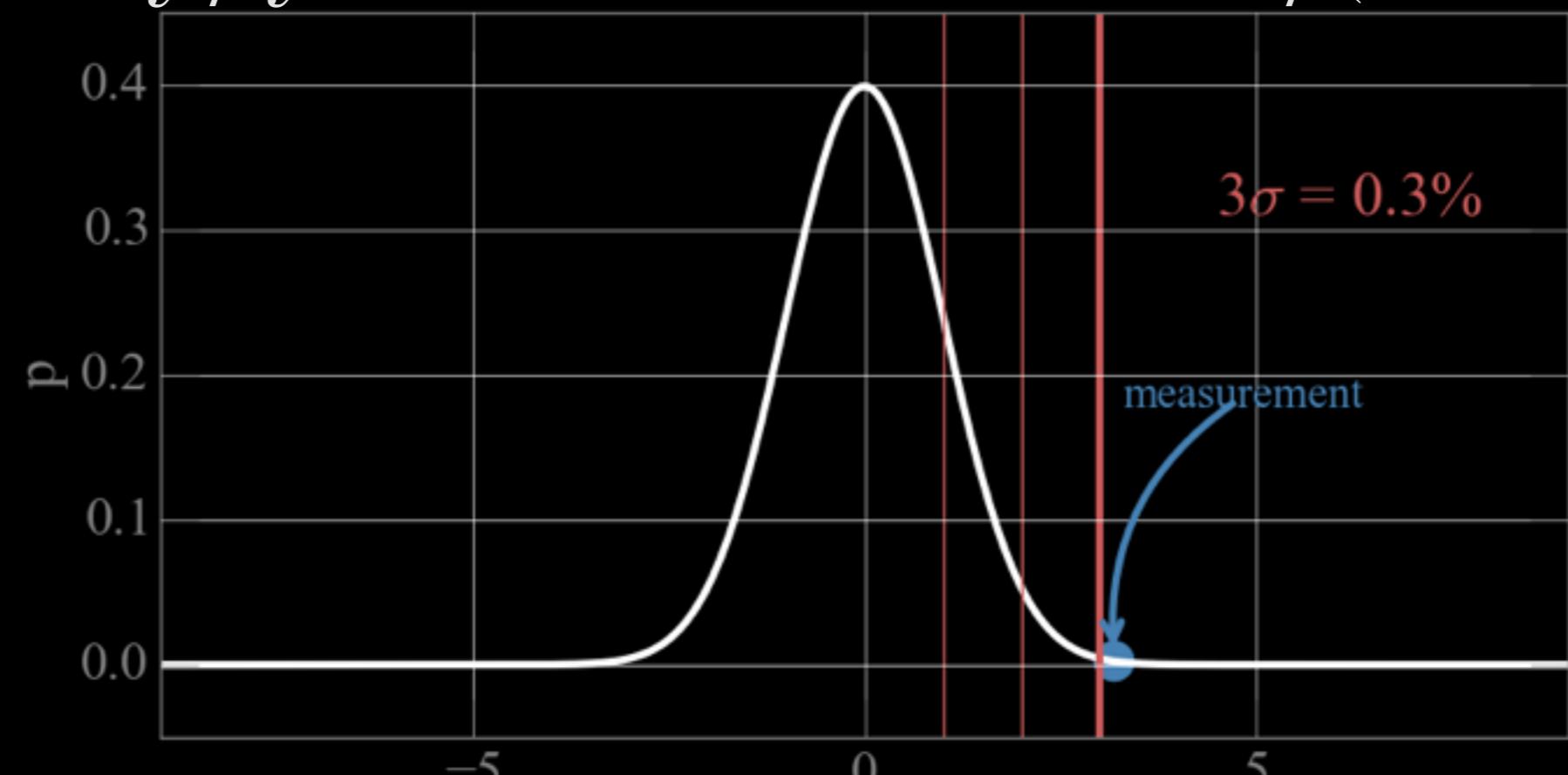
- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)
- find the probability p your measurement under $H_a: p(m | H_a)$



Rejecting the Null Hypothesis: what is the *p* value?

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

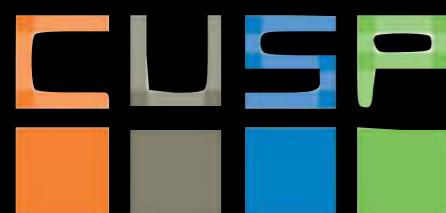
- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)
- find the probability p your measurement under $H_a: p(m | H_a)$



Rejecting the Null Hypothesis: what is the *p* value?

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

- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)
- find the probability p your measurement under $H_a: p(m | H_a)$
- if $p(H_a) - p(H_0) > \alpha$ the null hypothesis H_0 is falsified at the $1-\alpha$ confidence level

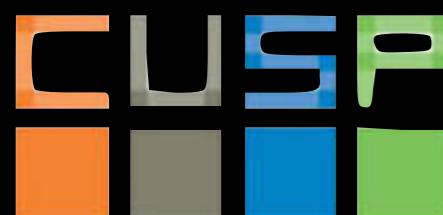


Rejecting the Null Hypothesis: what is the *p* value?

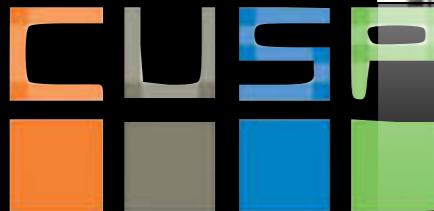
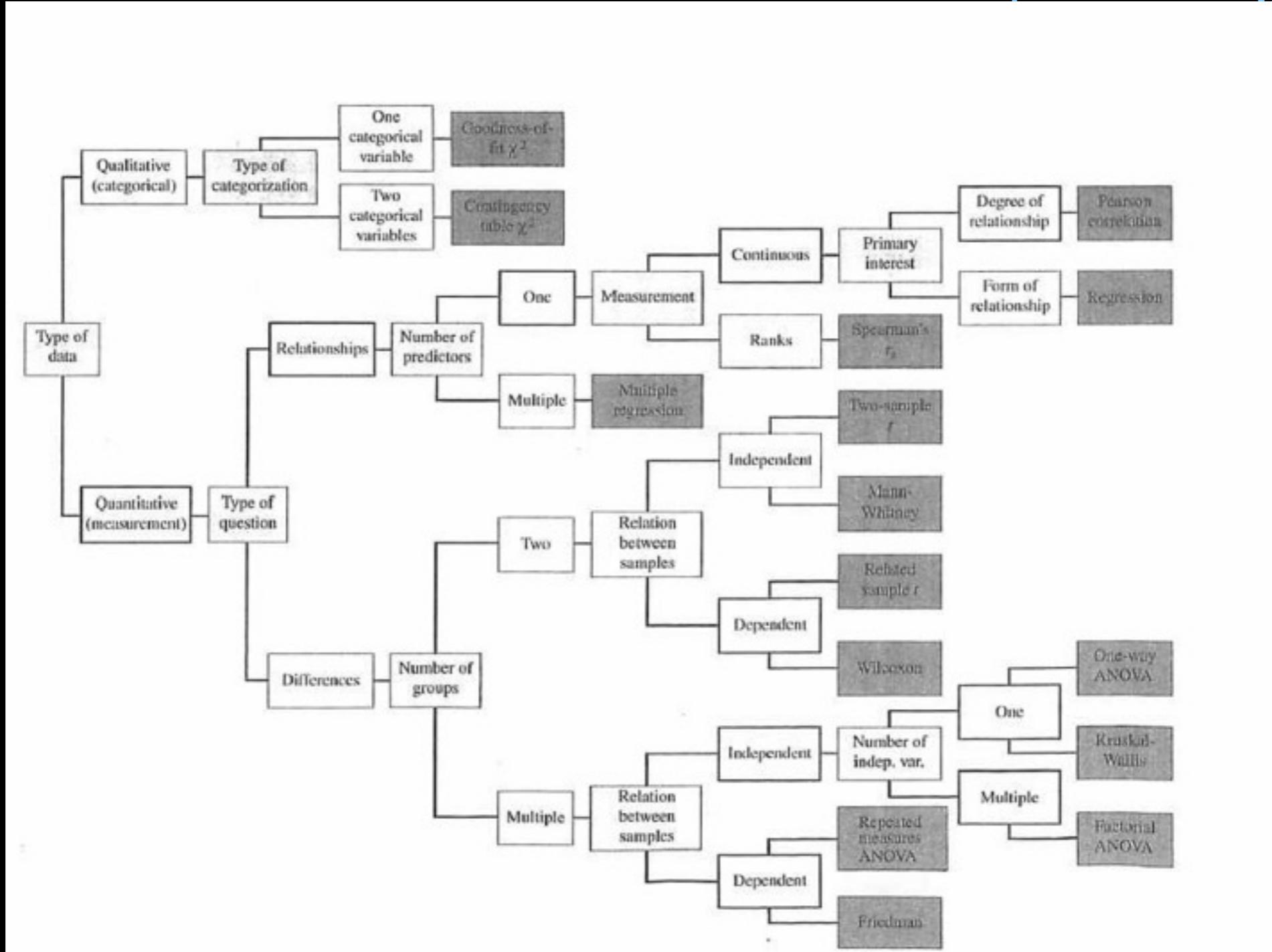
is the probability of getting a result at least as extreme as the one you observed just by chance lower than the significance level you established?

- decide what the significance threshold is: typically 5%
 $\alpha=0.05$
- choose a statistical test (T-test, Z-test, bayesian analysis...)
- find the probability p your measurement under $H_a: p(m | H_a)$
- if $p(H_a) - p(H_0) > \alpha$ the null hypothesis H_0 is falsified at the $1-\alpha$ 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!

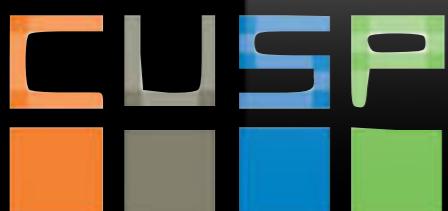


NULL HYPOTHESIS: the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,
significance level p=0.05

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

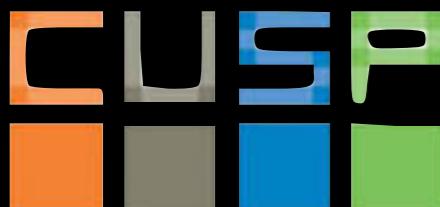
Outcome	Program Group	Control Group	Difference (Impact)	P-Value
Employment (Years 1-3) (%)				
Ever employed	83.8	70.4	13.4 ***	0.000
Ever employed in a CEO transitional job ^a	70.1	3.5	66.6 ***	0.000
Ever employed in an unsubsidized job	63.7	69.0	-5.3 *	0.078
Postprogram unsubsidized employment (Years 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		

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

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.



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) (%)	P₁	P₀		
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		

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

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).

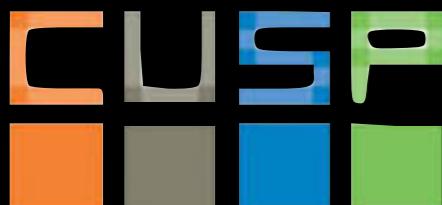
NOTES: Statistical significance levels are indicated as: *** = 1 percent; ** = 5 percent; * = 10 percent.

The p-value indicates the likelihood that the difference between the program and control groups arose by chance.

$$H_0: P_0 - P_1 > 0$$

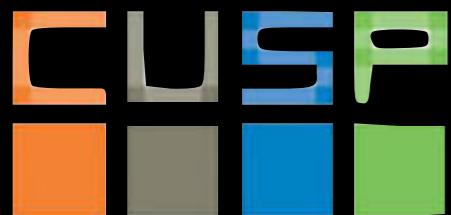
$$H_a: P_0 - P_1 \leq 0$$

$$\alpha=0.05$$



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>

