

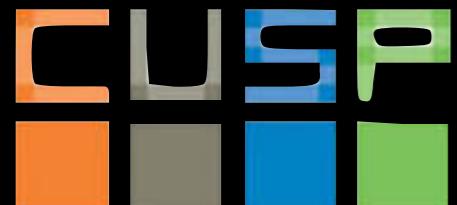
Urban Informatics

Fall 2015

dr. federica bianco fb55@nyu.edu

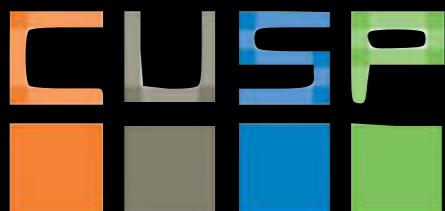


@fedhere



Recap:

- Good practices with data: falsifiability, reproducibility
- Basic data retrieving and munging: APIs, Data formats
- Basic statistics: distributions and their moments
- Hypothesis testing: p -value, statistical significance
- Systematic errors

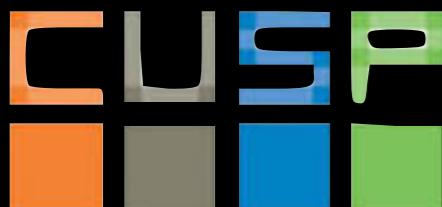


Recap:

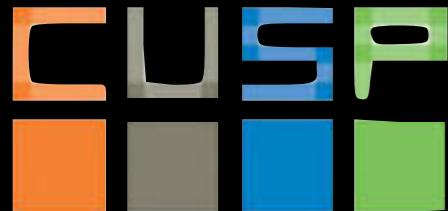
- Good practices with data: falsifiability, reproducibility
- Basic data retrieving and munging: APIs, Data formats
- Basic statistics: distributions and their moments
- Hypothesis testing: p -value, statistical significance
- Systematic errors

Today:

- Statistical error
- Goodness of fit tests
- Likelihood
- Linear Regression
- Predictive models



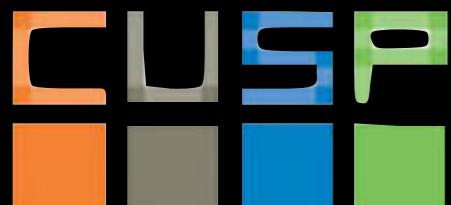
False Positives and False Negatives



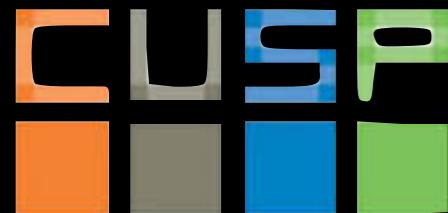
V: Likelihood and
Regression Models

False Positives and False Negatives

	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

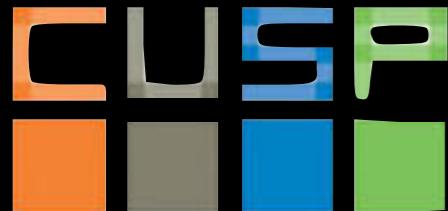


Errors and uncertainties.



V: Likelihood and
Regression Models

systematic errors



V: Likelihood and
Regression Models

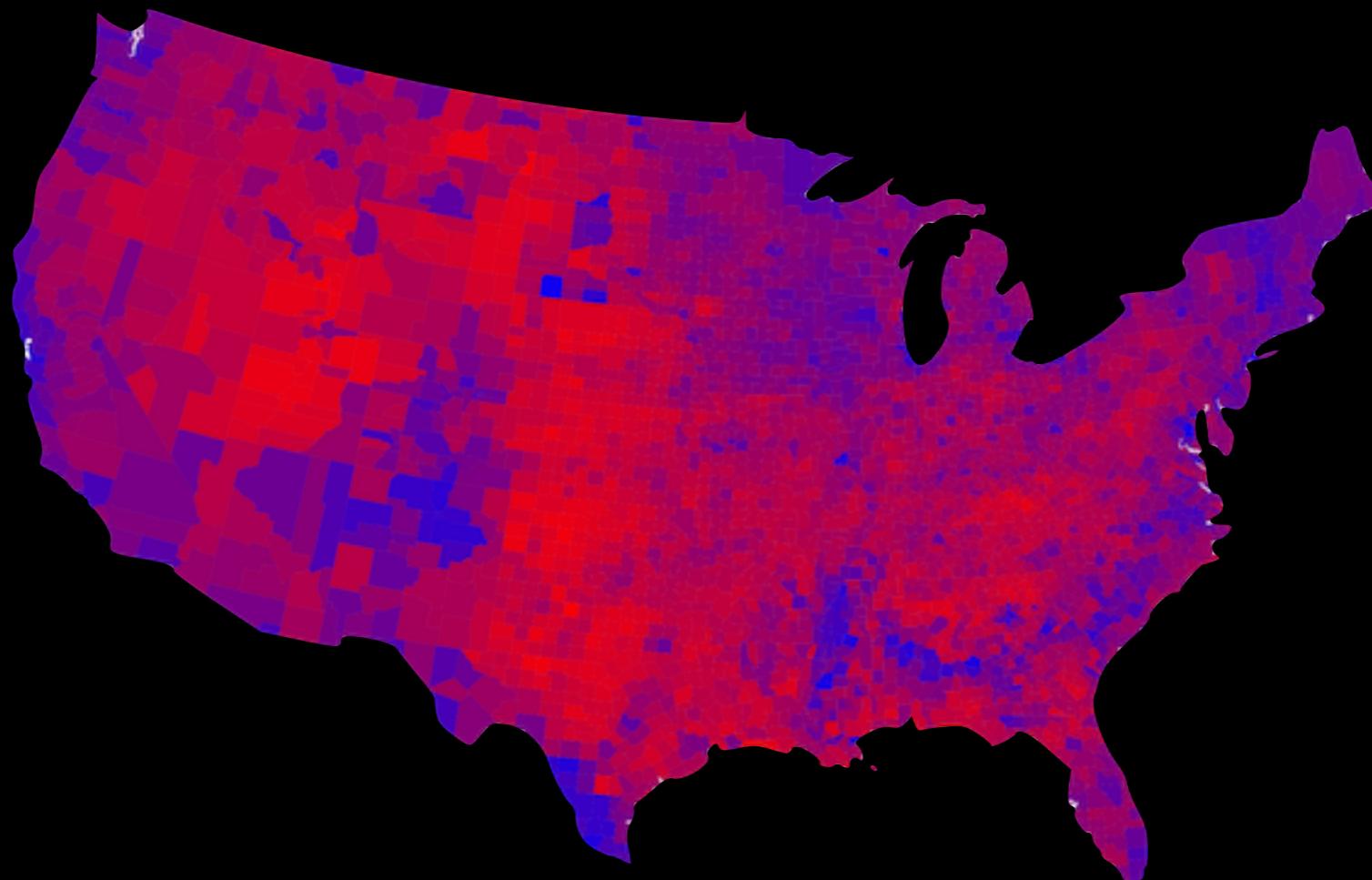
Errors and uncertainties.

- Systematic error
 - tendency to systematically underestimate/overestimate the average
difference between the *population* and the subset you test or *sample* because the sample is intrinsically different or the measurements are consistently off

Errors and uncertainties.

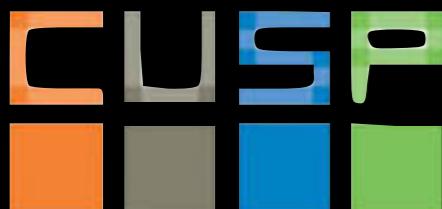


Where are the Real Errors in Political Polls?



<http://blogs.scientificamerican.com/guest-blog/where-are-the-real-errors-in-political-polls/>

Rasmussen Reports mostly finds its sample group through landline phones, which many people no longer use. For those who do not have landline phones, Rasmussen uses an online survey. There are a couple of problems with this methodology. First, *this means the company can only reach people who have landline phones or Internet access*. Yet, according to Nate Silver, the founder and editor of [FiveThirtyEight](#), 23 percent of adults do not have a landline, 4 percent don't answer their landline and 2 percent don't have a phone at all. So Rasmussen's method could definitely bias the poll towards the wealthier and older segments of the population that still uses landlines, both of which tend to vote Republican.



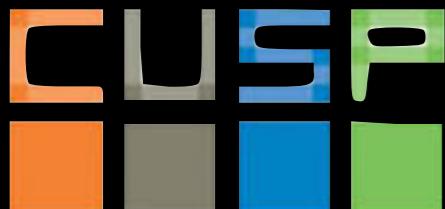
Bias in measurements: know your data

- Systematic error: SURVEY BIAS

UNBIASED survey:
average from *all* samples
equals *population* average

Undercoverage bias
Self selection bias
Social desirability bias

Small number statistic
Publication Bias
Data Dredging



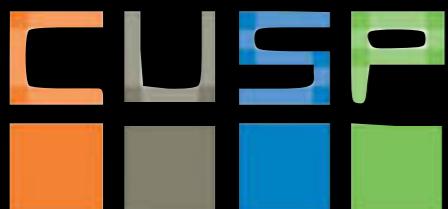
Errors and uncertainties.

- Systematic error
tendency to systematically underestimate/overestimate the average
difference between the *population* and the subset you test or *sample* because the sample is intrinsically different or the measurements are consistently off

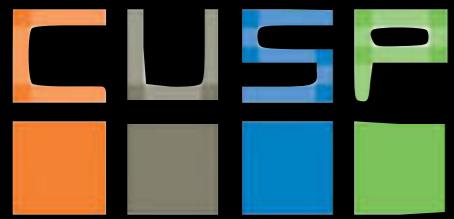
Solution: Good experimental design

Calibration (to assess systematics induced by your measurements)

Simulations (to assess the systematics induced by your analysis)



statistical errors



Errors and uncertainties.

- Systematic error
- Stochastic & Random error
 - unpredictable uncertainty in a measurement due to lack of sensitivity in the measurement or to stochasticity in a process

Errors and uncertainties.

- Stochastic & Random error
 - unpredictable uncertainty in a measurement due to lack of sensitivity in the measurement or
 - to stochasticity (inherent randomness) in a process



Errors and uncertainties.

- Stochastic & Random error



unpredictable uncertainty in a measurement
due to lack of sensitivity in the measurement or
to stochasticity in a process

$$2.0 \pm \varepsilon \text{ cm}, \varepsilon > 0.1 \text{ cm}$$

Repeated measurements
give ε (*bootstrap* data...)



Errors and uncertainties.

- Stochastic & Random error

Deterministic systems have no randomness in their evolution. *Chaos* is deterministic....

Stochastic processes can be *completely random*: the probability of any event is disjoint from that of the previous one
These are Poisson processes:

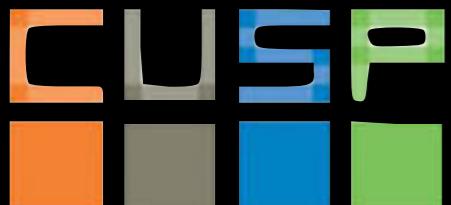
Errors and uncertainties.

- Stochastic & Random error

Deterministic systems have no randomness in their evolution. *Chaos* is deterministic....

Stochastic processes can be *completely random*: the probability of any event is disjoint from that of the previous one
These are Poisson processes:
they are described by a Poisson distribution.

A discrete distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event.

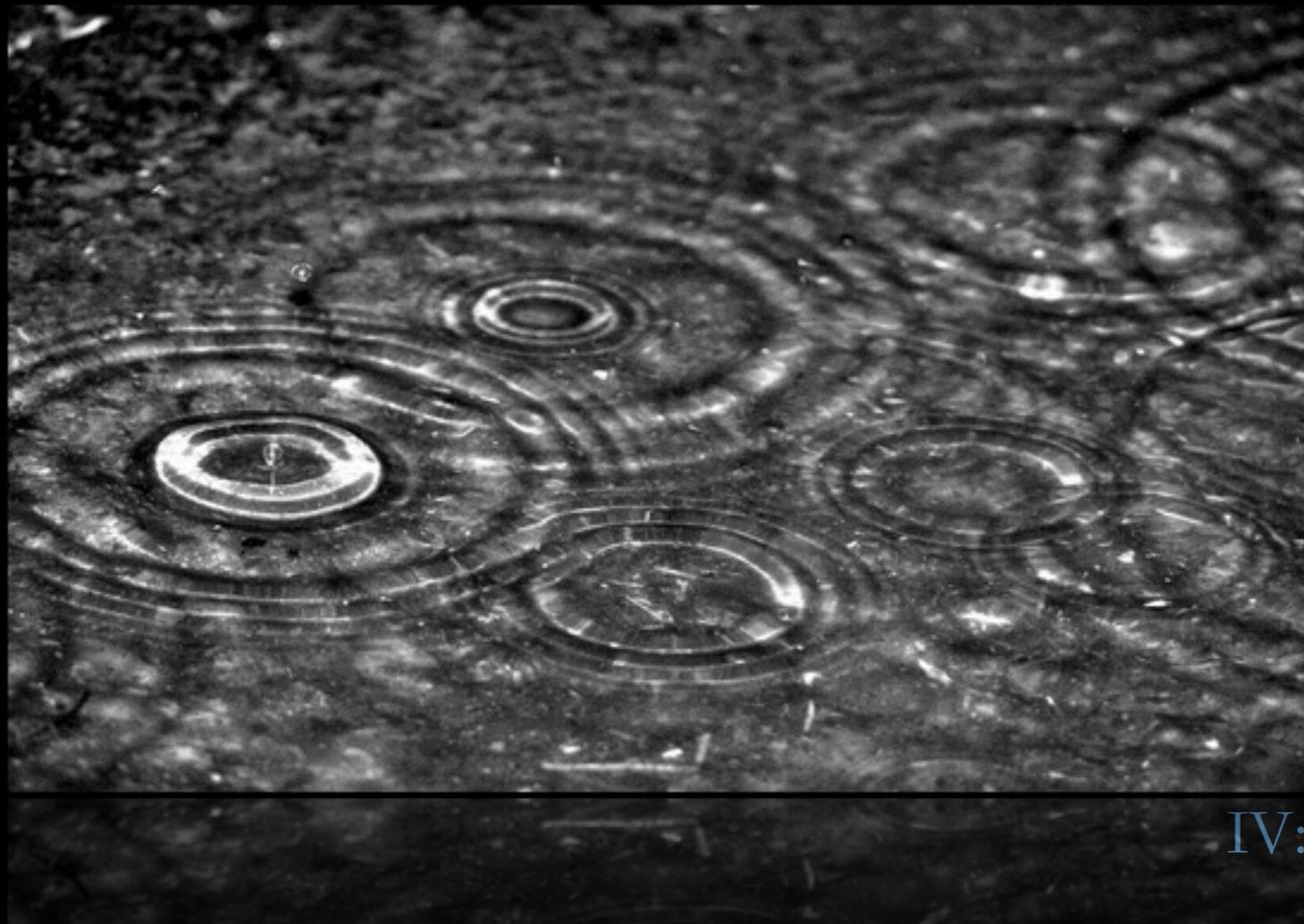


Errors and uncertainties.

- Stochastic & Random error

Poisson processes :

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



Errors and uncertainties.

- Stochastic & Random error

Poisson processes :

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



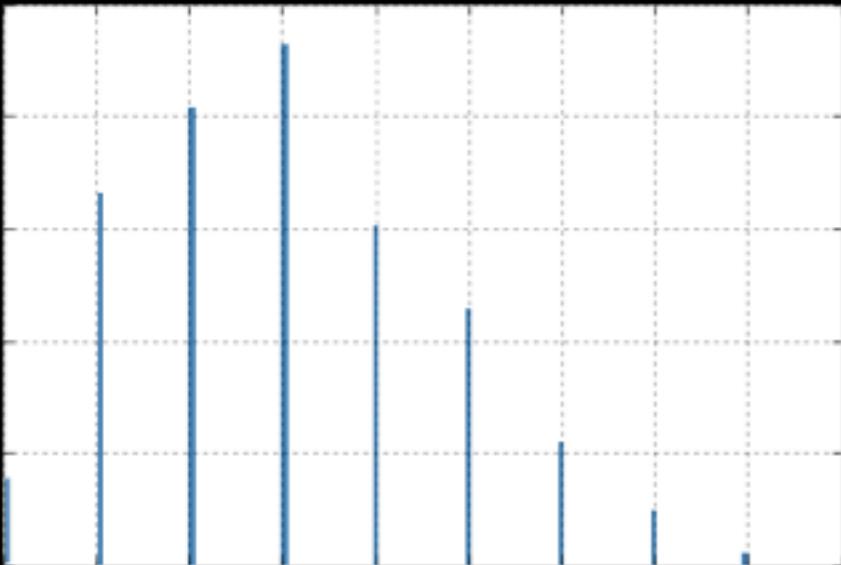
Binomial

discrete bivariate

2 parameters: n,p

support: [0 , ∞]

moments: np, \sqrt{npq} , >0



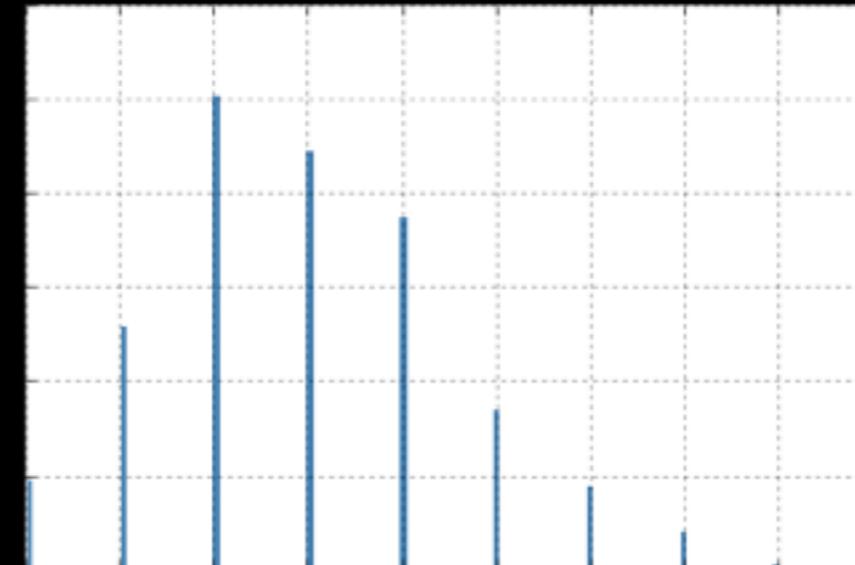
Poisson

discrete univariate

1 parameters: λ

support: [0 , ∞]

moments: $\lambda, \sqrt{\lambda}$, >0



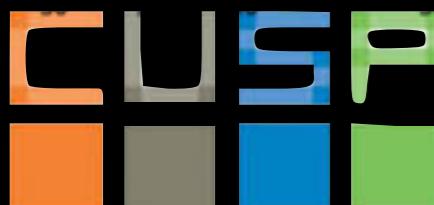
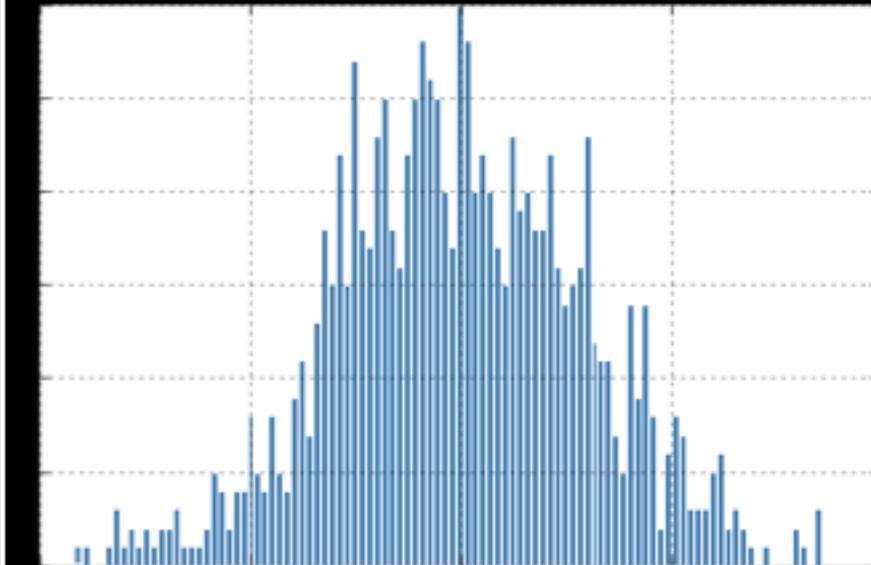
Gaussian

continuous bivariate

2 parameters: μ, σ

support:[- ∞ , ∞]

moments: $\mu, \sigma, 0$



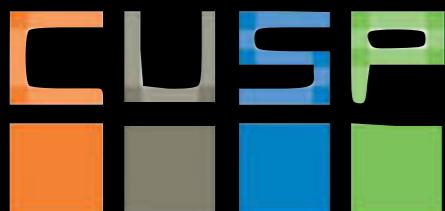
https://github.com/fedhere/UInotebooks/blob/master/binomial_gaussian_poisson.ipynb

IV: Statistical analysis

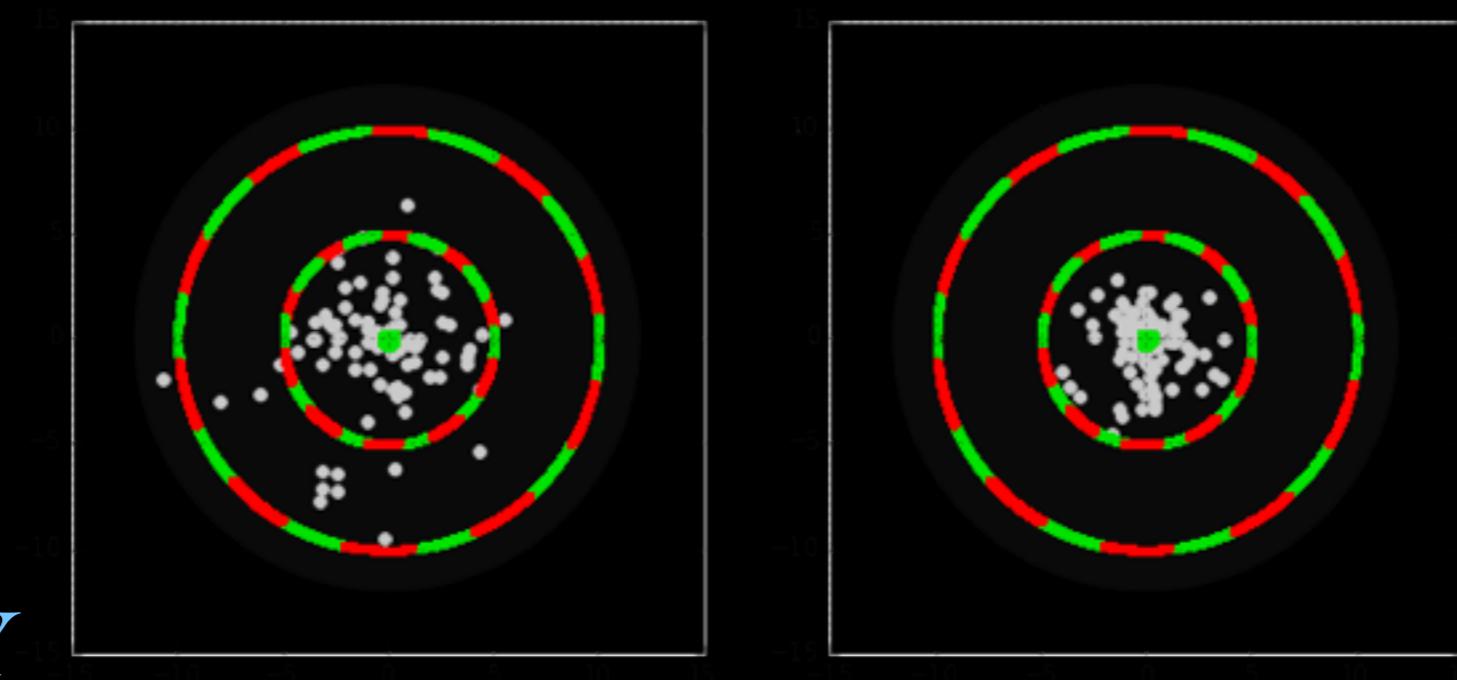
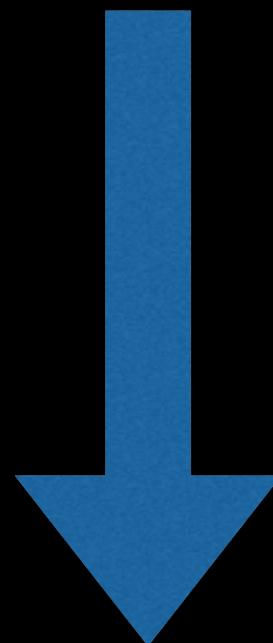
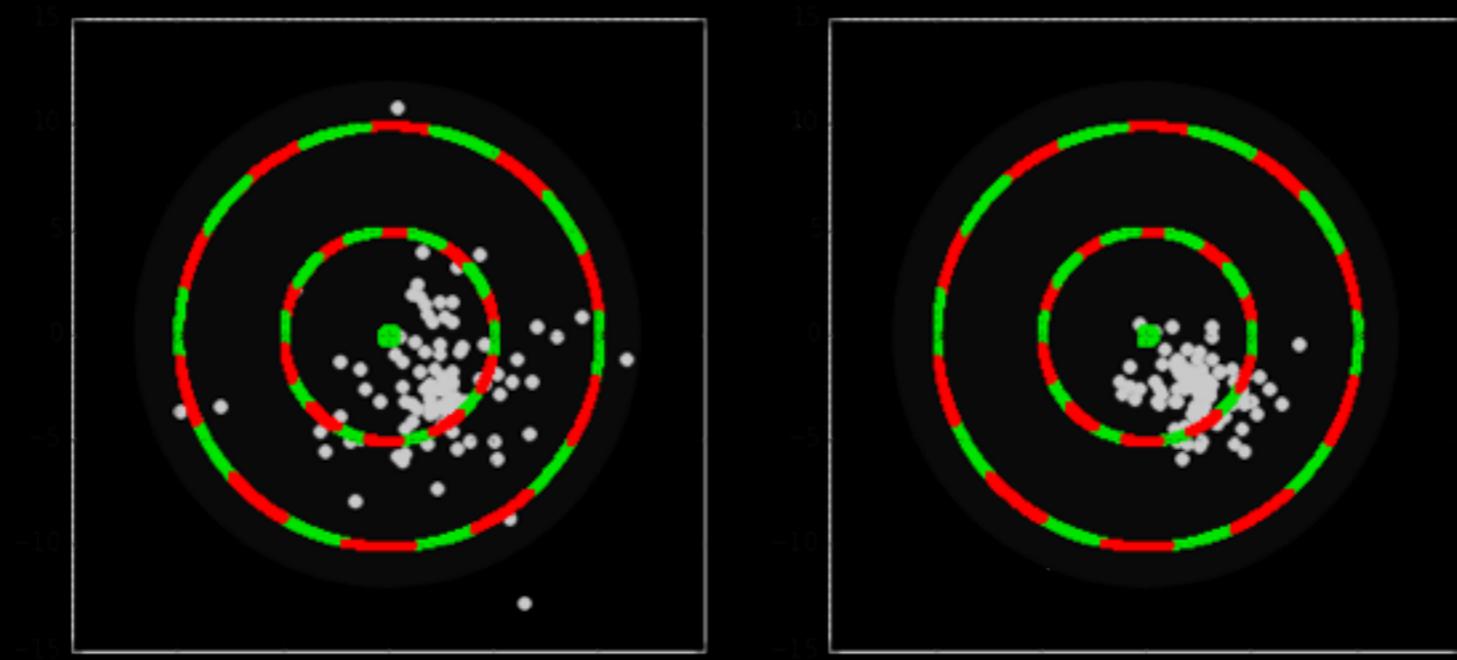
for large enough λ
a Poisson distribution $P(\lambda)$
parametrized by λ
tends to a
Gaussian distribution $N(\mu, \sigma)$
of mean $\mu = \lambda$ and standard deviation $\sigma = \sqrt{\lambda}$

$$P(\lambda) \xrightarrow{\lambda \rightarrow \infty} N(\lambda, \sqrt{\lambda})$$

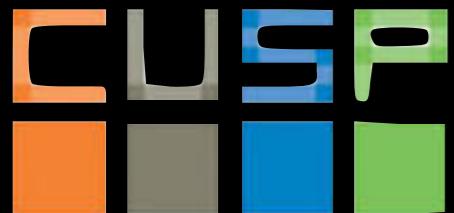
Systematic	Statistical
Biases the measurement in one direction	No preferred direction
Affects the sample regardless of the size	Shrinks with the sample size (typically as \sqrt{N})
Any distribution (usually we use Gaussian though)	Gaussian or Poisson



PRECISION



ACCURACY



IV: Statistical analysis

Error propagation

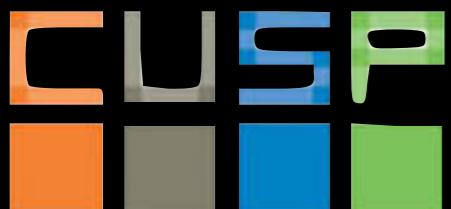
IID: Independent identically distributed:
add in quadrature for linear data operations

$$x_1 \pm \mathcal{E}(x_1)$$

$$x_2 \pm \mathcal{E}(x_2)$$

$$\bar{x} = \frac{x_1 + x_2}{2}$$

$$\mathcal{E}(\bar{x}) = \sqrt{\mathcal{E}(x_1)^2 + \mathcal{E}(x_2)^2}$$



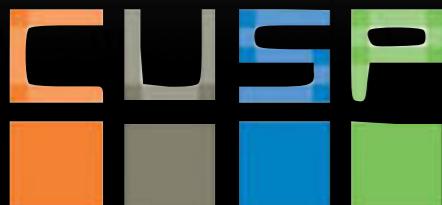
Function	Variance	Standard Deviation
$f = aA$	$\sigma_f^2 = a^2 \sigma_A^2$	$\sigma_f = a\sigma_A$
$f = aA + bB$	$\sigma_f^2 = a^2 \sigma_A^2 + b^2 \sigma_B^2 + 2ab \sigma_{AB}$	$\sigma_f = \sqrt{a^2 \sigma_A^2 + b^2 \sigma_B^2 + 2ab \sigma_{AB}}$
$f = aA - bB$	$\sigma_f^2 = a^2 \sigma_A^2 + b^2 \sigma_B^2 - 2ab \sigma_{AB}$	$\sigma_f = \sqrt{a^2 \sigma_A^2 + b^2 \sigma_B^2 - 2ab \sigma_{AB}}$
$f = AB$	$\sigma_f^2 \approx f^2 \left[\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 + 2 \frac{\sigma_{AB}}{AB} \right]$	$\sigma_f \approx f \sqrt{\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 + 2 \frac{\sigma_{AB}}{AB}}$
$f = \frac{A}{B}$	$\sigma_f^2 \approx f^2 \left[\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 - 2 \frac{\sigma_{AB}}{AB} \right]$ [11]	$\sigma_f \approx f \sqrt{\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 - 2 \frac{\sigma_{AB}}{AB}}$
$f = aA^b$	$\sigma_f^2 \approx (abA^{b-1}\sigma_A)^2 = \left(\frac{fb\sigma_A}{A} \right)^2$	$\sigma_f \approx abA^{b-1}\sigma_A = \left \frac{fb\sigma_A}{A} \right $
$f = a \ln(bA)$	$\sigma_f^2 \approx \left(a \frac{\sigma_A}{A} \right)^2$ [12]	$\sigma_f \approx \left a \frac{\sigma_A}{A} \right $
$f = a \log_{10}(A)$	$\sigma_f^2 \approx \left(a \frac{\sigma_A}{A \ln(10)} \right)^2$ [12]	$\sigma_f \approx \left a \frac{\sigma_A}{A \ln(10)} \right $
$f = ae^{bA}$	$\sigma_f^2 \approx f^2 (b\sigma_A)^2$ [13]	$\sigma_f \approx f(b\sigma_A) $
$f = a^{bA}$	$\sigma_f^2 \approx f^2 (b \ln(a)\sigma_A)^2$	$\sigma_f \approx f(b \ln(a)\sigma_A) $
$f = A^B$	$\sigma_f^2 \approx f^2 \left[\left(\frac{B}{A} \sigma_A \right)^2 + (\ln(A)\sigma_B)^2 + 2 \frac{B \ln(A)}{A} \sigma_{AB} \right]$	$\sigma_f \approx f \sqrt{\left(\frac{B}{A} \sigma_A \right)^2 + (\ln(A)\sigma_B)^2 + 2 \frac{B \ln(A)}{A} \sigma_{AB}}$

$\chi = \frac{\partial \chi}{\partial A} \approx \chi \left[\left(\frac{\partial \chi}{\partial A} \right)_{A_0 B} + (\chi(A_0 B))_{A_0 B} + \frac{\partial \chi}{\partial B} \right]_{A_0 B}$ $\chi \approx |\chi| \sqrt{\left(\frac{\partial \chi}{\partial A} \right)_{A_0 B}^2 + (\chi(A_0 B))_{A_0 B}^2 + \frac{\partial \chi}{\partial B}^2}$

https://en.wikipedia.org/wiki/Propagation_of_uncertainty#Linear_combinations

$$\chi = \sigma_{\rho \chi} \quad \frac{\partial \chi}{\partial \rho} \approx \chi \left(\rho \mu(\sigma) \chi \right)_{\rho}$$

$$\chi \approx |\chi| \left(\rho \mu(\sigma) \chi \right)$$



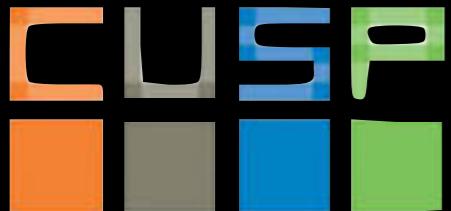
Covariance matrix

$$\xrightarrow{\hspace{1cm}} \mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_n\}$$

$$f_k = \sum_i^n A_{ki} x_i$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \dots & \dots & \dots & \dots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 \end{pmatrix}$$

$$\Sigma(f) = A \Sigma^x A^\top$$



Covariance matrix



$$\textbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_n\}$$

$$f_k = \sum_i^n A_{ki} x_i$$

IF
Independent
variables

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \dots & \dots & \dots & \dots \\ \sigma_{m1} & \sigma_{m2} & \dots & \sigma_m^2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_m^2 \end{pmatrix}$$

$$\Sigma(f) = A \Sigma^x A^\top \quad \Sigma(f)_{ij} = \sum_k^n A_{ik} \Sigma_k A_{jk}$$

Reporting Your Results

It is essential that the systematic error be reported separately from the imprecision part of the reported value

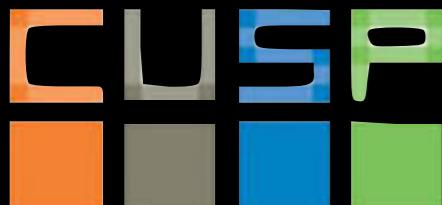
Statistical Concepts and Procedures
by United States. National Bureau of Standards 1969

Keep statistical, systematic errors separate. Report results as something like:

$$x = [965 \pm 30(\text{stat}) \pm 12(\text{sys})] \text{ number of car accidents}$$

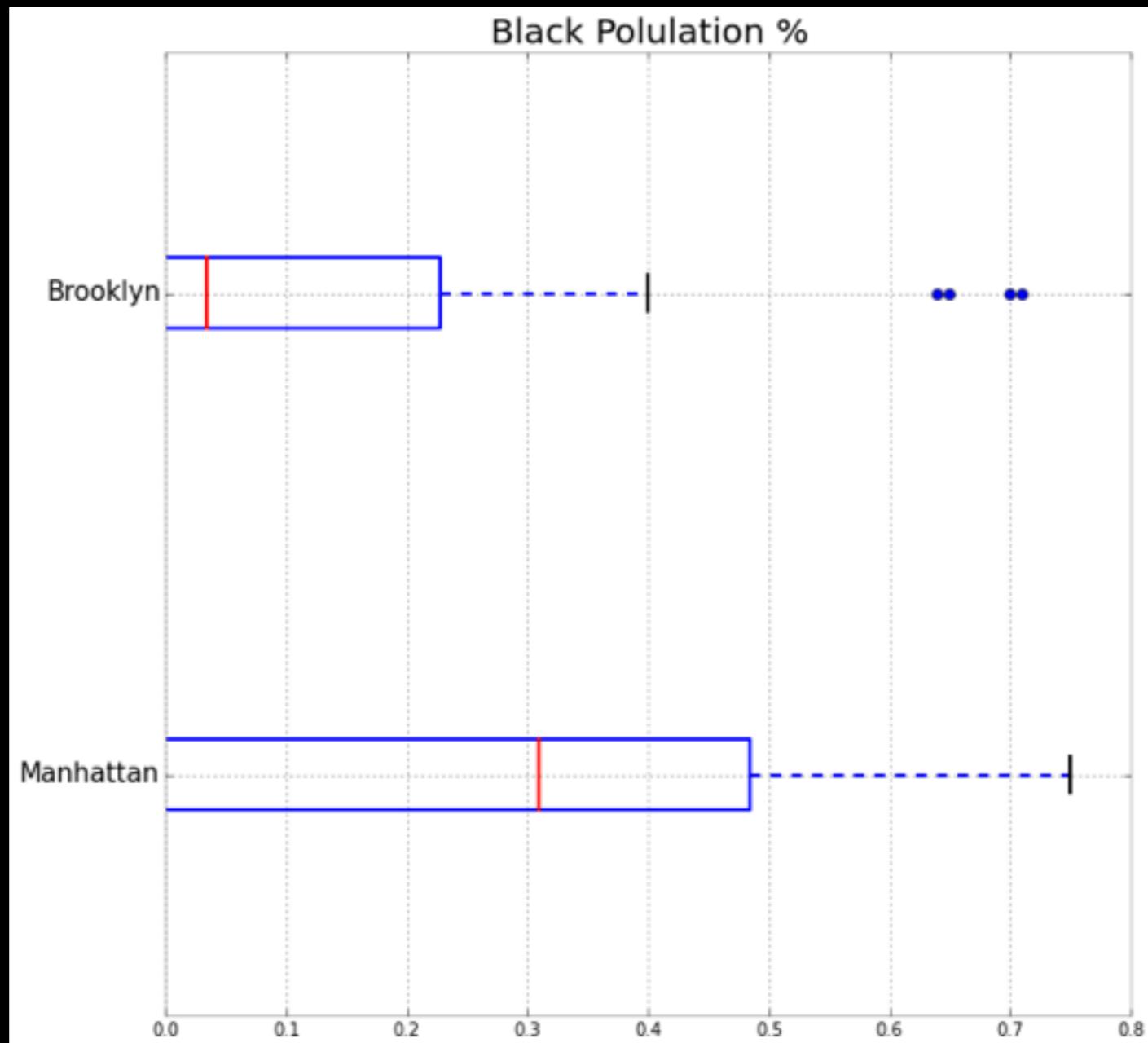
Add in quadrature (note that this assumes Gaussian distribution)
compare with known values $32 = \sqrt{30^2 + 12^2}$:

$$x = [965 \pm 32(\text{total})] \text{ number of car accidents}$$

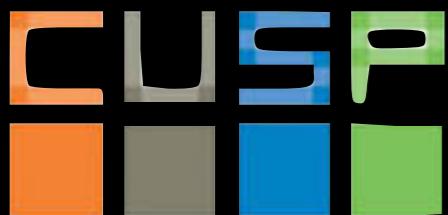


Reporting Your Results

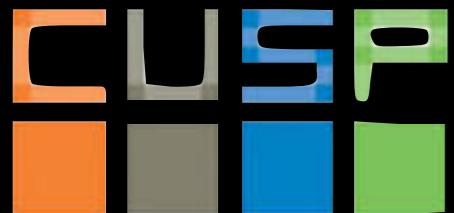
Percentage of Black population by Borrow
(Manhattan vs Brooklyn)



jupyter
[https://github.com/
fedhere/
UInotebooks/blob/
master/
black_percentage.ip
ynb](https://github.com/fedhere/UInotebooks/blob/master/black_percentage.ipynb)



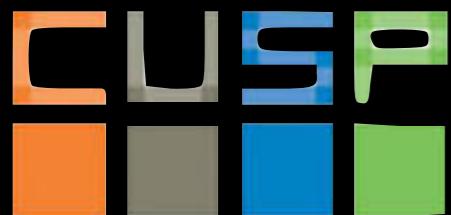
Goodness of fit



You have some data, and an idea of how it should look: a *model*

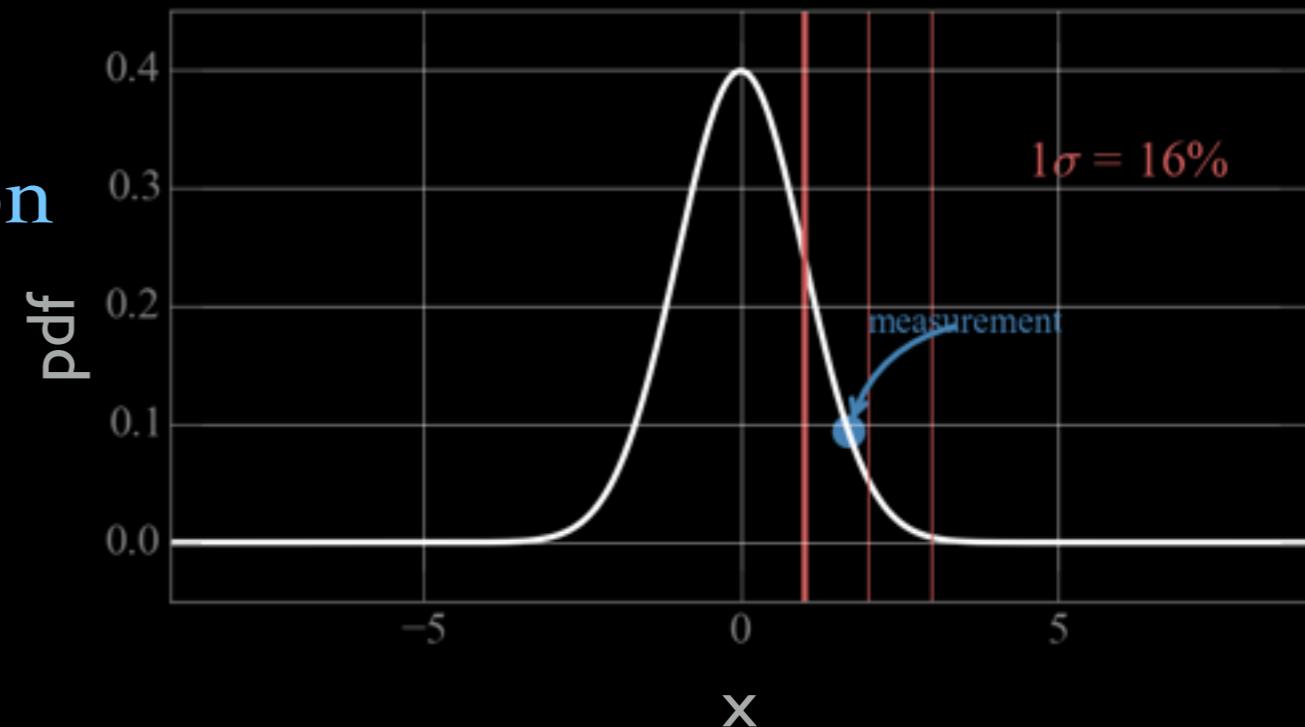
Is it a good model?

Goodness of fit



Probability Distribution Function

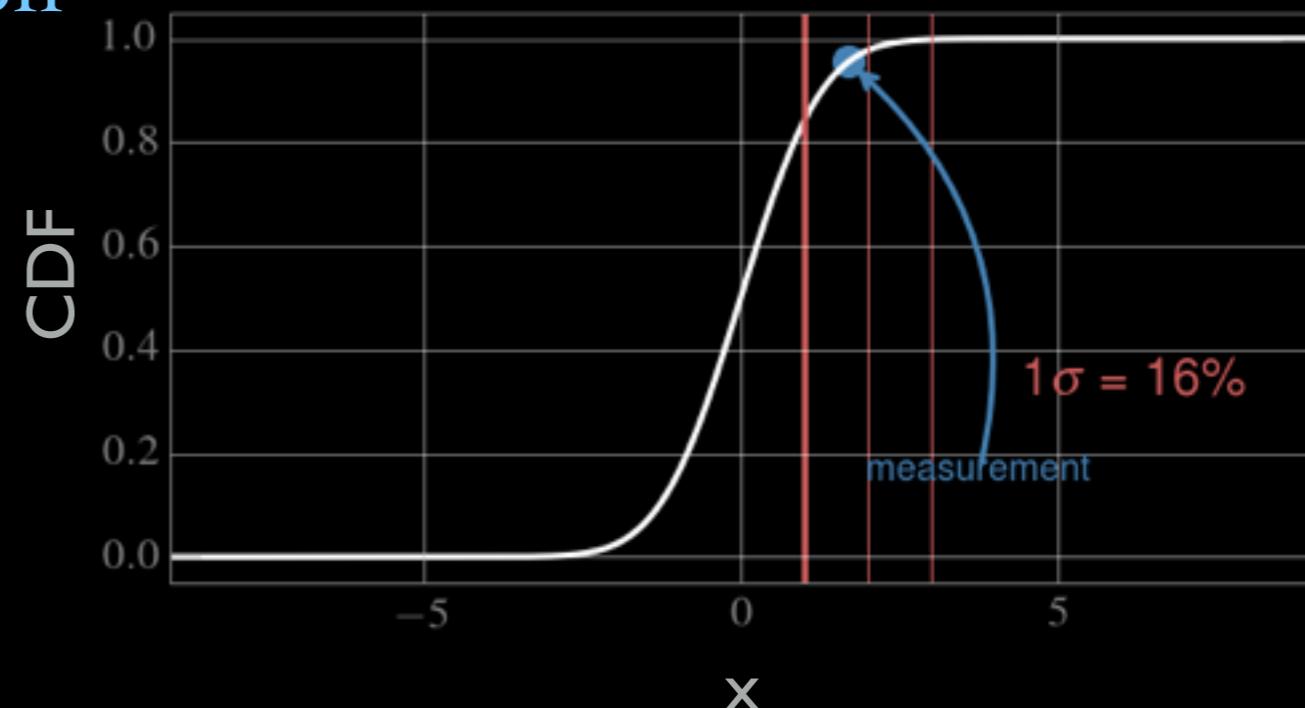
$$f_{x_0}(x) \sim p(x=x_0)$$

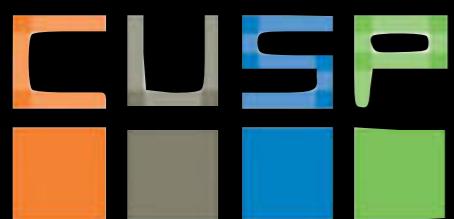
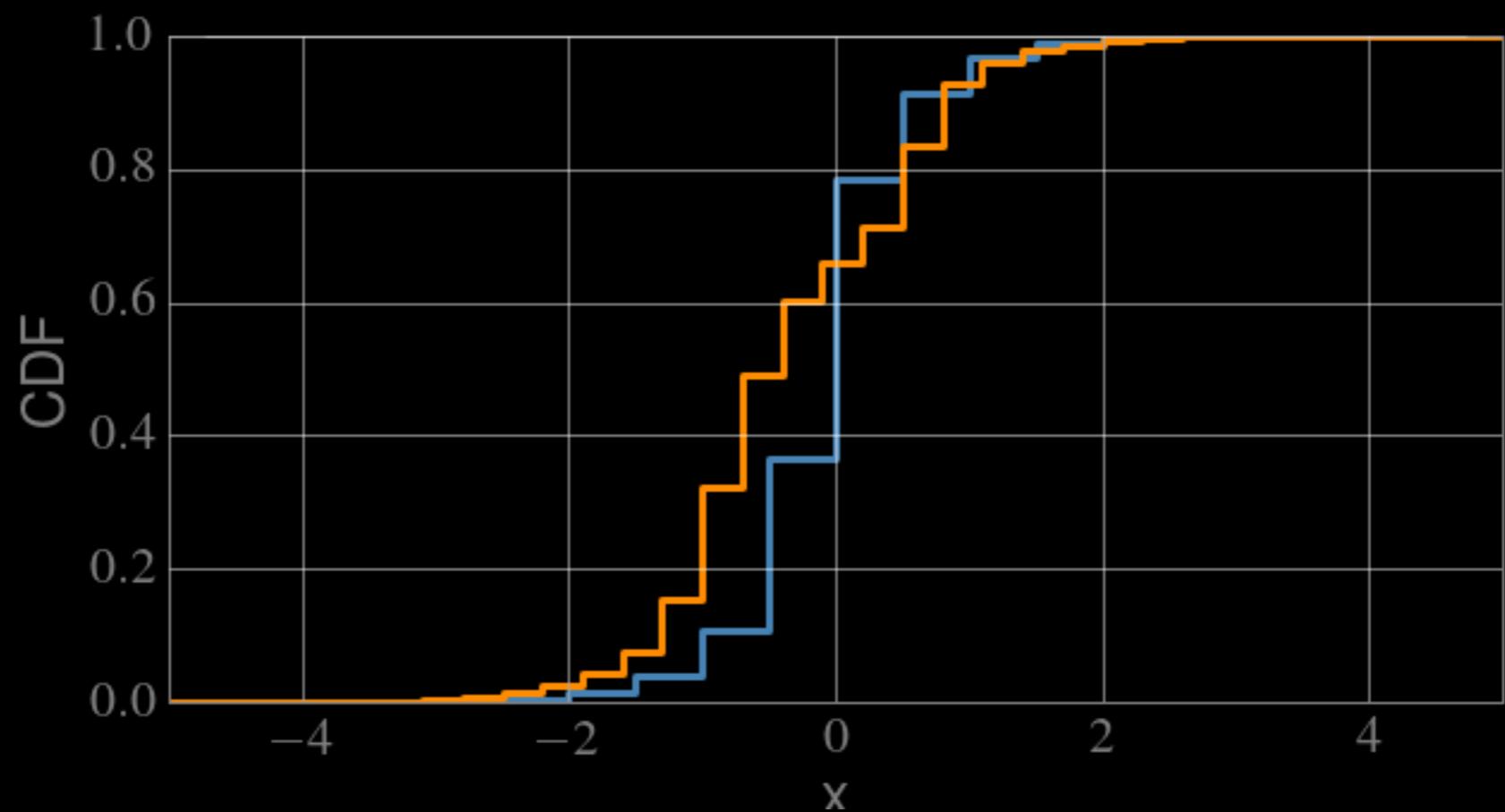
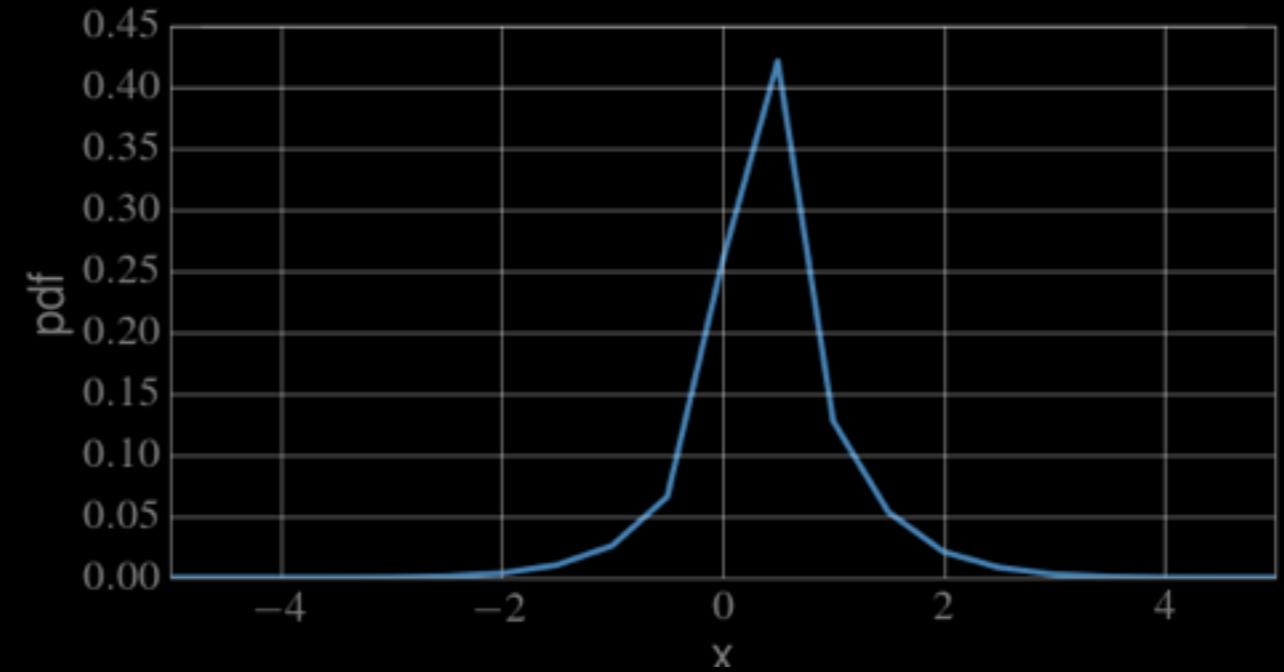
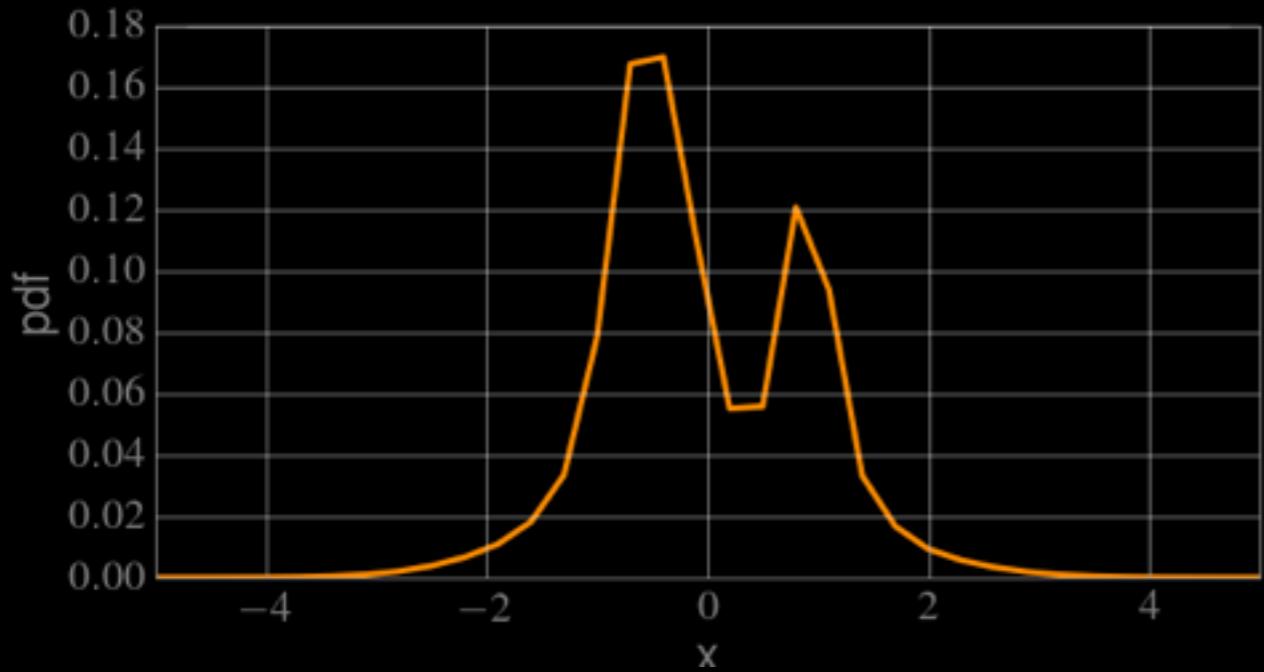


$$f_{x_0}(x) \sim p(x > x_0 - dx) \cap p(x < x_0 + dx)$$

Cumulative Distribution Function

$$F_{x_0}(x) = P(x < x_0)$$





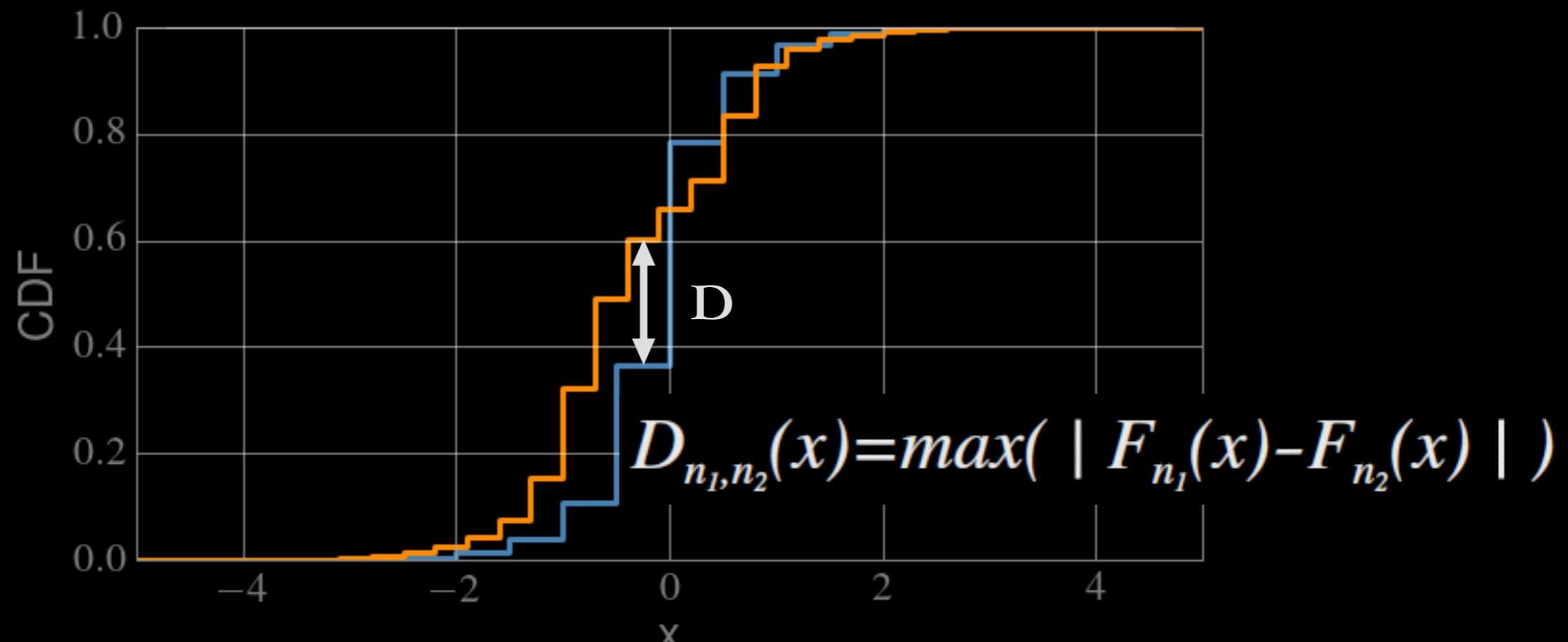
Two sample Kolmogorov Smirnoff test:

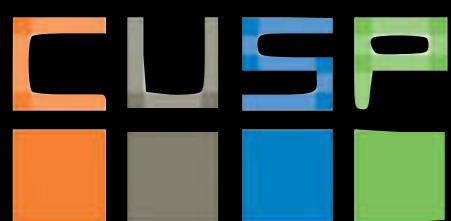
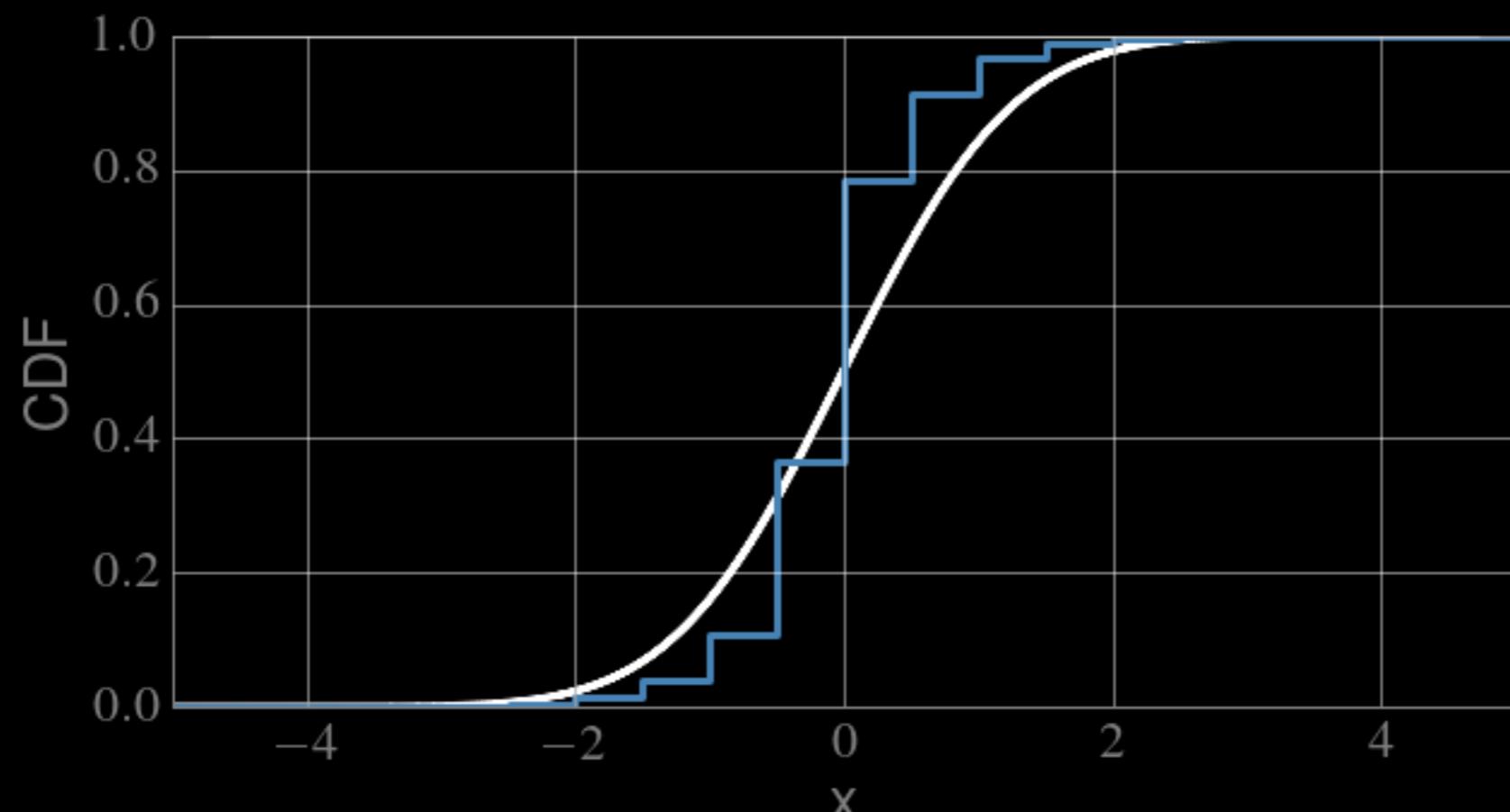
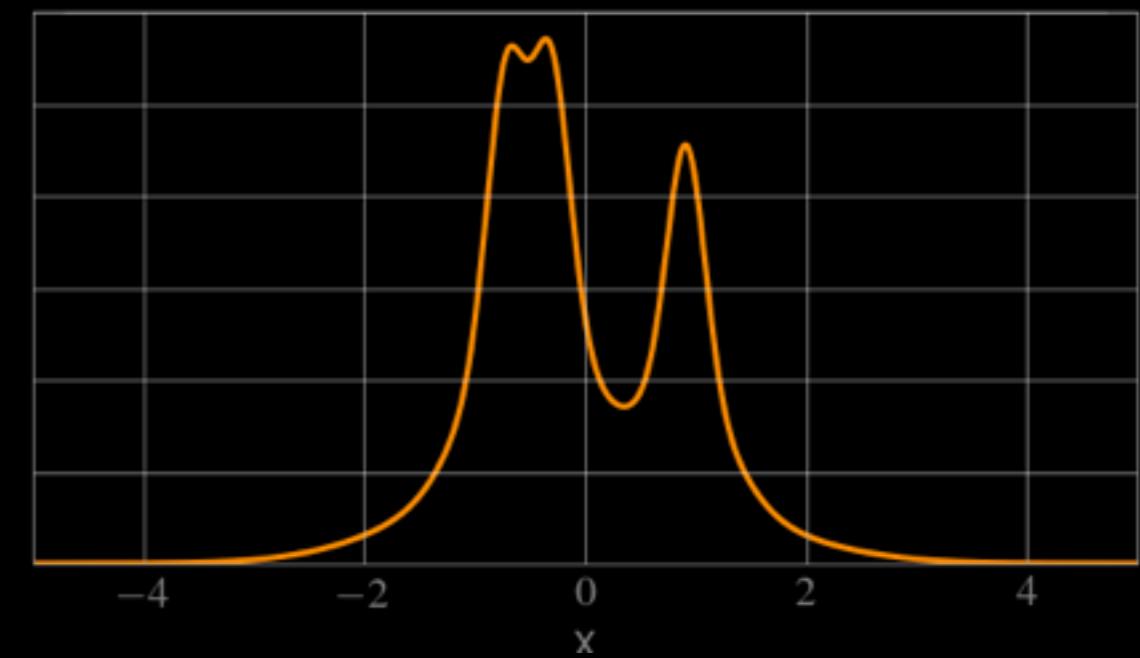
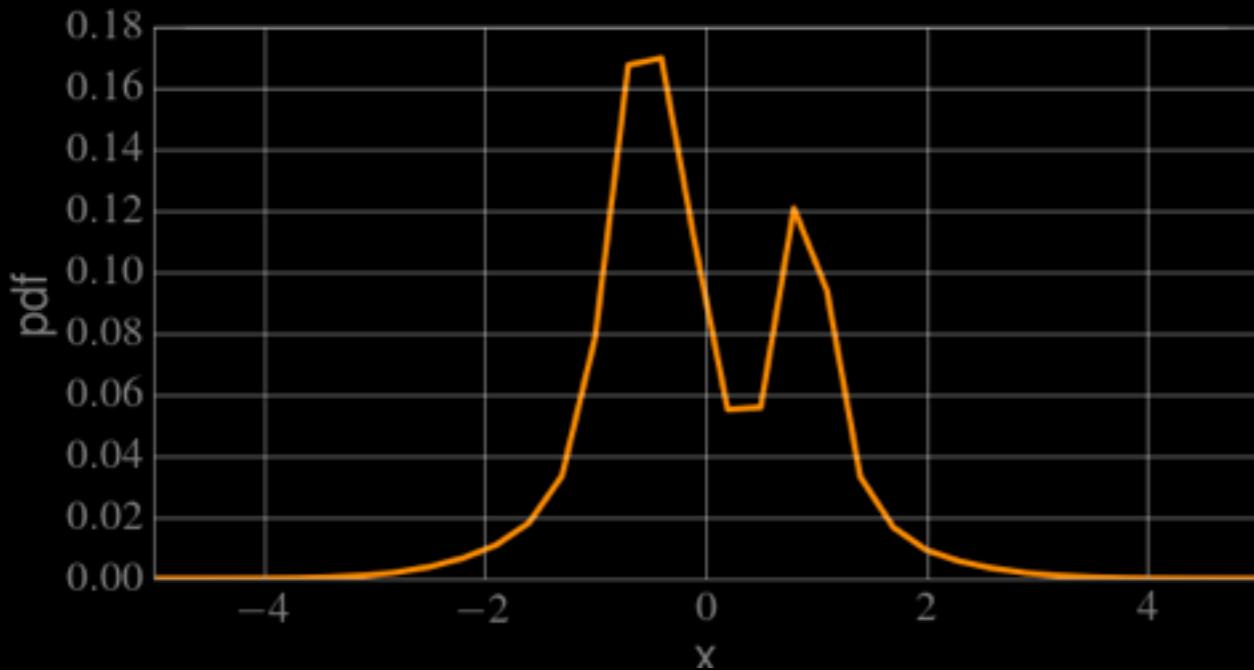
null hypothesis H_0 : the samples come from the same parent distribution

H_0 is rejected at level α if $D(n_1, n_2) > c(\alpha) \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$

with $c(\alpha)$ given by a table

NOTE: it ONLY works in 2D where the Euclidian distance is uniquely defined!

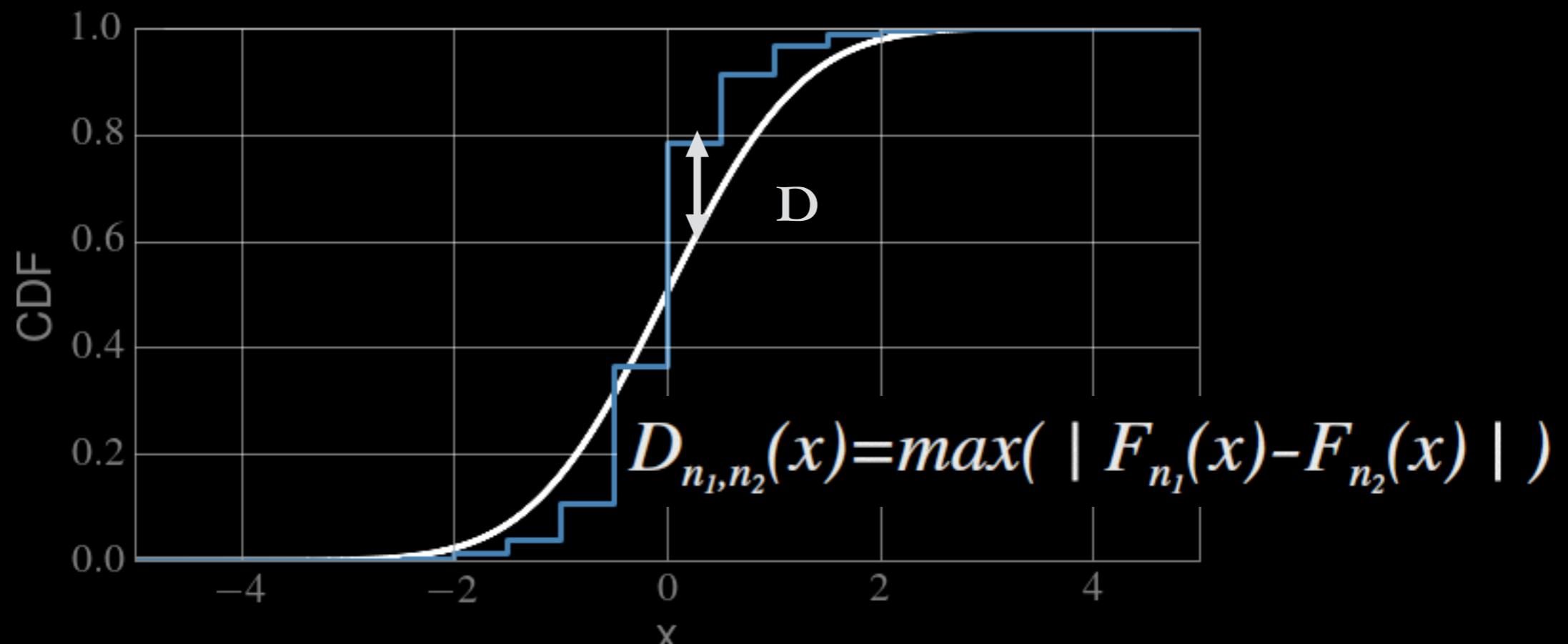


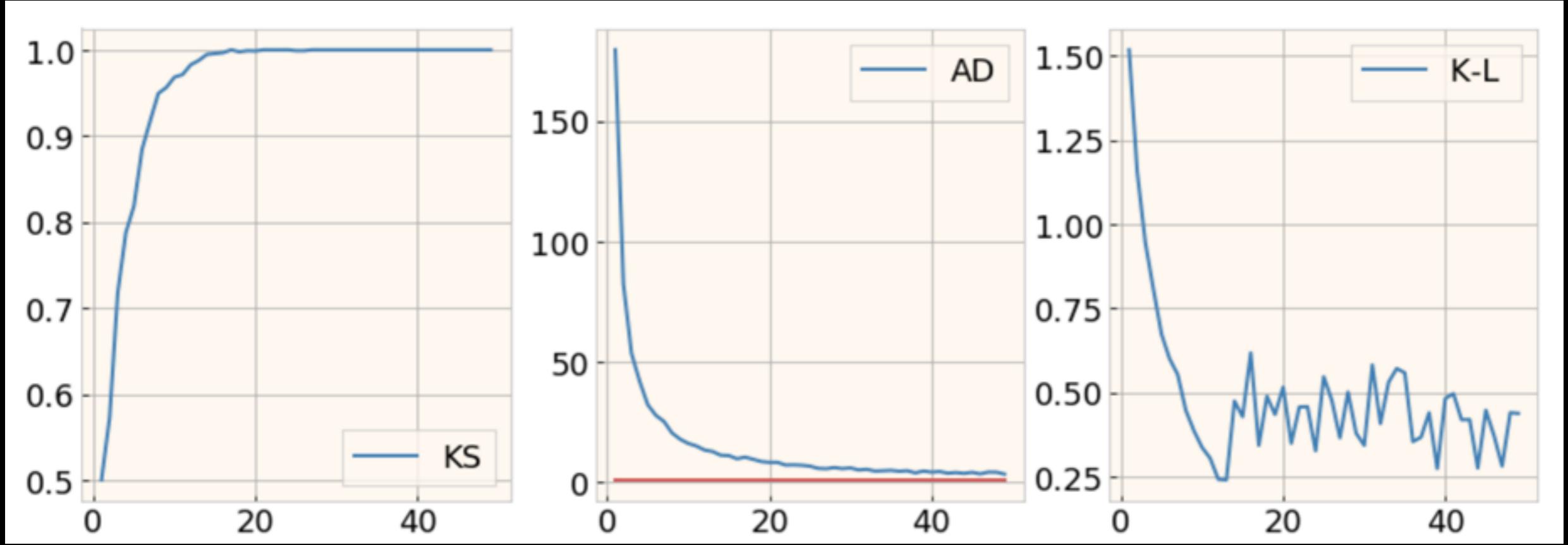


Goodness-of-fit Kolmogorov Smirnoff test:

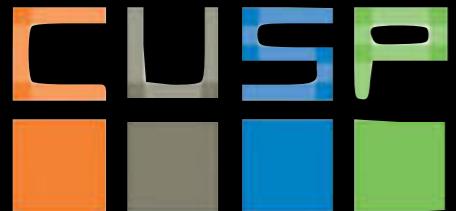
null hypothesis H_0 : the sample does comes from the model distribution

H_0 is rejected at level α if $\sqrt{n} D_n > K_\alpha$ where $P(K \leq K_\alpha) = 1 - \alpha$





jupyter



https://github.com/fedhere/UInotebooks/blob/master/oh_my_goodness_of_fit.ipynb

IV: Statistical analysis

Compare Tests for Correlation and Goodness of fit:

The following are 3 tests that assess correlation between 2 samples:

- Pearson's test e.g. Estimate number of MTA Bus passengers at different hours
- Spearman's test (morning, afternoon, or in time chunks as 7:30-10:30, 10:30-1:30, 1:30-3, 3:6, 6:9, you can do it per bus line, per origin or destination neighborhood...)
- K-S test

The following are 5 tests that can be used to assess the goodness of fit of a model

- K-S
- Pearson's Chi squared
- Anderson-Darling e.g. Estimate number of MTA Bus passengers per bus line within an interval of time: are the passengers randomly distributing on busses.
- K-L Divergence
- Likelihood ratio

In HW3 you used 2 out of these tests to assess if 2 samples are related (measure their correlation, or decide if they come from the same parent distributions)

Now in HW4 use 2 of the goodness of fit tests to see if a dataset comes from a normal distribution, or from another distribution (where possible) of your choice.

Compare Tests for Correlation and Goodness of fit:

The following are 3 tests that assess correlation between 2 samples:

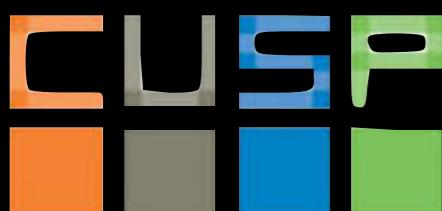
- Pearson's test e.g. Age distribution of male vs female Citibikes riders. Age
- Spearman's test distribution in different seasons. Age distribution for long/short
- K-S test trips

The following are 5 tests that can be used to assess the goodness of fit of a model

- K-S
 - Pearson's Chi squared
 - Anderson-Darling
 - K-L Divergence
 - Likelihood ratio
- e.g. Estimate Age of riders: could be Gaussian, could be lognormal, power law, some bimodal distribution...

In HW3 you used 2 out of these tests to assess if 2 samples are related (measure their correlation, or decide if they come from the same parent distributions)

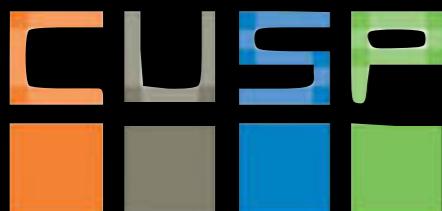
Now in HW4 use 2 of the goodness of fit tests to see if a dataset comes from a normal distribution, or from another distribution (where possible) of your choice.



Homework: 1. Compare Tests for Goodness of fit (real data)

Test whether a gaussian model for the age distribution of citibike drivers is a sensible model, or if you can find a better fit with another distribution. Use 2 tests: KS, AD, KL, chisq to do this. Test at least 2 distributions.

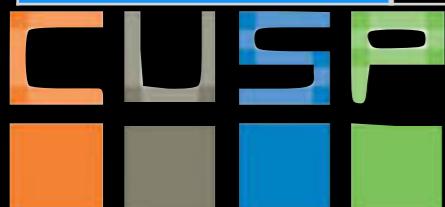
Optional (extra credit): Divide your sample geographically: by Borrow + split Manhattan in an Uptown and a Downtown sample (use your discretion to do so) and see if you notice any differences in how the age distribution can be modeled.



Tests Cheat Sheet:

2 (+) samples comparison

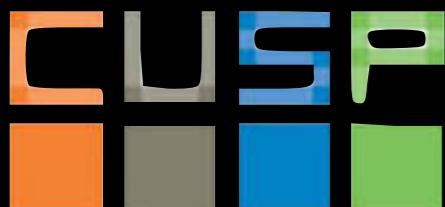
	metric (statistic)	compare to	
KS	$D_{n_1, n_2}(x) = \max(F_{n_1}(x) - F_{n_2}(x))$	$c(\alpha) \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$	Non parametric 2 samples only
K-sample Anderson-Darling	$ADK = \frac{n-1}{n^2(k-1)} \sum_{i=1}^k \frac{1}{n(i)} \left(\sum_{j=1}^L h_j \frac{(nF_{ij} - n_i H_j)^2}{H_j(n-H_j) - nh_j/4} \right)$	• AK table	Non parametric, N samples
Pearson's	$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$	The interpretation of a correlation coefficient depends on the context and purpose	-1 anticorrelated 0 uncorrelated 1 correlated .
Spearman's	$\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{n(n^2 - 1)}$	t test $t = r \sqrt{\frac{n-2}{1-r^2}}$	ranked data only p-value from t-test, Fisher's transformation +z score, permutation test



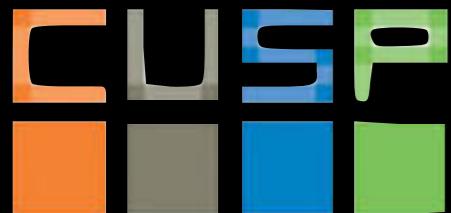
Tests Cheat Sheet:

goodness of fit

	metric (statistic)	compare to	
KS	$D_{n_1, n_2}(x) = \max(F_n(x) - F(x))$	$\frac{K_\alpha}{\sqrt{n}}$	power in the core only
Pearson's chi square	$\chi^2_{red} = \frac{\chi^2}{df} = \frac{1}{df} \sum \frac{(O-E)^2}{\sigma^2}$	scipy.stats.chisquare(f_obs, f_exp=None, ddof=0, axis=0)[0]	
Anderson-Darling	$A = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x) (1-F(x))} dF(x)$	scipy.stats.anderson(x, dist='norm')	power in the tails
K-L divergence	$D_{KL} = - \int_x p(x) \log(q(x)) + p(x) \log(p(x))$	scipy.stats.entropy(pk, qk=<not None>)	relates to information entropy
Likelihood ratio	$\frac{L(\text{model 1} \text{data})}{L(\text{model 2} \text{data})}$		suitable to bayesian analysis

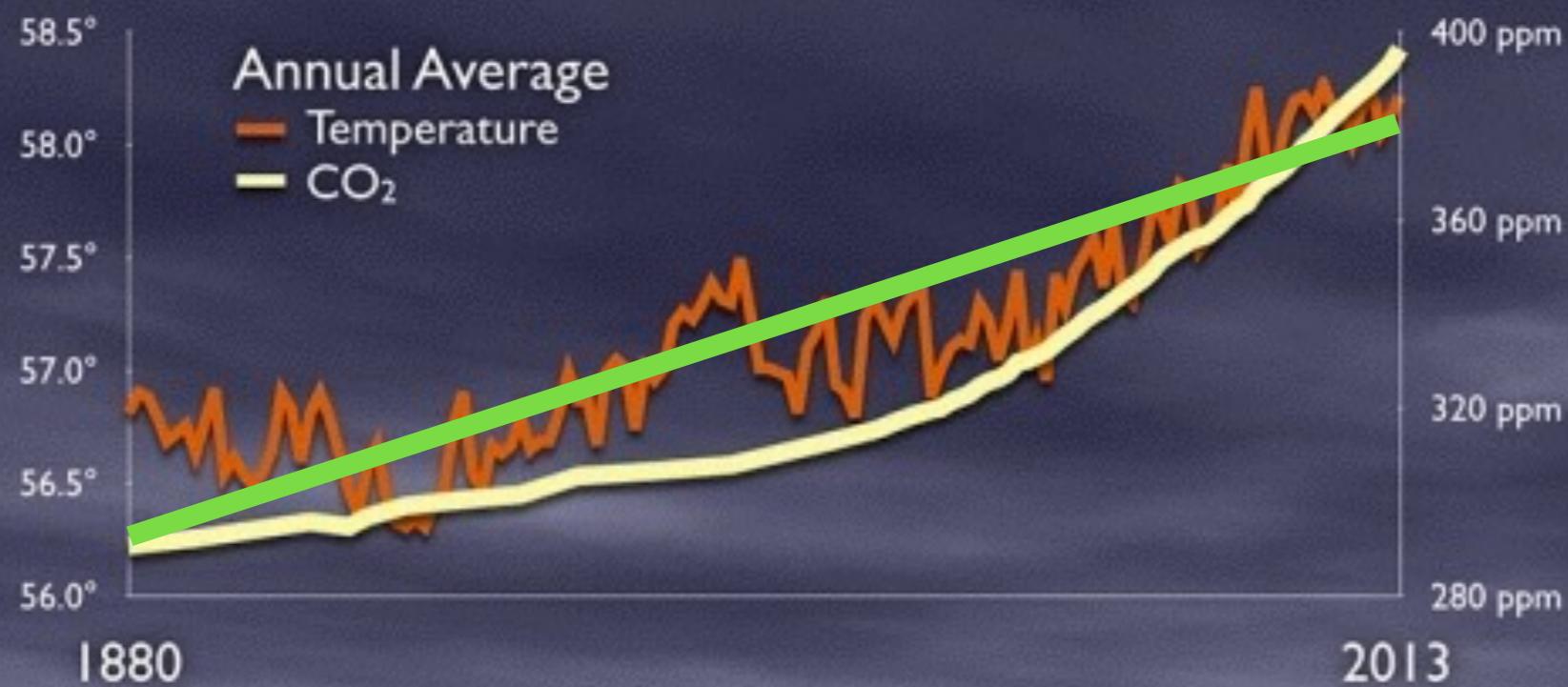


Predictive



V: Likelihood and
Regression Models

Global Temperature and CO₂

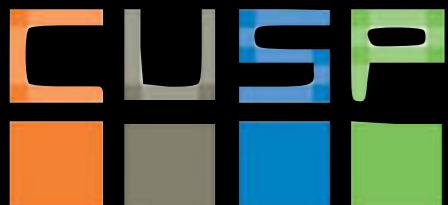


Source: National Climate Assessment 2014

CLIMATE CO₂ CENTRAL

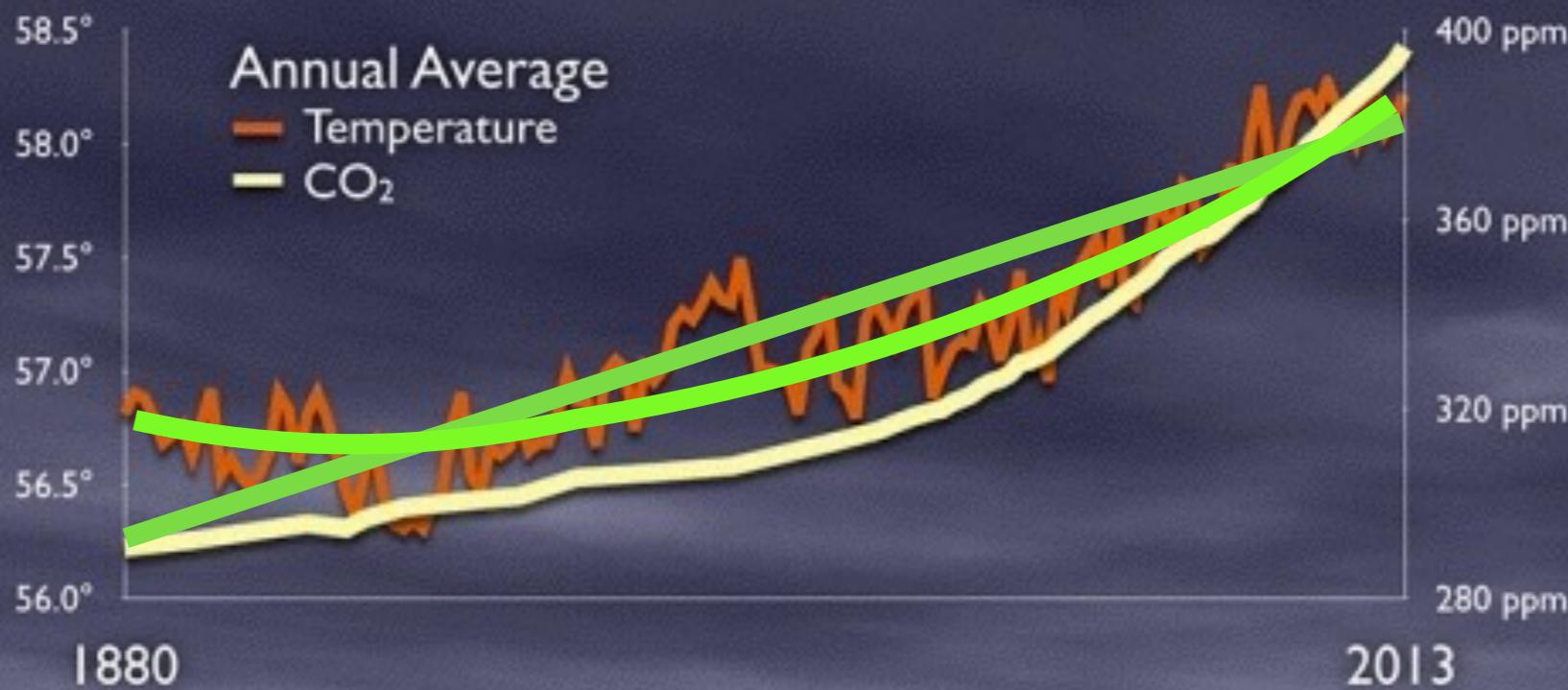
1880 2013
1905 Annualized Global Mean Temperature (°C)

CLIMATE CO₂ CENTRAL



V: Likelihood and
Regression Models

Global Temperature and CO₂

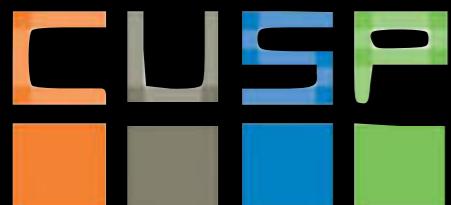


2013 Annual Average Global Temperature (Source: GISS)

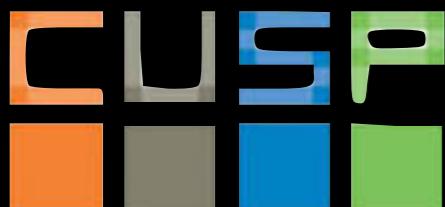
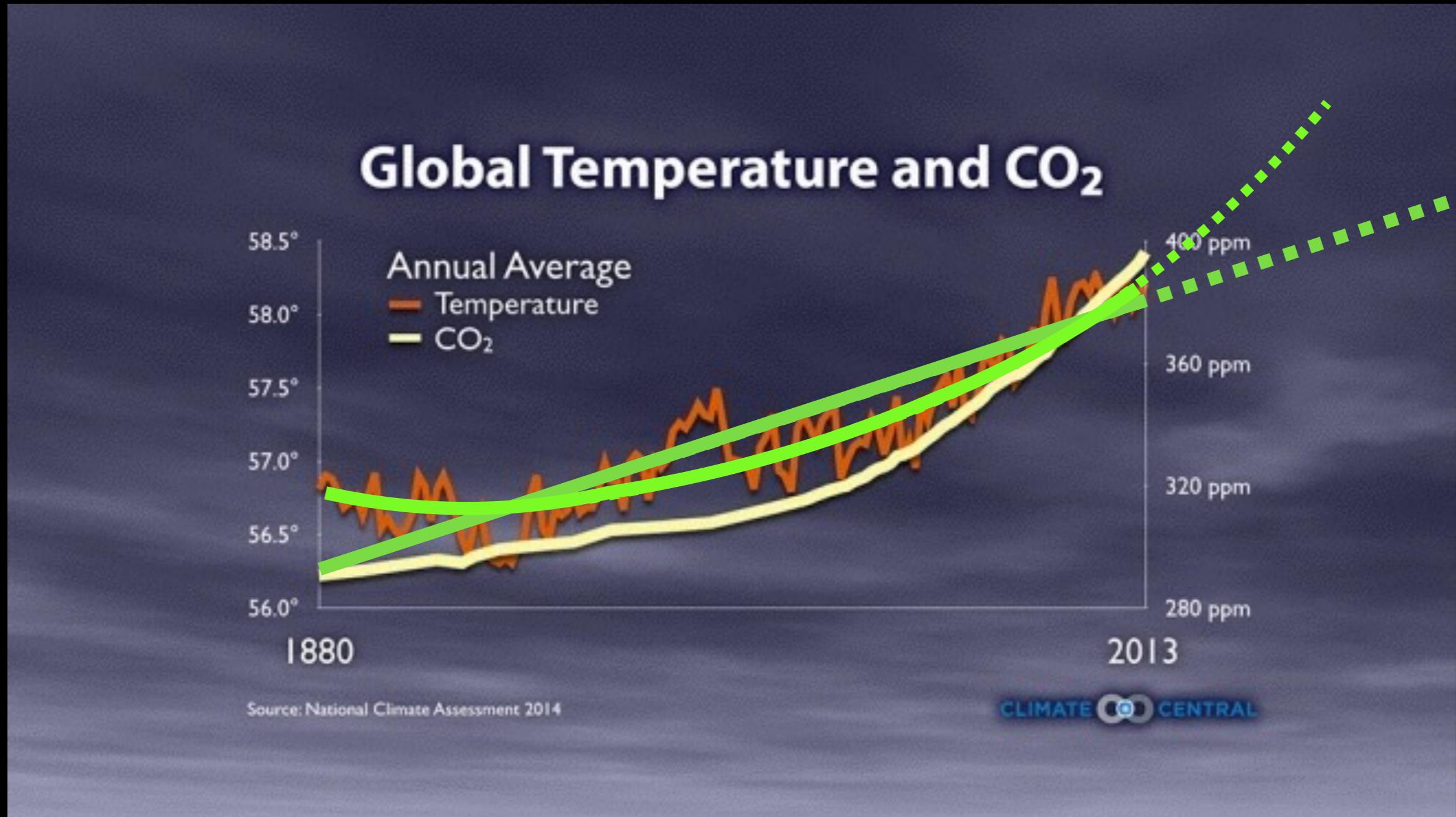
1880

2013 Annual Average Global Temperature (Source: GISS)

2013



V: Likelihood and
Regression Models



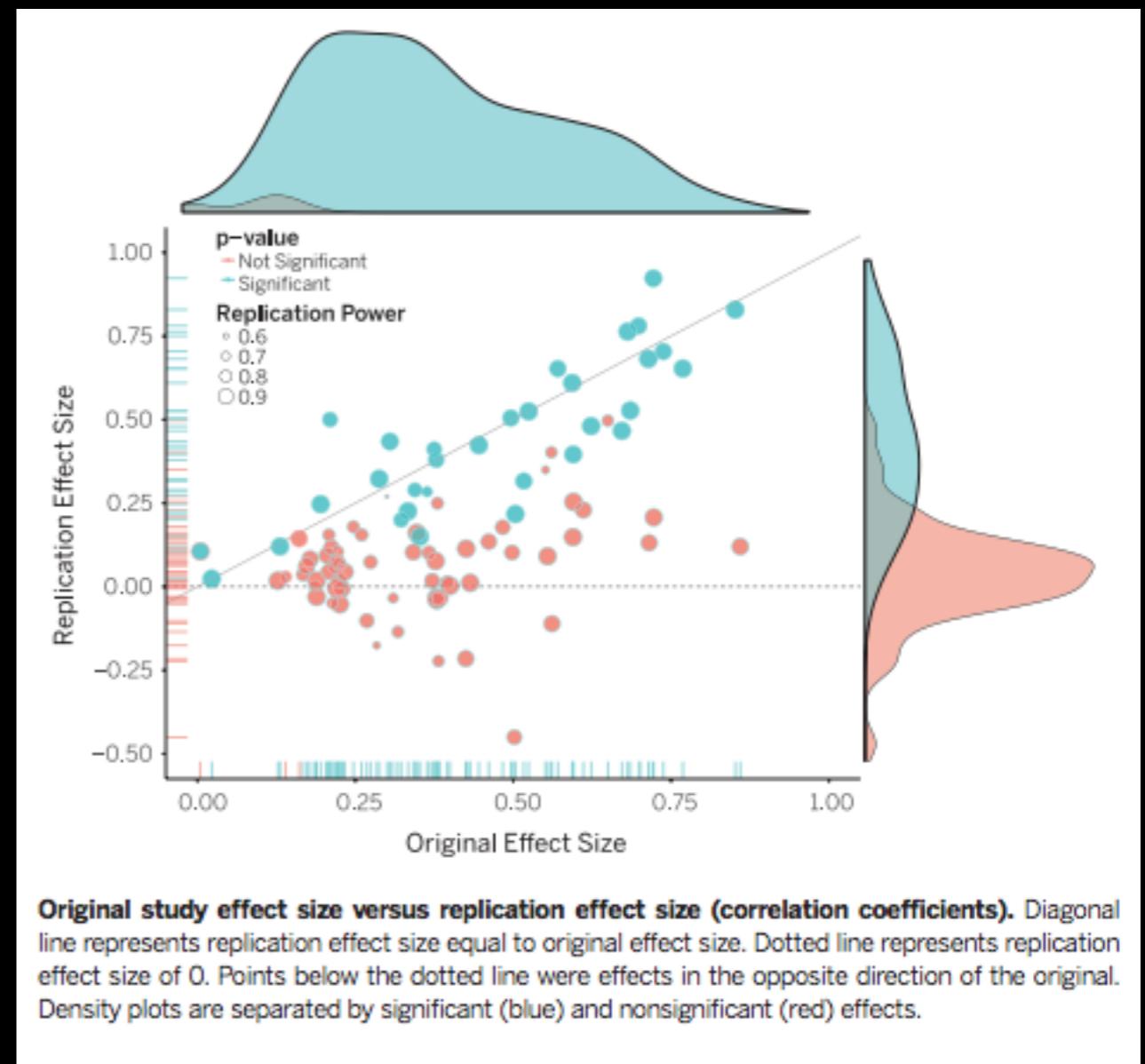
V: Likelihood and
Regression Models

RESEARCH ARTICLE SUMMARY

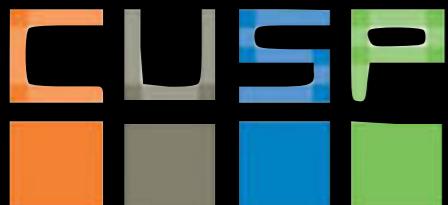
PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

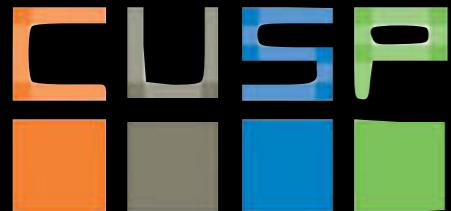


<http://www.sciencemag.org/content/349/6251/aac4716.full.pdf>



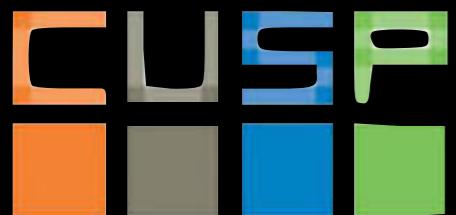
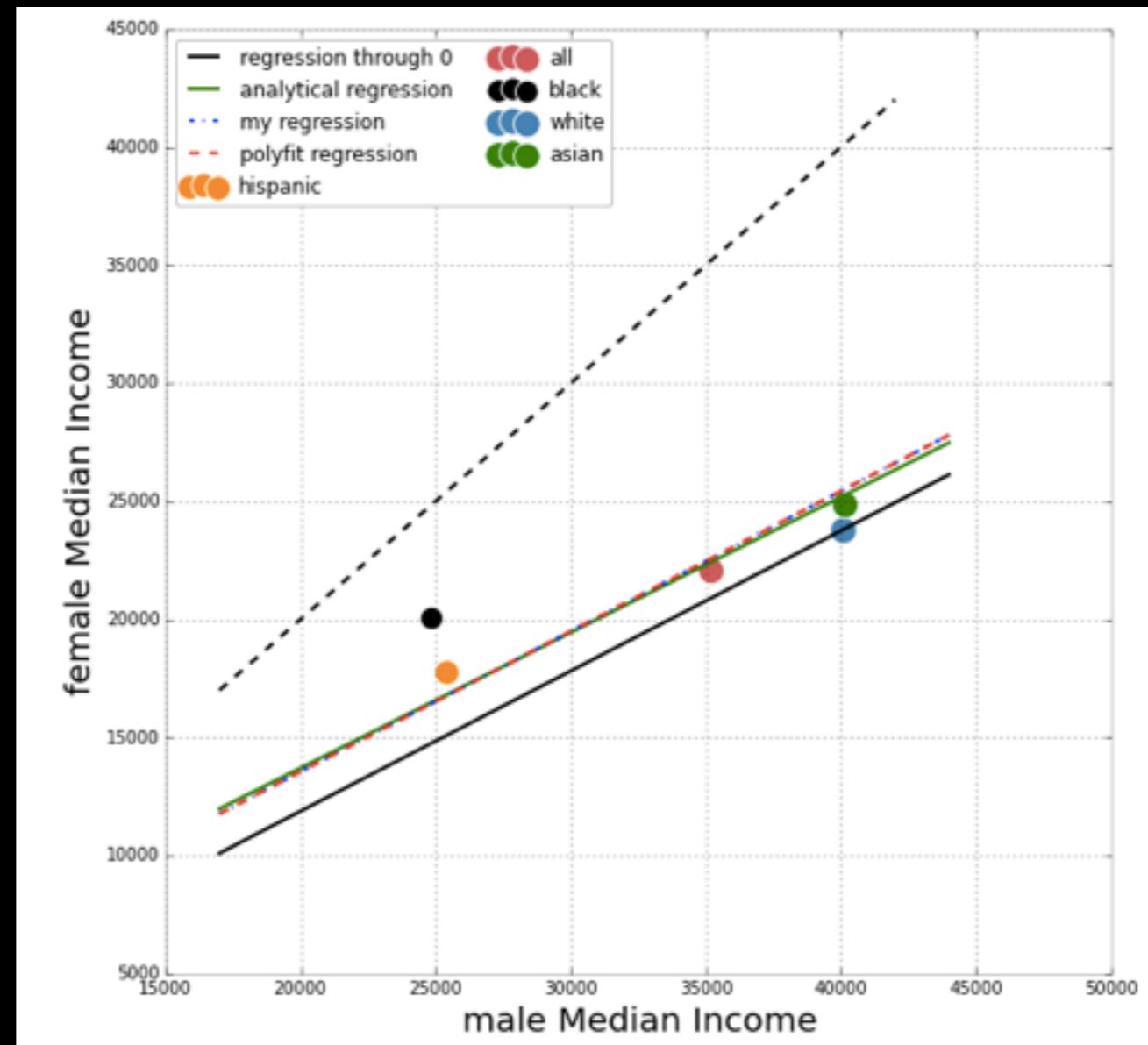
V: Likelihood and
Regression Models

Why?



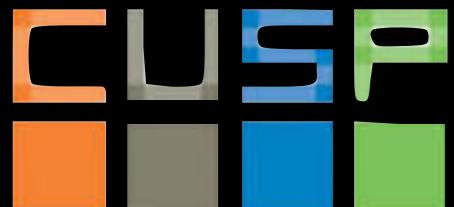
V: Likelihood and
Regression Models

jupyter



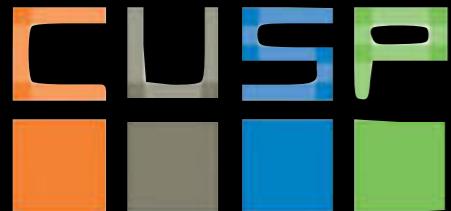


[https://github.com/fedhere/UInotebooks/blob/master/OLS/
line_fit_and_residuals.ipynb](https://github.com/fedhere/UInotebooks/blob/master/OLS/line_fit_and_residuals.ipynb)

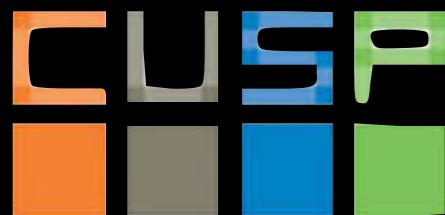
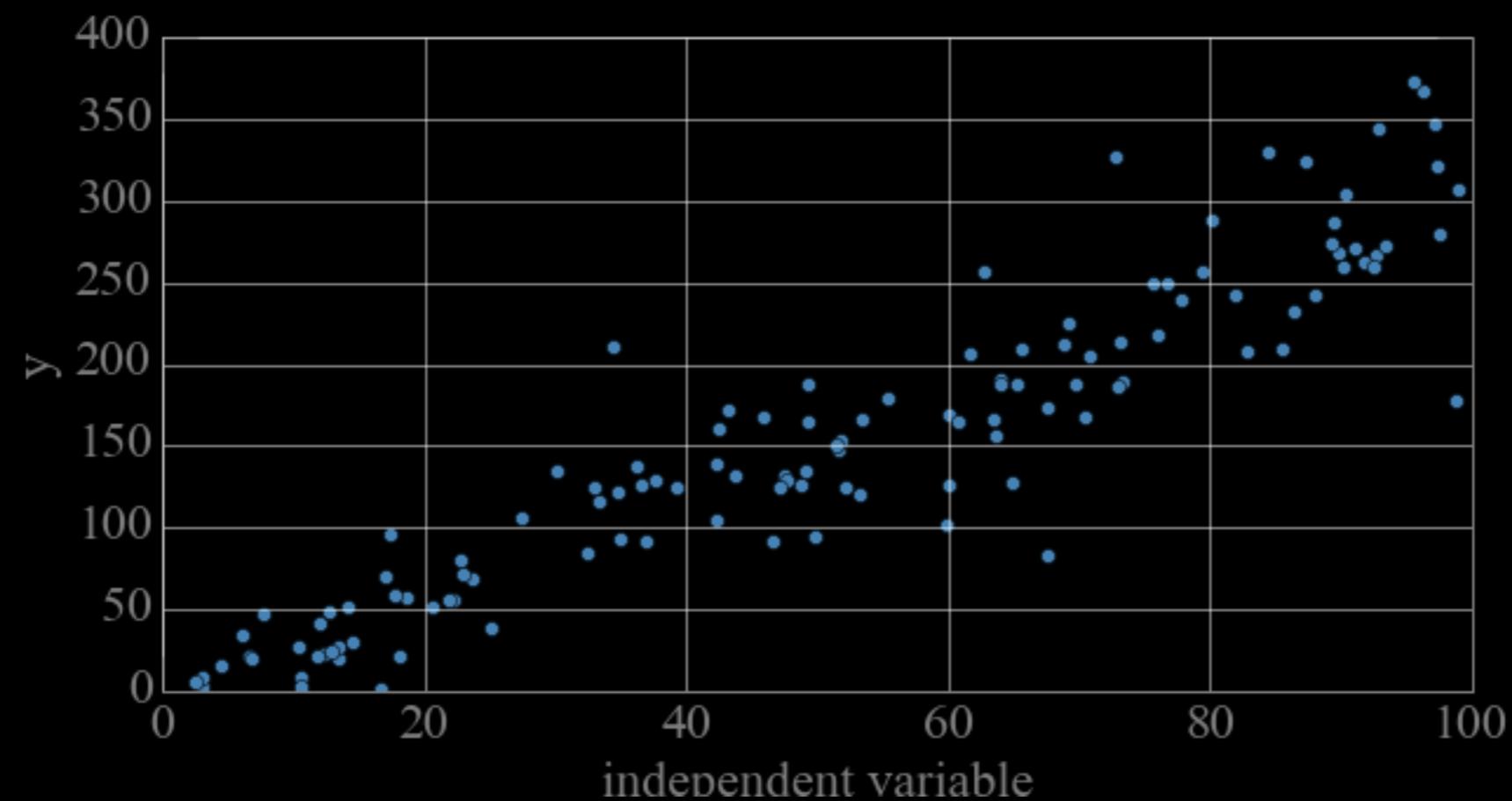


V: Likelihood and
Regression Models

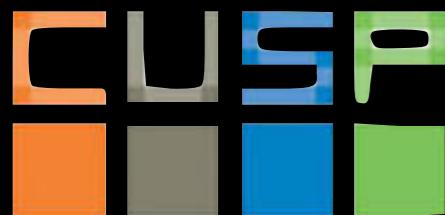
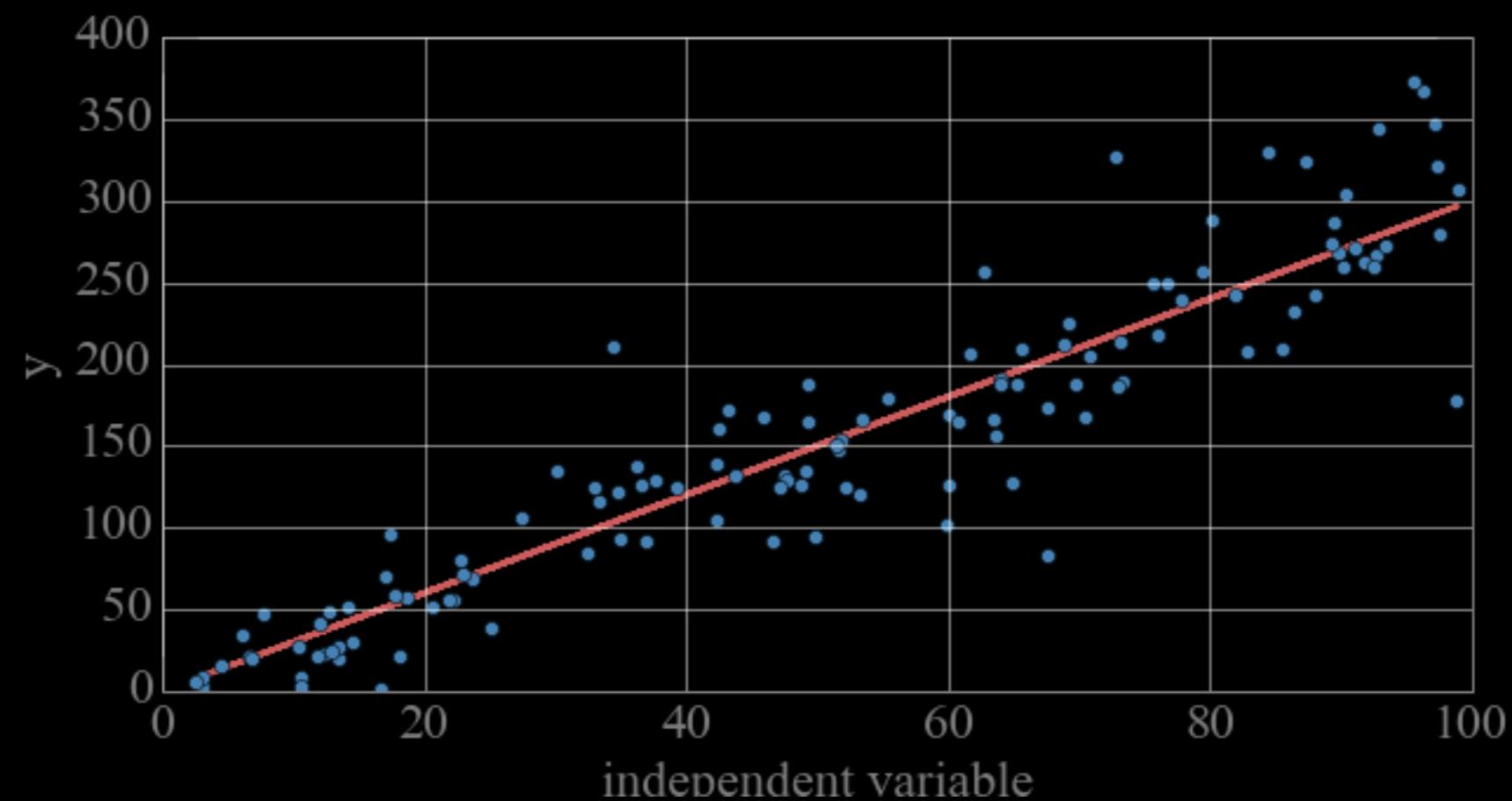
How?



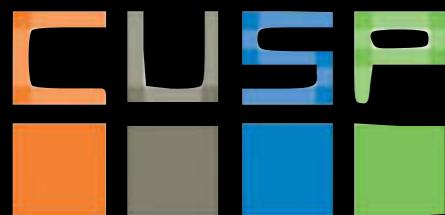
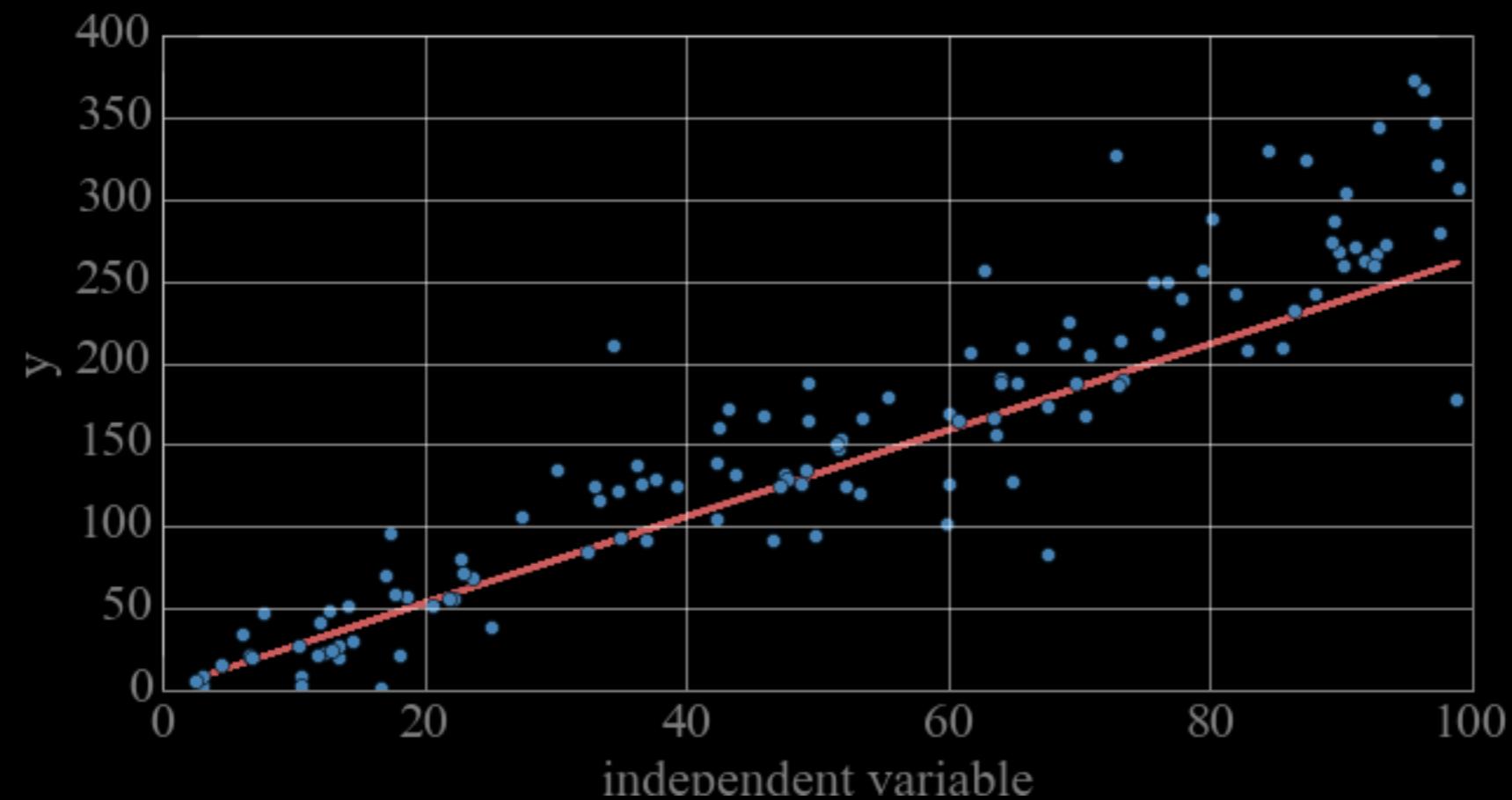
V: Likelihood and
Regression Models

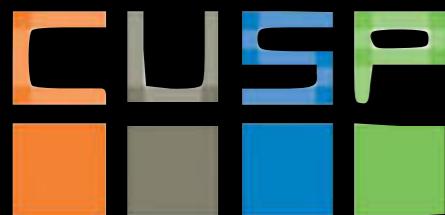
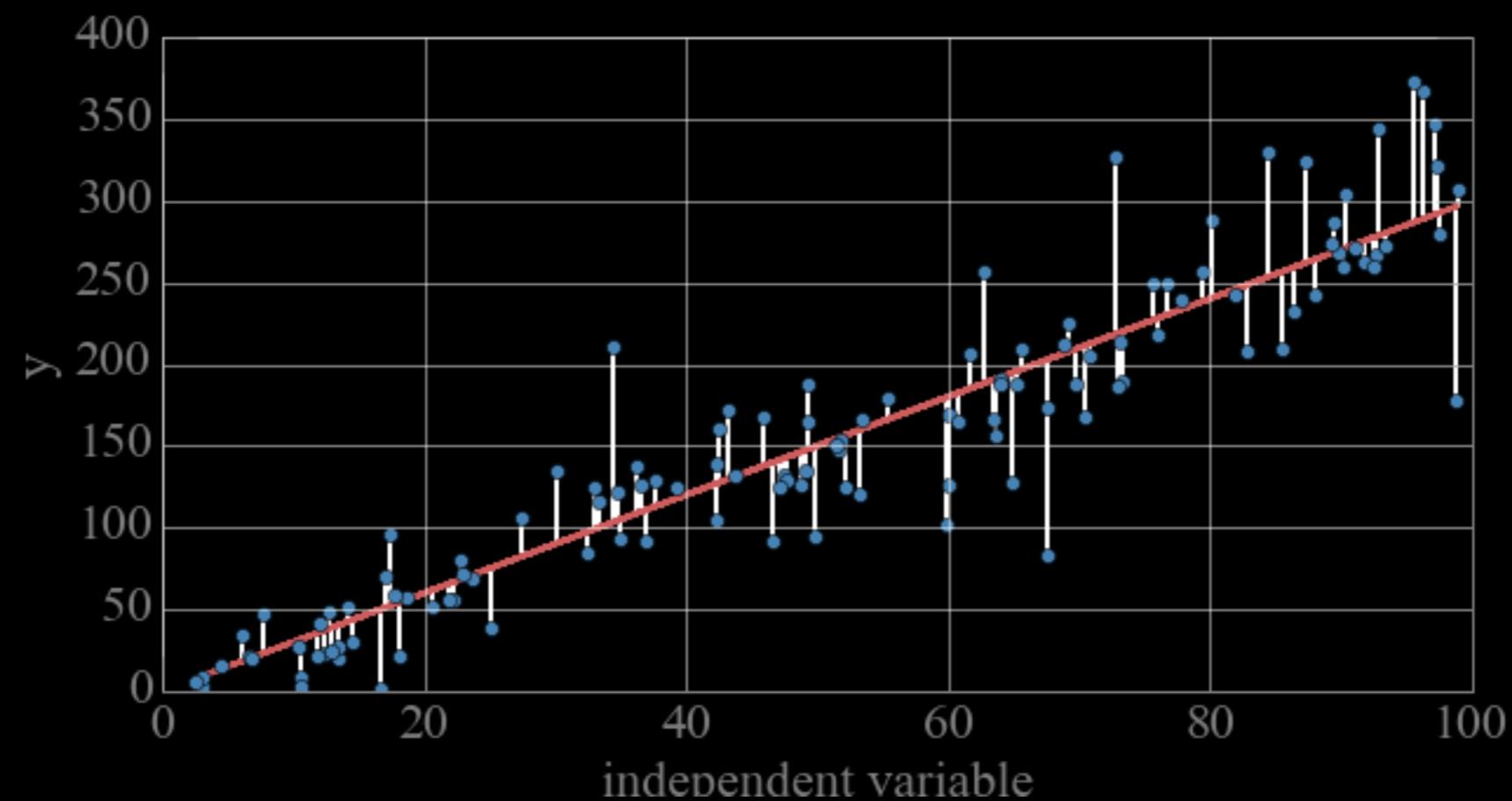


V: Likelihood and
Regression Models

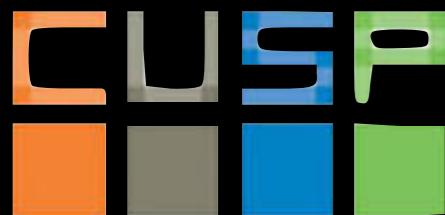
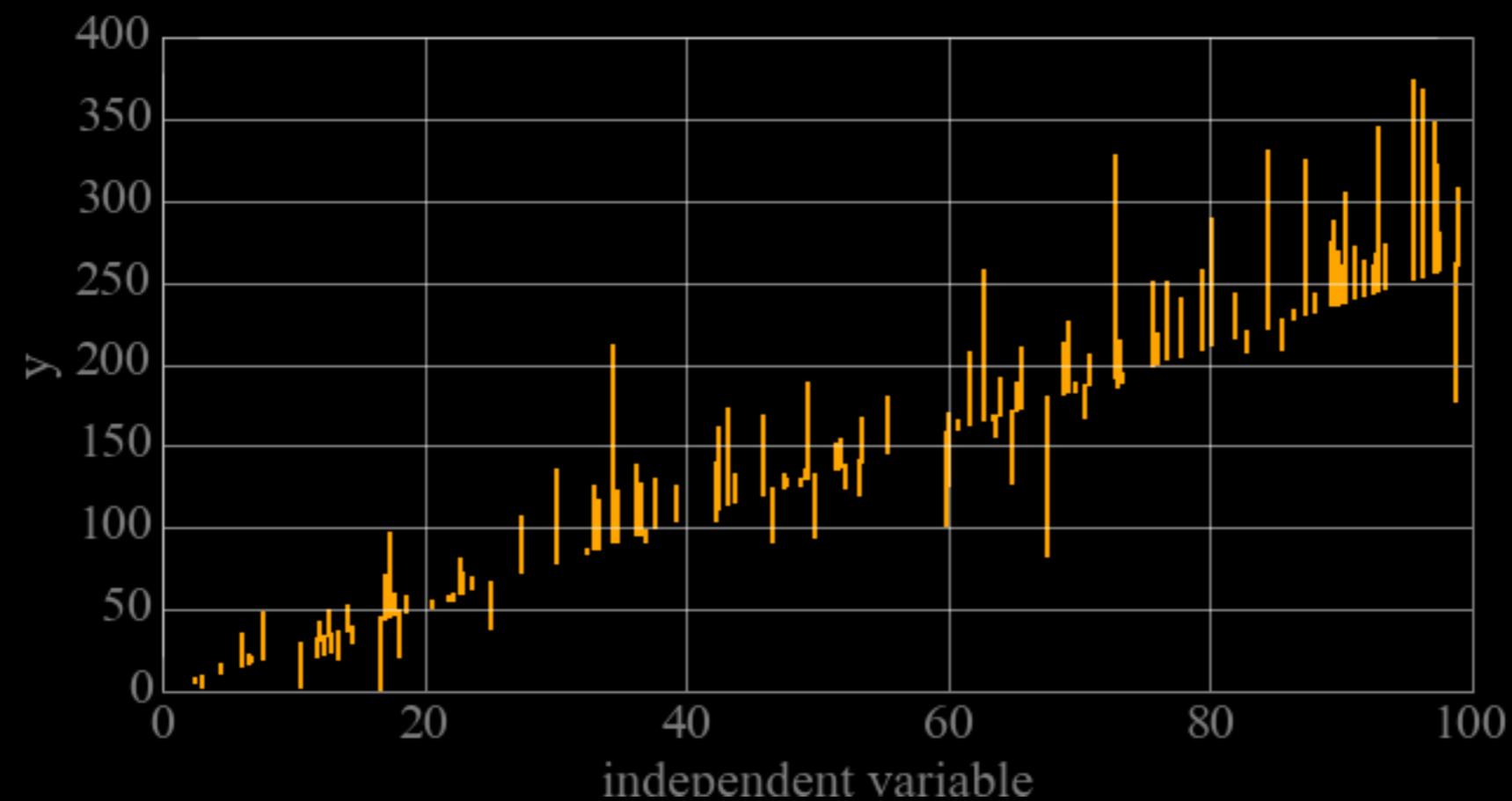


V: Likelihood and
Regression Models

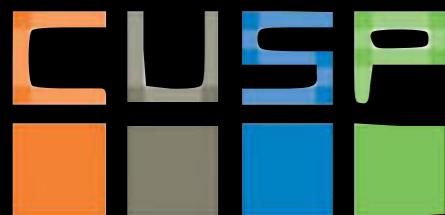
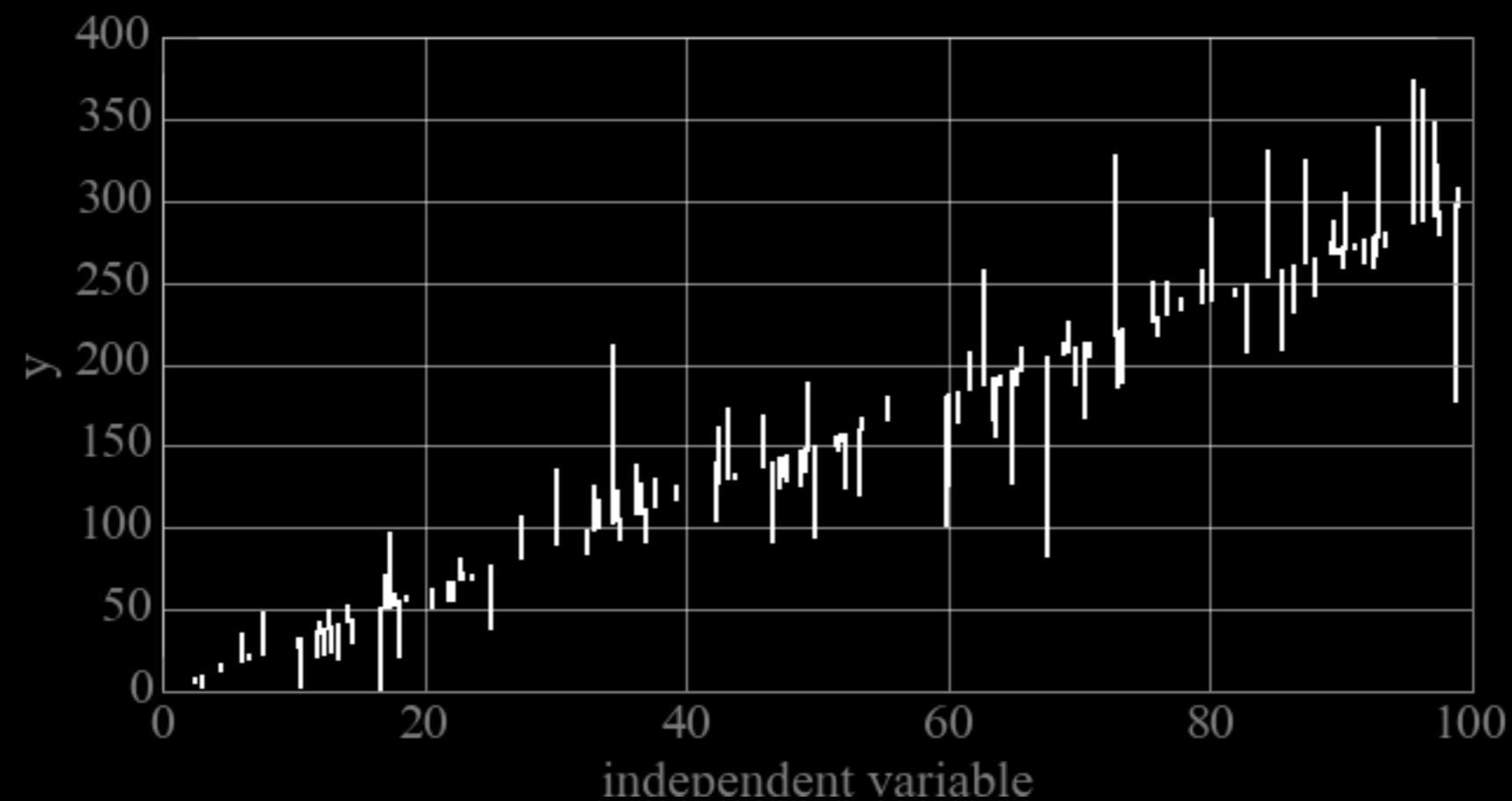




V: Likelihood and
Regression Models



V: Likelihood and
Regression Models

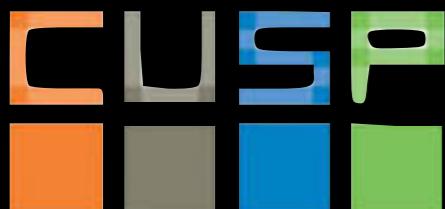
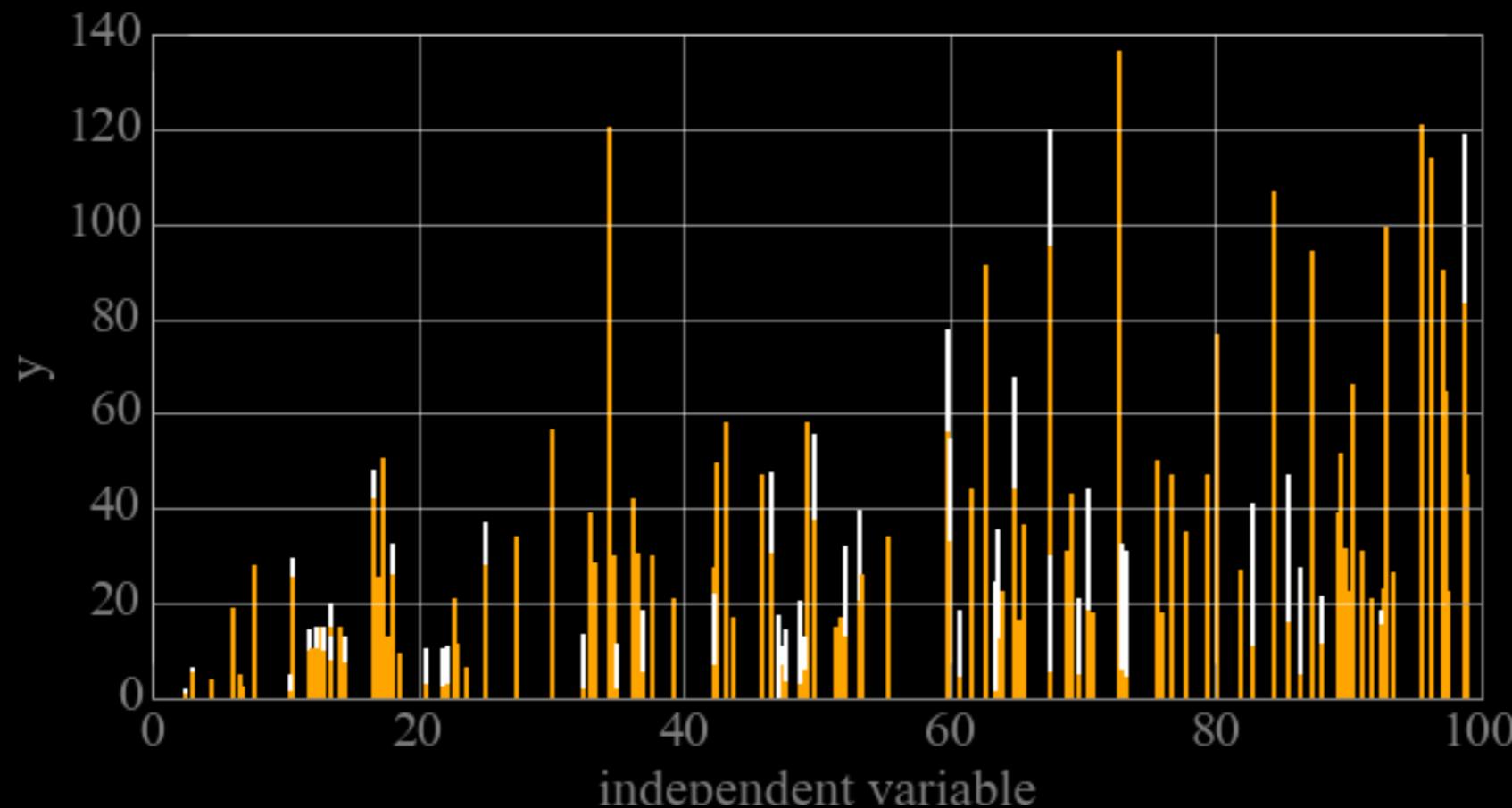


V: Likelihood and
Regression Models

Sum of residuals squared

$$\sum_i (|y_i - (mx_i + b)|)$$

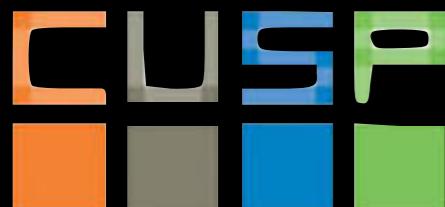
11655.34 12155.24



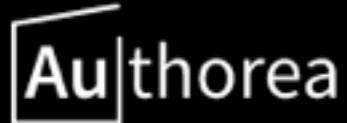
Homework: 2. linear regression and prediction

GENDER INCOME GAP

you may know that it is estimated that women earn about 78% of men in the same job position. Can we test that on NYC income data? Can we turn that into a prediction: if you get hired at a certain stipend as a man, what should you expect to make as a woman? (or from the point of view of a job employer, perhaps not one with a very strong moral compass, what should I offer to a woman job candidate, given what I would offer a man for the same job?)



Assigned reading

[ABOUT](#)[EXPLORE](#)[HELP](#)[SIGN UP](#)[LOGIN](#)

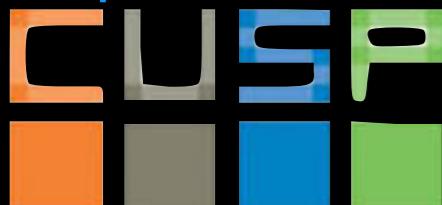
Accelerating Discovery

Authorea is a new kind of research editor.

Write, cite, collaborate, host, and publish all in one place.

NAMEE-MAIL ADDRESSPASSWORD[SIGN UP](#)

10 Simple Rules for the Care and Feeding of Scientific Data
<https://www.authorea.com/users/3/articles/3410/> show article



V: Likelihood and
Regression Models

MUST KNOWS:

- Systematic and Statistical Errors
- Precision vs Accuracy
- Errors are generally added in quadrature
- Goodness of fit testing (why, how, few tests)
- Least square fits (OLS)

Resources:

Sarah Boslaugh, Dr. Paul Andrew Watters, 2008

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.

Introduction to Statistics (XVIII)

http://onlinestatbook.com/Online_Statistics_Education.epub

<http://onlinestatbook.com/2/index.html>

