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

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

Today:

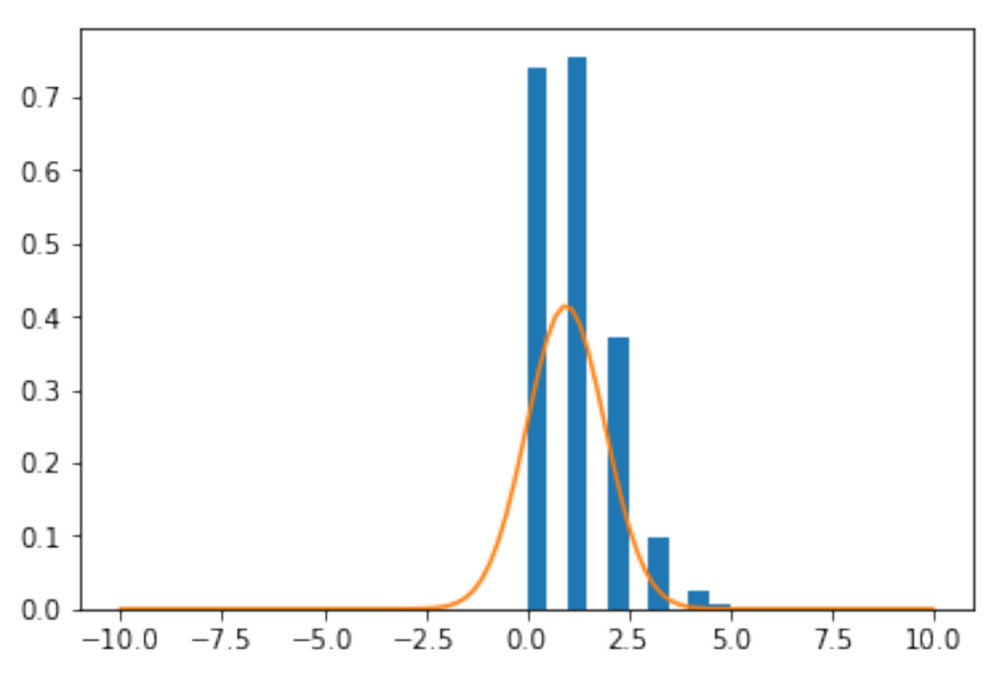
- Goodness of fit tests
- Models
- Systematic errors
- Statistical error
- Linear Regression

Goodness of fit

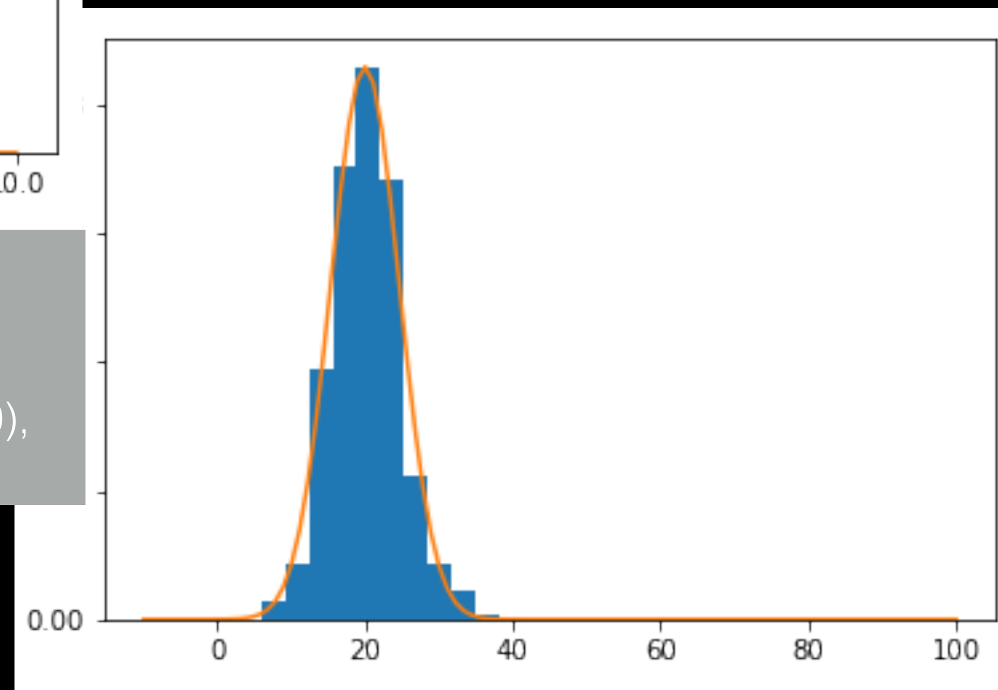
You have some data, and an idea of how it should look: a *model*

Is it a good model?

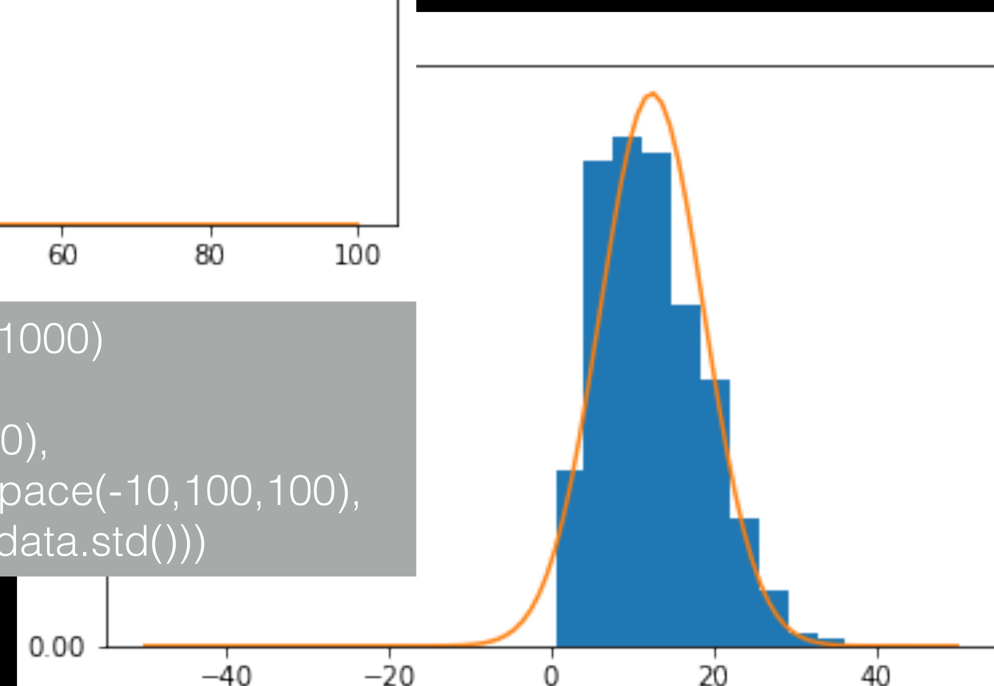
Goodness of fit




```
data = np.random.poisson(5, 1000)
pl.hist(data, normed=True)
pl.plot(np.linspace(-10,100,100),
        sp.stats.norm.pdf(np.linspace(-10,100,100),
        loc=data.mean(), scale=data.std()))
```




```
data = np.random.poisson(5, 1000)
pl.hist(data, normed=True)
pl.plot(np.linspace(-10,100,100),
        sp.stats.norm.pdf(np.linspace(-10,100,100),
        loc=data.mean(), scale=data.std()))
```

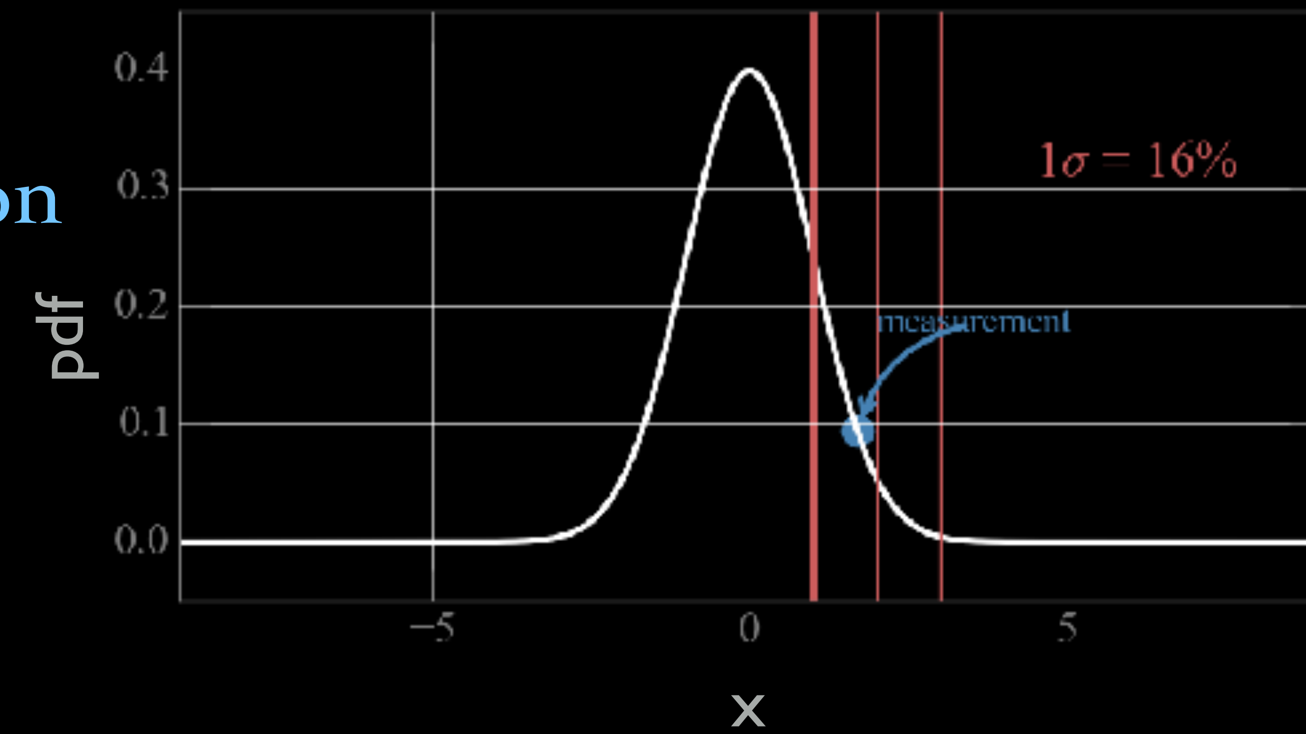



```
data = np.random.rayleigh(10, size=1000)
pl.hist(data, normed=True)
pl.plot(np.linspace(-50,50,100),
        sp.stats.norm.pdf(np.linspace(-50,50,10
        0), loc=data.mean(), scale=data.std()))
```

I have a distribution.
It is useful to be able to associate it to a math equation. **That math equation is a *model.***
How can I tell if I have the "right" model?

Probability Distribution Function

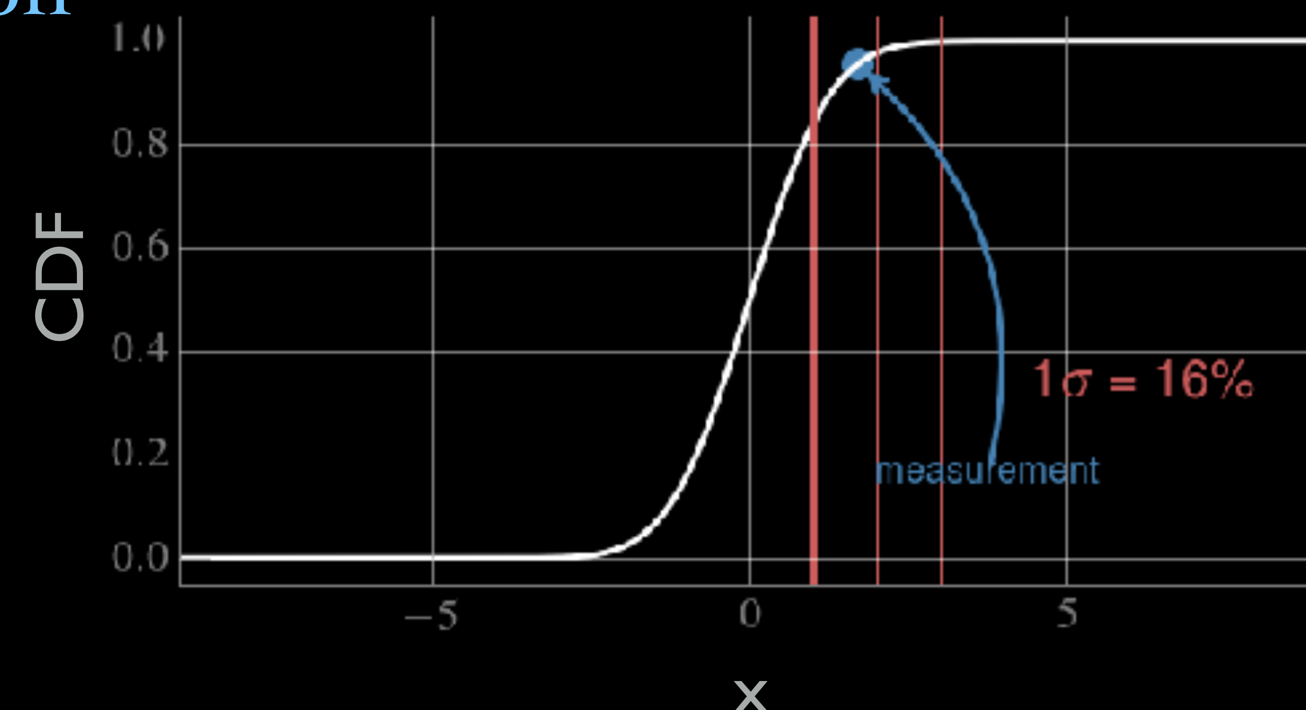
$$f_{x_0}(x) \sim p(x = x_0)$$

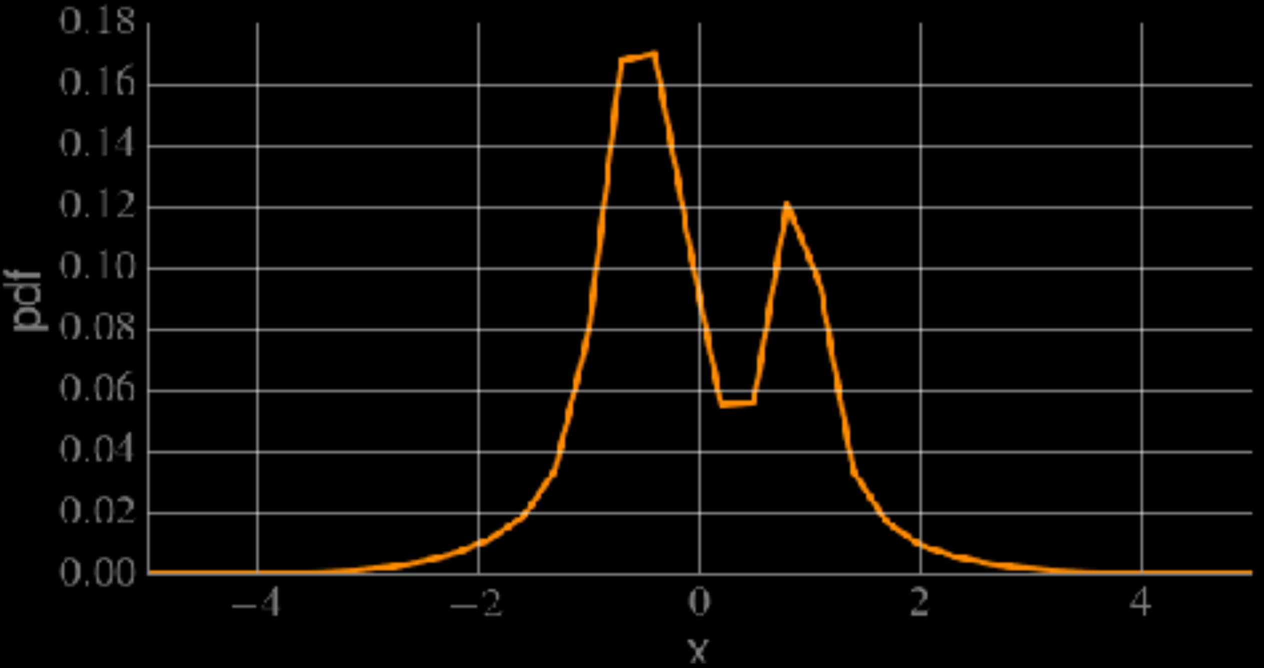


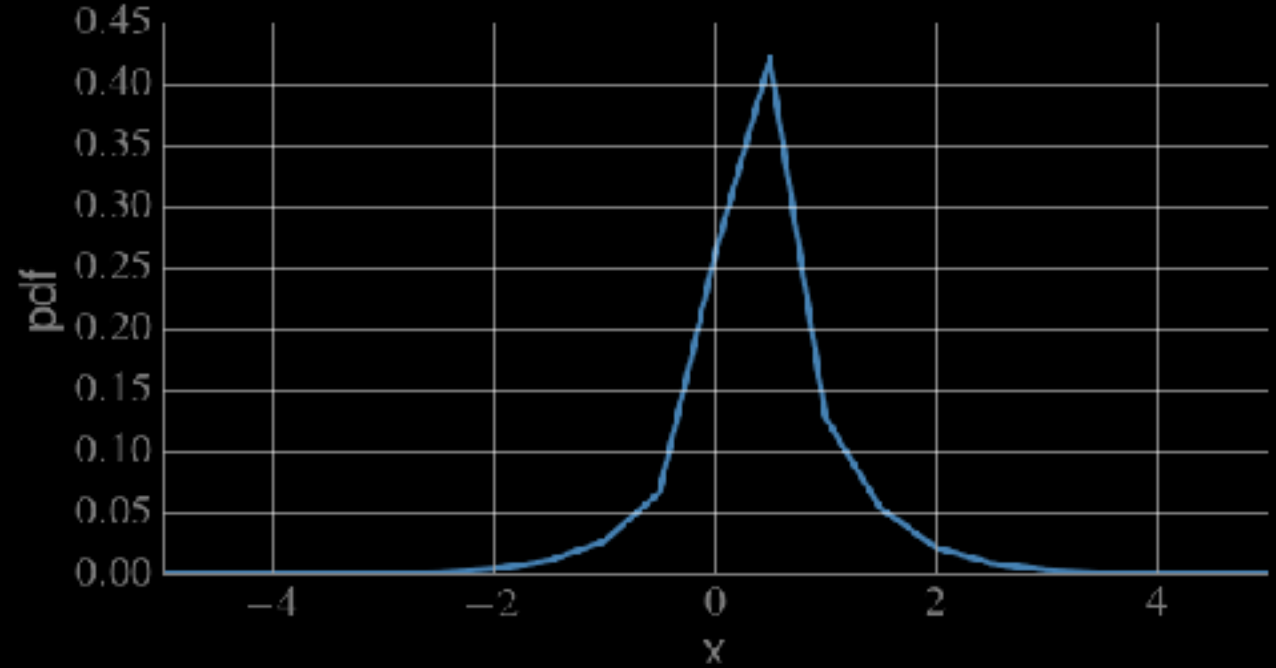
$$f_{x_0}(x) \sim p(x > x_0 - dx) \bigcap p(x < x_0 + dx)$$

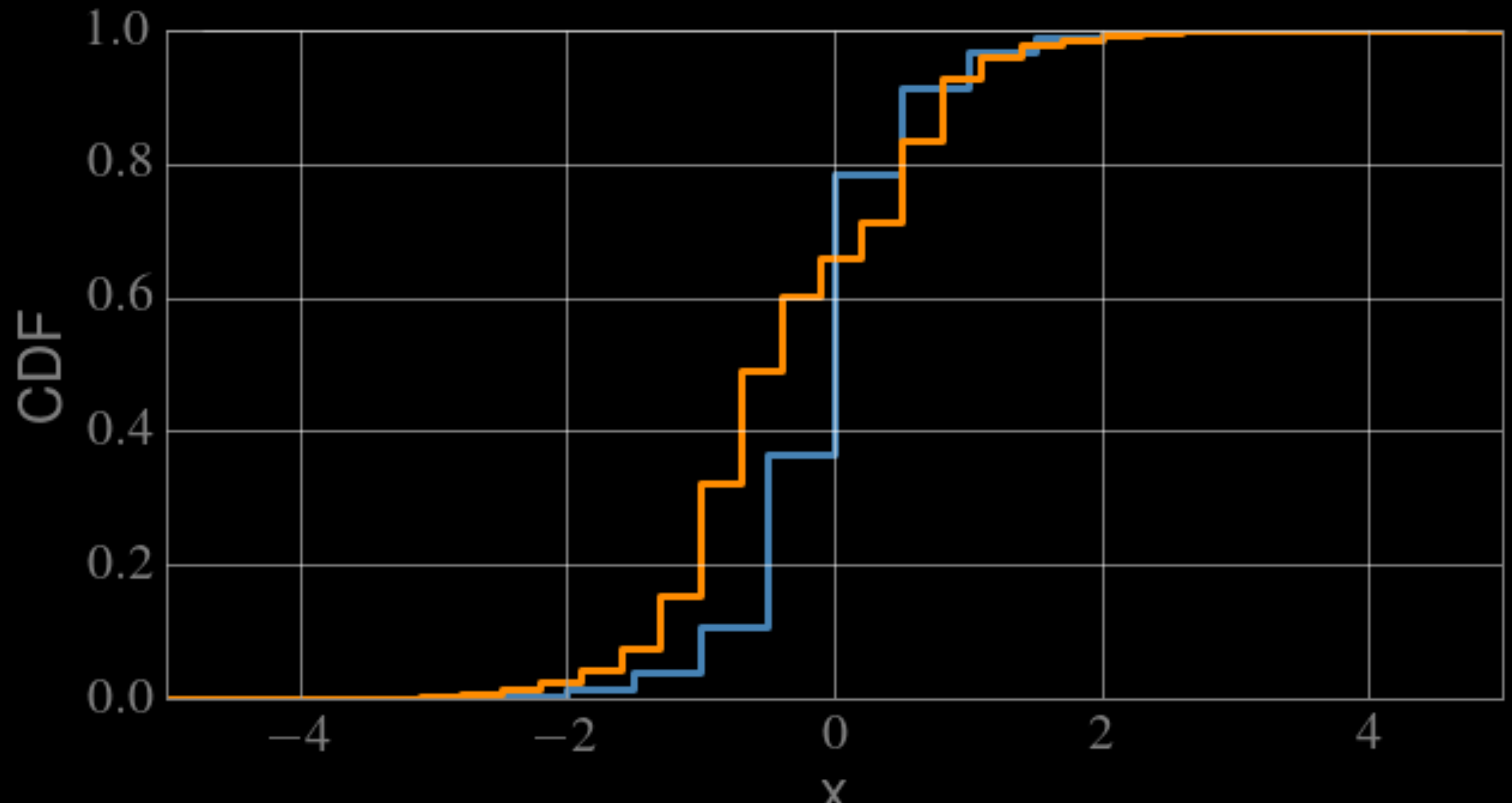
Cumulative Distribution Function

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




pdf
x
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
−4
−2
0
2
4


pdf
x
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
−4
−2
0
2
4


CDF
x
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4

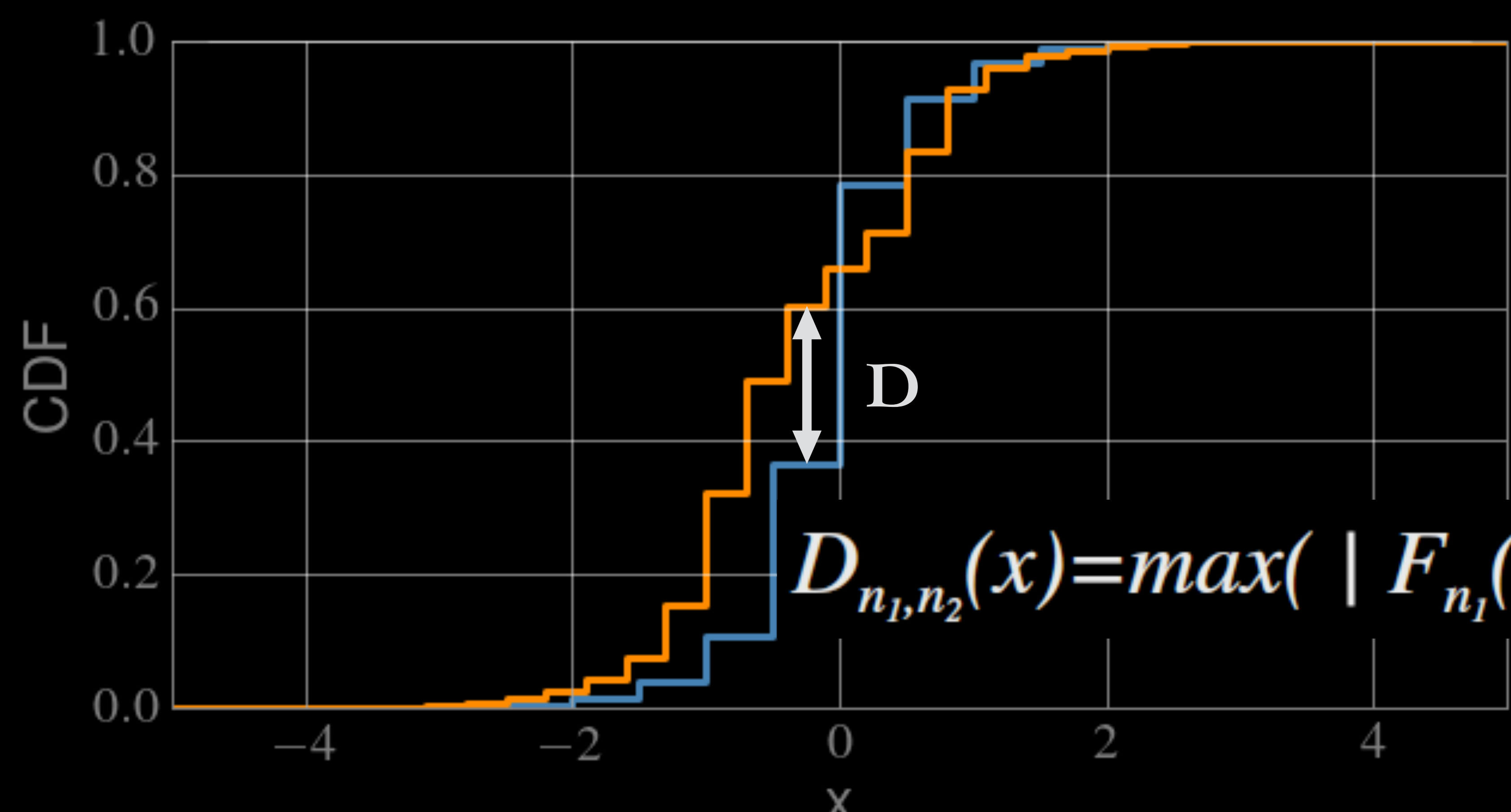
Two sample Kolmogorov Smirnoff test:

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

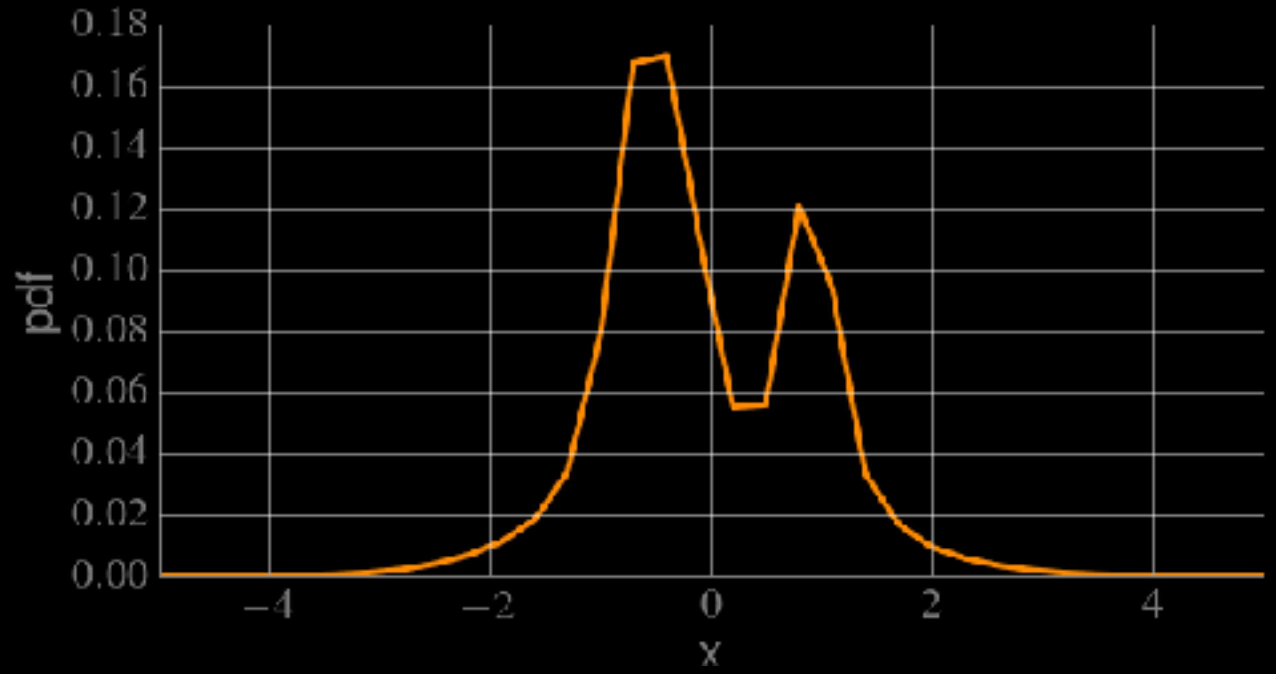
H_0 is rejected at level α if $D(n_1,n_2) > c(\alpha)\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

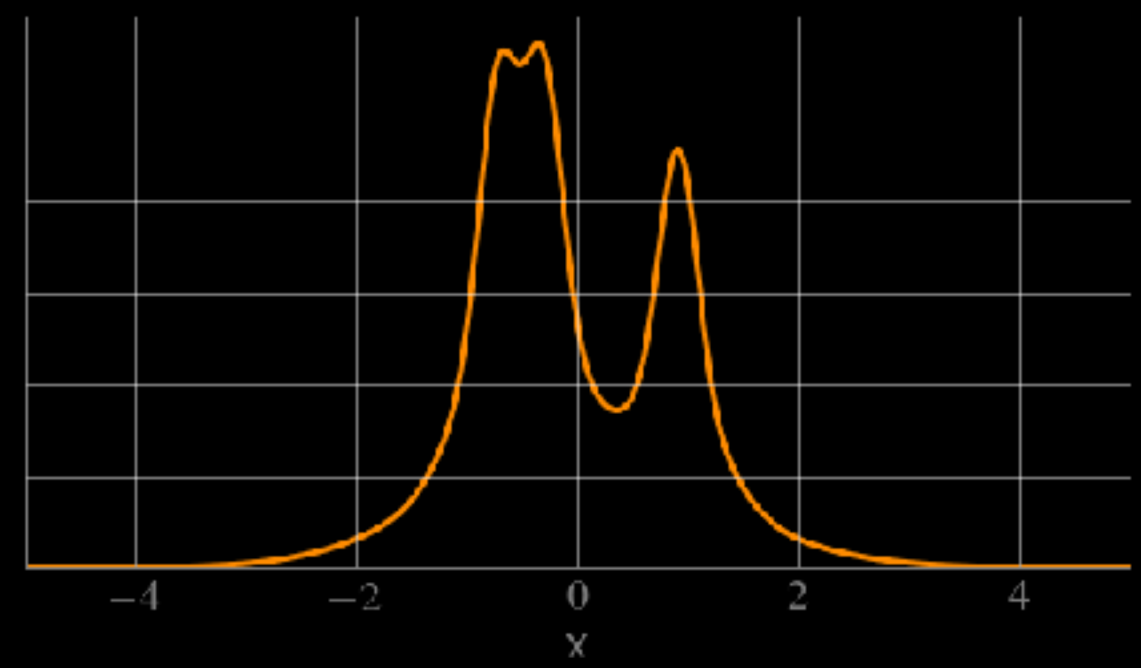
with $c(\alpha)$ given by a table

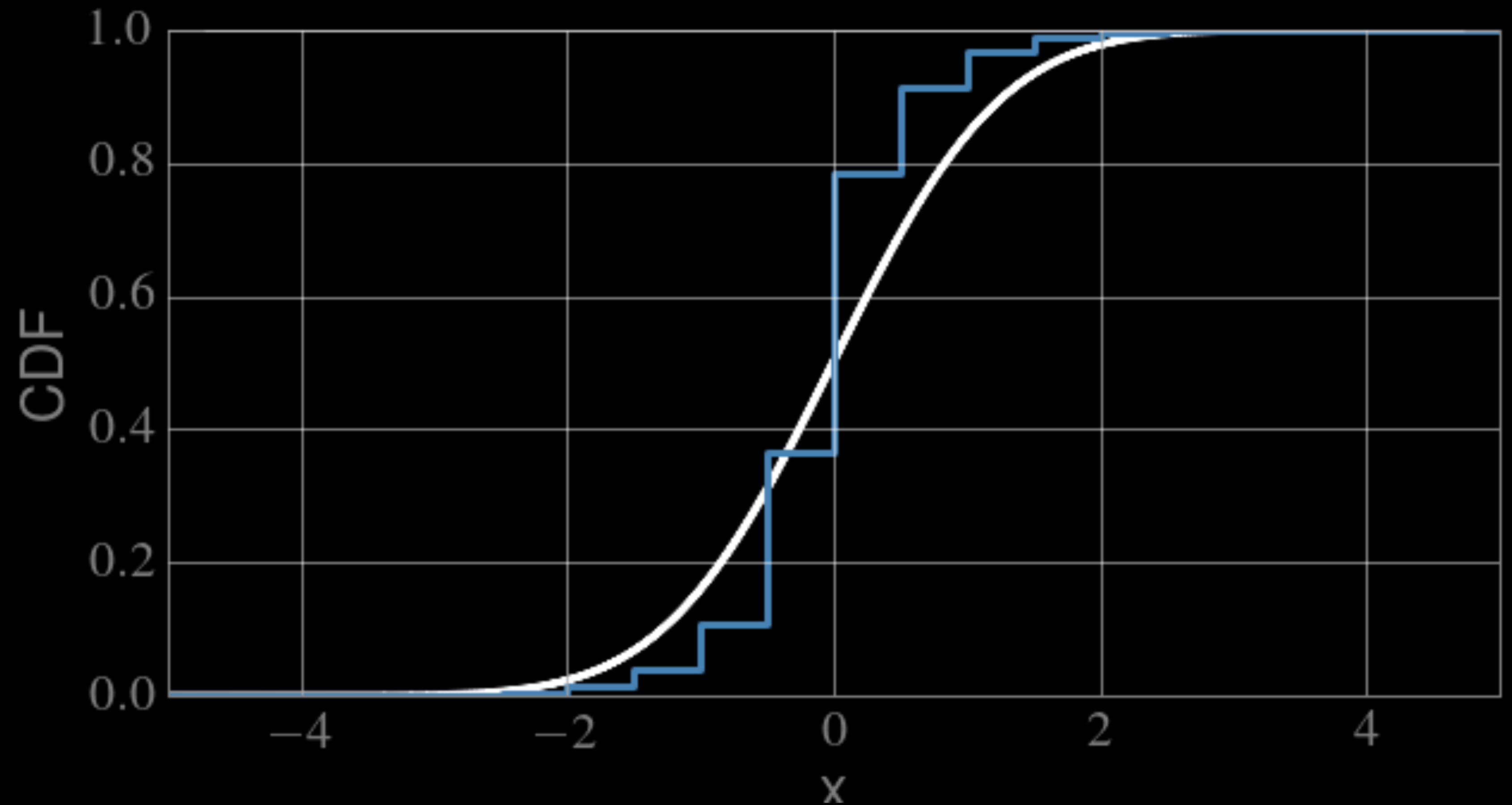
NOTE: it ONLY woks in 2D where the Euclidian distance is uniquely defined!



$$D_{n_1,n_2}(x) = max(\ |F_{n_1}(x) - F_{n_2}(x)|\ )$$


pdf
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−4
−2
0
2
4
x

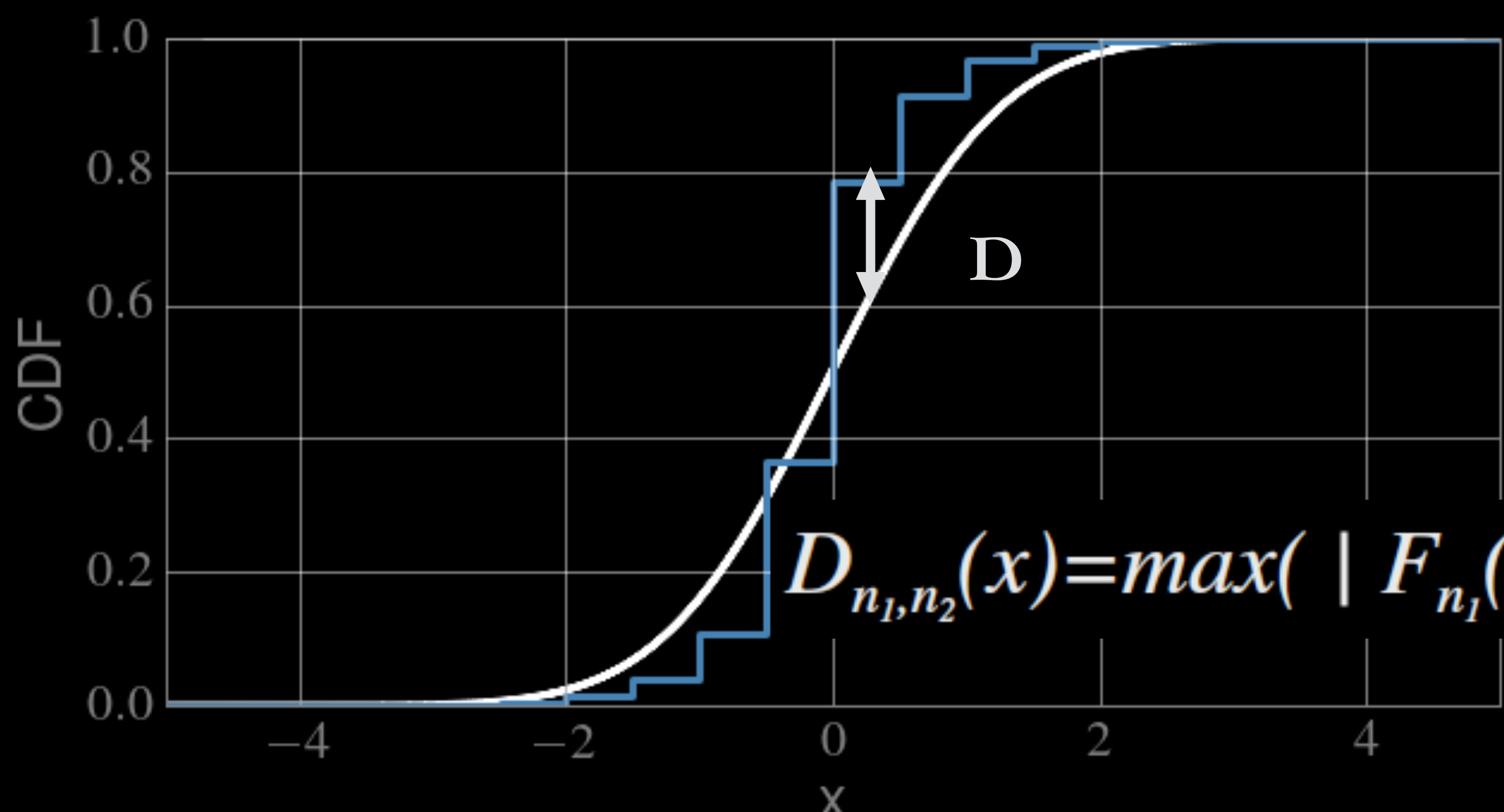

−4
−2
0
2
4
x


CDF
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4
x

Goodness-of-fit Kolmogorov Smirnoff test:

null hypothesis H_0: the sample does comes from the <u>*model*</u> distribution

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



$$D_{n_1,n_2}(x) = max(\ |\ F_{n_1}(x) - F_{n_2}(x)\ |\ )$$

Tests Cheat Sheet:
2 (+) samples comparison

	metric (statistic)	compare to	
KS	$D_{n_1,n_2}(x)=max(\mid F_{n_1}(x)-F_{n_2}(x) \mid)$	$c(\alpha)\sqrt{\frac{n_1+n_2}{n_1 n_2}}$	*Non parametric* *2 samples only*
K-sample Anderson-Darling	$ADK=\frac{n-1}{n^2(k-1)}\sum_{i=1}^{k}\frac{1}{n(i)}\left(\sum_{j=1}^{L}h_j\frac{(nF_{ij}-n_iH_j)^2}{H_j(n-H_j)-nh_j/4}\right)$	•AK table	*Non parametric,* *N samples*
Pearson's	$r_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right)$	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	$D_{n_1,n_2}(x) = max(\mid F_n(x) - F(x) \mid)$	$\frac{K_\alpha}{\sqrt{n}}$	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)[	
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	$\frac{L\ (model\ 1 \mid\ data)}{L\ (model\ 2 \mid\ data)}$		suitable to bayesian analysis

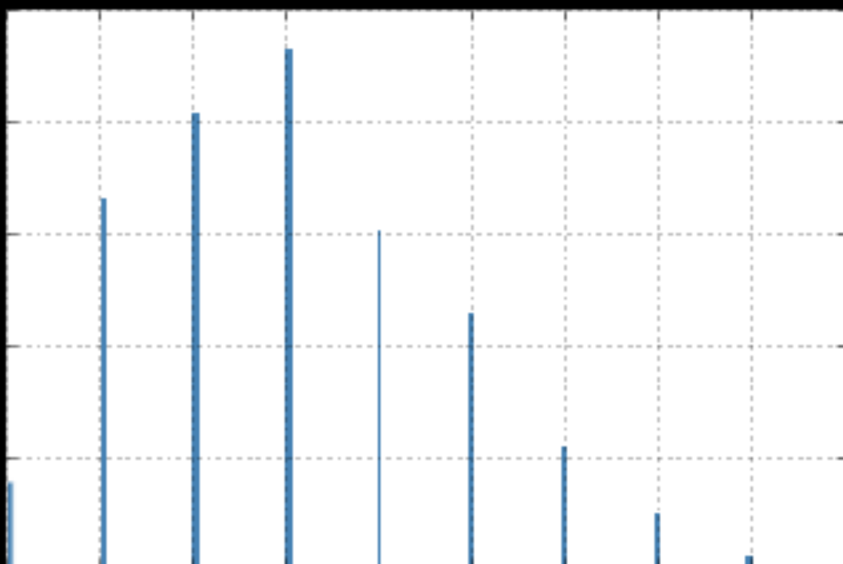
Binomial

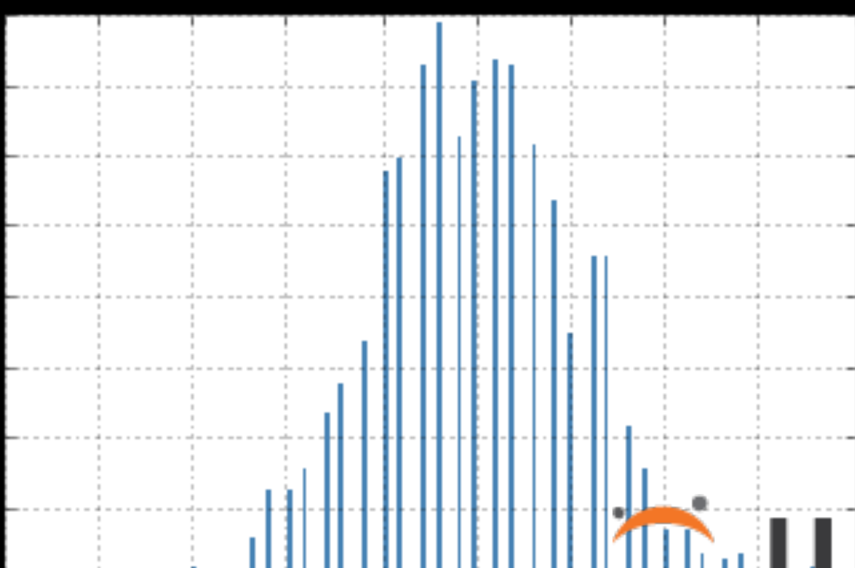
discrete bivariate

2 parameters: n,p

support: [0 , ∞]

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





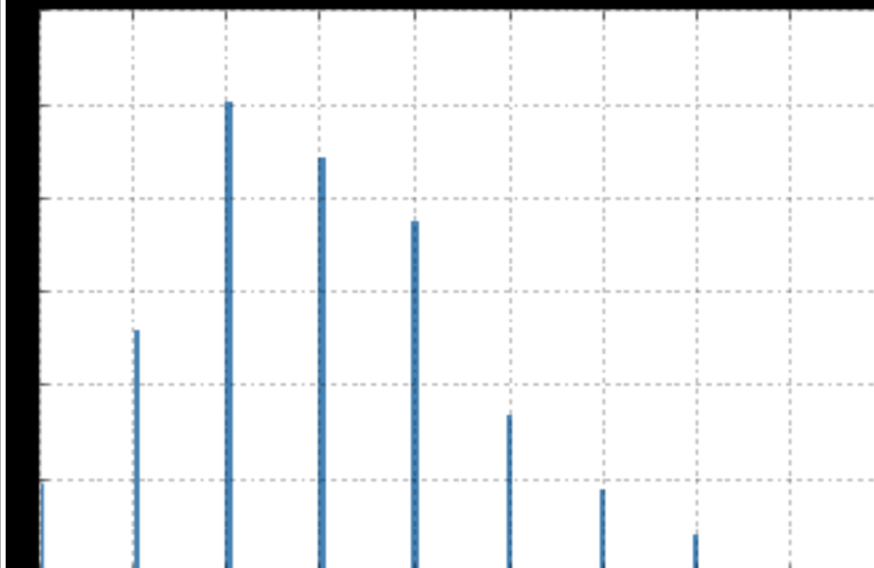
Poisson

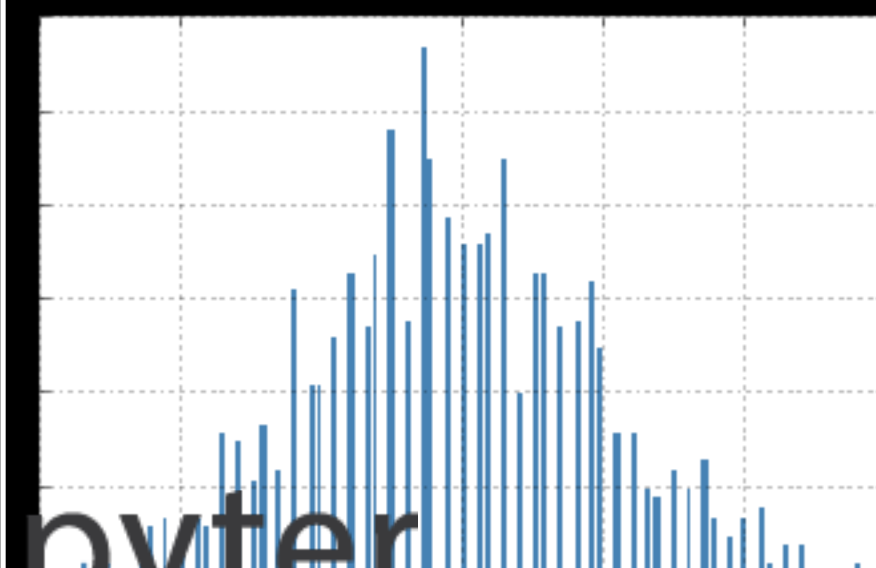
discrete univariate

1 parameters: λ

support: [0 , ∞]

moments: λ, $\sqrt{\lambda}$,>0





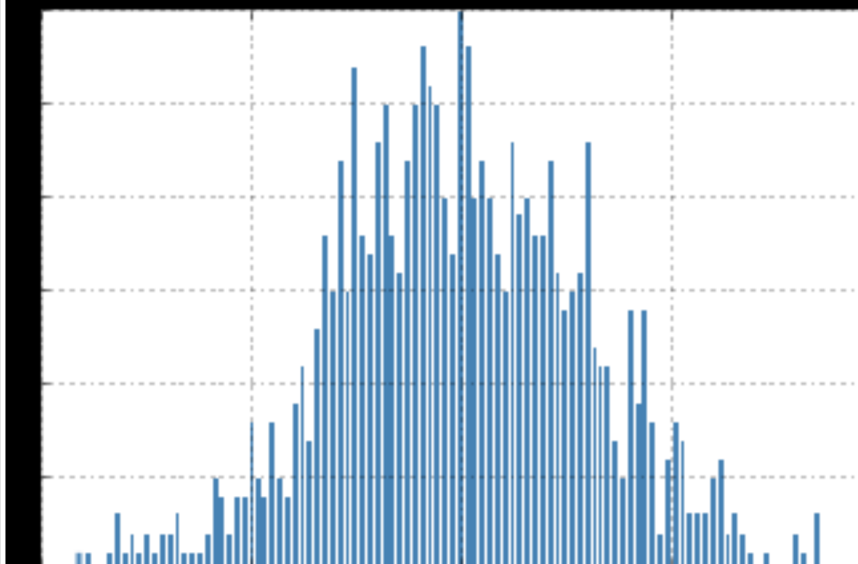
Gaussian

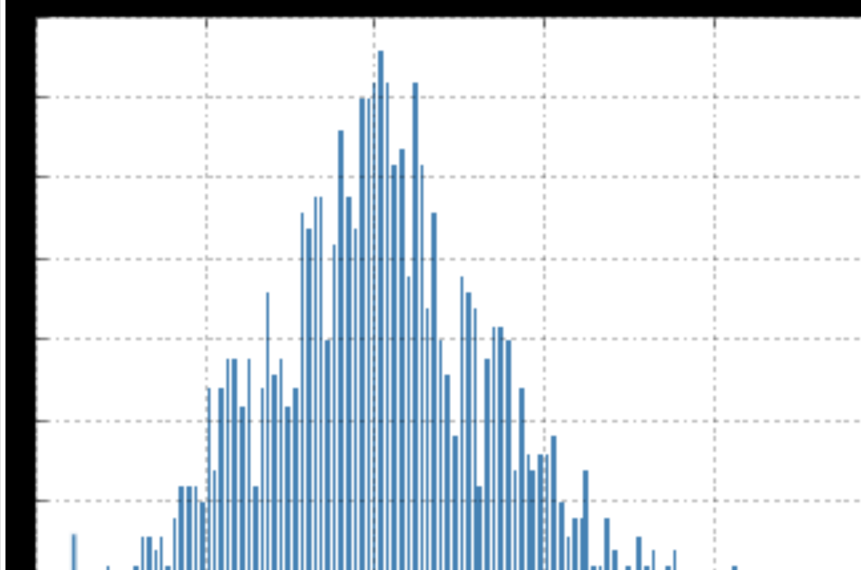
continuous bivariate

2 parameters: μ, σ

support:[- ∞ , ∞]

moments: μ, σ, 0

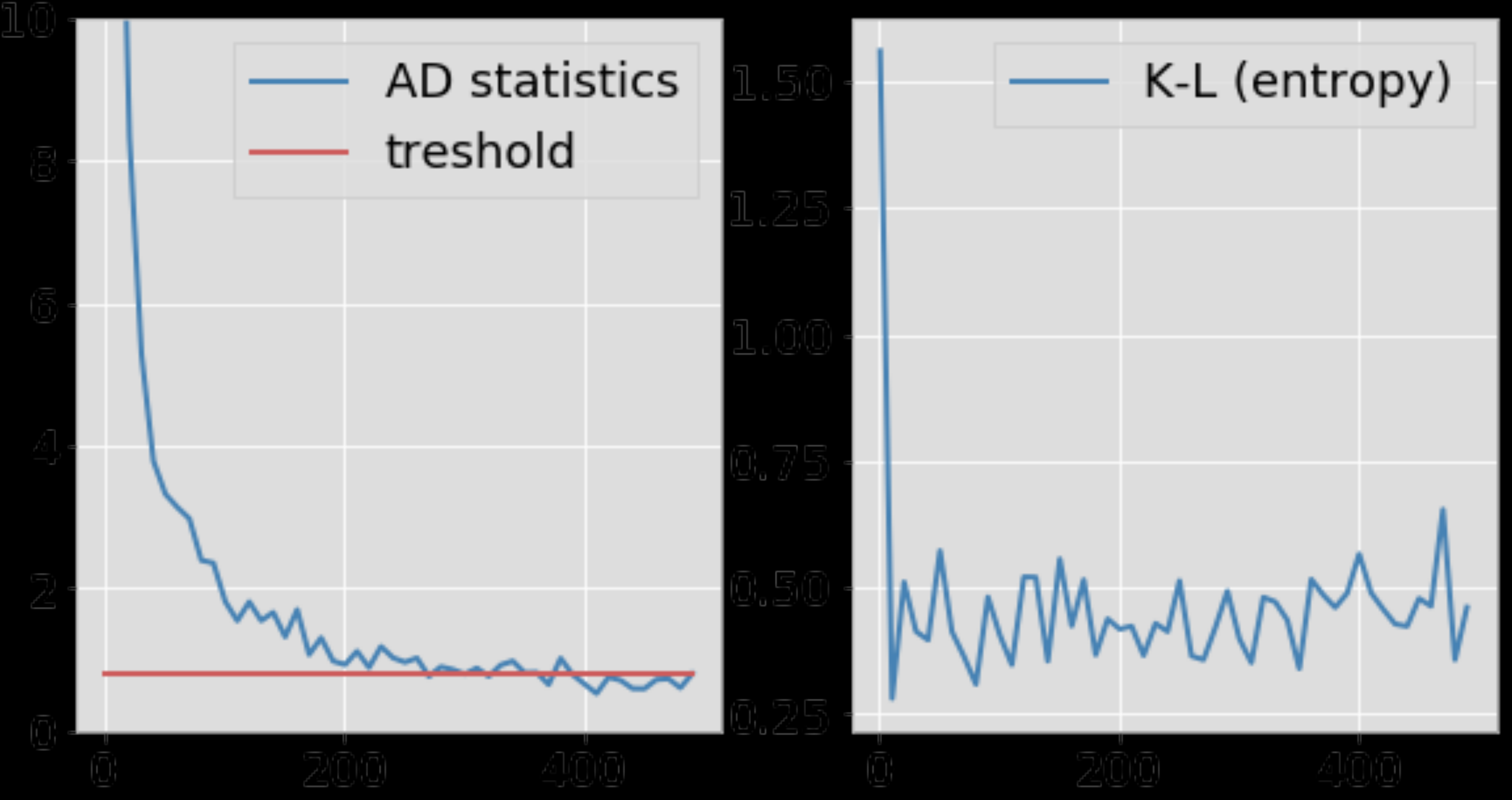




http://localhost:8892/notebooks/distributions/poisson%20vs%20gaussian.ipynb

Homework 1:

show that as the parameters *nxp* for a Binomial distribution and λ for a Poisson distribution increase, they become increasingly hard to distinguish from the Gaussian (Normal) distribution



Homework: 1. Test the *Z* test

(all simulated data)

Generate N samples from a Gaussian distribution with a chosen mean μ and standard deviation σ: $\mathcal{N}(\mu, \sigma)$ and calculate the mean of each sample (all samples should have the same size *n*).

Assess the validity of the *Z*-test: If the samples are drawn from the distribution you are testing the *z*-values you calculate should follow a $\mathcal{N}(0,1)$ distribution (a Gaussian with mean 0 and standard deviation 1). Show that the distribution of *z* -statistics (find the formula in a statistics book or in last week's slides) that you calculated (one for each sample) is indeed consistent with $\mathcal{N}(0,1)$.

Homework: 2. 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 chosen from: KS, AD, KL, chisq to do this. Test at least 2 distributions.

Optional (extra credit): Divide your sample geographically: by Borrow (Manhattan vs Brooklyn) and see if you notice any differences in how the age distribution can be modeled. You can do this with the chisq test: is the chisq better for the fit to Manhattan vs Brooklyn?

Compare Tests for Correlation and Goodness of fit:

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

- Pearson's test
- Spearman's test
- K-S test

e.g. Estimate number of MTA Bus passengers at different hours (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…)

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 number of MTA Bus passengers per bus line within an interval of time: are the passengers randomly distributing on busses.

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
- Spearman's test
- K-S test

e.g. Age distribution of male vs female Citibikes riders. Age distribution for long/short trips. Trip duration distribution in different seasons.

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: is it Gaussian distributed, or lognormal, power law, some bimodal distribution…

In HW4 Assignment 2 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 HW5 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.

CUSP

Errors and uncertainties.

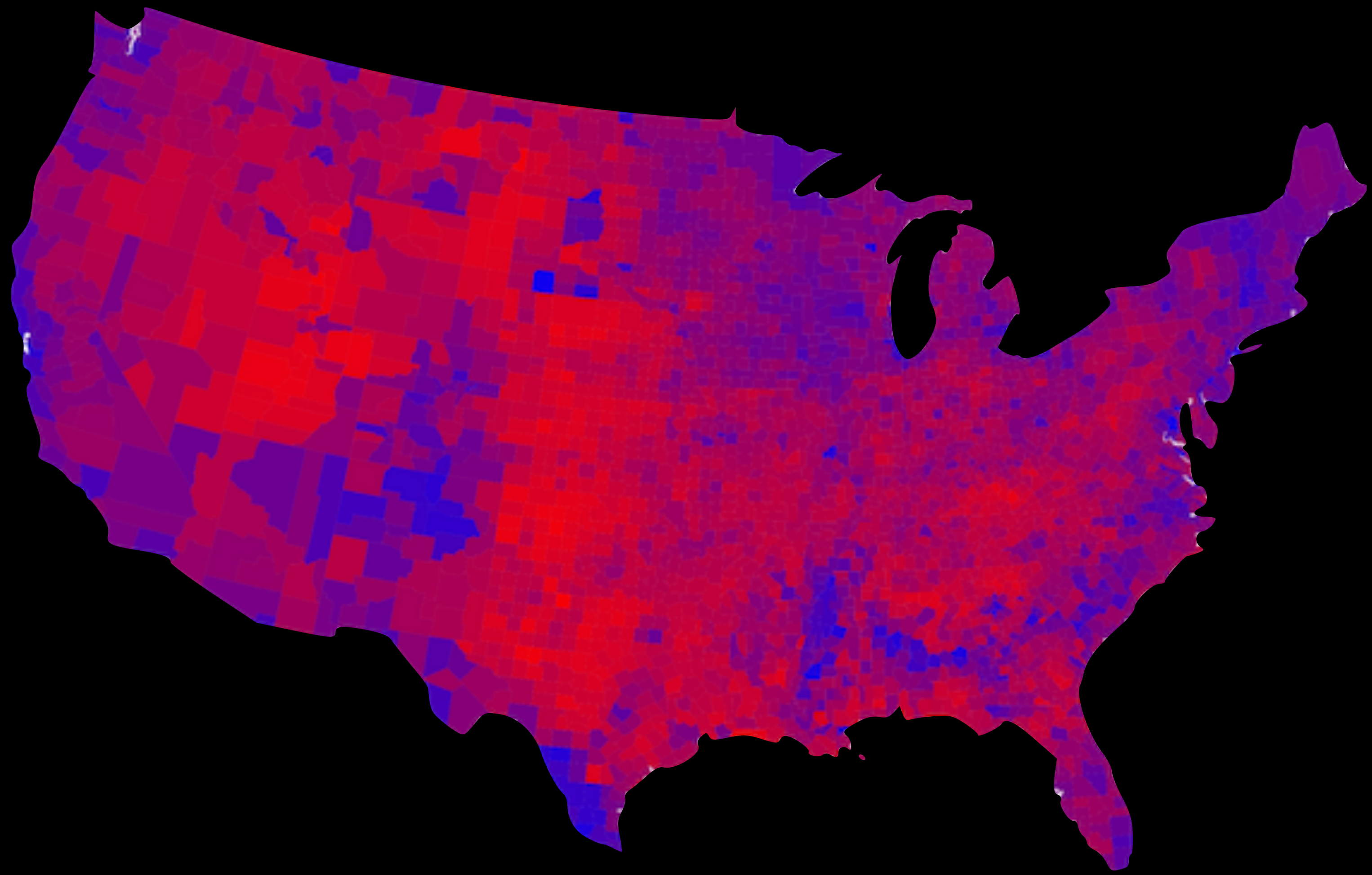
systematic errors

Errors and uncertainties.

- Systematic error

  tendency to *systematically*
  underestimate OR 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

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.

www.nyc.gov/html/doh/html/data/pat-methods.shtml#4

Apps | Bookmarks | gmail | pedometer | http://www.weather.c | New folder | SEGN in literature | nyu | trave | copp

MENTAL & BEHAVIORAL HEALTH

HEALTHY ENVIRONMENT

EMERGENCY PREPAREDNESS

DATA & STATISTICS

- Surveys
- Tools
- Your Neighborhood
- Health Data Publications

HEALTH CARE PROVIDERS

INFORMATION FOR:

- Licenses and Permits
- Press
- Public Testimony
- Vendors and Contractors
- Hurricane Sandy Health
- 9/11 Health

HEALTH DEPARTMENT:

- About Us
- Take Care New York
- Board of Health/Health Code
- Public Meetings Archive
- Official Notices
- Publications
- Career Opportunities

Physical Activity and Transit Survey

Physical Activity and Transit Survey: Methodology

Target Population | Health Topics | Sampling | Limitations | Sample Size, Response and Cooperation Rates | Data Analysis

The Physical Activity and Transit (PAT) survey was conducted in 2010 and 2011 by the New York City Department of Health and Mental Hygiene. Data were collected to measure the level and context of physical activity in New York City and to improve understanding of what motivates individuals to be physical active including opportunities for activity, perceptions of safety and security, and other neighborhood factors. For more information on the PAT, please visit:

- PAT Overview
- PAT Device Methodology
- PAT Public Use Datasets

TARGET POPULATION
The target population of the PAT was adults aged 18 and older who were able to walk more than 10 feet and who lived in one of the five boroughs of New York City. Of the 3908 adults who completed the initial telephone screener, 3811 were mobile and completed the full survey.

HEALTH TOPICS
The PAT asked approximately 125 questions, covering the following: a modified version the Global Health and Physical Activity Questionnaire (GPAQ) designed by the World Health Organization on physical activity in the work, recreation and transportation domains. Also included were questions on chronic disease, diet, alcohol and tobacco, neighborhood conditions, and mental health. The survey also asked a multiple demographic variables to facilitate weighting and comparisons among different groups of New Yorkers.

SAMPLING
The PAT was conducted using a fully overlapping dual frame design, using randomly generated landline and cellular telephone samples. (Roughly 25% were completed on a cell phone.) To provide equal statistical power for borough-level comparisons, a similar number of participants were interviewed in each borough of New York (Bronx, Brooklyn, Manhattan, Queens and Staten Island). All data were then weighted to adjust for the probability of selection and differential nonresponse and sum to Census estimates of the number of people living in each borough.

Interviewing was done by Abt-SRBI, a survey research company based in New York City. Interviews were conducted in English, Spanish, Russian and Chinese (Cantonese and Mandarin). Data collection for wave 1 occurred in September – November of 2010; wave 2 was conducted in March – November of 2011. The average length of the survey was 35 minutes.

back to top

LIMITATIONS
The sampling methodology did not capture adults who could not be reached by either landline or cellular telephone. The PAT also excluded adults living in institutional group housing, such as incarcerated persons or those living in college dormitories.

back to top

http://www.nyc.gov/html/doh/html/data/pat-methods.shtml#4

Errors and uncertainties.

Errors and uncertainties.

Errors and uncertainties.

projection induces a *systematic underestimation*

Bias in measurements: know your data

Bias in measurements: know your data

Undercoverage bias
Self selection bias
Social desirability bias
Publication Bias
Data Dredging

Bias in measurements: know your data

Undercoverage bias

the surveyed segment of the population is lower in a sample than it is in the population. This can happen because the frame used to obtain the sample is incomplete or not representative of the population.

Bias in measurements: know your data

Undercoverage bias
<u>Self selection bias</u>
Social desirability bias
Publication Bias
Data Dredging

Bias in measurements: know your data

Self selection bias

people willing to answer a survey about climate are more likely concerned citizens that care abut the climate

Bias in measurements: know your data

<u>Self selection bias</u>

higher test scores observed among students who participate in a test preparation courses, but due to self-selection, people who *choose* to take the course may be more motivated, have more support…

Bias in measurements: know your data

Undercoverage bias
Self selection bias
Social desirability bias
Publication Bias
Data Dredging

Bias in measurements: know your data

Social desirability bias

tendency of survey respondents to answer in a manner that may be viewed favorably: over-reporting "good behavior", under-reporting undesirable behavior (e.g. drug+alcohol use).

CUSP

Bias in measurements: know your data

Organizational Research Methods
00(0) 1-34

DOI: 10.1177/1094428113493120
orm.sagepub.com
SAGE

Random and Systematic Error Effects of Insomnia on Survey Behavior

Larissa K. Barber[1], Christopher M. Barnes[2], and Kevin D. Carlson[2]

Abstract

Insomnia is a prevalent experience among employees and survey respondents. Drawing from research on sleep and self-regulation, we examine both random (survey errors) and systematic (social desirability) effects of research participant insomnia on survey responses. With respect to random effects, we find that insomnia leads to increased survey errors, and that this effect is mediated by a lack of self-control and a lack of effort. However, insomnia also has a positive systematic effect, leading to lower levels of social desirability. This effect is also mediated by self-control depletion and a lack of

http://www.researchgate.net/publication/244478619_Random_and_Systematic_Error_Effects_of_Insomnia_on_Survey_Behavior

Bias in measurements: know your data

Undercoverage bias
Self selection bias
Social desirability bias
<u>Publication Bias</u>
Data Dredging

Bias in measurements: know your data

Publication Bias

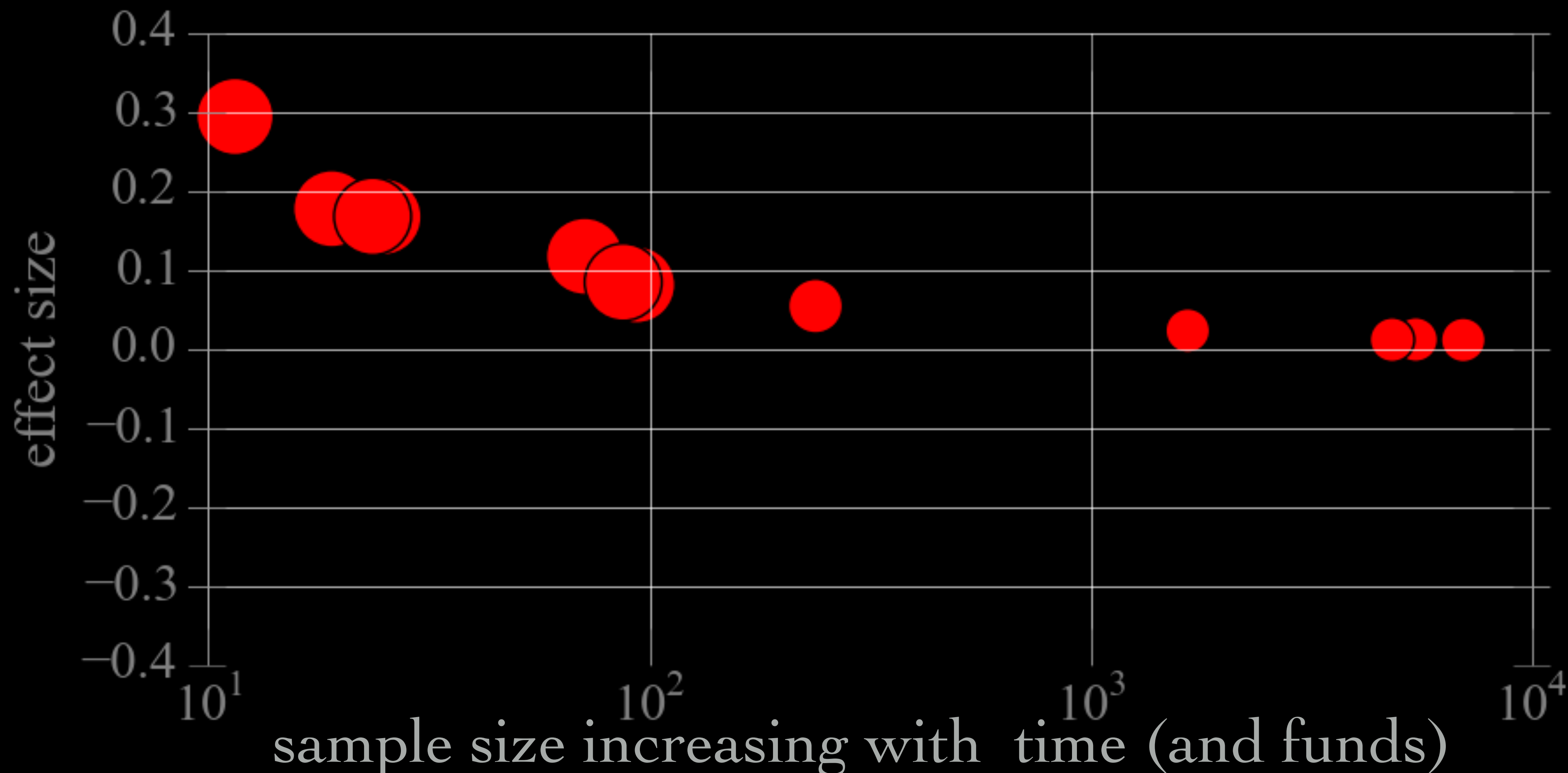
Bias in measurements: know your data

His team investigated the fate of 221 sociological studies conducted between 2002 and 2012, which were recorded by Time-sharing Experiments for the Social Sciences (TESS), a US project that helps social scientists to carry out large-scale surveys of people's views.

Only 48% of the completed studies had been published. So the team contacted the remaining authors to find out whether they had written up their results, or submitted them to a journal or conference. They also asked whether the results supported the researchers' original hypothesis.

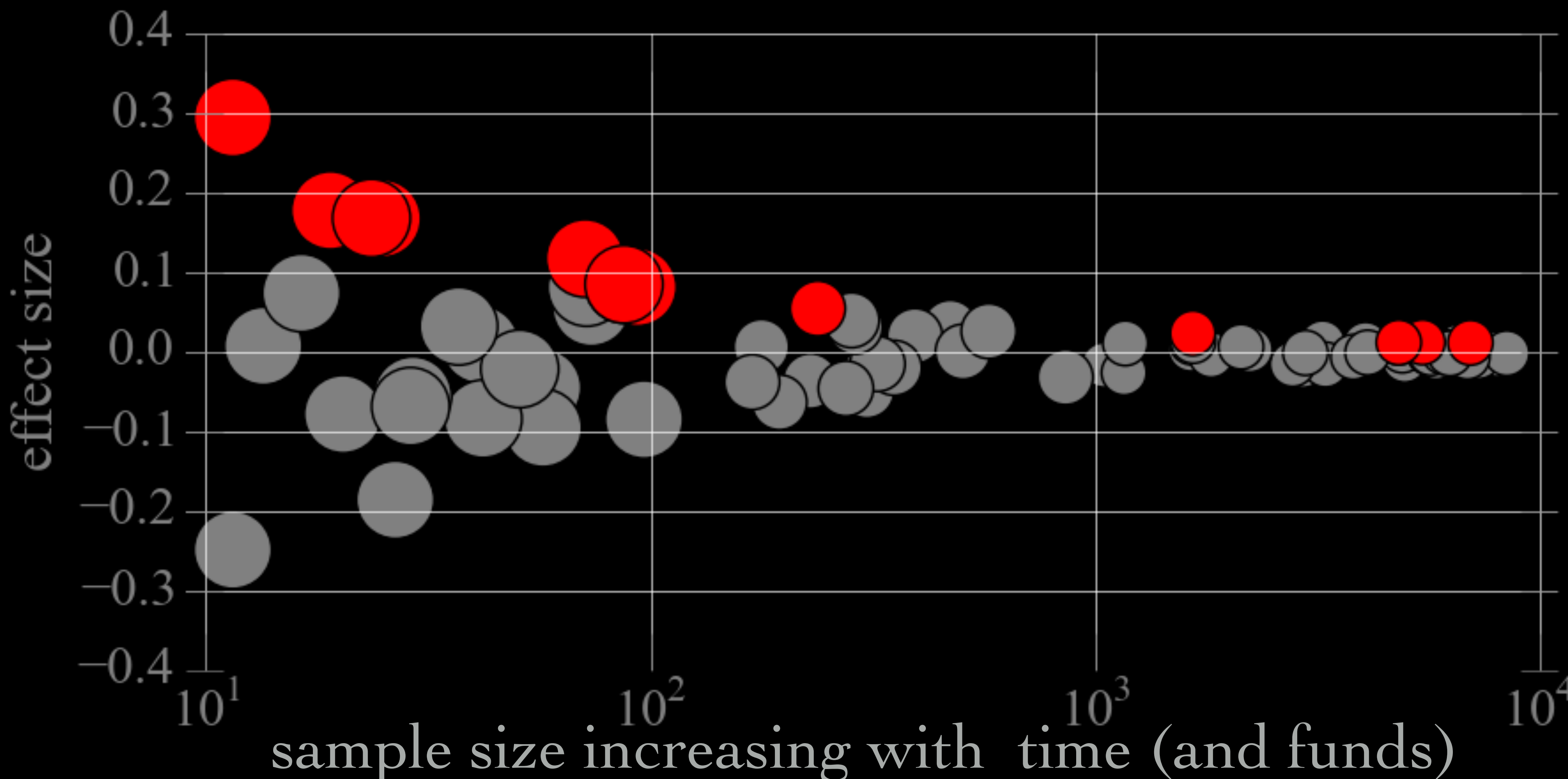
Of all the null studies, just 20% had appeared in a journal, and 65% had not even been written up. By contrast, roughly 60% of studies with strong results had been published. Many of the researchers contacted by Malhotra's team said that they had not written up their null results because they thought that journals would not publish them, or that the findings were neither interesting nor important enough to warrant any further effort.

Publication Bias



<u>Publication Bias</u>

Bill Howe,
UW e-science center



<u>Publication Bias</u>

Bill Howe,
UW e-science center

Less significant results are less likely to be published

Bias in measurements: know your data

EconPapers
Economics at your fingertips

EconPapers Home
About EconPapers

Working Papers
Journal Articles
Books and Chapters
Software Components

Authors

JEL codes
New Economics Papers

Advanced Search

Publication Bias in Measuring Climate Sensitivity

Share

Dominika Rečková and Zuzana Iršová (zuzana.irsova@ies-prague.org)

No 2015/14, Working Papers IES from Charles University Prague, Faculty of Social Sciences, Institute of Economic Studies

Abstract: We present a meta-regression analysis of the relation between the concentration of carbon dioxide in the atmosphere and changes in global temperature. The relation is captured by "climate sensitivity", which measures the response to a doubling of carbon dioxide concentrations compared to pre-industrial levels. Estimates of climate sensitivity play a crucial role in evaluating the impacts of climate change and constitute one of the most important inputs into the computation of the social cost of carbon, which reflects the socially optimal value of a carbon tax. Climate sensitivity has been estimated by many researchers, but their results vary significantly. We collect 48 estimates from 16 studies and analyze the literature quantitatively. We find evidence for publication selection bias: researchers tend to report preferentially large estimates of climate sensitivity. Corrected for publication bias, the bulk of the literature is consistent with climate sensitivity lying between 1.4 and 2.3C.

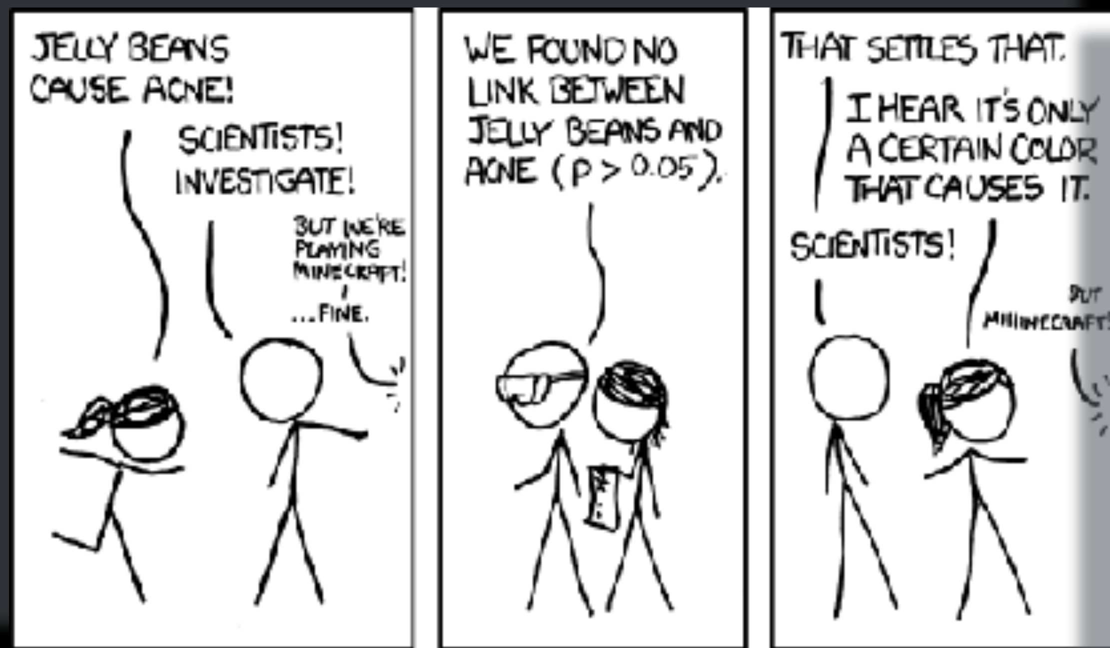
Publication Bias

http://ies.fsv.cuni.cz/default/file/download/id/28421

Bias in measurements: know your data

Undercoverage bias
Self selection bias
Social desirability bias
Publication Bias
<u>Data Dredging</u>

Bias in measurements: know your data



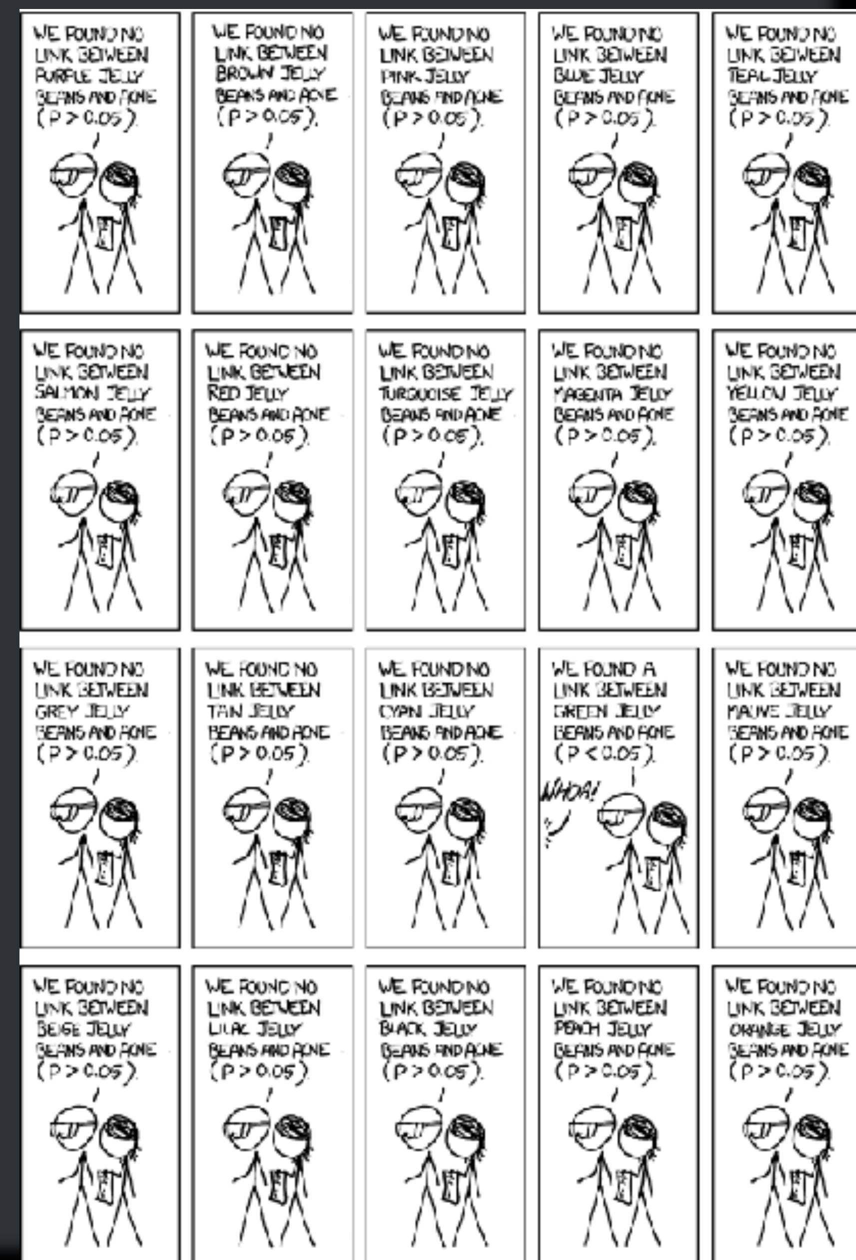
https://xkcd.com/882/

Data Dredging

Bias in measurements: know your data

each test has a
probability $p<=0.05$
of Type I error
significance 95%

20 tests are preformed



Data Dredging

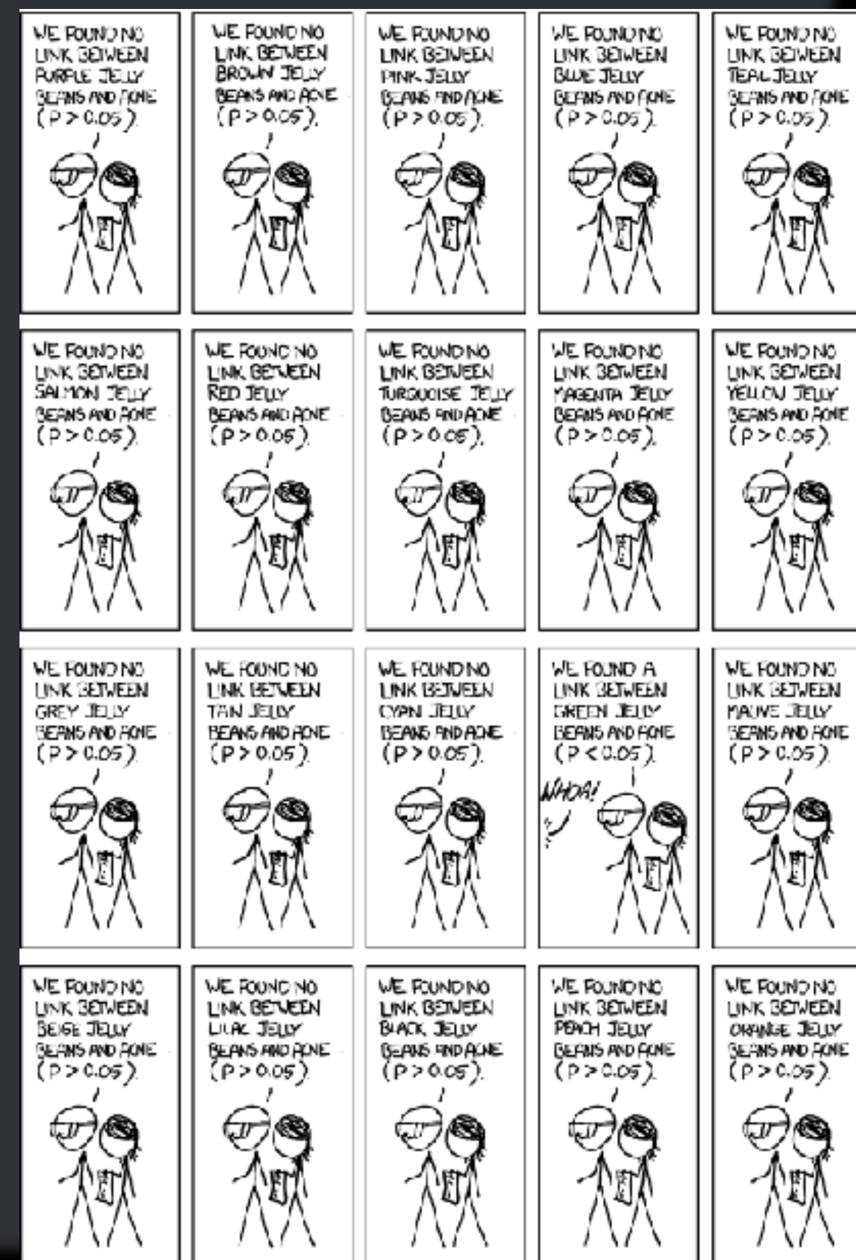
Bias in measurements: know your data

each test has a probability p<=0.05 of Type I error significance 95%

20 tests are preformed

assume independence:
if $p_i = 0.05$ for each i=1..20

total significance= $1-(1-0.05)^{20}$
$p_{tot} = (1-0.05)^{20}$



Data Dredging

Bias in measurements: know your data

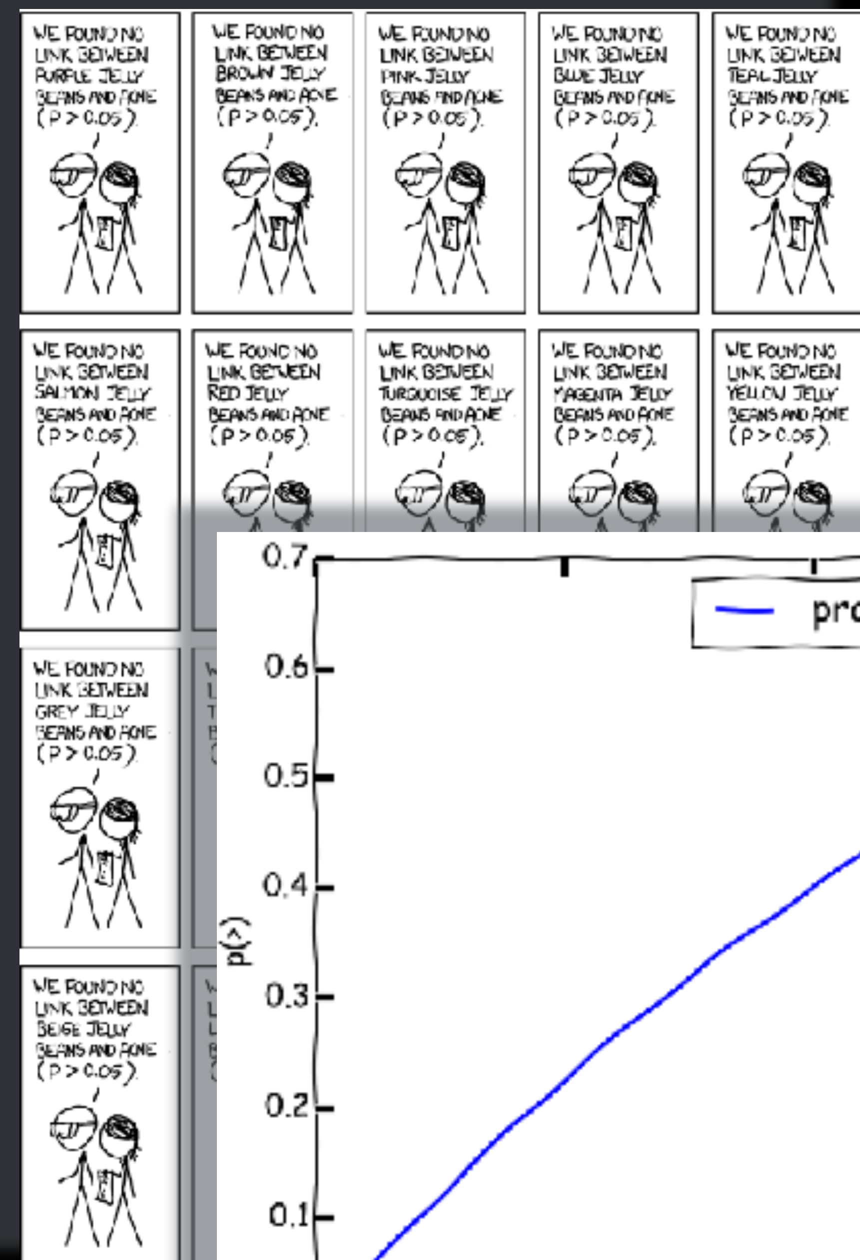
each test has a probability $p<=0.05$ of Type I error significance 95%

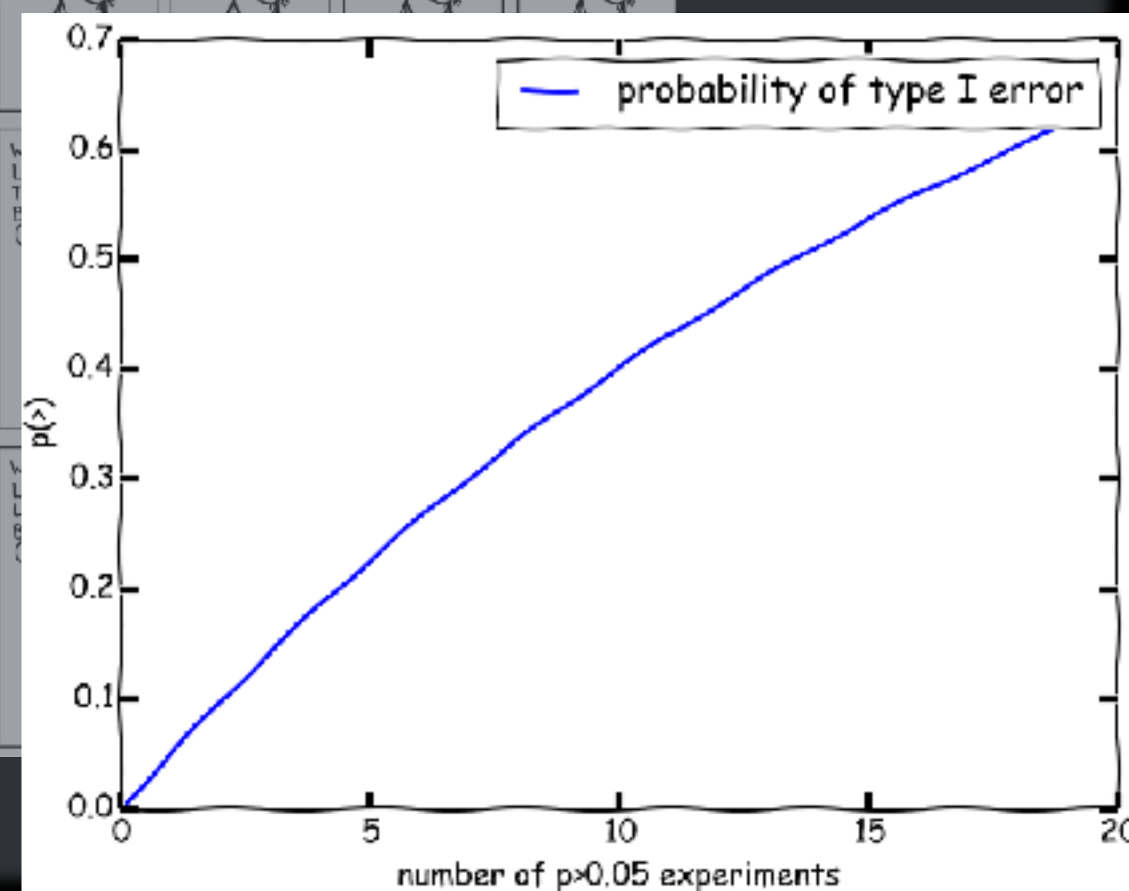
20 tests are preformed

assume independence: if $p_i = 0.05$ for each i=1..20

total significance= $1-(1-0.05)^{20}$

$p_{tot} = (1-0.05)^{20}$





Data Dredging

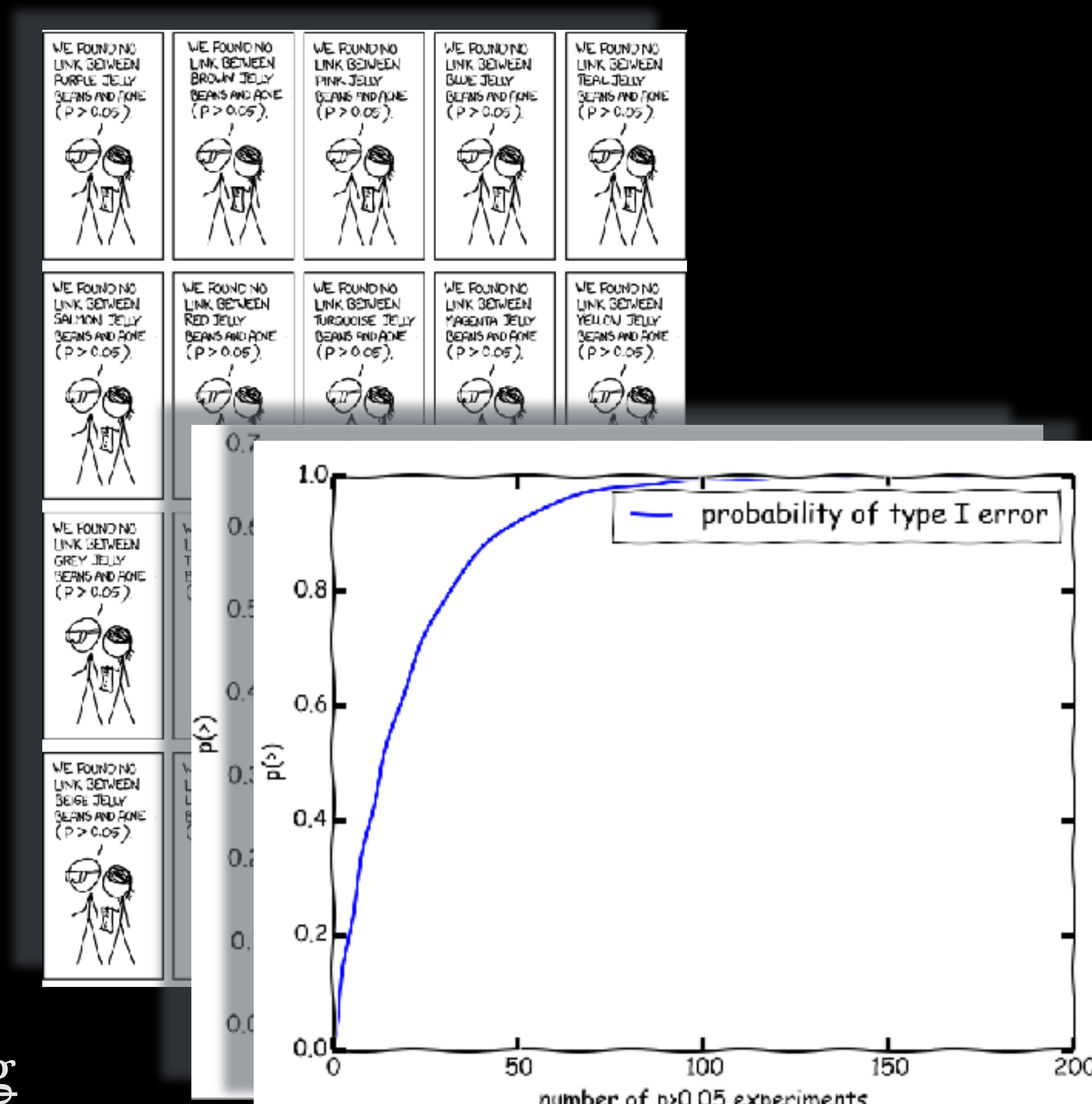
Bias in measurements: know your data

each test has a
probability p<=0.05
of Type I error
significance 95%

20 tests are preformed

assume independence:
if $p_i = 0.05$ for each i=1..20

total significance=
$1-(1-0.05)^{20}$
$p_{tot} = (1-0.05)^{20}$



<u>Data Dredging</u>

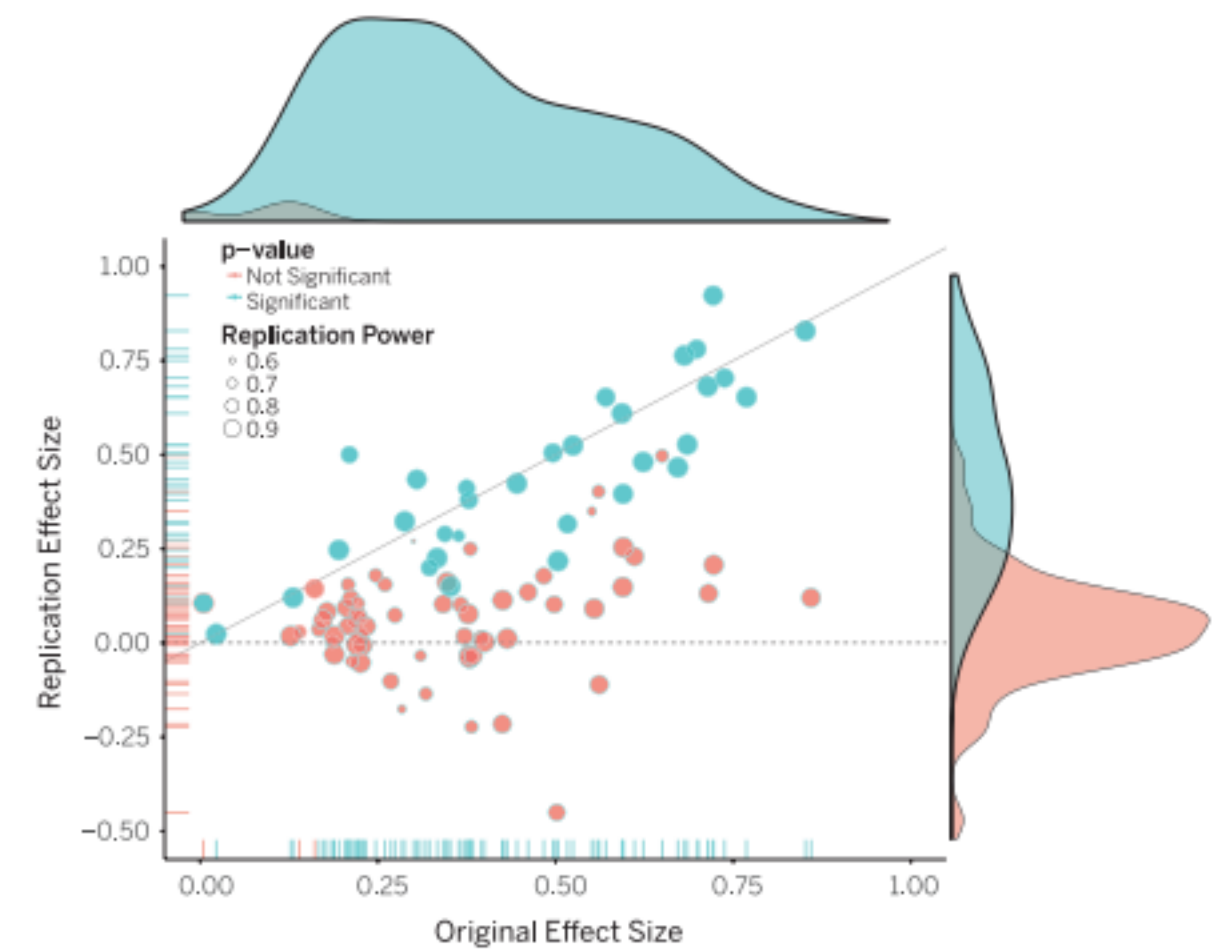
RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

Science,
August 28, 2015



Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

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

Errors and uncertainties.

- Systematic error
  tendency to *systematically*
  underestimate OR 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)

CUSP

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
  Publication Bias
  Data Dredging

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 randomeness) in a process

4.0 +/- 0.1 cm

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 +/- ε cm, ε > 0.1 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.

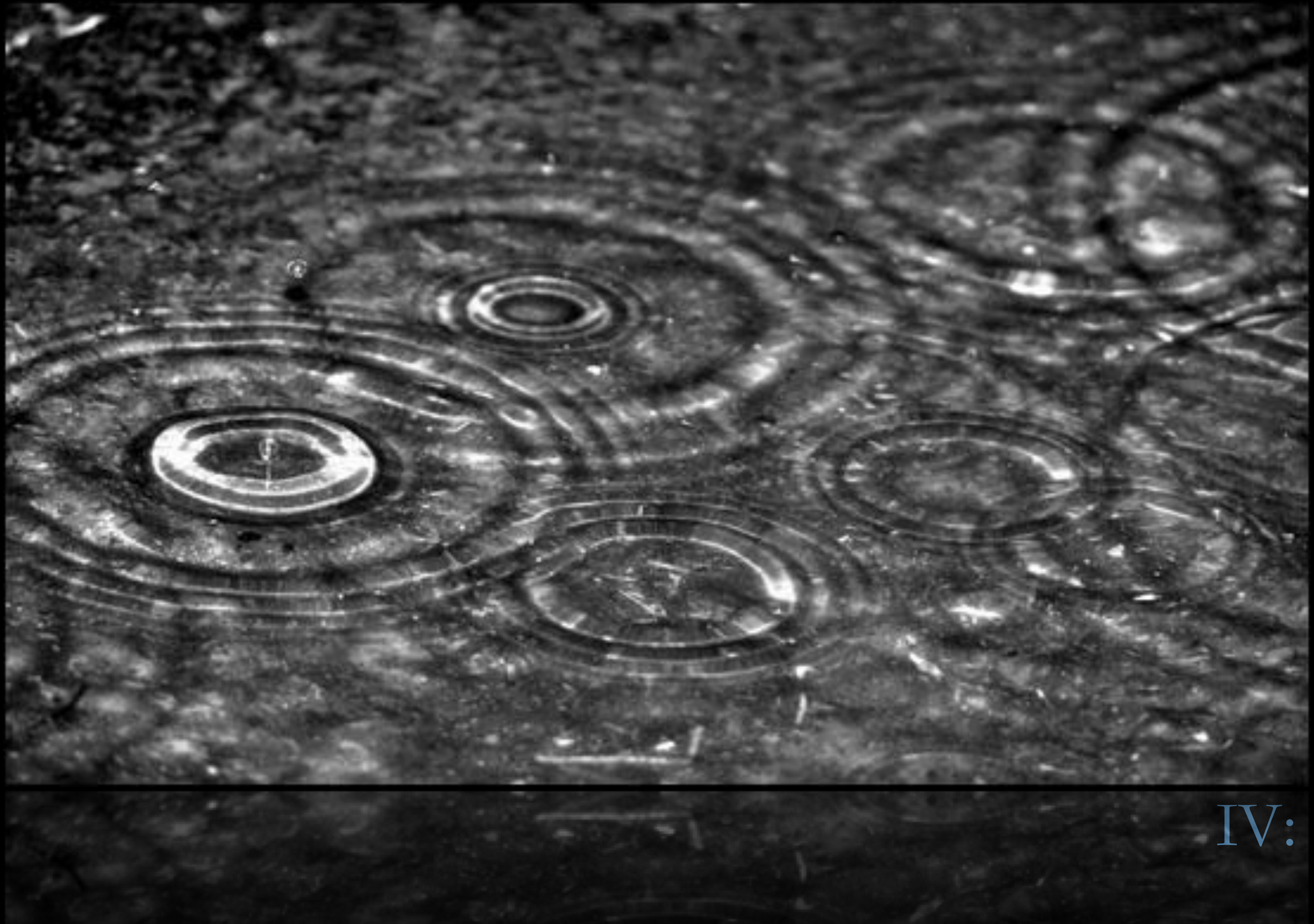
A discrete distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event.

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

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

$$P(\lambda) \xrightarrow[\lambda->\infty]{} N(\lambda, \sqrt{\lambda})$$

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

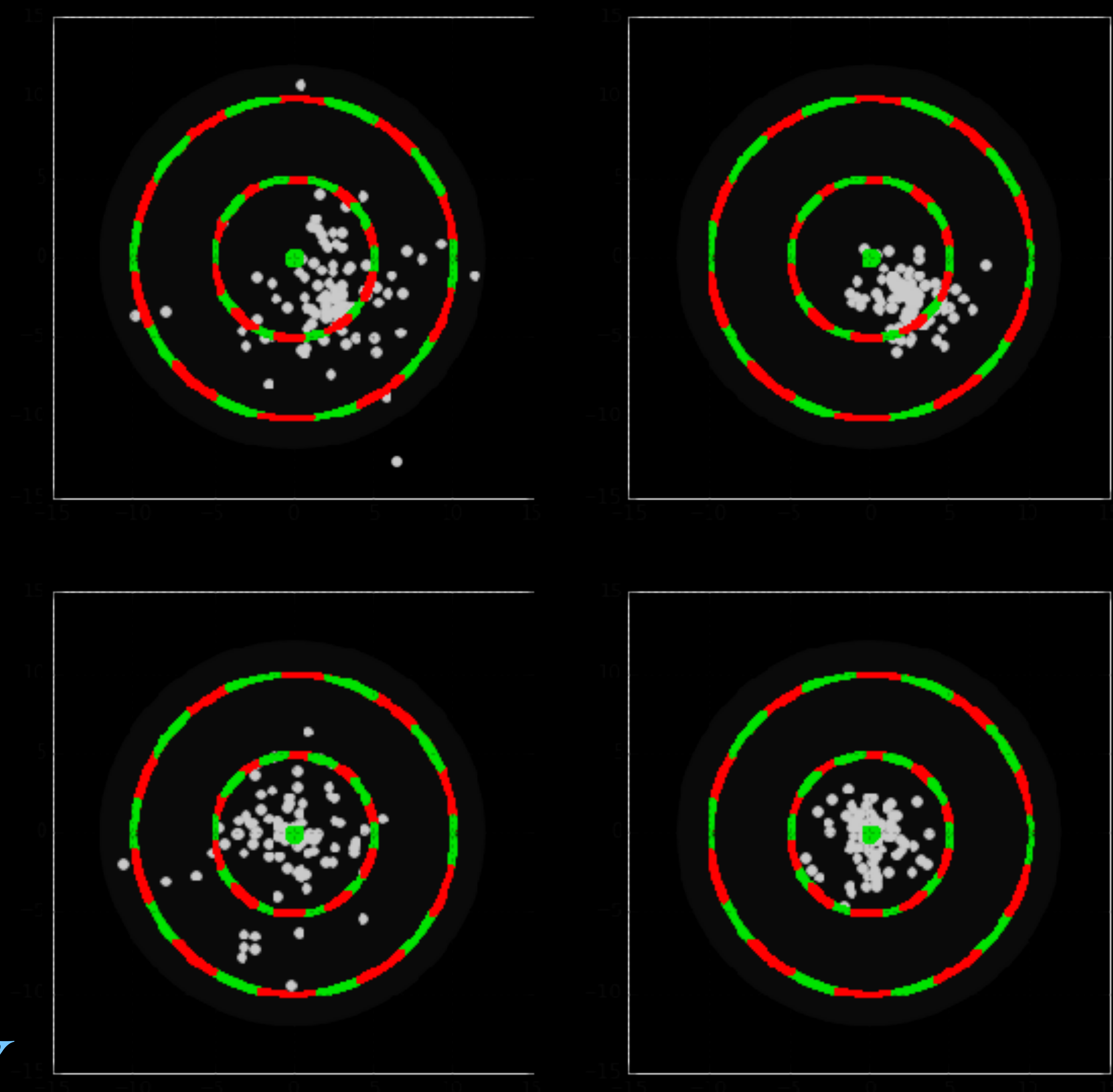
False Positives and False Negatives

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

CUSP

PRECISION



ACCURACY

Error propagation

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

$$x_1 \pm \mathcal{E}(x_1)$$
$$x_2 \pm \mathcal{E}(x_2)$$

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

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

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

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

Covariance matrix

$$\vec{x} = \{x_1, x_2, x_3, \dots, x_n\}$$

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

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

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

Covariance matrix

$$\vec{x} = \underline{\{x_1, x_2, x_3, \ldots, x_n\}}$$

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

IF
<u>Independent variables</u>

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

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

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

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

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:

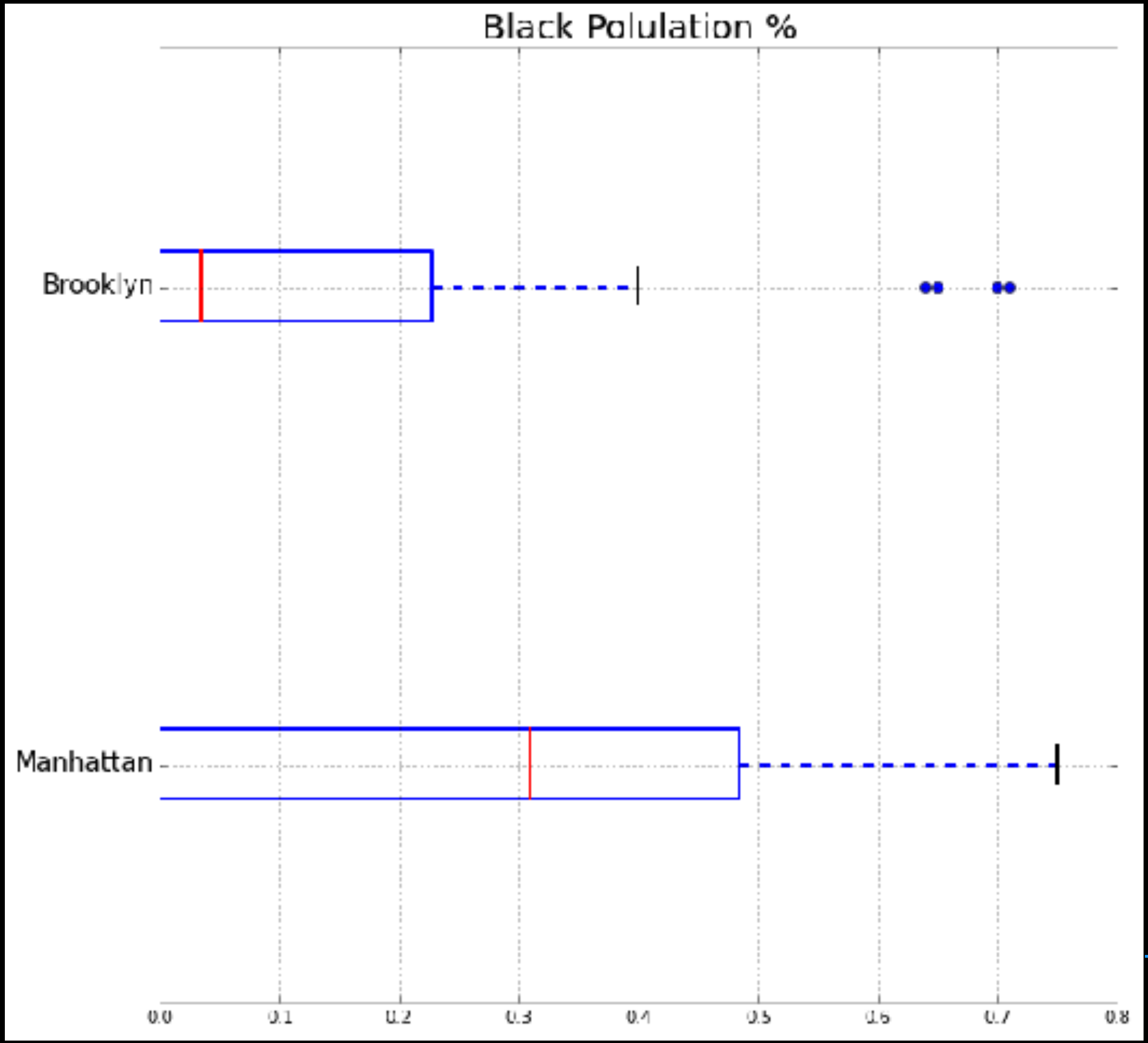
$x = [965 \pm 30(\text{stat}) \pm 12(\text{sys})]$ number of car accidents

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

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

Reporting Your Results

Percentage of Black population by Borrow (Manhattan vs Brooklyn)



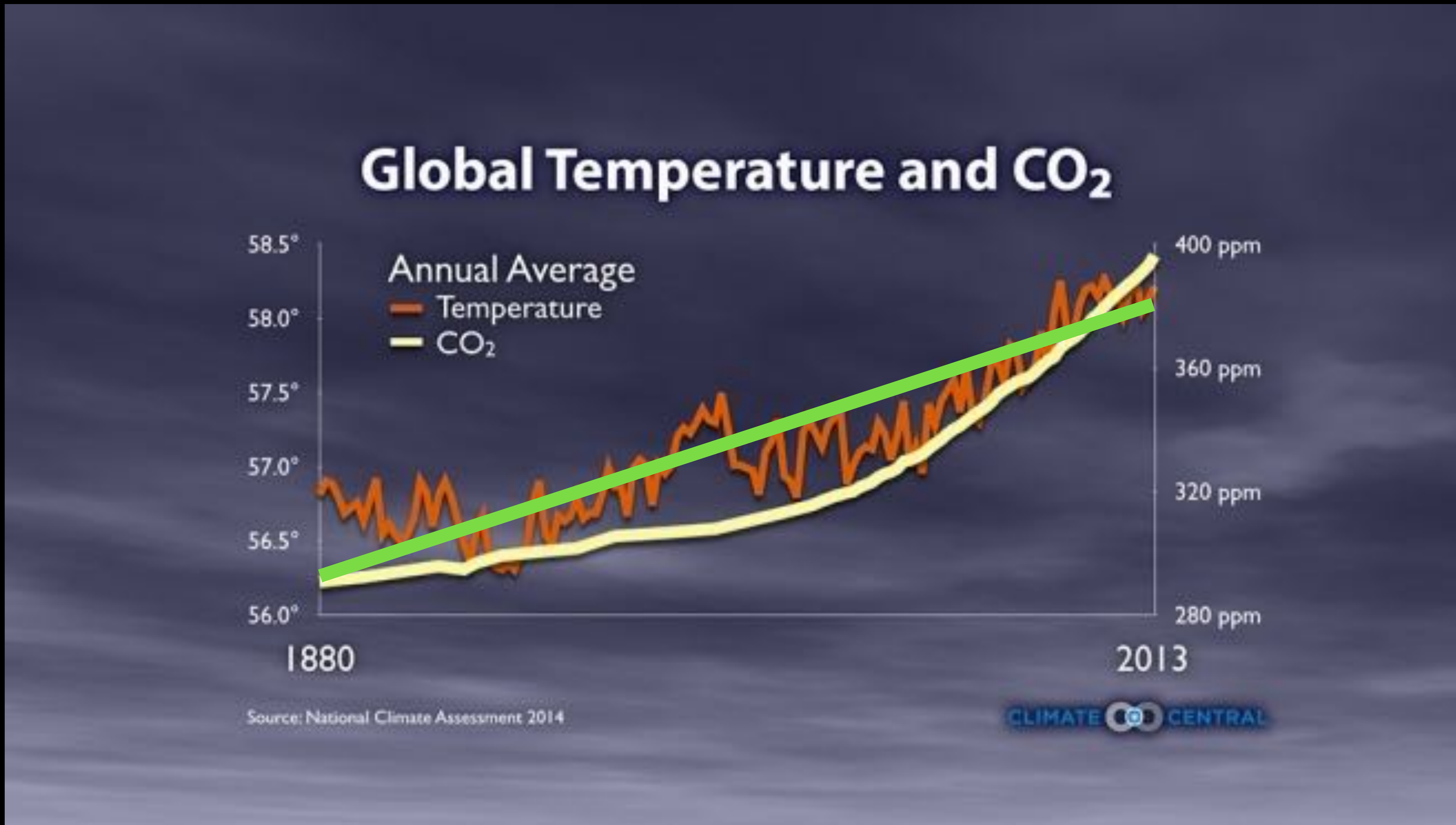
jupyter

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

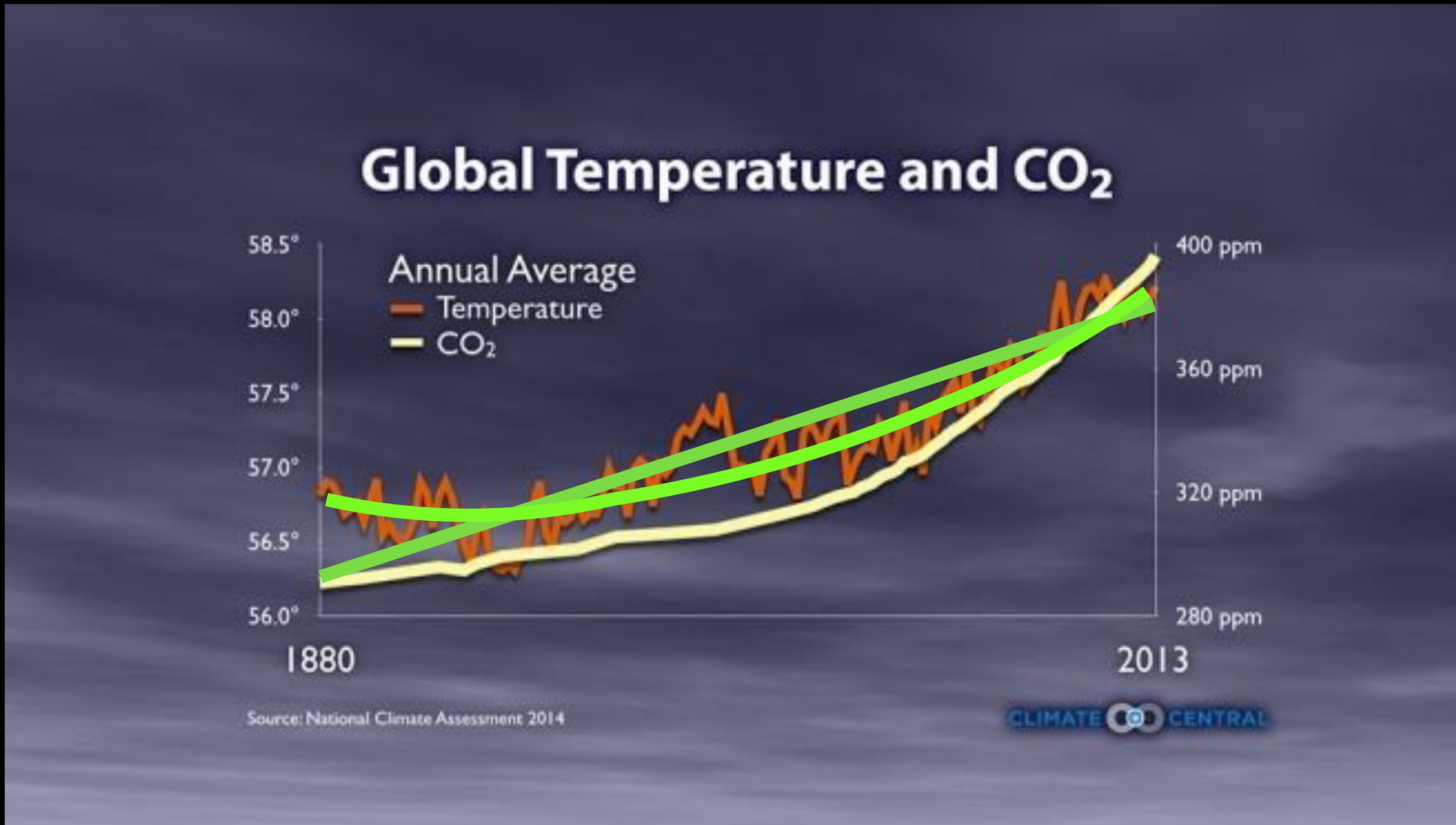
CUSP

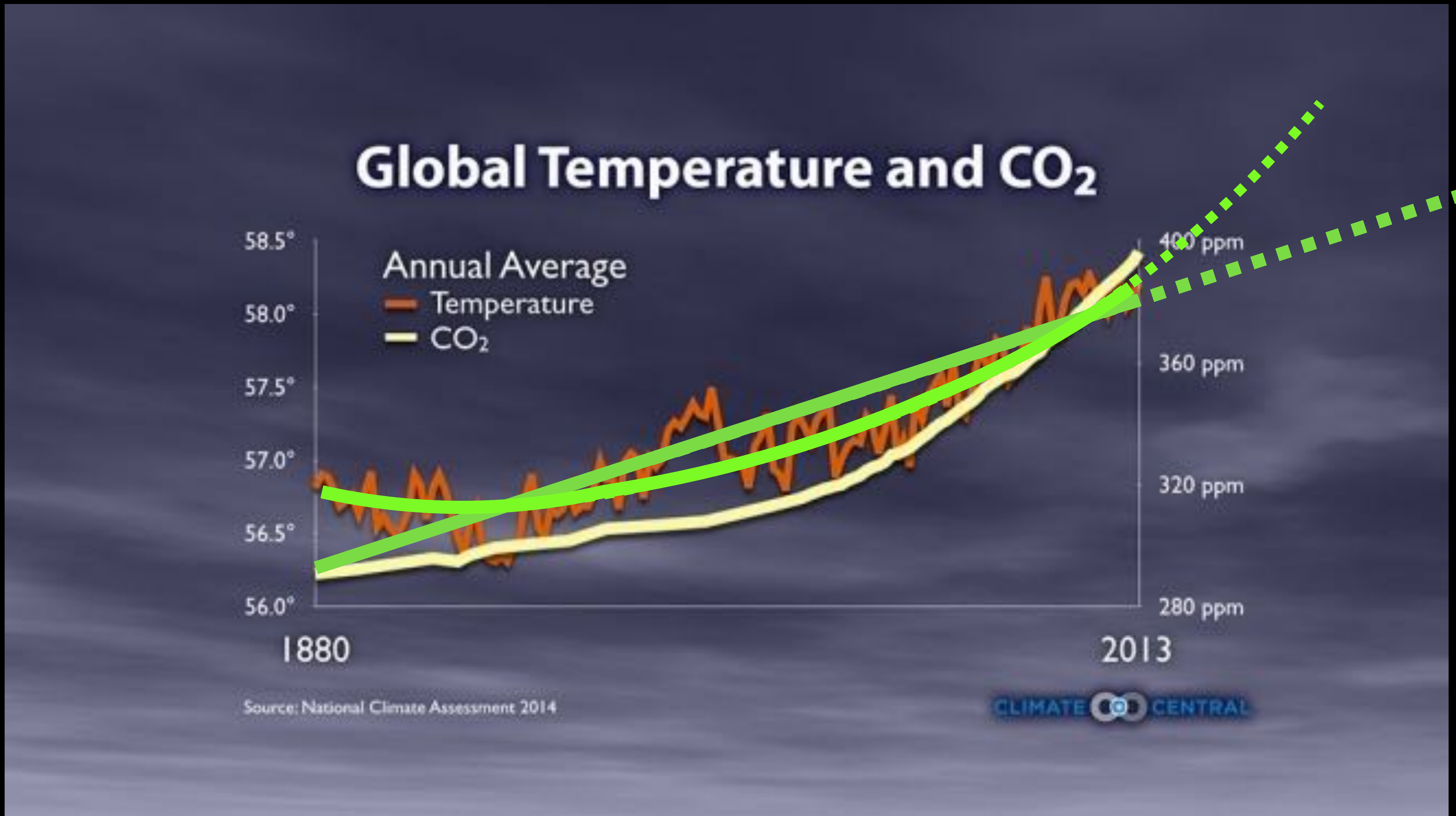
Why?

Predictive

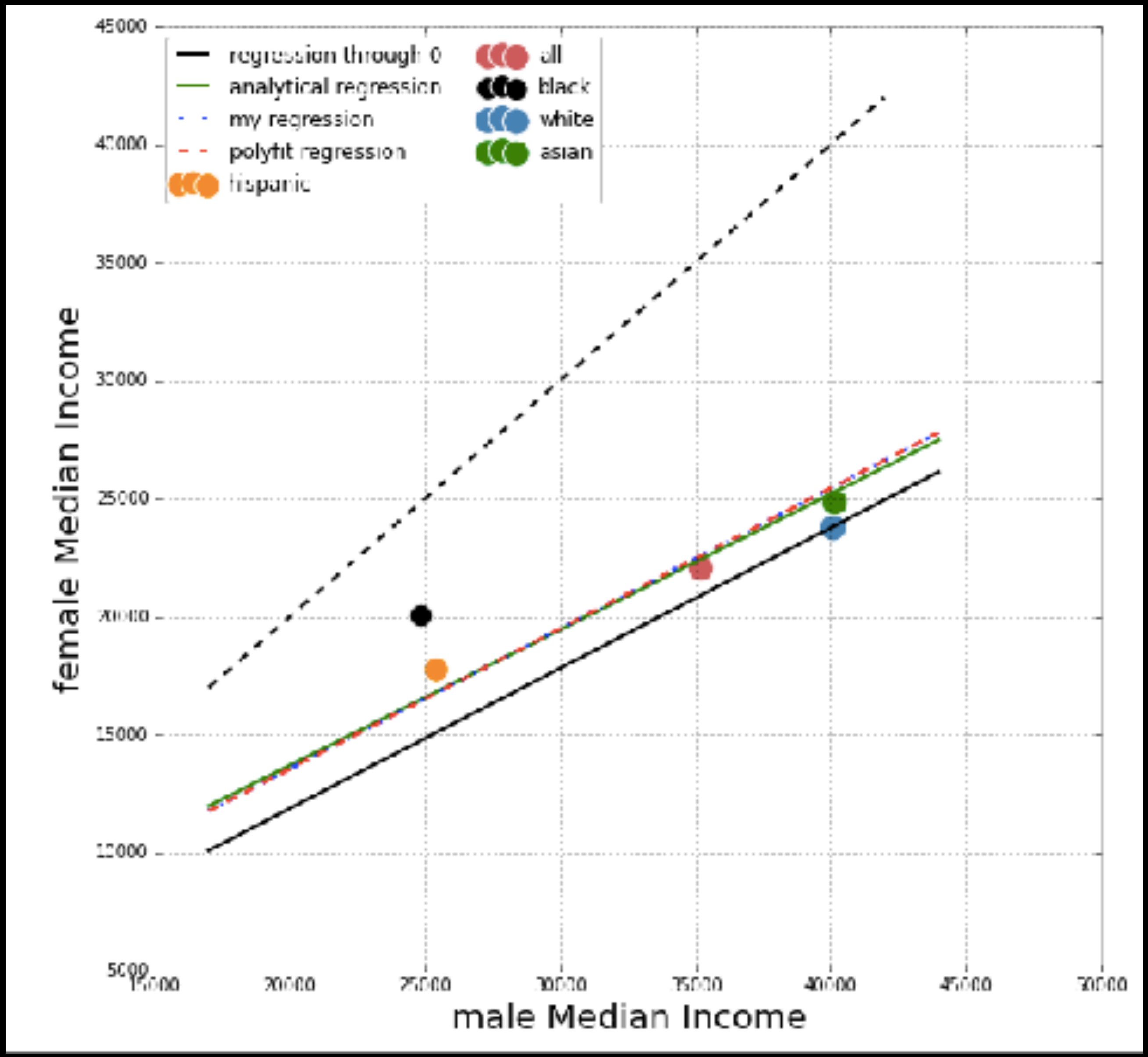


Global Temperature and CO_2
Annual Average
Temperature
CO_2
58.5°
58.0°
57.5°
57.0°
56.5°
56.0°
400 ppm
360 ppm
320 ppm
280 ppm
1880
2013
Source: National Climate Assessment 2014
CLIMATE CENTRAL


Global Temperature and CO_2
Annual Average
Temperature
CO_2
58.5°
58.0°
57.5°
57.0°
56.5°
56.0°
400 ppm
360 ppm
320 ppm
280 ppm
1880
2013
Source: National Climate Assessment 2014
CLIMATE CENTRAL



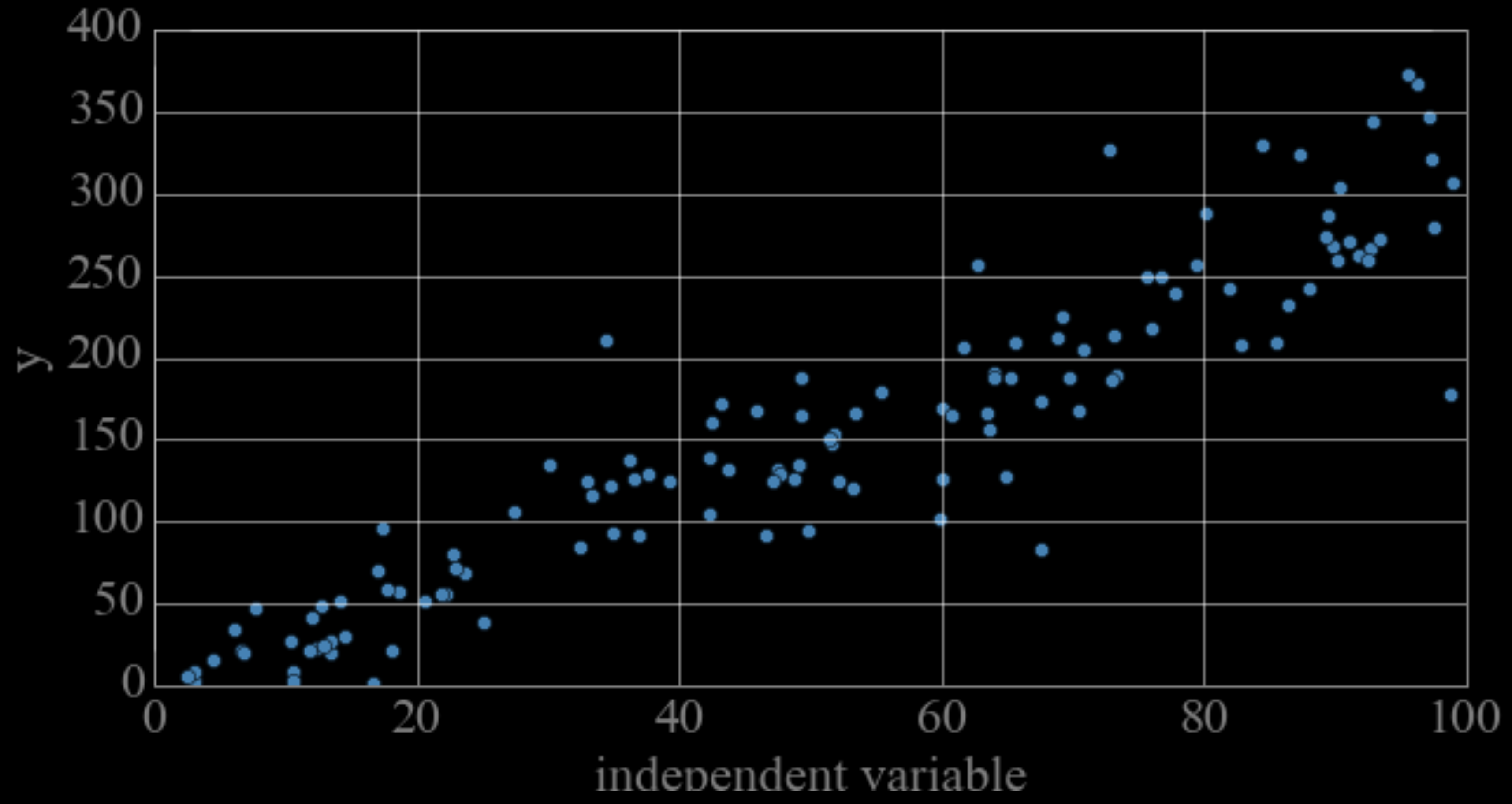
Global Temperature and CO_2
Annual Average
Temperature
CO_2
58.5°
58.0°
57.5°
57.0°
56.5°
56.0°
400 ppm
360 ppm
320 ppm
280 ppm
1880
2013
Source: National Climate Assessment 2014
CLIMATE CENTRAL


regression through 0
analytical regression
my regression
polyfit regression
hispanic
all
black
white
asian
45000
40000
35000
30000
25000
15000
female Median Income
male Median Income

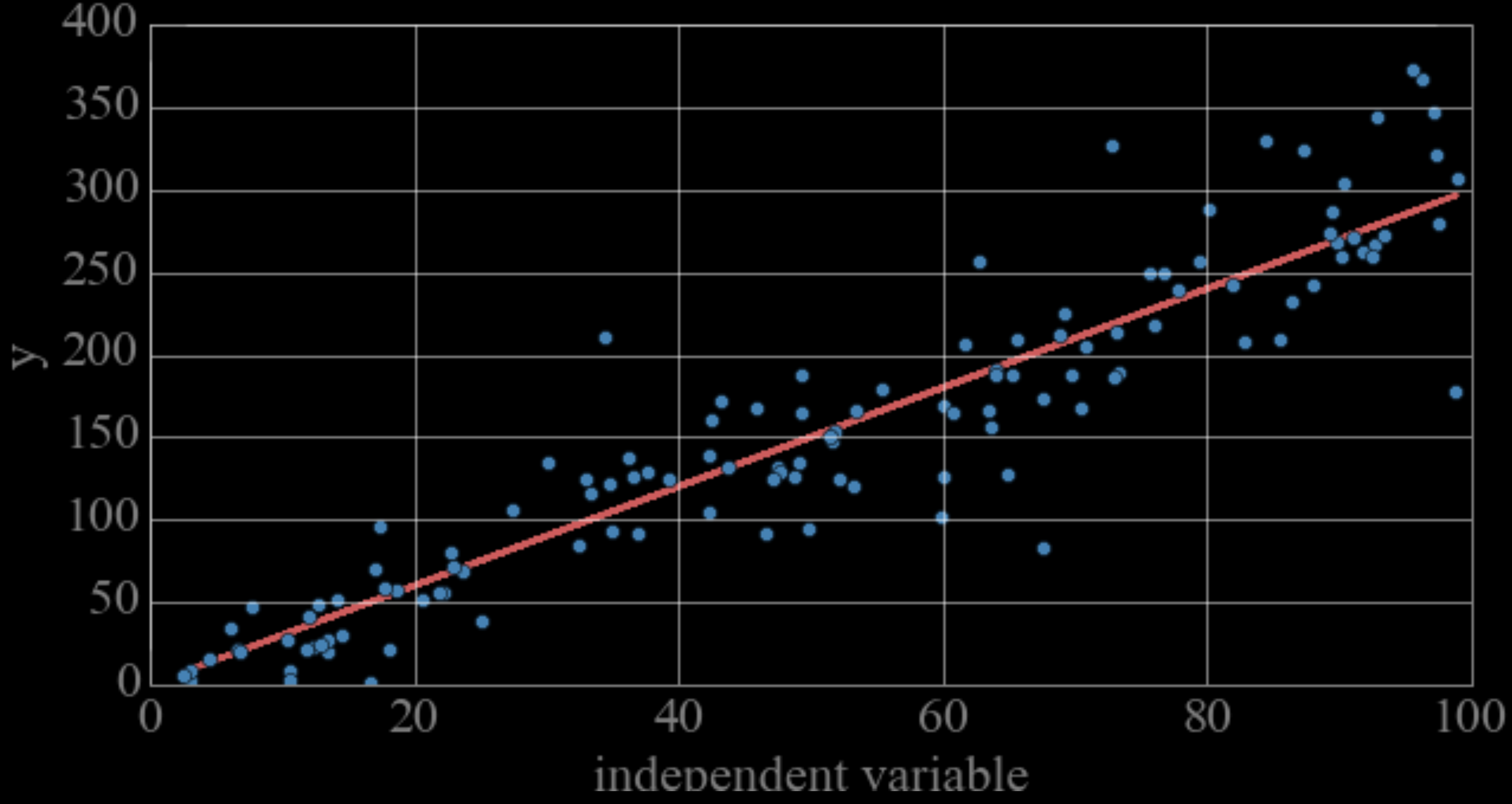
jupyter

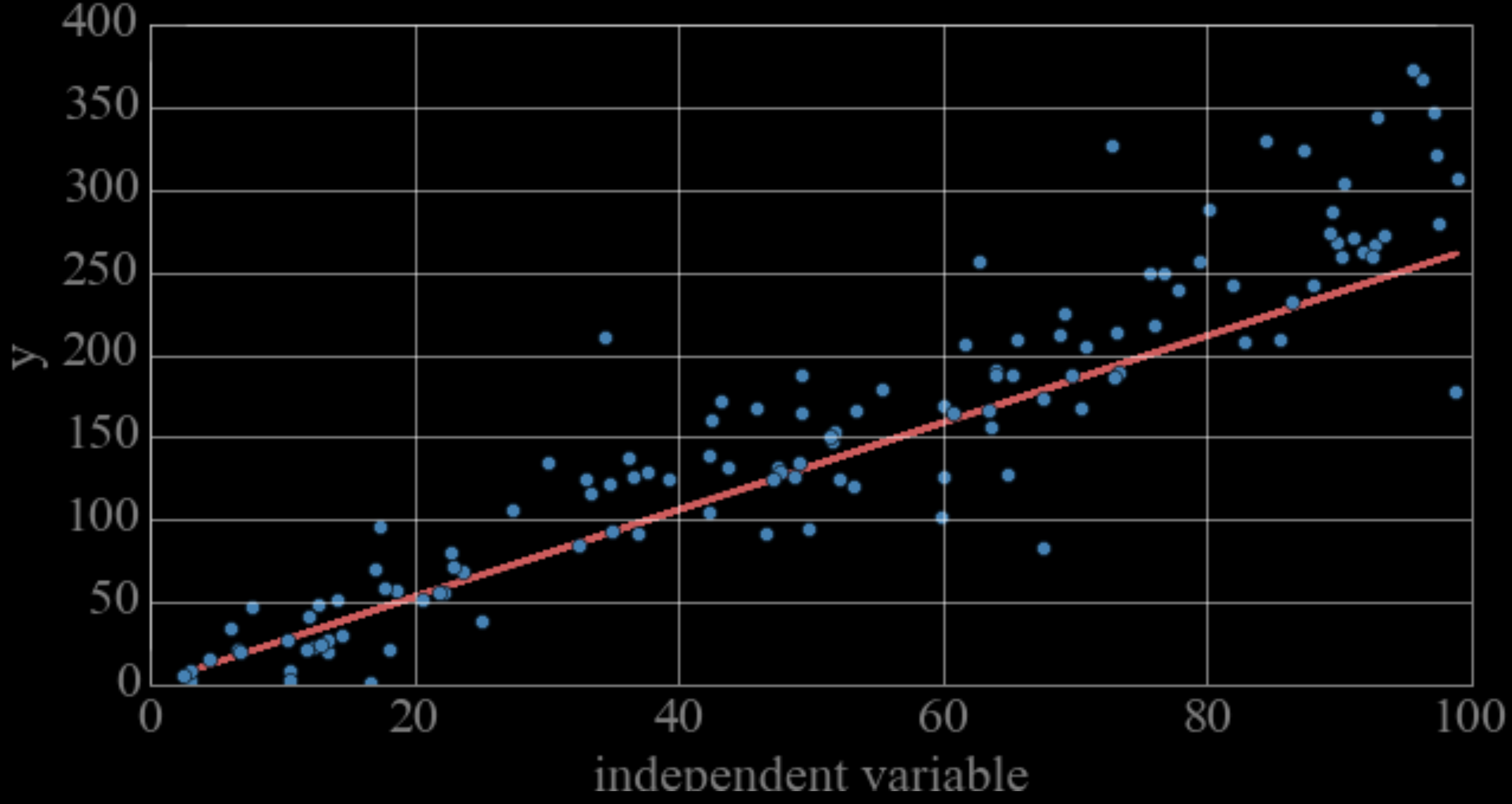
https://github.com/fedhere/UInotebooks/blob/master/OLS/line_fit_and_residuals.ipynb

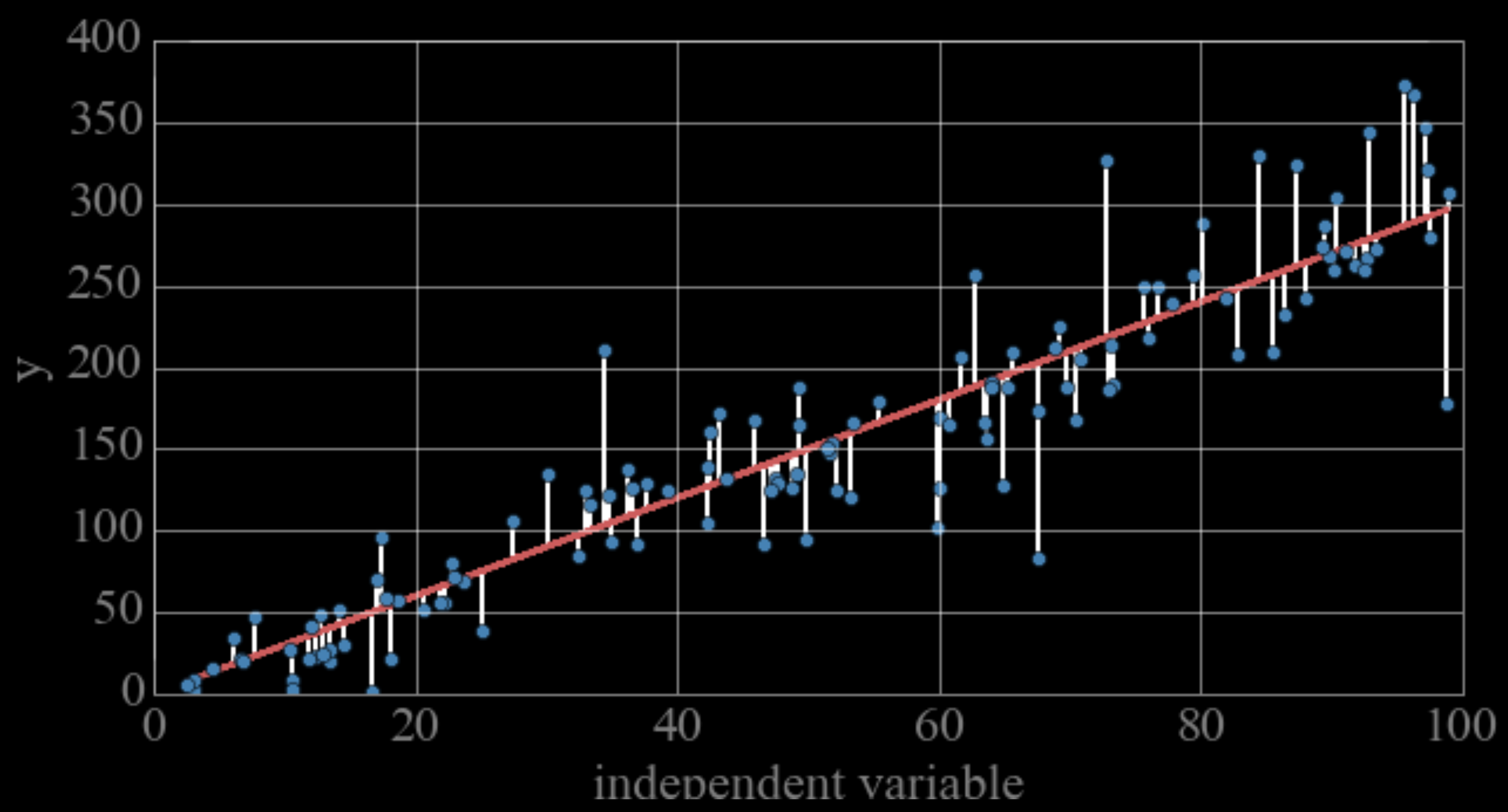
How?

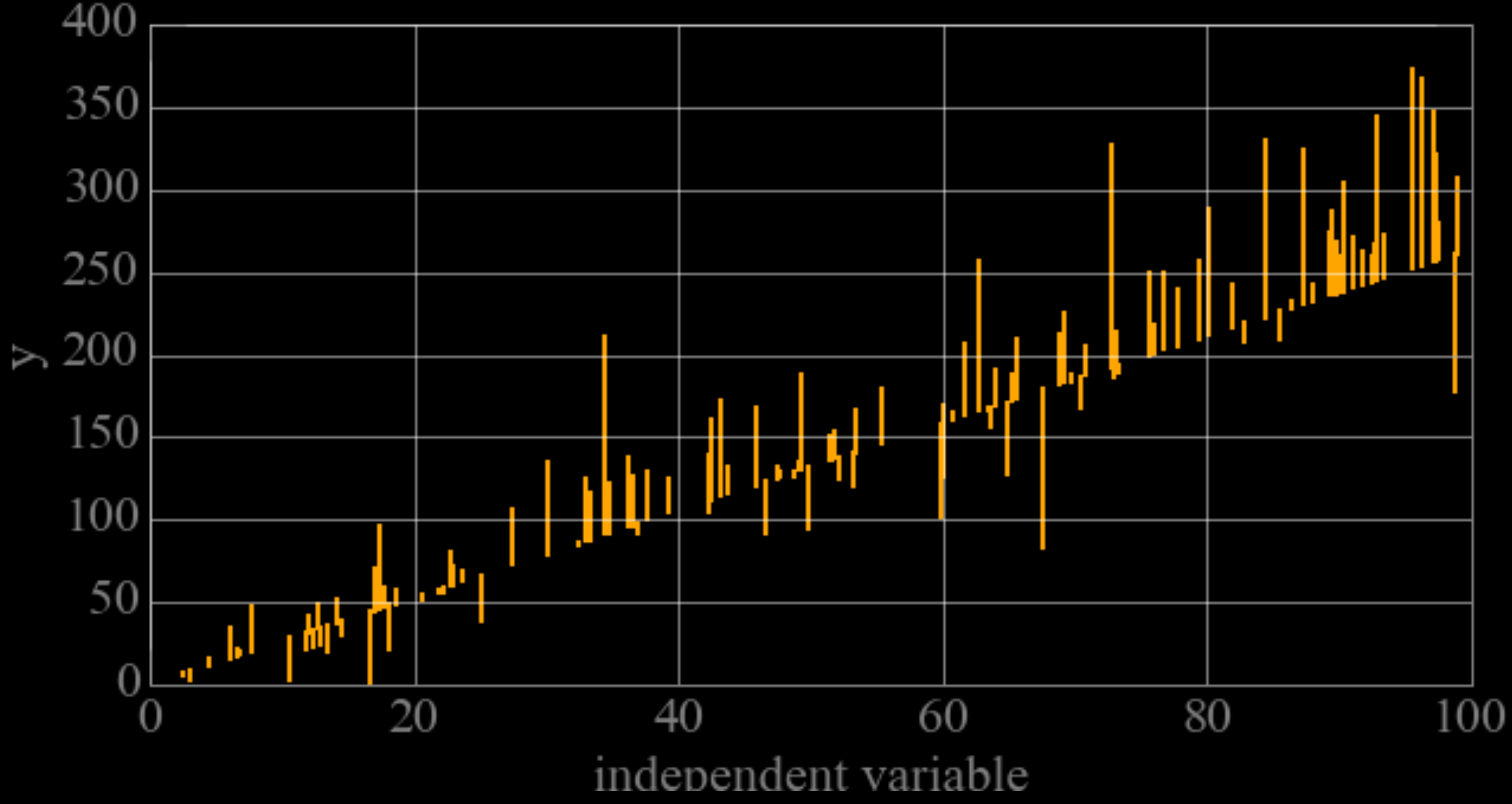


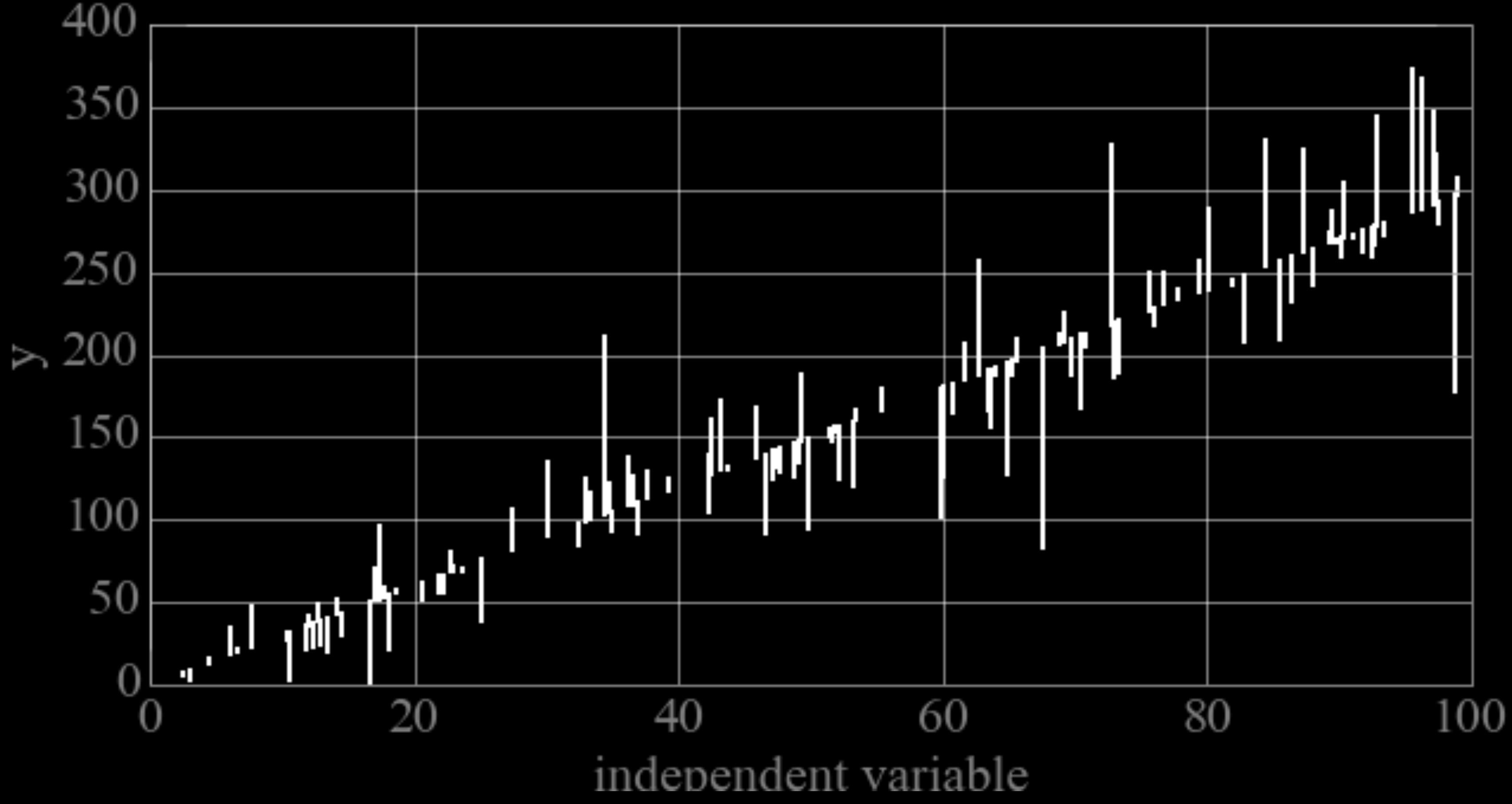
400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable


400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable


400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable


400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable

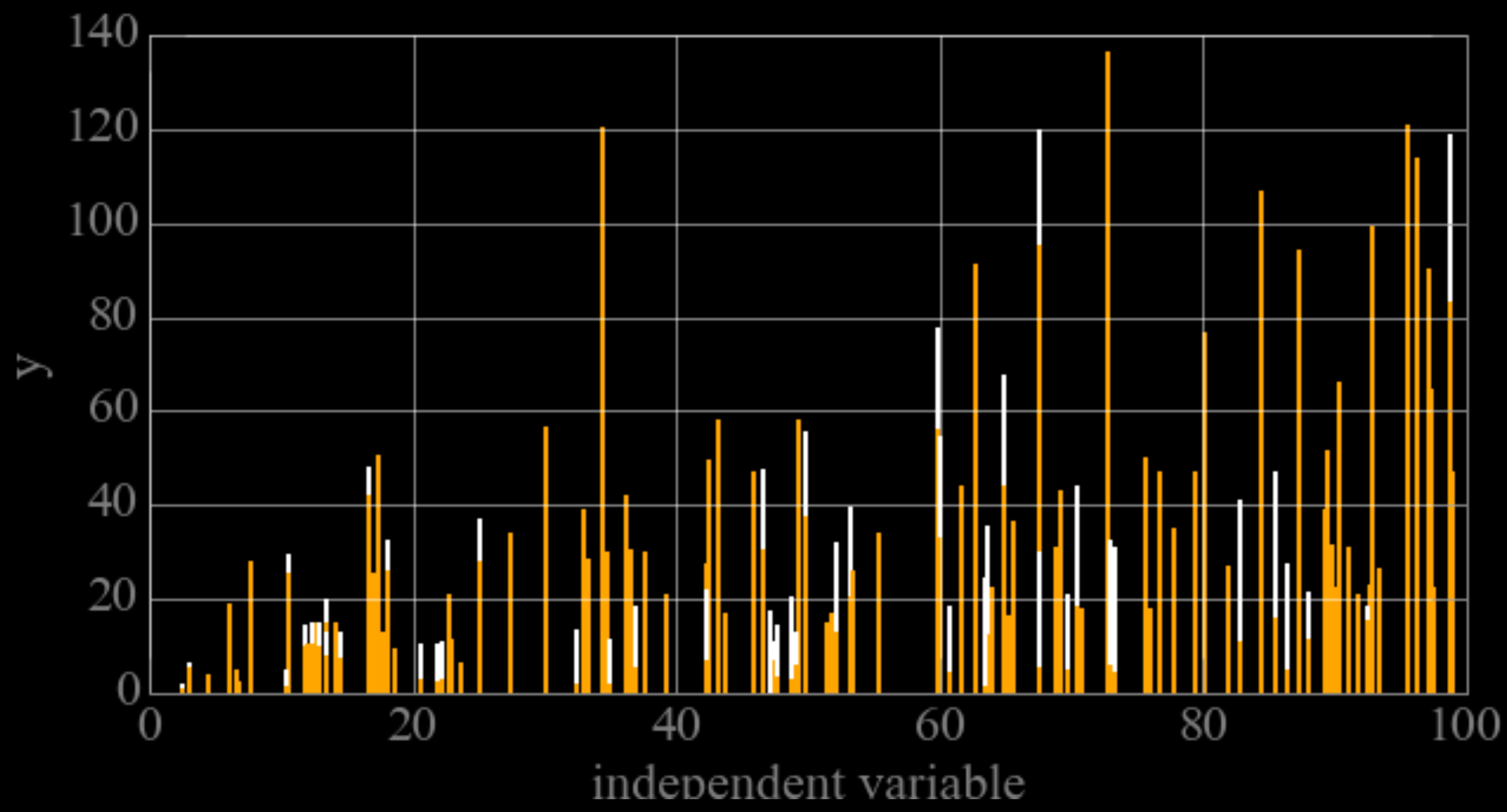

400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable



400
350
300
250
200
150
100
50
0
y
0
20
40
60
80
100
independent variable

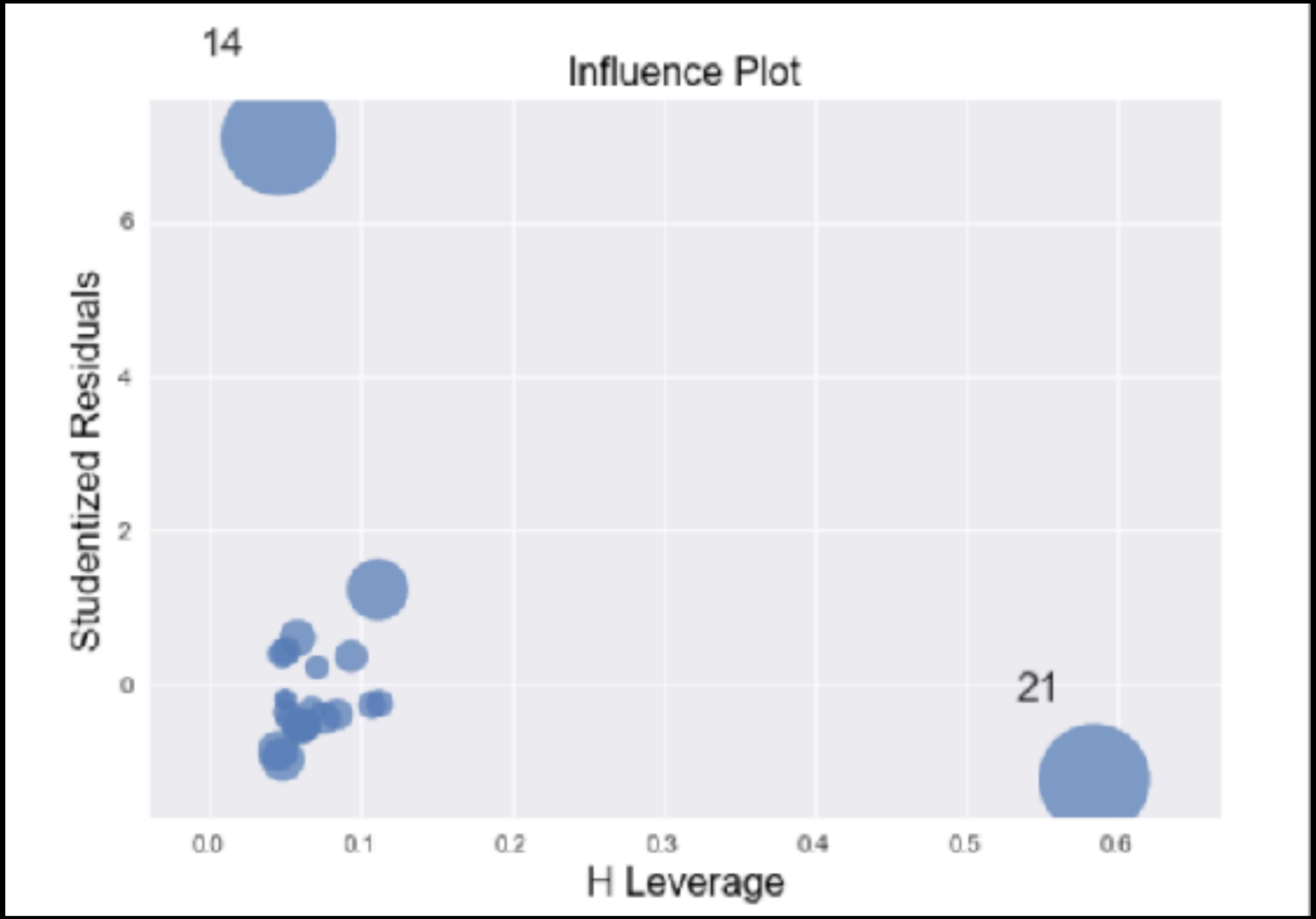
Sum of residuals squared $\sum_i (|y_i - (mx_i + b)|)$

11655.34 12155.24



Assignment 3: regression

investigate linear relationships between fire arm possession, homicides by fire arms, and mass shootings for different countries, considering also the country GDP



Example of liner regression 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 men, 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

POINTS OF SIGNIFICANCE

Analyzing outliers: influential or nuisance?

Some outliers influence the regression fit more than others.

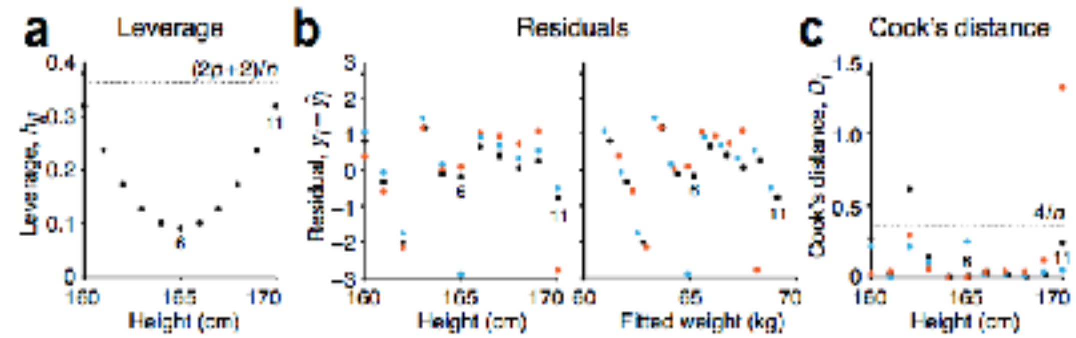


Figure 2 | The leverage, residual and Cook's distance of an observation are used to assess the robustness of the fit. (**a**) The leverage of an observation tells us about its potential to influence the fit and increases as the square of the distance from the predictor value to its mean. Shown are leverage

https://www.nature.com/nmeth/journal/v13/n4/pdf/nmeth.3812.pdf

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
Introduction to General Linear Regression (Chap 12 in most versions)
https://books.google.com/books/about/Statistics_in_a_Nutshell.html?id=ZnhgO65Pyl4C

PennState online statistics course: Lesson 35: Simple Linear Regression
https://onlinecourses.science.psu.edu/stat414/node/262

David M. Lane et al.
Introduction to Statistics (XVIII)
regression : Chapter 14
http://onlinestatbook.com/Online_Statistics_Education.epub
http://onlinestatbook.com/2/index.html

Error Analysis from UPenn physics labs
these are prepared for the phyiscs undergrads labs, which I taught while in gradschool. The examples are really physics, so more relevant to remote sensing than social science type projects, but the error propagation etc works the same way
http://virgo-physics.sas.upenn.edu/uglabs/lab_manual/Error_Analysis.pdf