Urban Informatics

Fall 2018

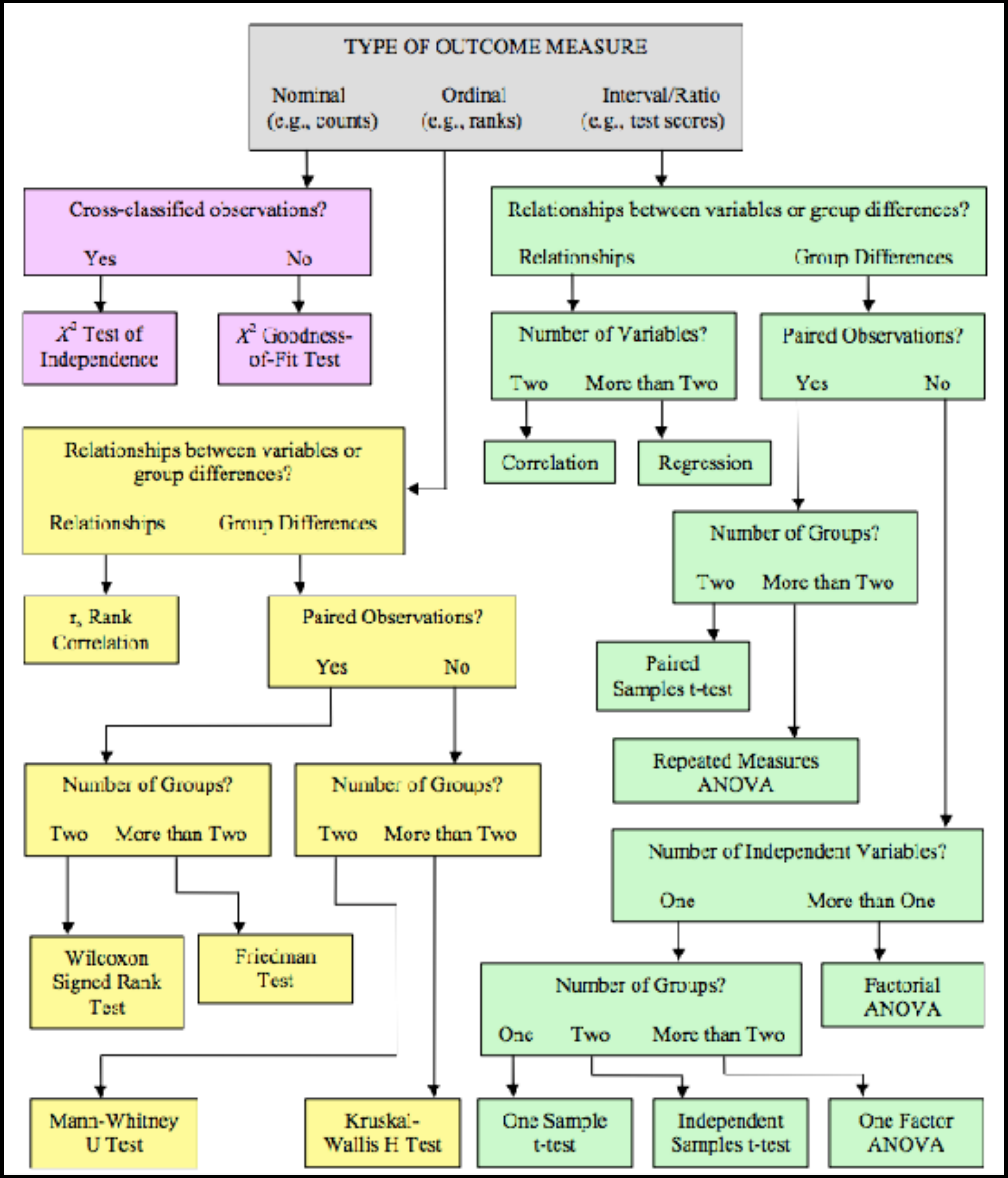
dr. federica bianco fbianco@nyu.edu

@fedhere

CUSP

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



IV: Statistical analysis

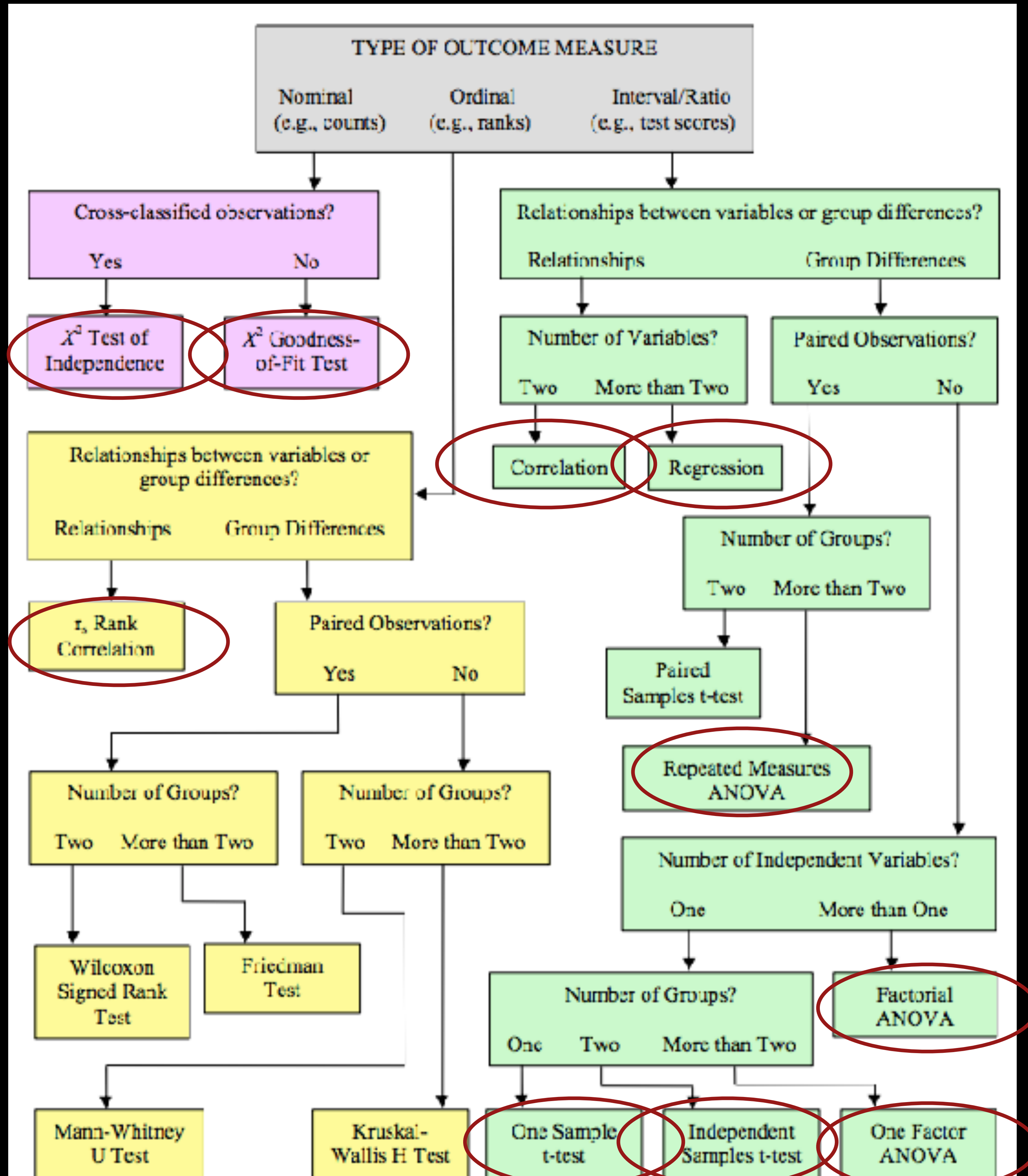
Recap:

- 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:

- Statistical and Systematic errors
- Goodness of fit tests

CUSP



IV: Statistical analysis

hypothesis testing

did we detect a phenomenon?

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?

CUSP

A simple (too simple?) answer

did we detect a phenomenon?

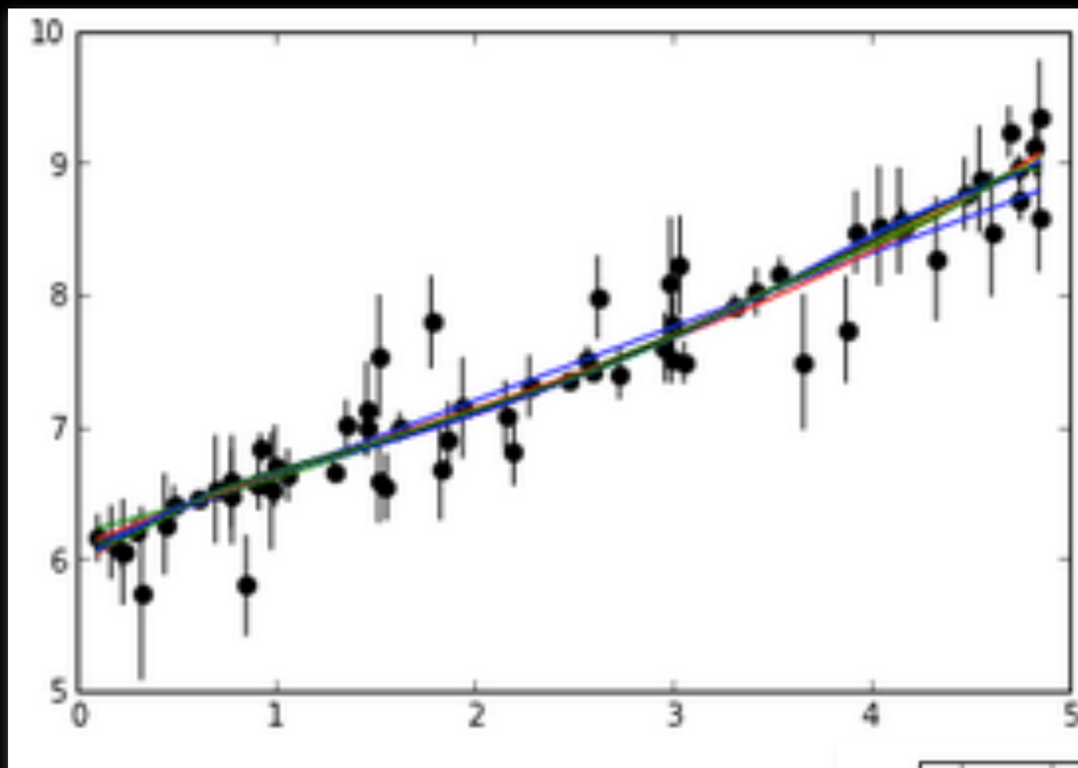
p-value a measure of the probability that the result you observed could have been observed by chance under the *Null hypothesis*

Effect size in place of test statistics

Estimate the strength of, for example, an apparent relationship, rather than assigning a significance level. The effect size does not directly determine the significance level, or vice versa.

A more complex and complete answer…

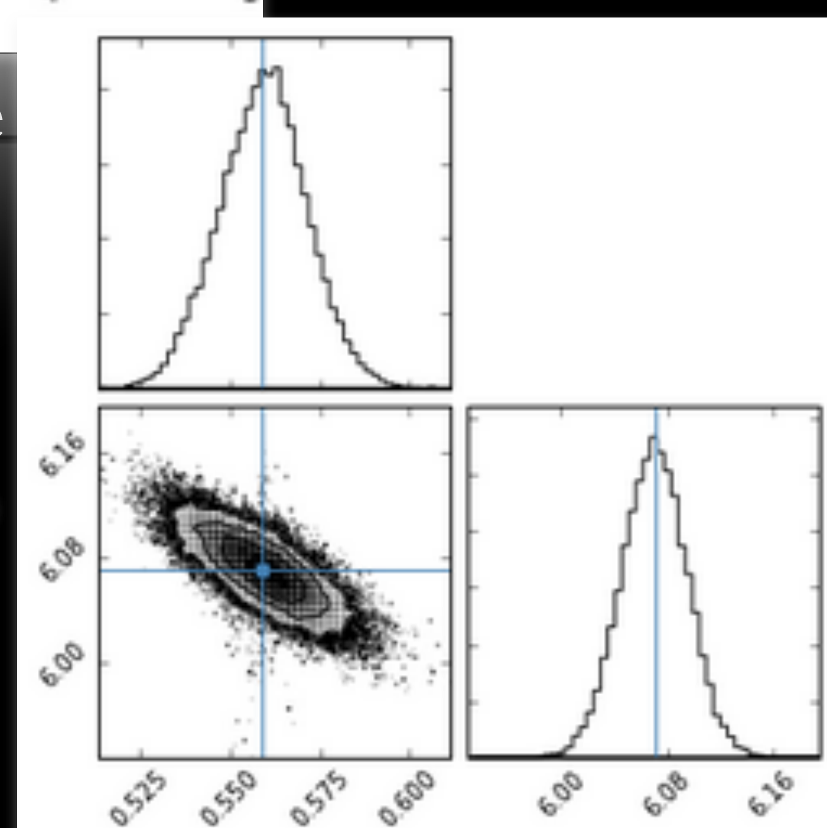
independent variable



dependent variable

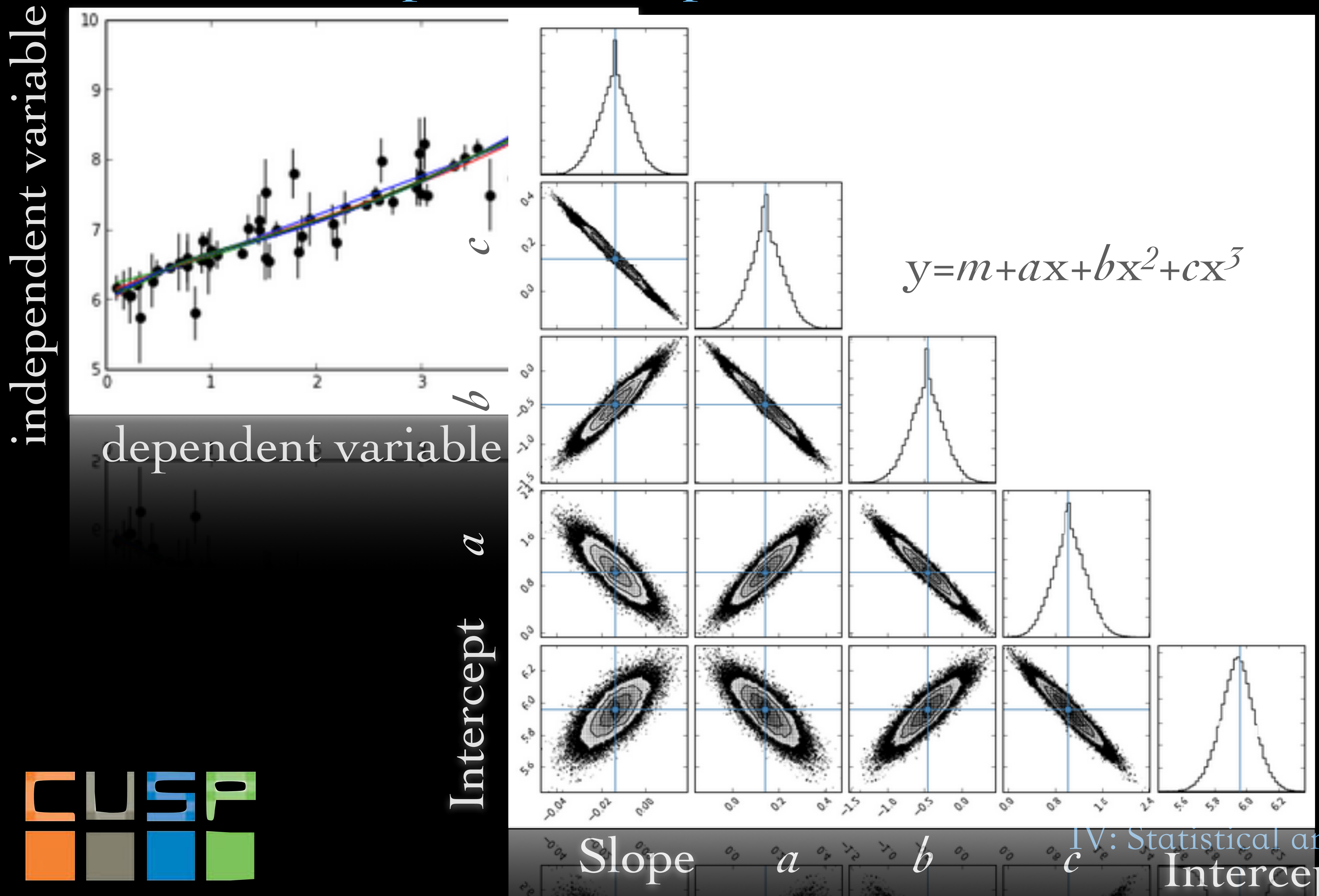
$$y=m+ax$$

Intercept



Slope Intercept

A more complex and complete answer…



Errors and uncertainties.

Errors and uncertainties.

- Systematic error
- Statistical error (Stochastic & Random)

systematic errors

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

Apps | Bookmarks | gmail | pedometer | http://www.weather.c | New folder | SESN 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

- Systematic error: SURVEY BIAS

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

different surveys/experiments give different results
because they have different systematic errors or bias

Bias in measurements: know your data

- Systematic error: SURVEY BIAS

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

Undercoverage bias

Bias in measurements: know your data

- Systematic error: SURVEY BIAS

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

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

- Systematic error: SURVEY BIAS

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

Self selection bias

Bias in measurements: know your data

- Systematic error: SURVEY BIAS

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

Self selection bias

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

- Systematic error: SURVEY BIAS

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

<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…
people willing to answer a survey about climate are likely more concerned and responsible citizens

Bias in measurements: know your data

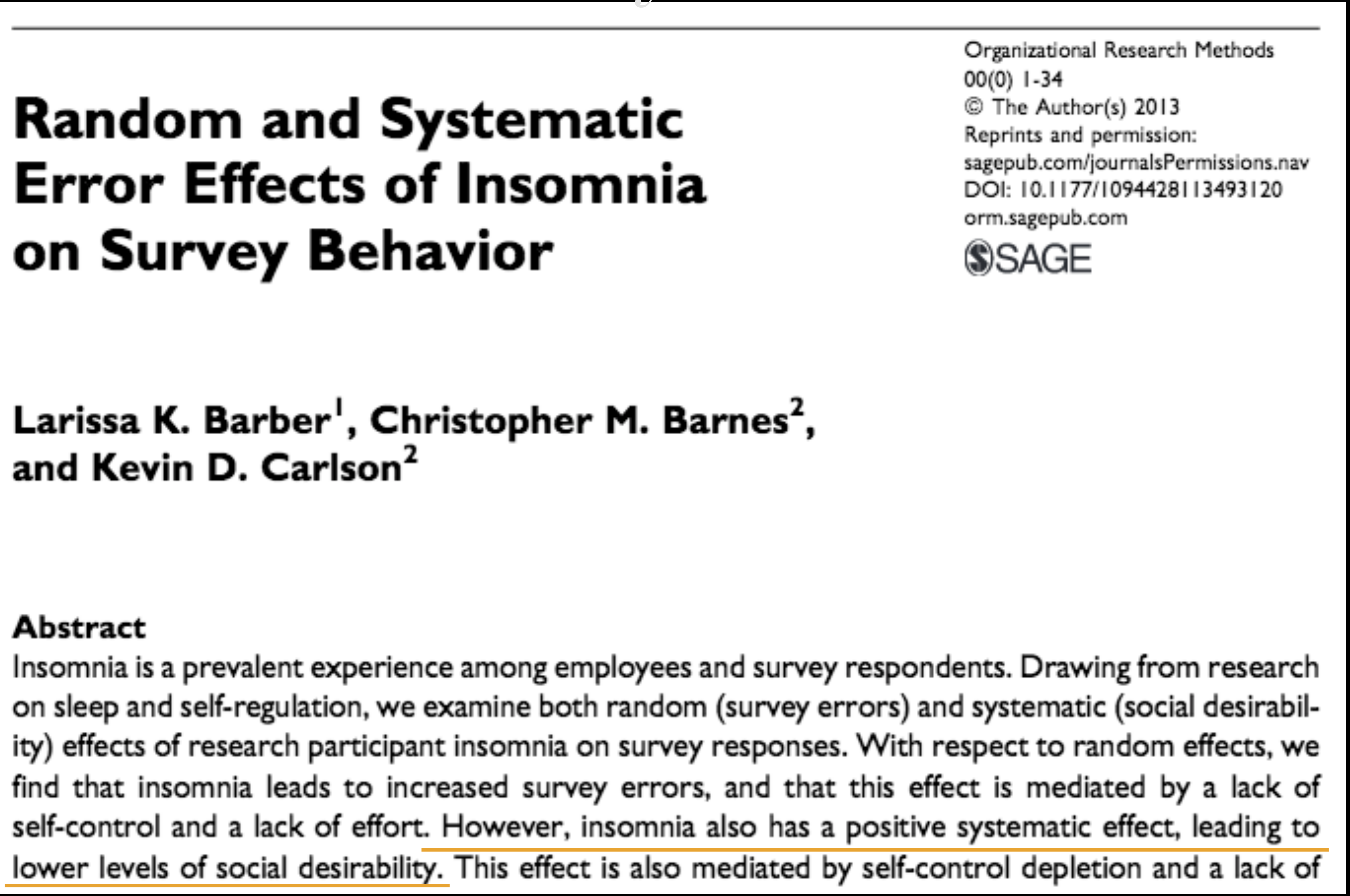
- Systematic error: SURVEY BIAS

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

Social desirability bias

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

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

- Systematic error: SURVEY BIAS

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

Undercoverage bias
Self selection bias
Social desirability bias

Other systematic errors
Publication Bias
Data Dredging

Bias in measurements: know your data

- Systematic error: PUBLICATION BIAS


nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | F

News & Comment > News > 2015 > August > Article

NATURE | NEWS

Social sciences suffer from severe publication bias

Survey finds that 'null results' rarely see the light of the day.

Mark Peplow

28 August 2014

Publication Bias

Bias in measurements: know your data

- Systematic error: PUBLICATION BIAS

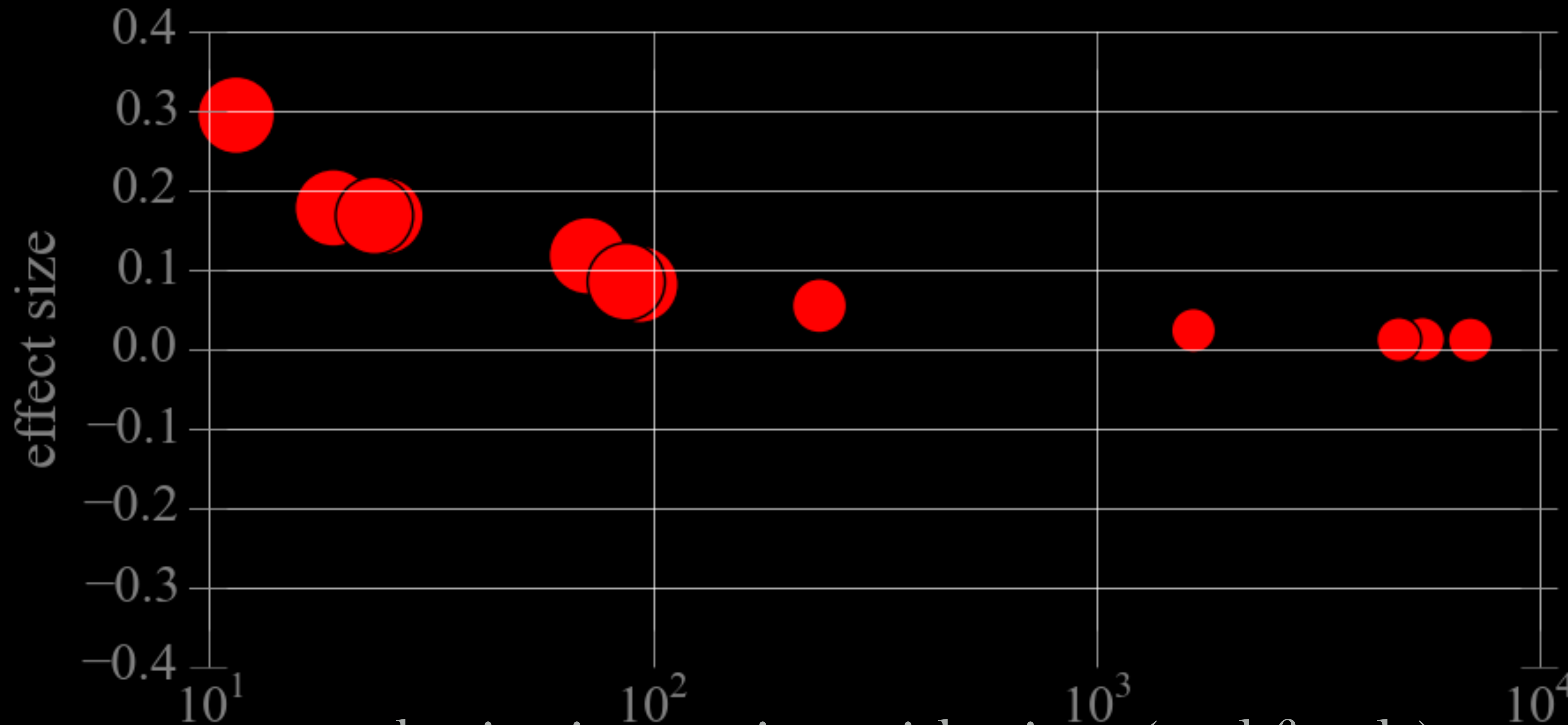
His team investigated the fate of 221 sociological studies conducted between 2002 and 2012, which were recorded by Time-sharing Experiments for the Social Sciences (TESS), a US project that helps social scientists to carry out large-scale surveys of people's views.

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

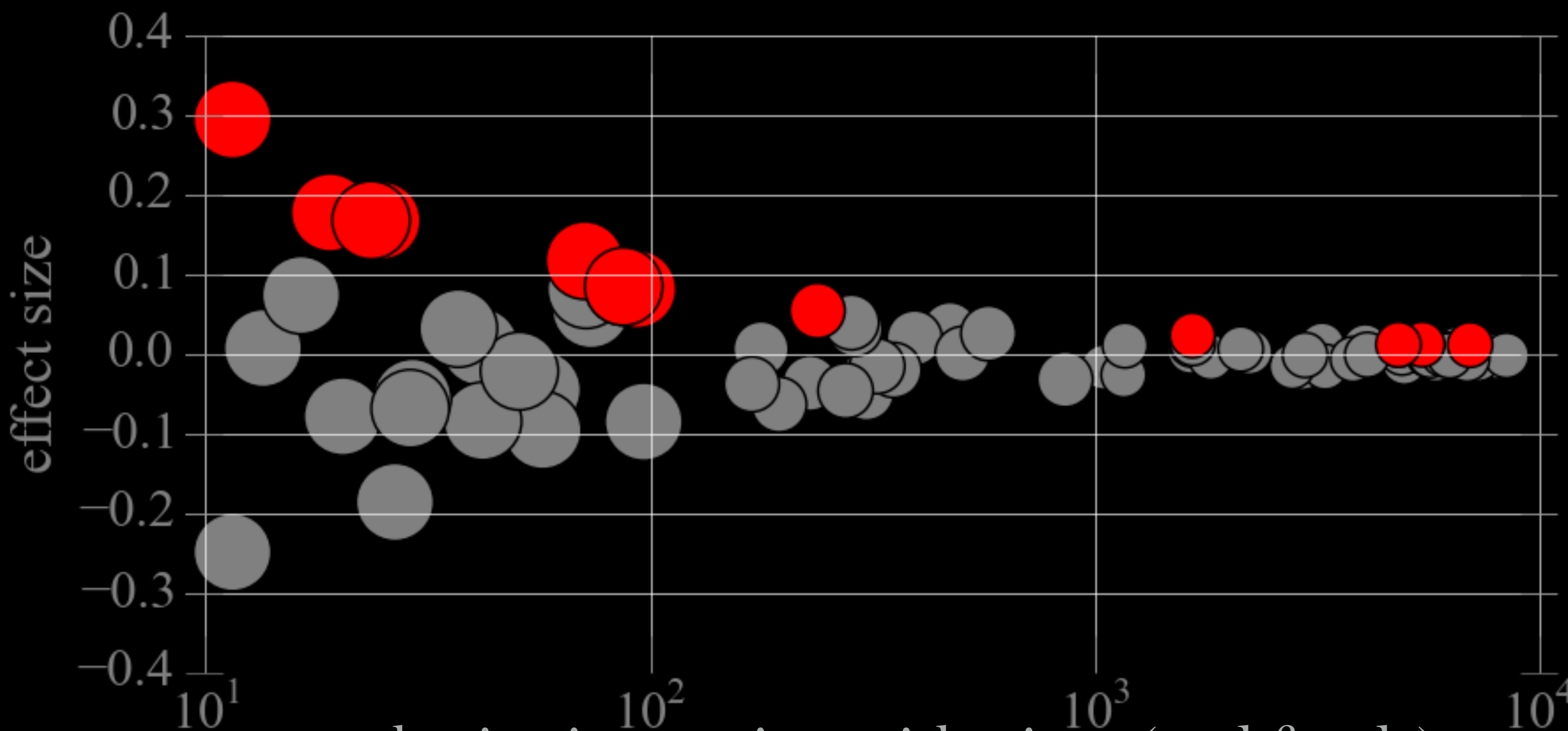
Bias in measurements: know your data



sample size increasing with time (and funds)

Publication Bias

Bias in measurements: know your data



sample size increasing with time (and funds)

Publication Bias

Bias in measurements: know your data

- Systematic error: PUBLICATION BIAS

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

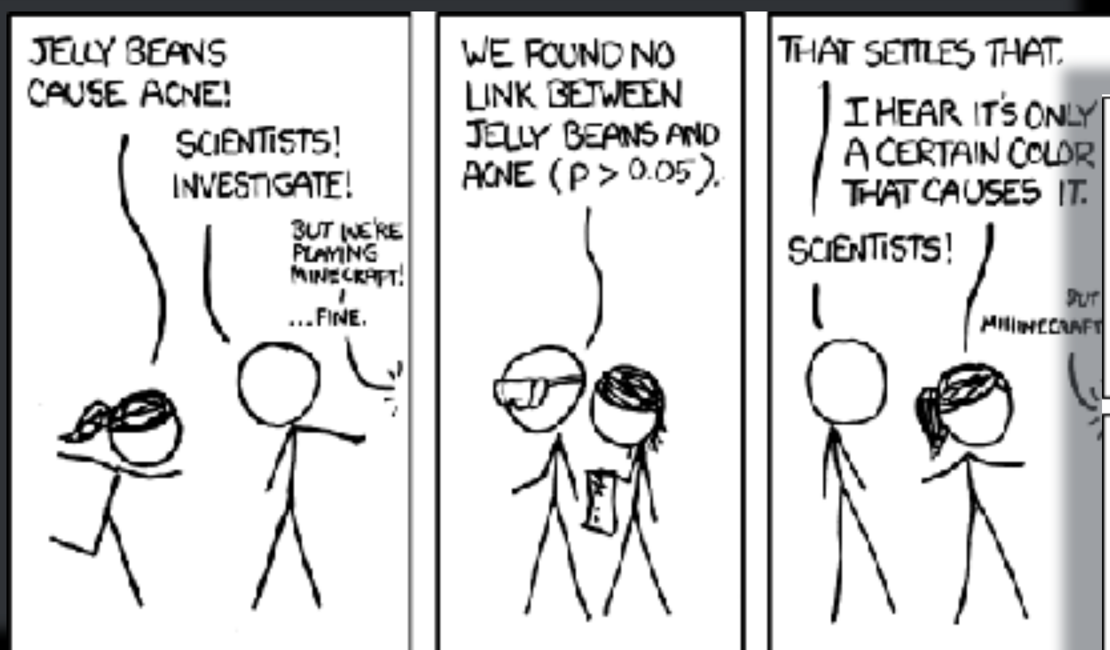
- Systematic error:

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

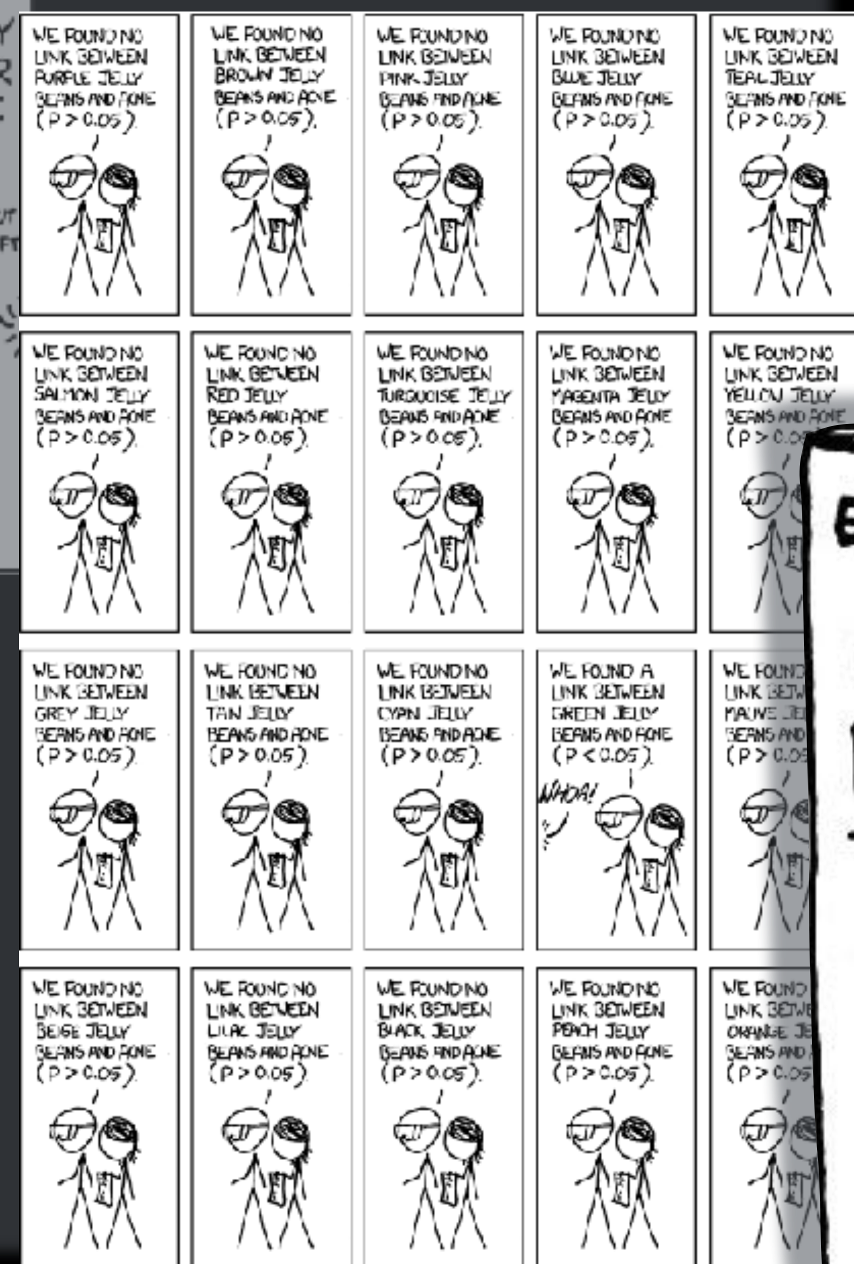
  Undercoverage bias
  Self selection bias
  Social desirability bias

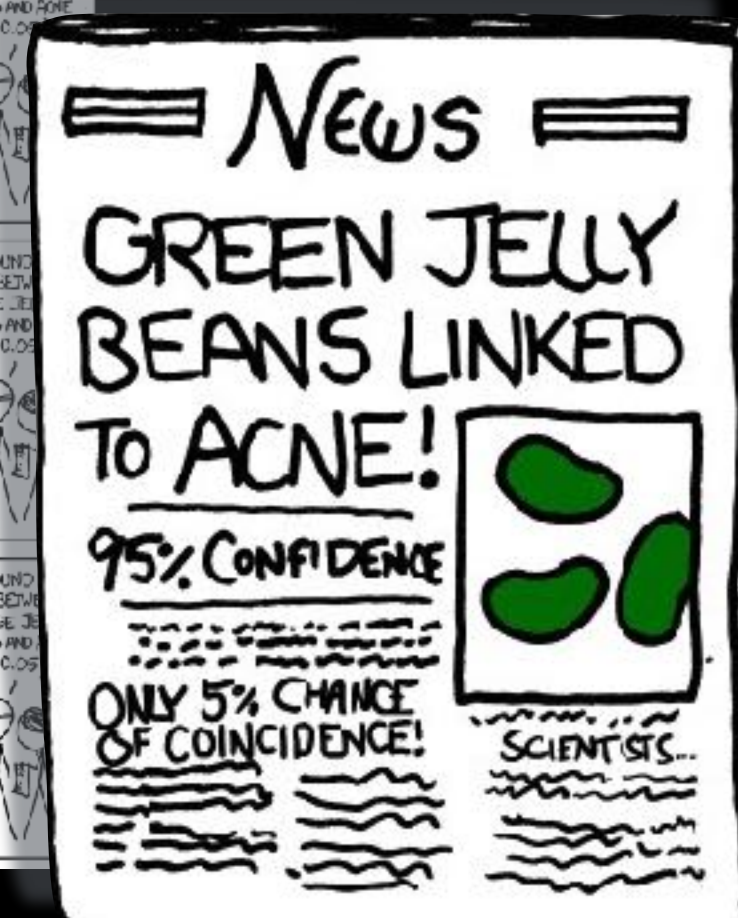
  Other systematic errors
  Publication Bias
  Data Dredging

Bias in measurements: know your data



https://xkcd.com/882/



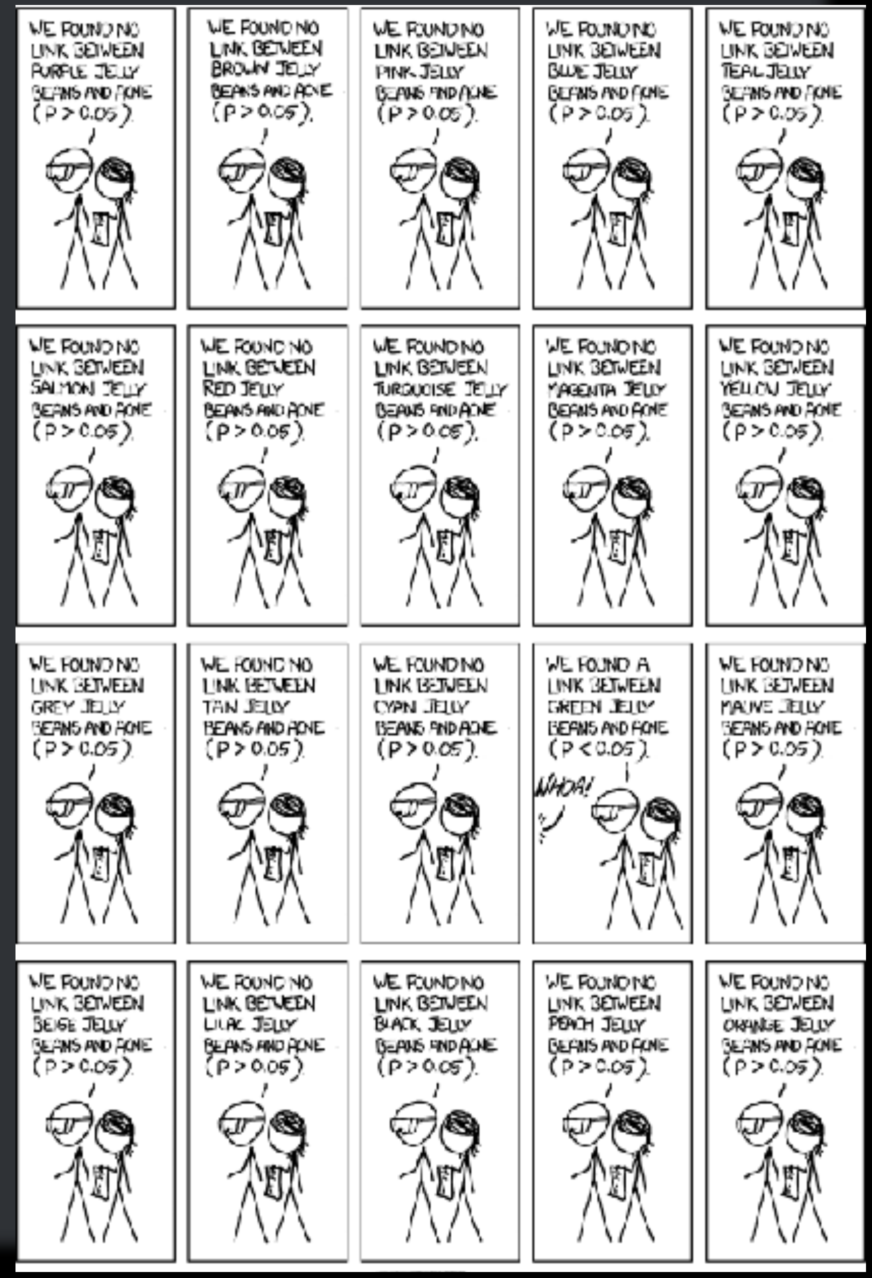


Data Dredging

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

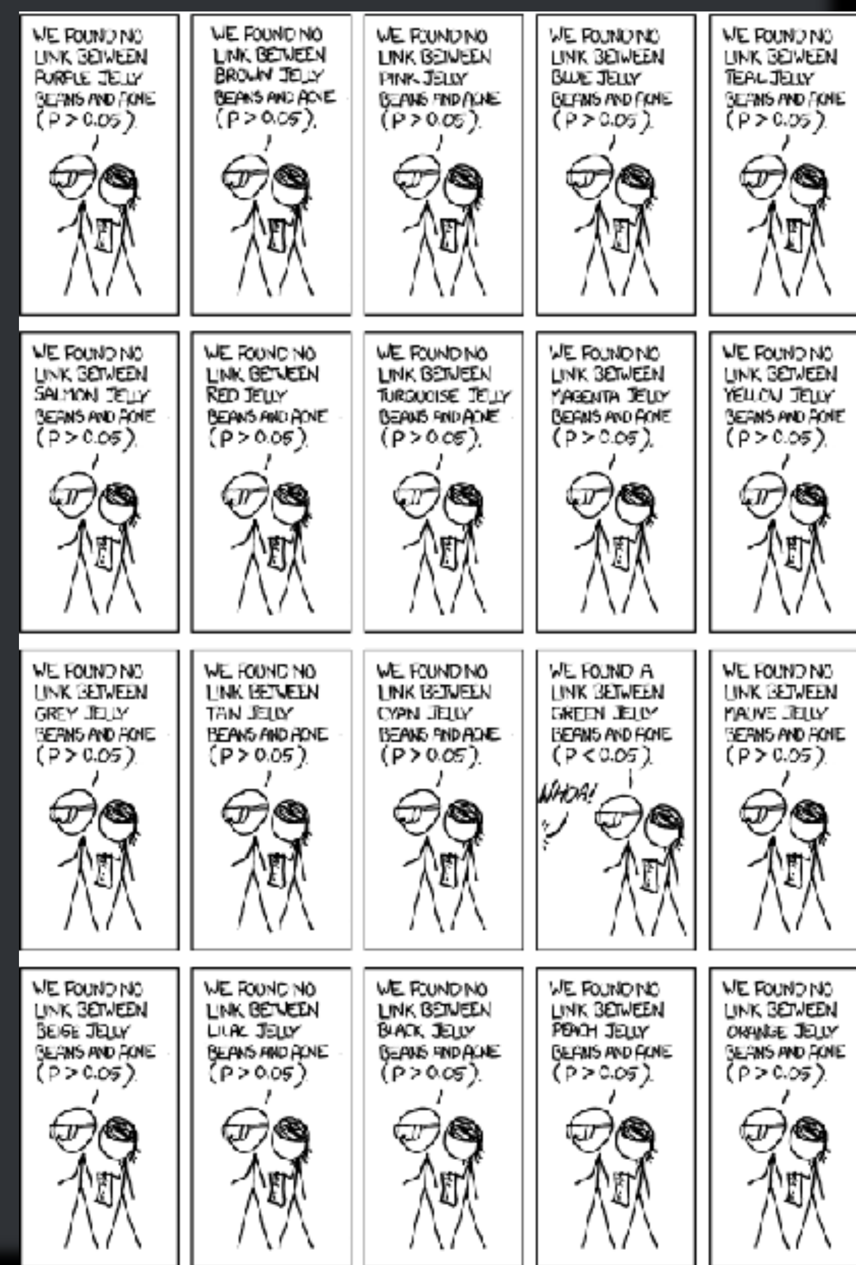
Bias in measurements: know your data

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

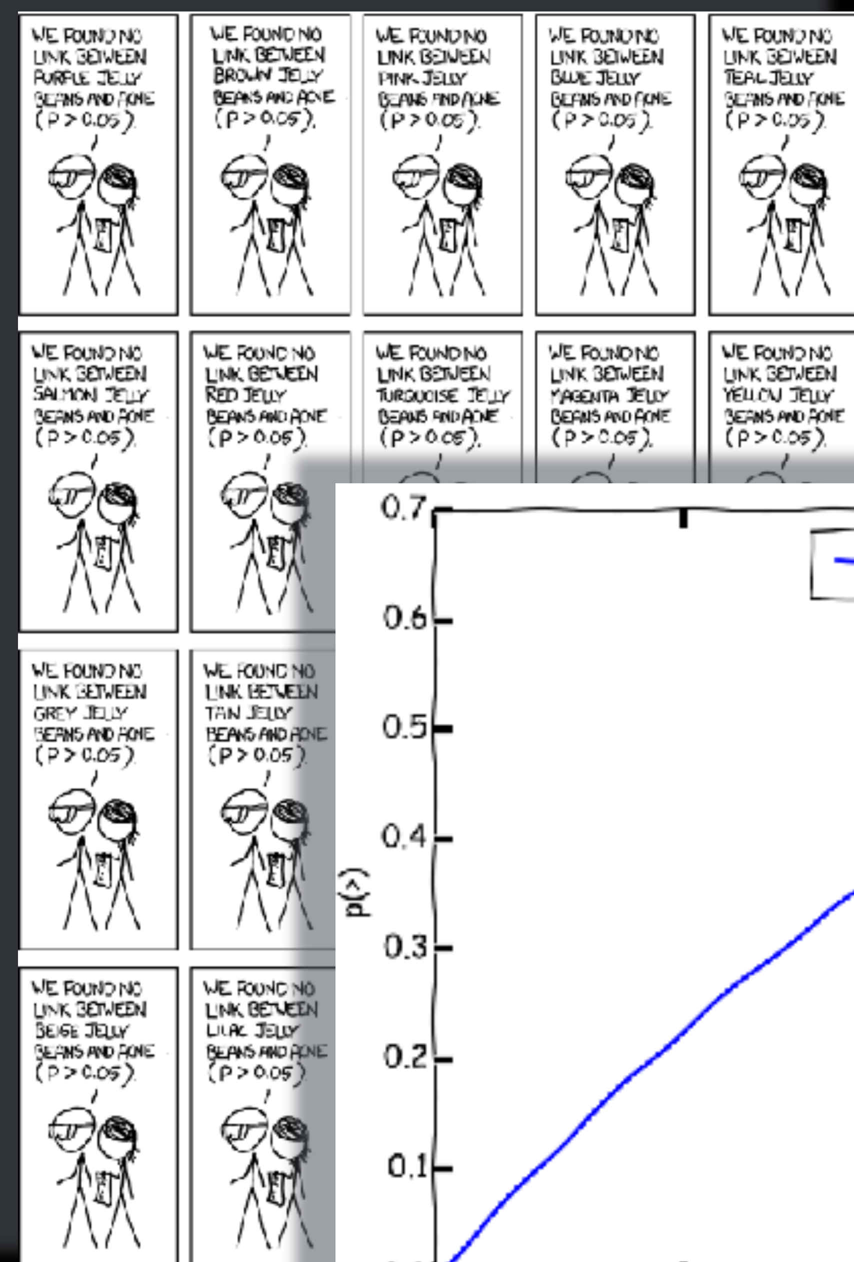
Bias in measurements: know your data

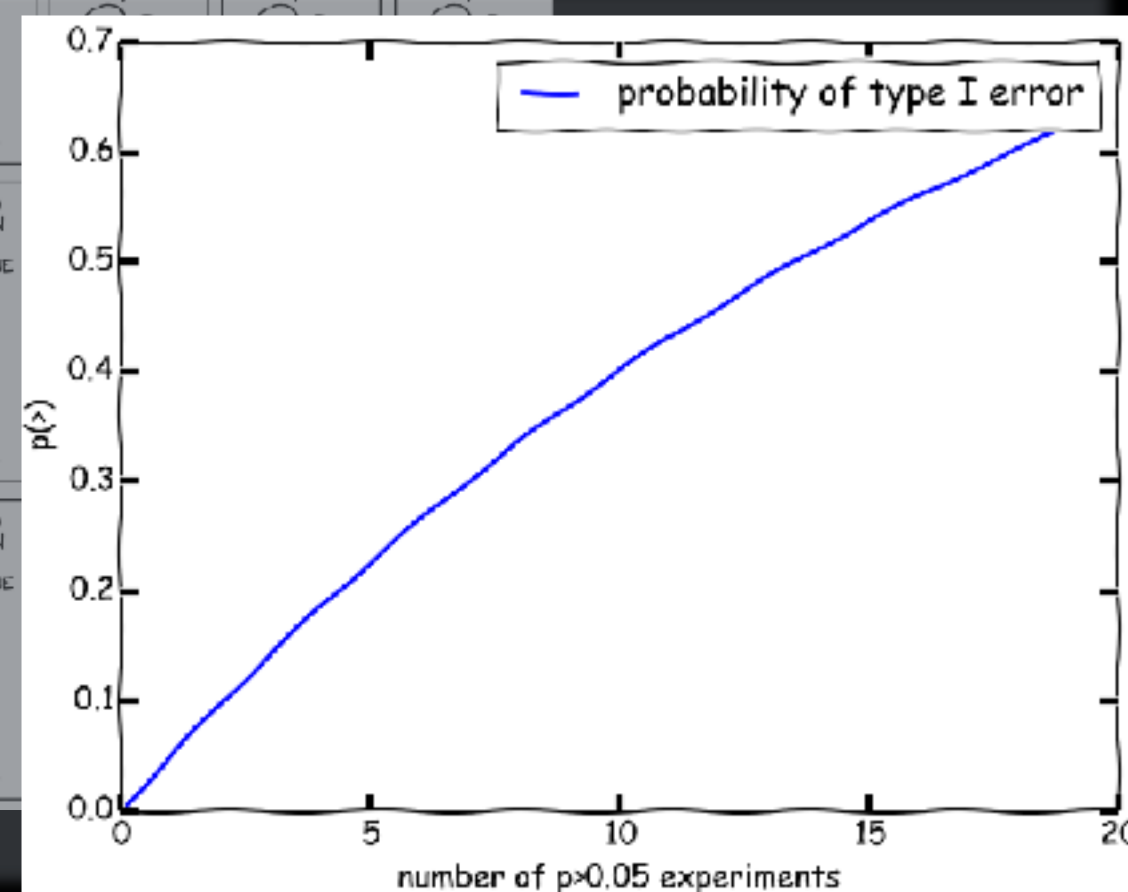
each test has a
probability $p<=0.05$
of Type I error
significance 95%

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

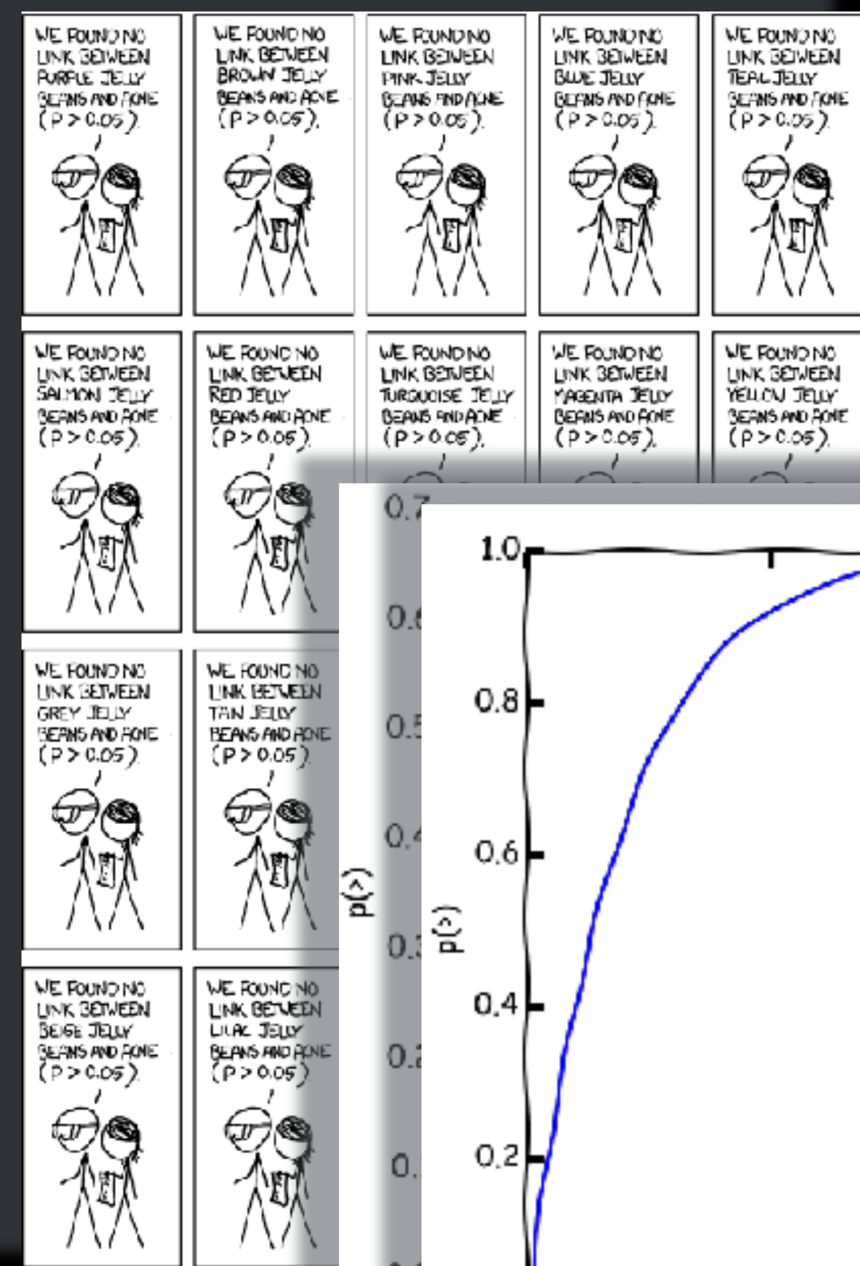
each test has a probability $p <= 0.05$ of Type I error significance 95%

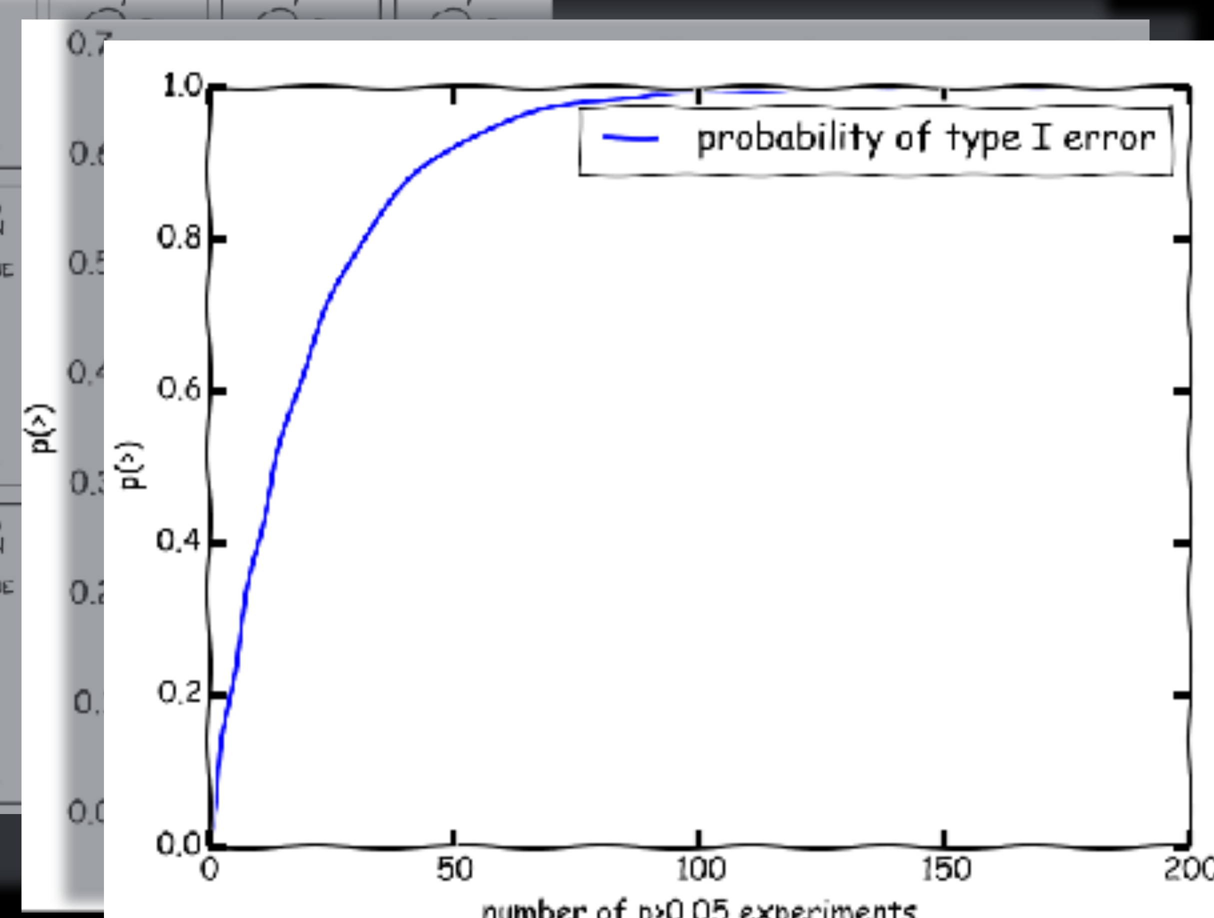
20 tests are preformed

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

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

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





Data Dredging

Systematic errors, biases, and reproducibility

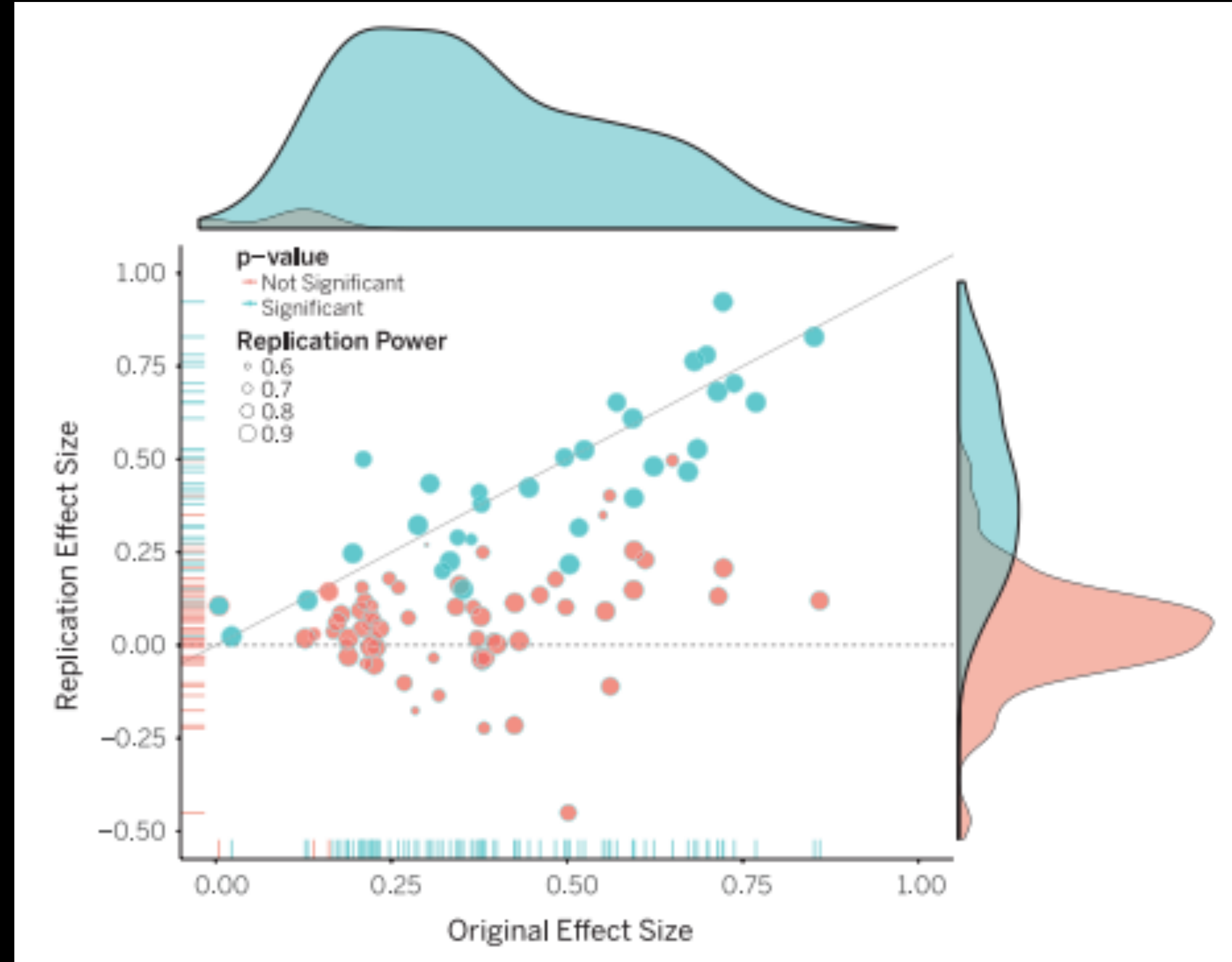
RESEARCH ARTICLE SUMMARY

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
  <u>tendency to *systematically* underestimate/overestimate the average</u> 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.

- Systematic error

- Stochastic & Random error

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

4.0 +/- 0.1 cm

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

2.0 +/- ε cm, ε > 0.1 cm

Repeated measurements give ε (*bootstrap* data…)

Errors and uncertainties.

- Systematic error
- 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.

- Systematic error
- 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.

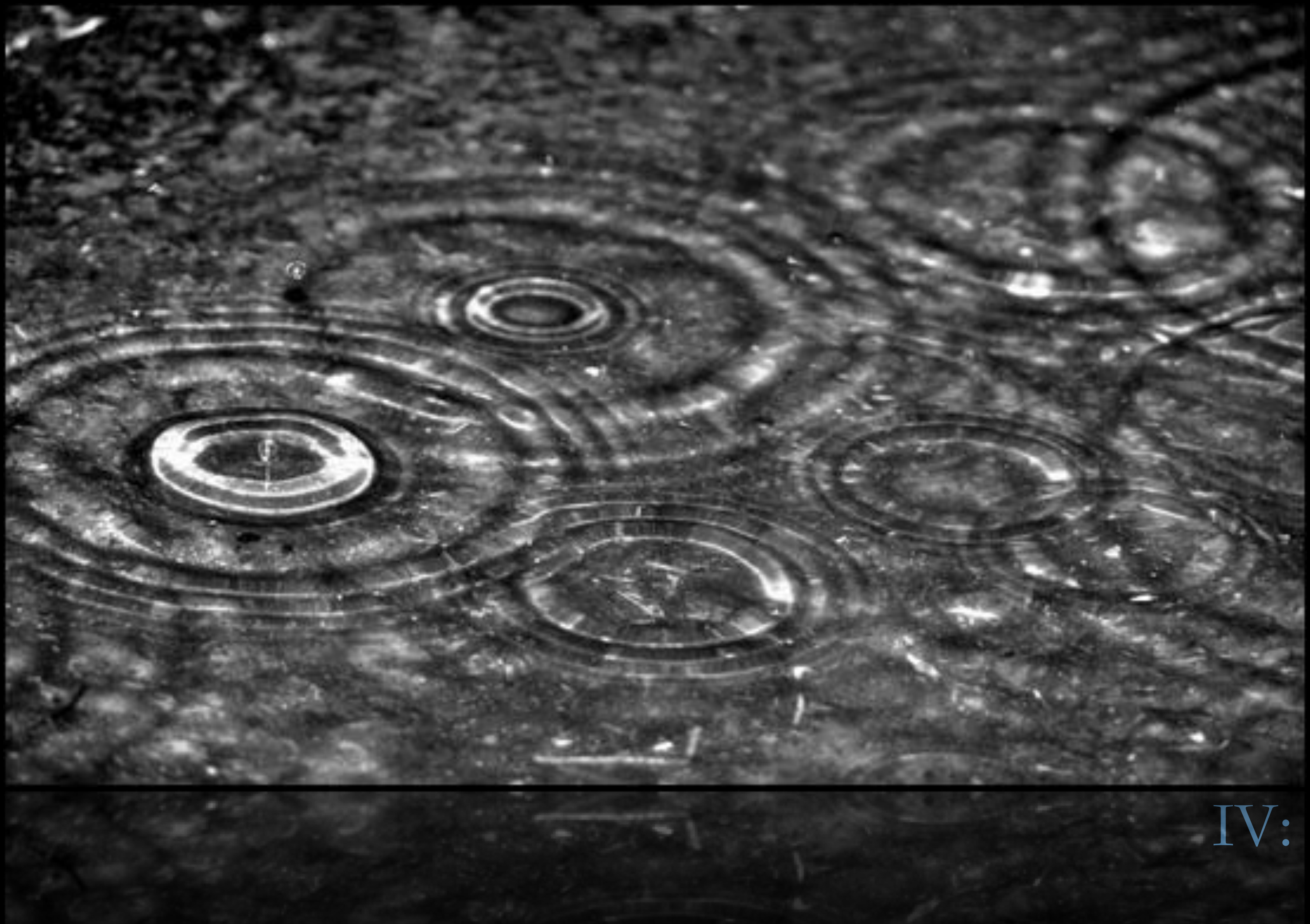
A discrete distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event.

Errors and uncertainties.

- Systematic error
- Stochastic & Random error

  Poisson processes :

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



Errors and uncertainties.

- Systematic error

- Stochastic & Random error

  Poisson processes :

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

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 $\sigma = \sqrt{\lambda}$

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

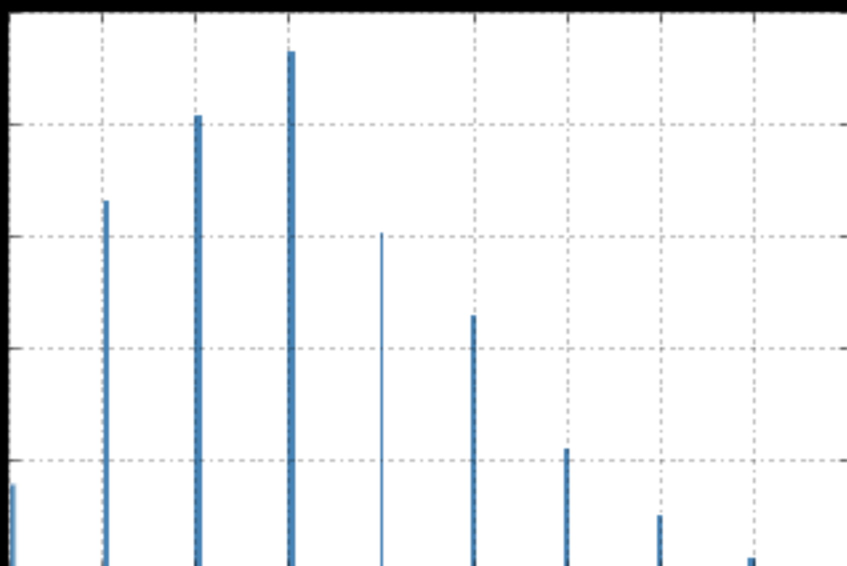
Binomial

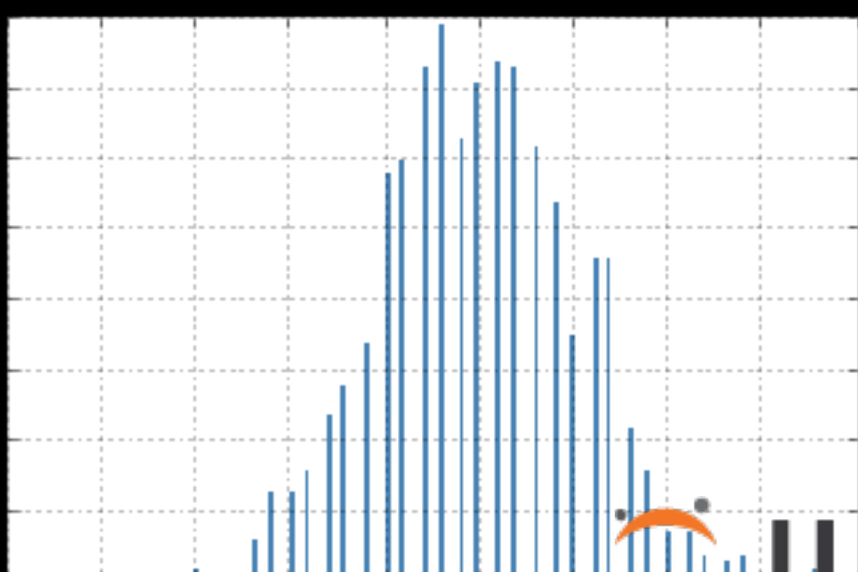
discrete bivariate

2 parameters: n,p

support: [0 , ∞]

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





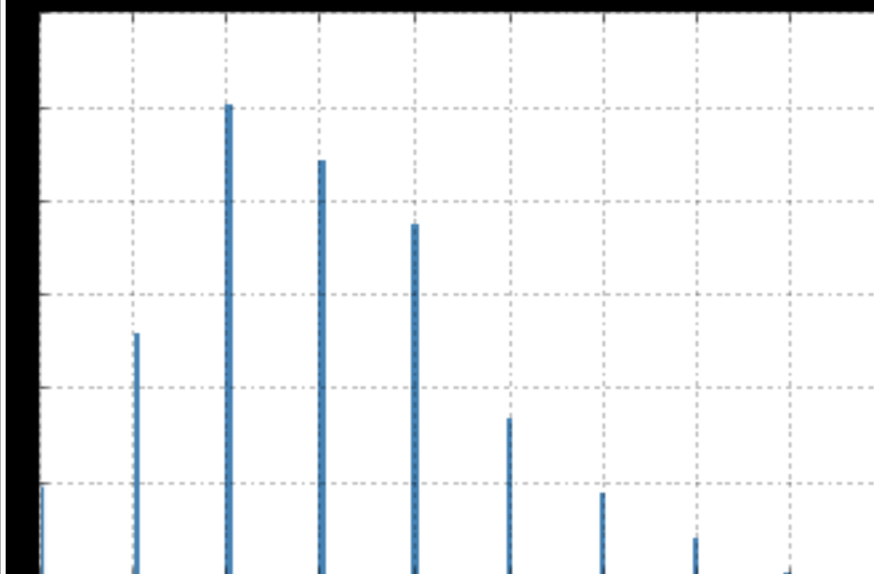
Poisson

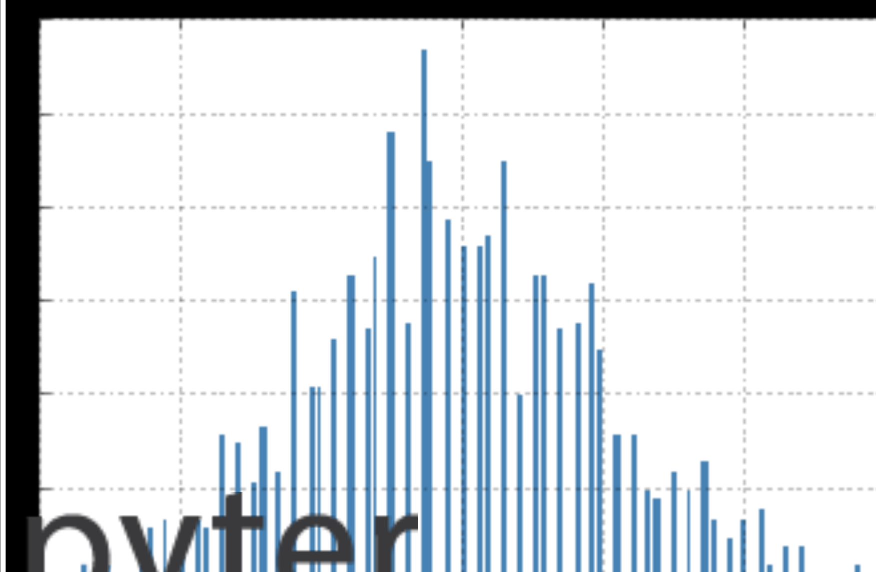
discrete univariate

1 parameters: λ

support: [0 , ∞]

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





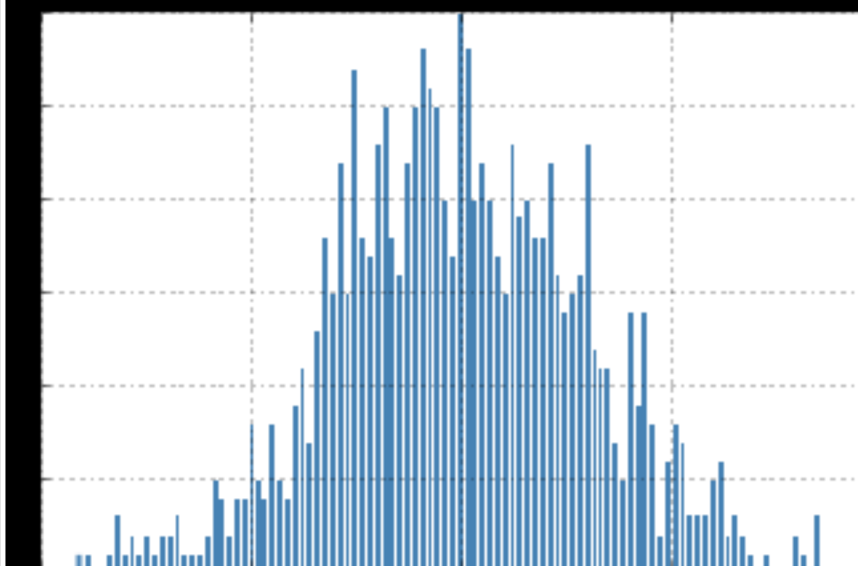
Gaussian

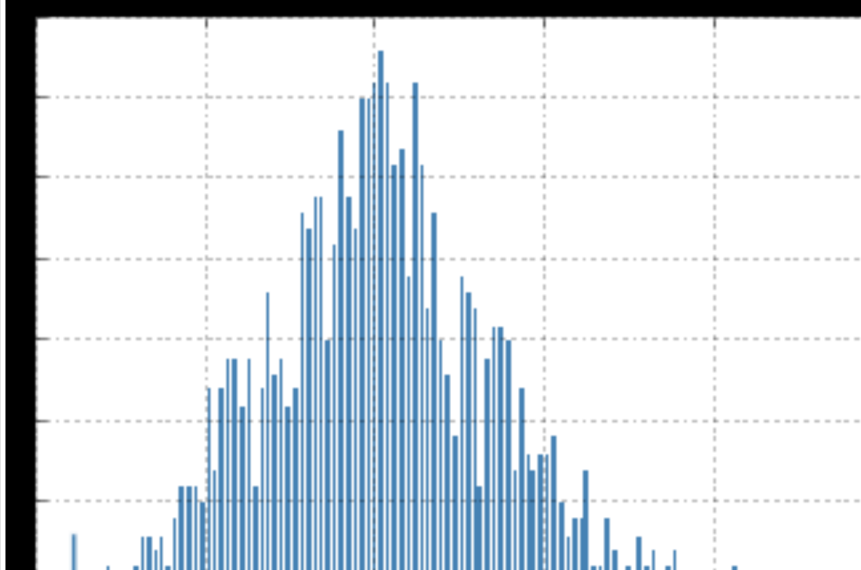
continuous bivariate

2 parameters: μ, σ

support:[- ∞ , ∞]

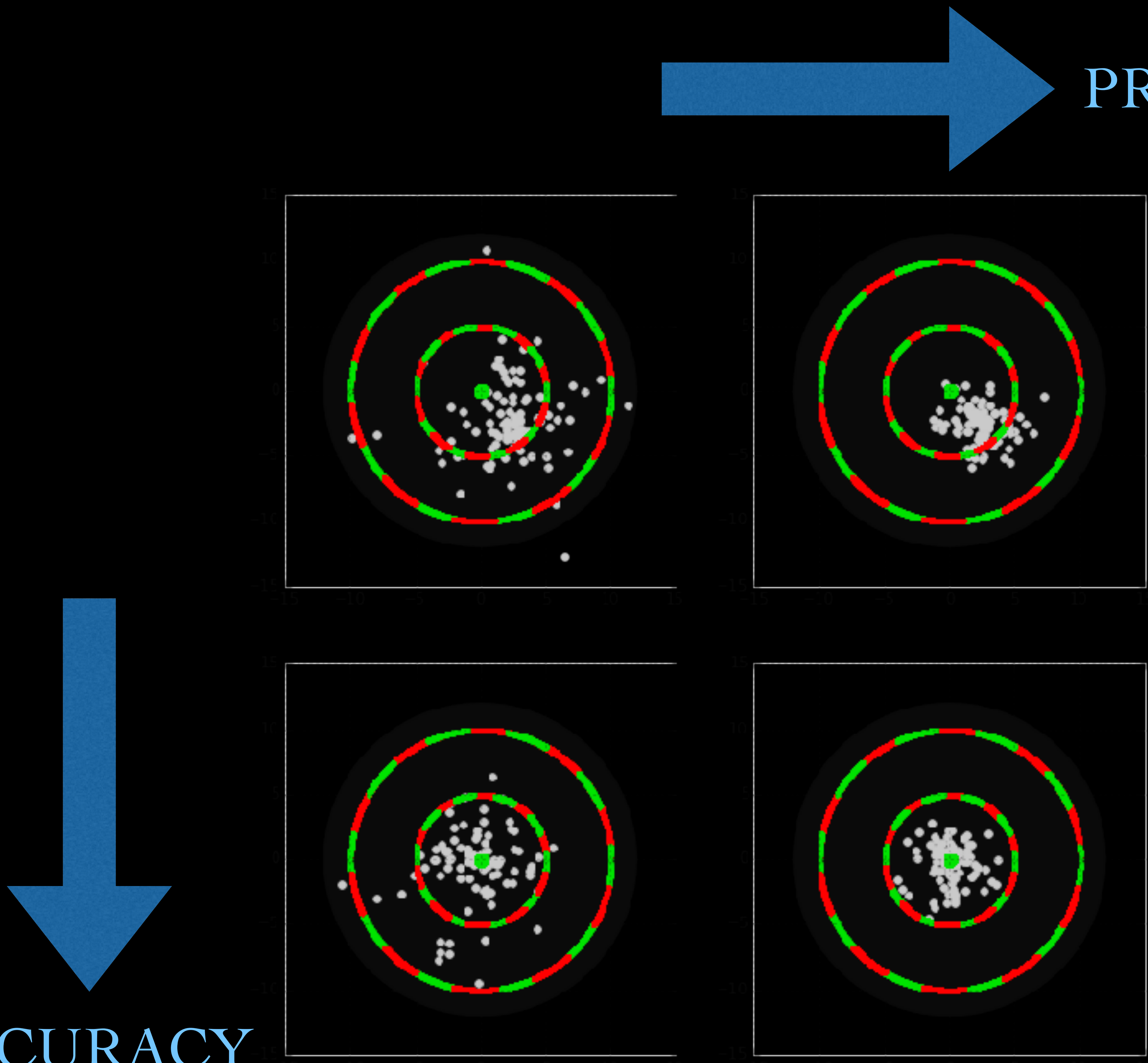
moments: μ, σ, 0





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

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

Error propagation

IID: <u>Independent</u> identically distributed:
add in quadrature for linear data operations

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

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

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

CUSP

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

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

Covariance matrix

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

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

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

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

Covariance matrix

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

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

IF
Independent
variables

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

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

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

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

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:

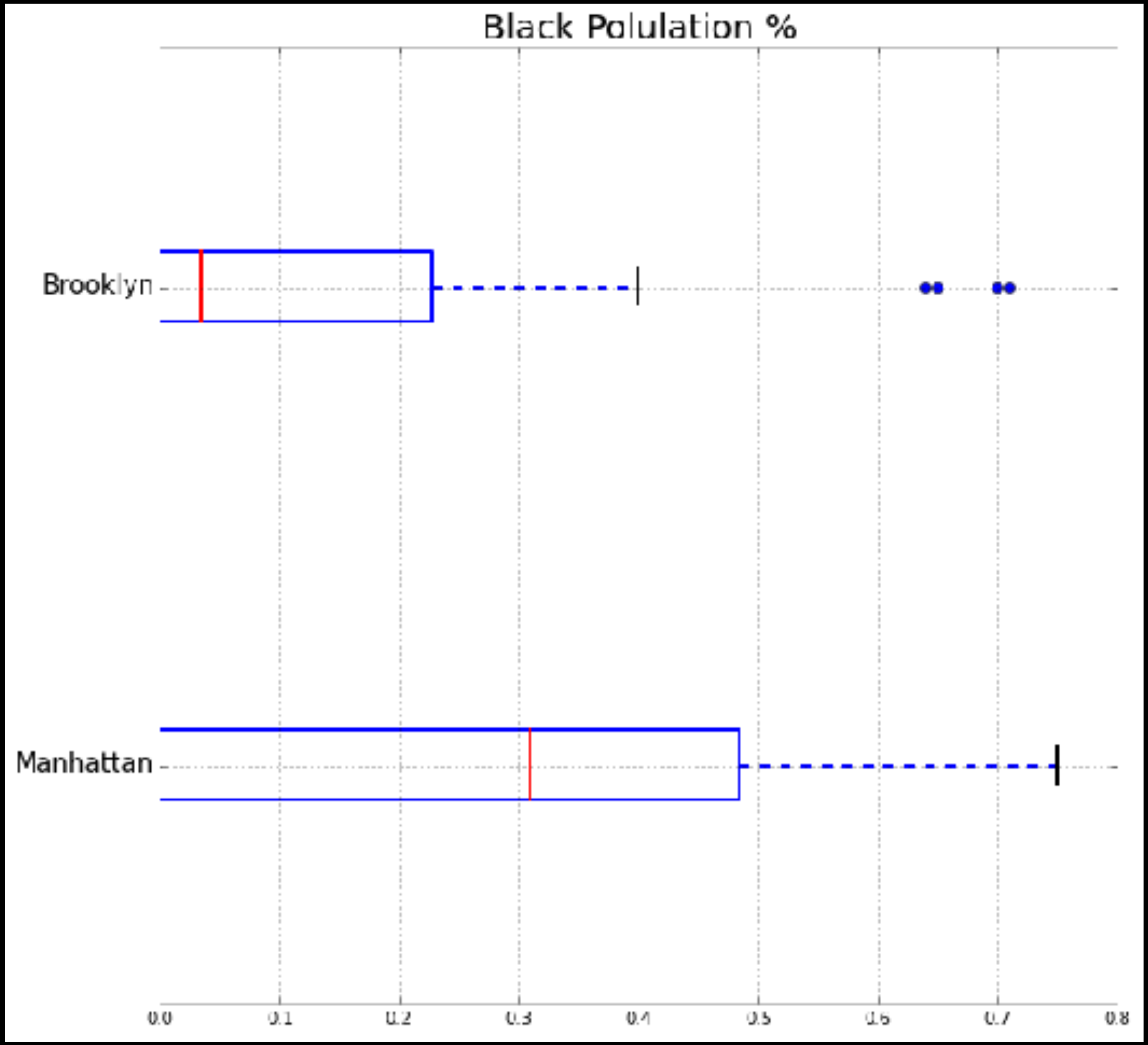
$x = [965 \pm 30(\text{stat}) \pm 12(\text{sys})]$ number of car accidents

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

http://localhost:8889/notebooks/black_percentage.ipynb#

Reporting Your Results

x_test = [965 ± 32(total)] number of car accidents
x_population = [932 ± 29(total)] number of car accidents

Test statistic = (Statistic - Parameter) / (statistics Standard error)

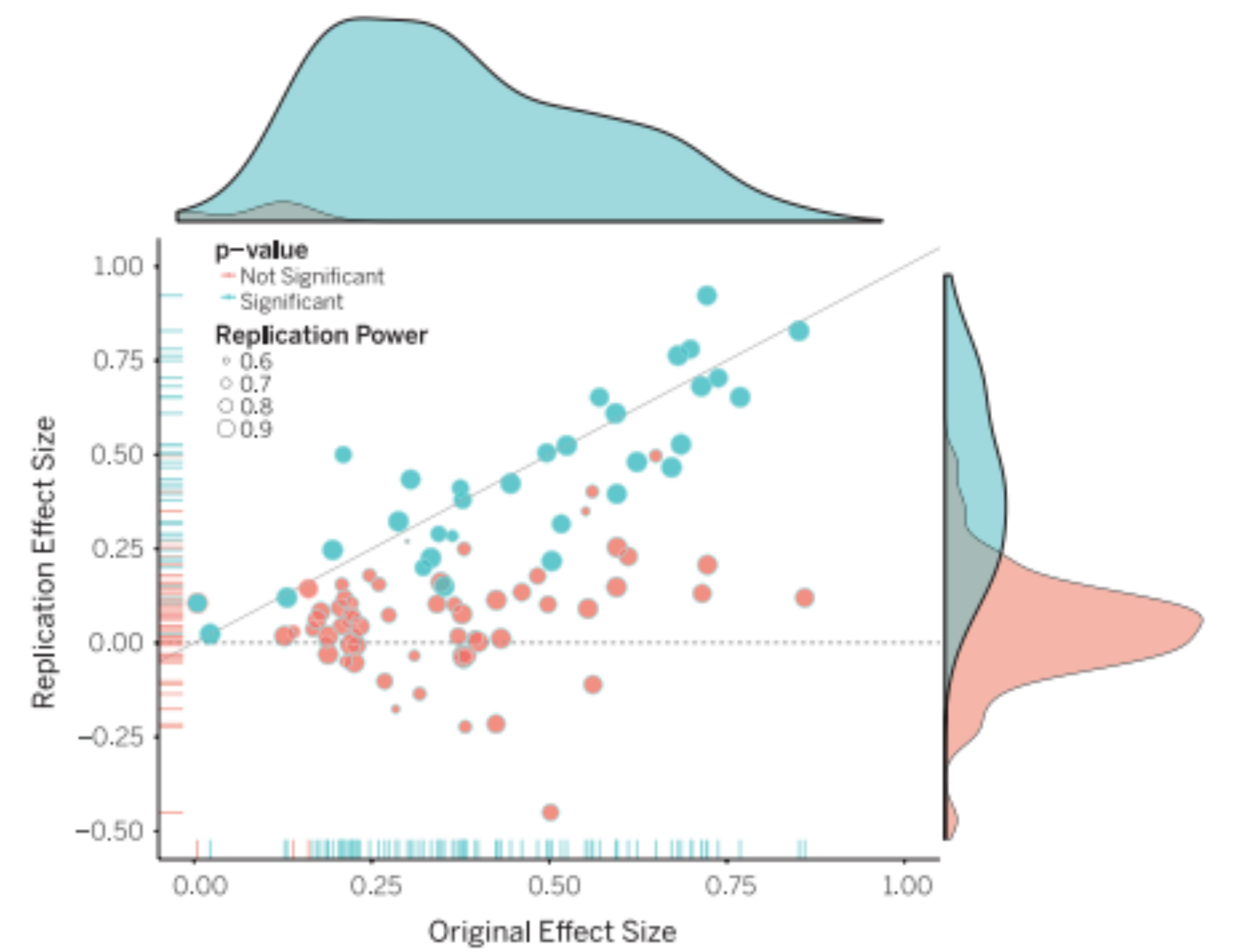
Homework: READING

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*



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

Homework: 1. Compare Tests for Correlation

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

- **Pearson's test** (answer: are the 2 samples correlated?)
- **Spearman's test** (answer: are the 2 samples correlated?)
- **K-S test** (answer: are the samples likely to come from the same parent distribution?)

Use: age of bikers for 2 genders, age of bikers in day vs night and assess the correlation/independence of the 2 samples in each case. State your result in words.

Homework: 2. Compare Tests for Goodness of fit

(synthetic data)

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, you do not need to do this yet!)**

Use KS, K-L divergence, and one more test (AD or Chisq) to quantify the difference between a binomial & Gaussian distribution and a Poisson & Gaussian distribution as a function of the parameters of the first distribution (np for binomial, λ for poisson)

For each test plot the relevant parameter (the K-L parameter, Anderson-Darling statistics, p-value for KS, Chi-sq parameter), against the distribution parameter (np, λ)

Homework: 3. 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 (from the previous exercise) to do this. Test at least 2 distributions.

Divide your riders sample by seasons: SpringSummer vs FallWinter and see if you notice any changes.

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.

MUST KNOWS:

- Statistical errors
- Systematic errors
- Undercoverage, SelfSelection, Social desirability
  publication bias, data dredging
- Precision vs accuracy

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

Reckova & Irsova
Publication Bias in Measuring Climate Sensitivity
IES Working Paper: 14/2015
http://salserver.org.aalto.fi/vanhat_sivut/Opinnot/Mat-2.4108/pdf-files/emet03.pdf