Estimating Associations

Michele Coscia

First Year Project #1

February 15th, 2022

Lecture Plan

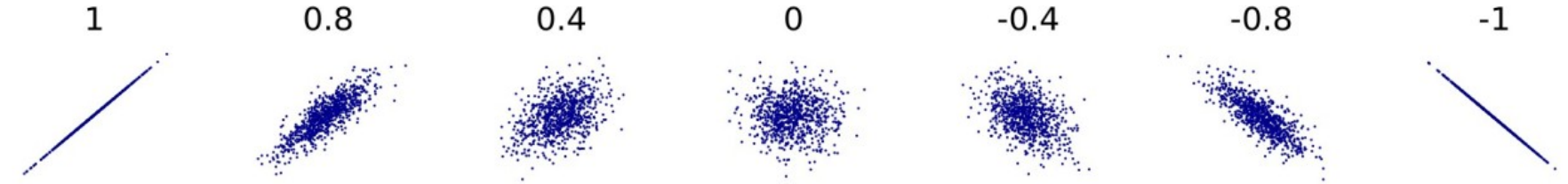
1) (February 8th) Intro
2) (February 10th) Geospatial Basics
3) **(Today) Estimating Associations**
4) (February 17th) Multivariate Regression
5) (February 22nd) Interventions
6) (February 24th) Project Run Through
7) (March 1st) Q&A – Open Supervision
8) (March 3rd) Q&A – Open Supervision

Outline

- Pearson & Spearman correlations
- Non-linear associations
- P-values
- Multiple Hypotheses Testing

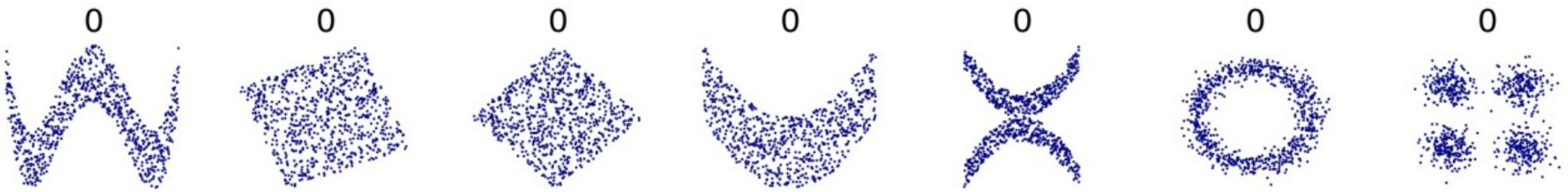
Pearson

Pearson Correlation



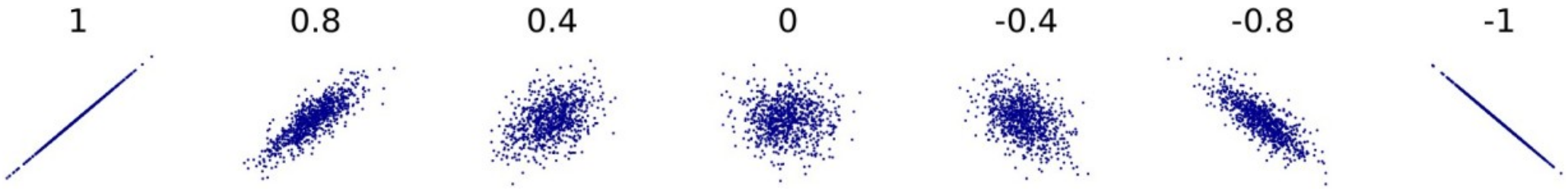
The higher the tighter the relationship

Slopes don't matter



Correlation = 0 does not imply no pattern

The higher the tighter the relationship

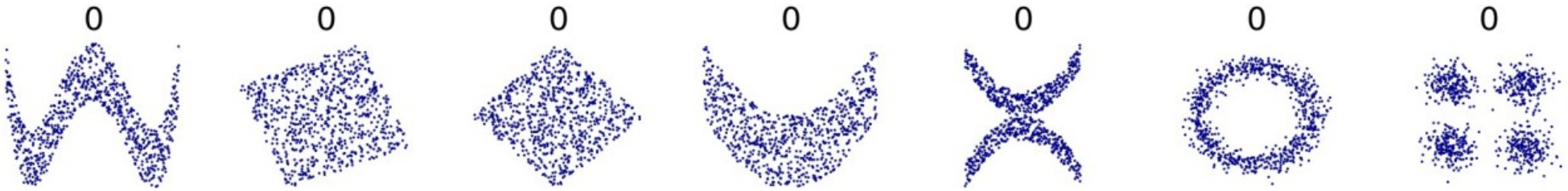


- 1 = Every time variable X increases, so does variable Y
- -1 = Every time variable X increases, variable Y decreases
- 0 = Knowing that variable X increases tells you nothing about variable Y
- Million $$$ question: is 0.4 a high or low correlation?

Slopes don’t matter

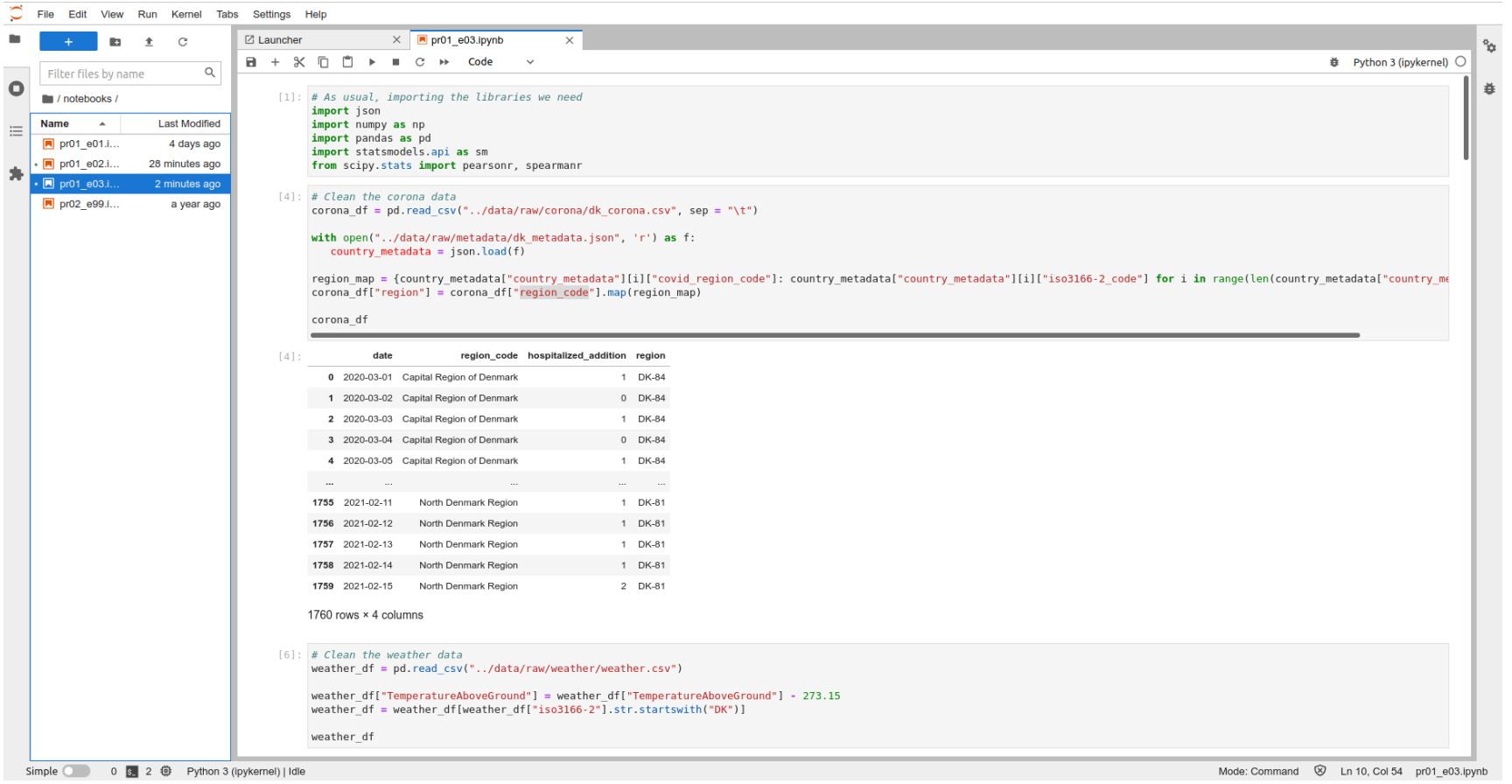
- Correlation strength != Scale of the effect
- Significance != Scale of the effect
- E.g. X & Y are two raters and Y always rates half the score than X
- The correlation is the relation between X and Y if they were to be standardized

The curse of linearity



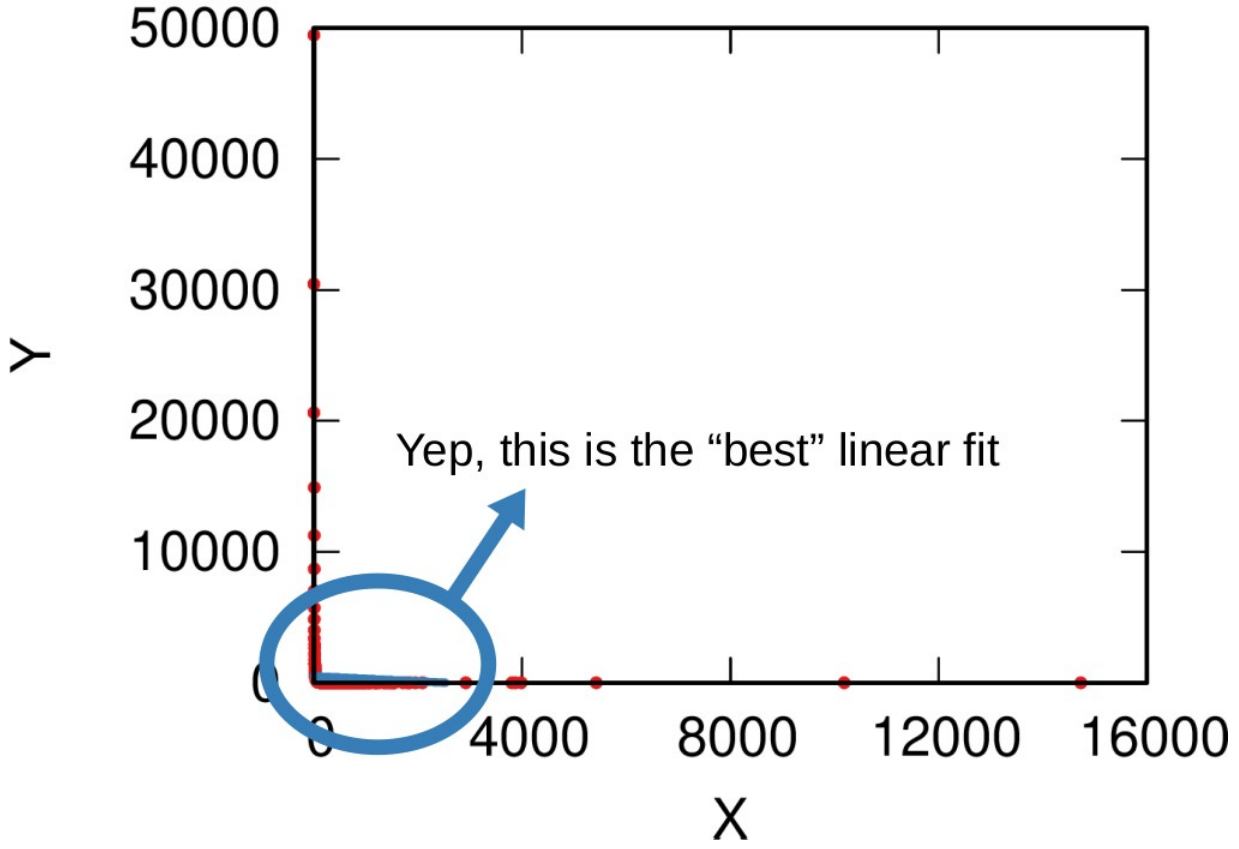
- There are TONS of interesting relationships that are not linear
- Pearson (and linear regression) are blind to this
- In this case, Pearson is an underestimation of the relation
- These examples are extreme, but most likely? Diminishing returns and/or skewed data

Let’s give it a try



Non-linear Associations

What if your data looks like this?

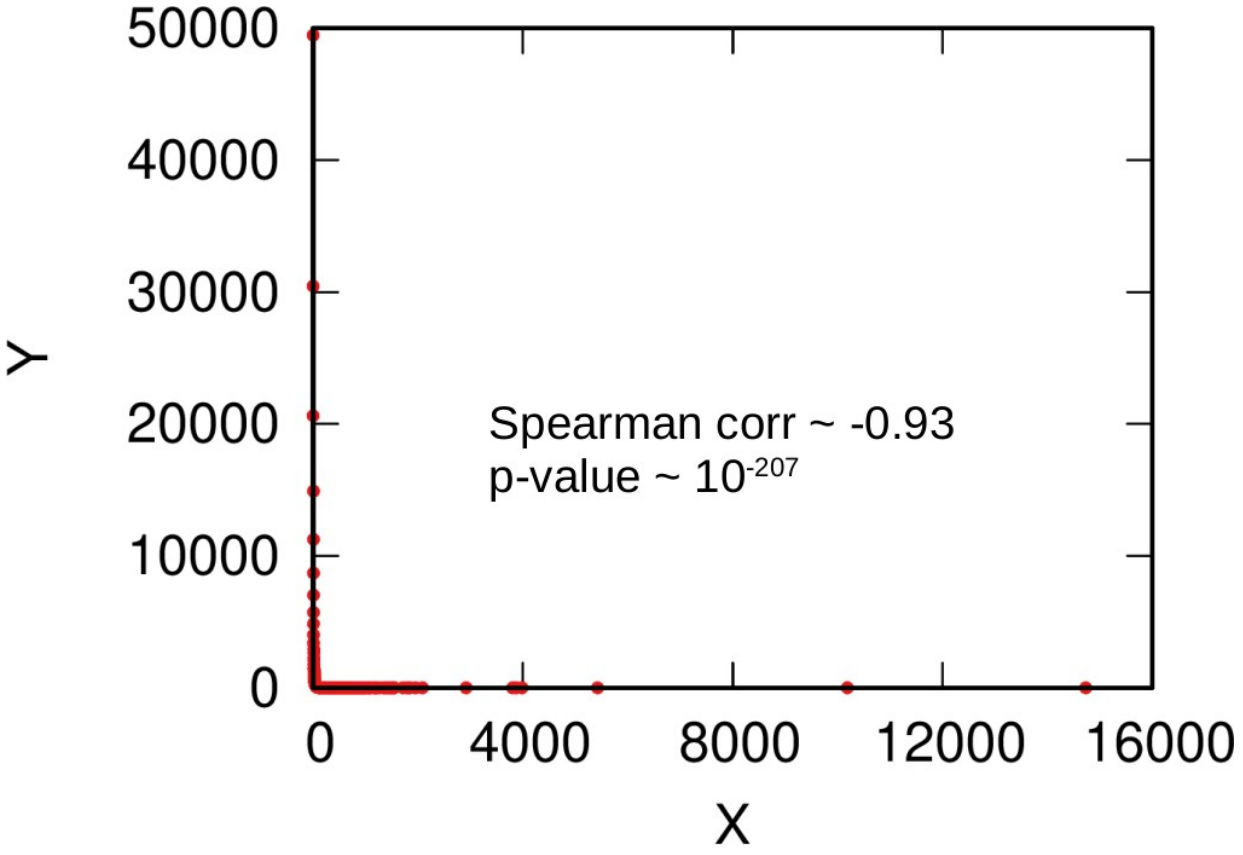


Pearson correlation ~ -0.05

p-value ~ 0.2

Ugh!

Two Solutions

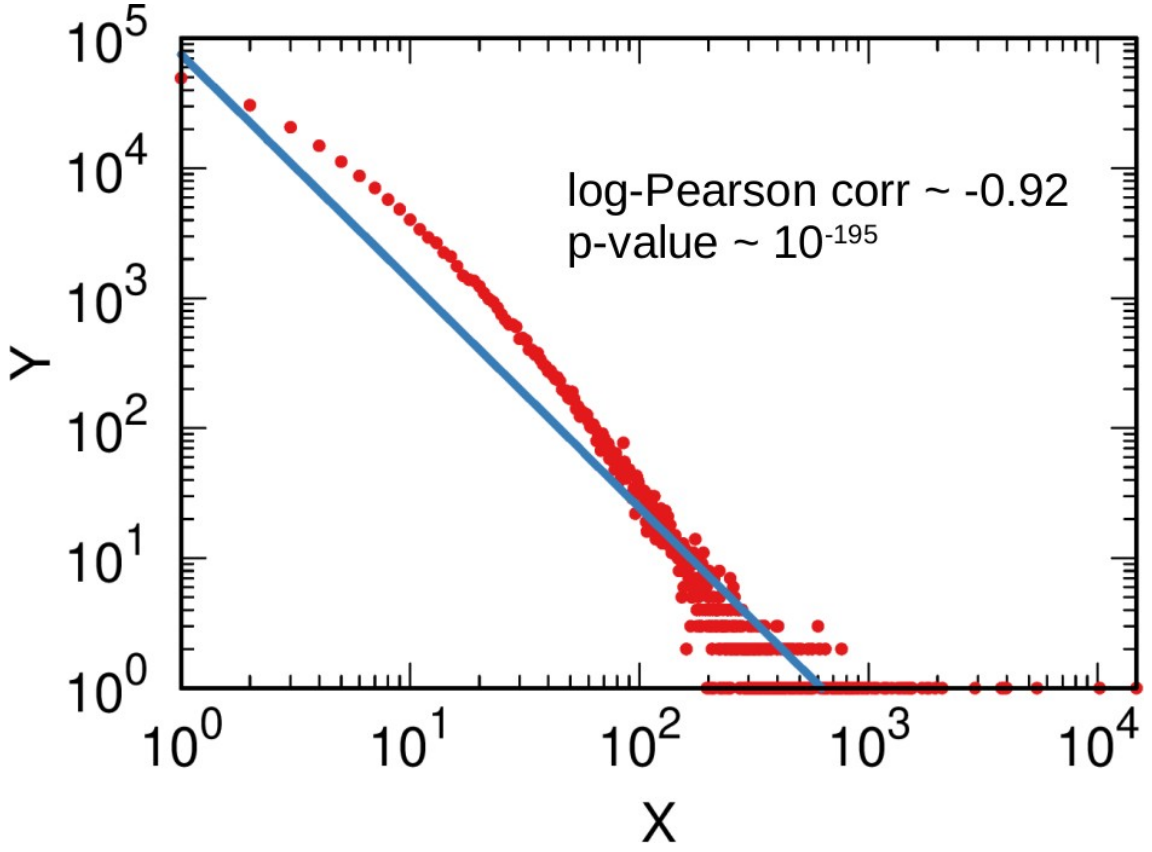


Solution #1: Spearman rank correlation

The values of X & Y don't matter

We only care that the n-th ranked value of X is also ranked n-th in Y

Two Solutions

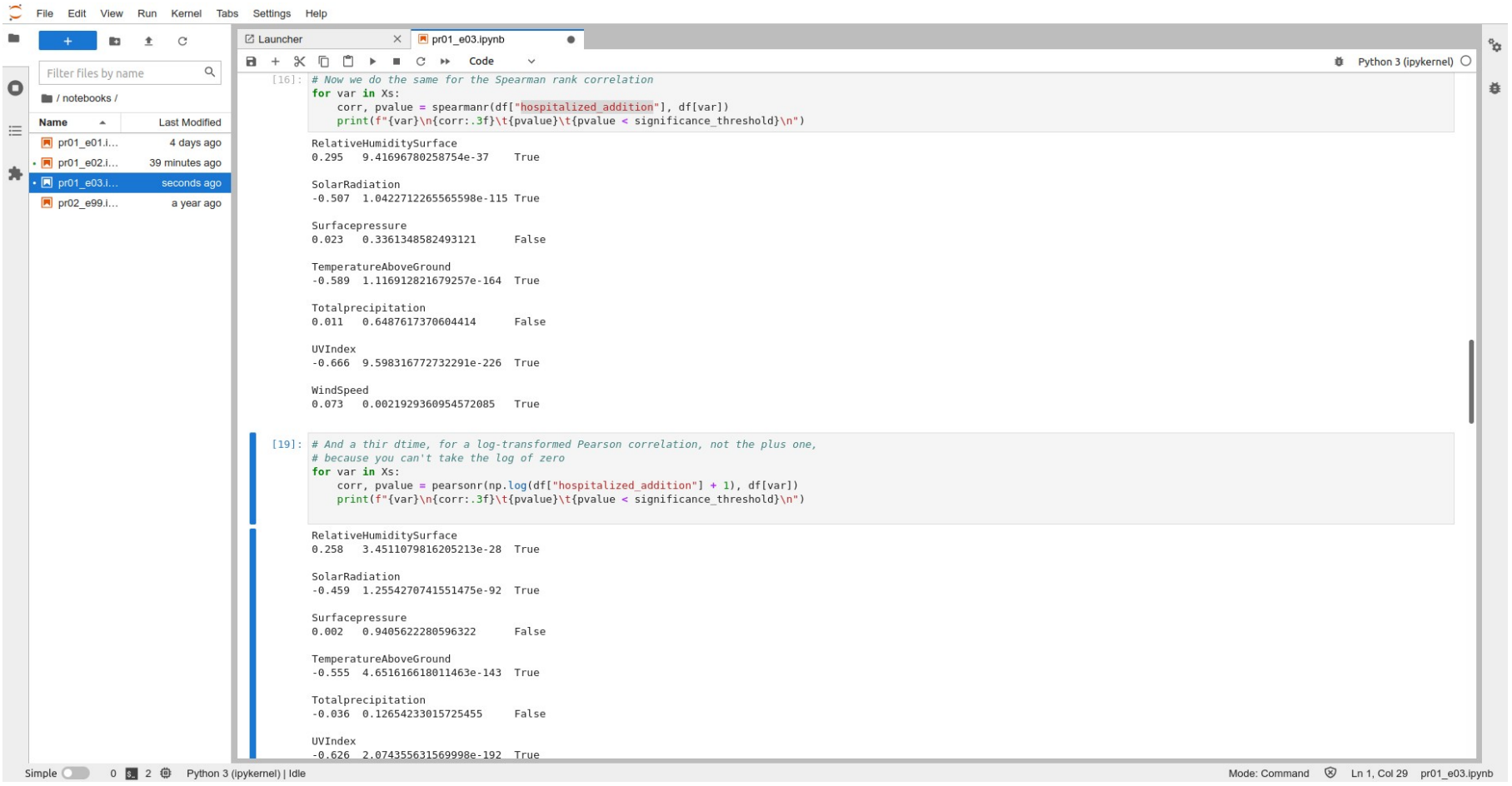


Solution #2: get rid of skewedness

Your best friend: the logarithm

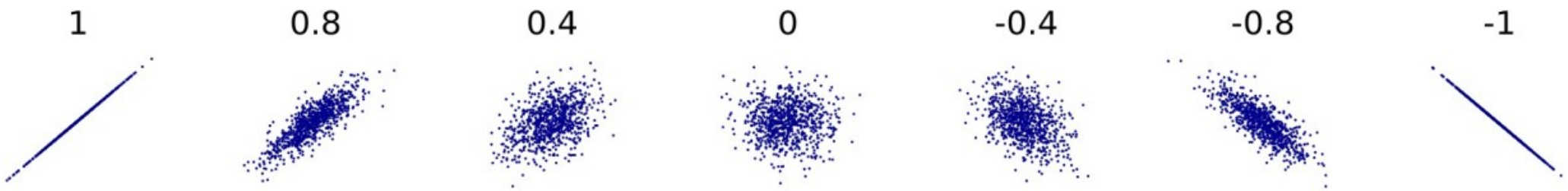
We don’t care about the precise values, only about the order of magnitude

Let’s try them out



p-values

Significance: Basics

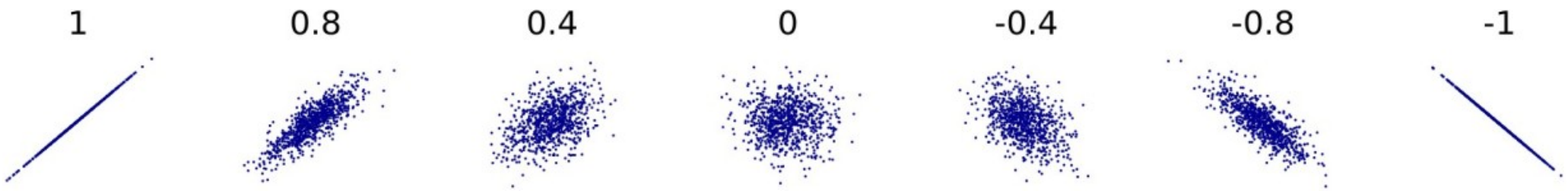


- p-value, jargon version: the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.
- Translation: how likely it is to see this correlation value if there is no relation between X and Y

P-values: Rules of Thumb

- 0.05: loose, used in social science (weak, noisy effects)
- 0.01: social science with many observations
- 0.001: better benchmark for big data science
- 5 sigma ($3*10^{-7}$): physics

Significance: Basics



- The higher the coefficient, the less likely the null hypothesis can generate the data
- More observations can strengthen your confidence even for low correlations
- (We'll come back to this later)

What are we doing?

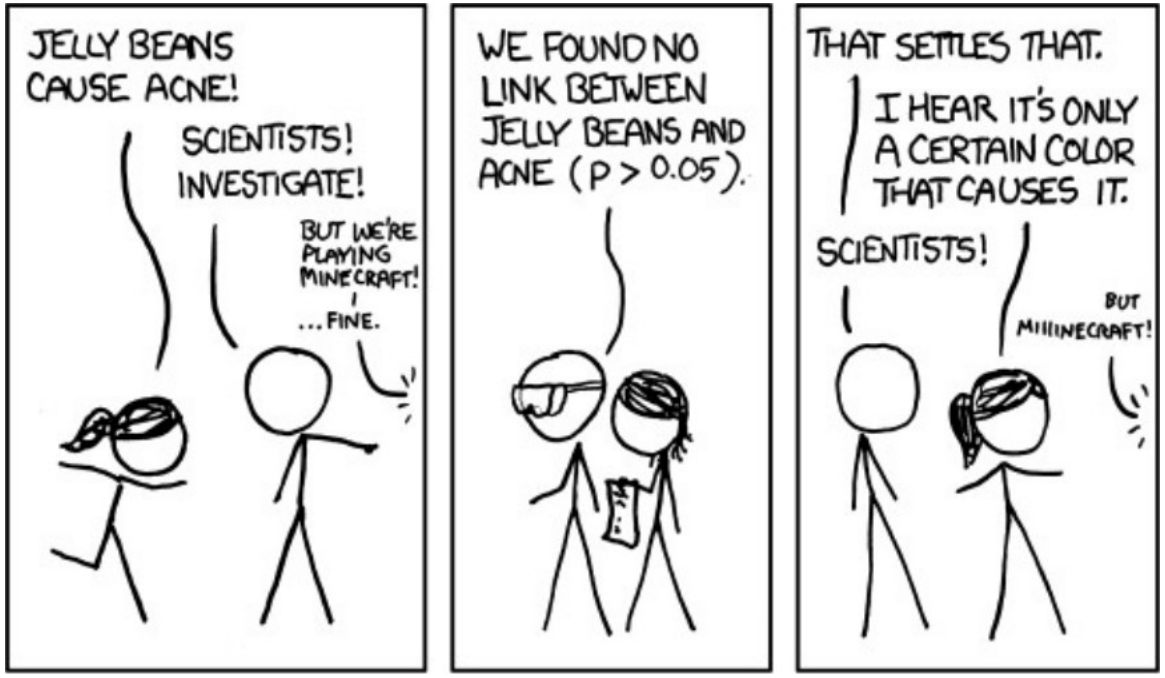
- h1: there is a linear relationship between X and Y
- h0: there is not a linear relationship between X and Y
- h0 is the null hypothesis

Skewed Data

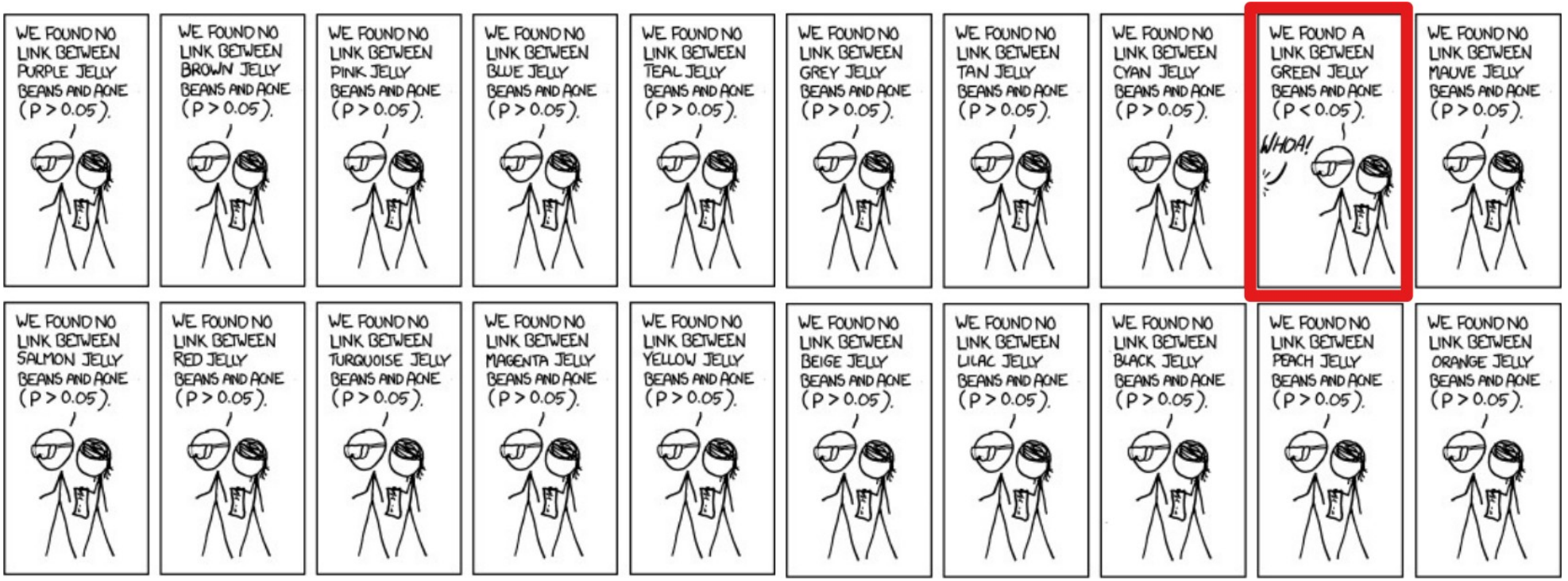
- Inflated p-values because **h0** is correct!
- But we really care about a **different** h1!
- i.e. there is a **non-linear** relationship between X and Y
- Spearman & log-log make different h1s

Multiple Hypotheses Testing

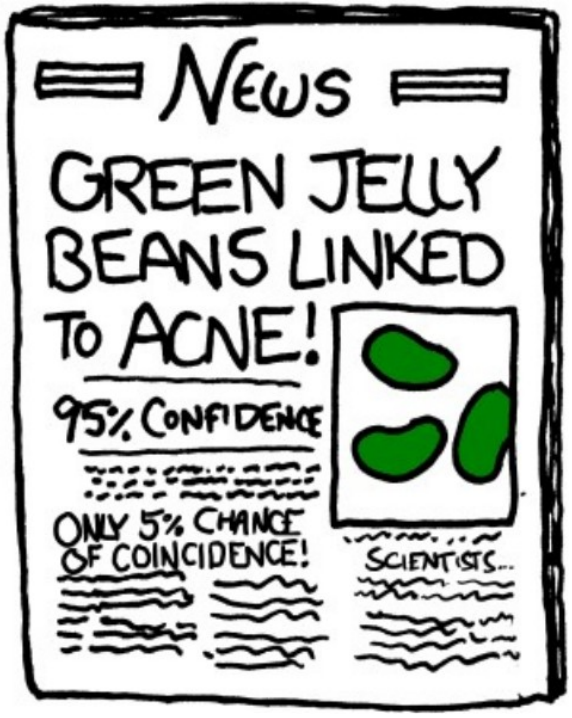
A rule of thumb for not fooling yourself with p-values



A rule of thumb for not fooling yourself with p-values



A rule of thumb for not fooling yourself with p-values



Right?

WRONG!

If you run X tests, you expect one of them to have 1/X p-value by pure chance!

(That's literally what "p-value" means!)

Bonferroni Correction

- If 20 tests will generate a p-value ~ 0.05 by chance…
- …and 0.05 is the p-value threshold I chose to determine significance…
- Then my real threshold should be lower than 0.05!
  - Your h0 is that **none** of h1 to h20 is true!
- Specifically it should be 0.05/20

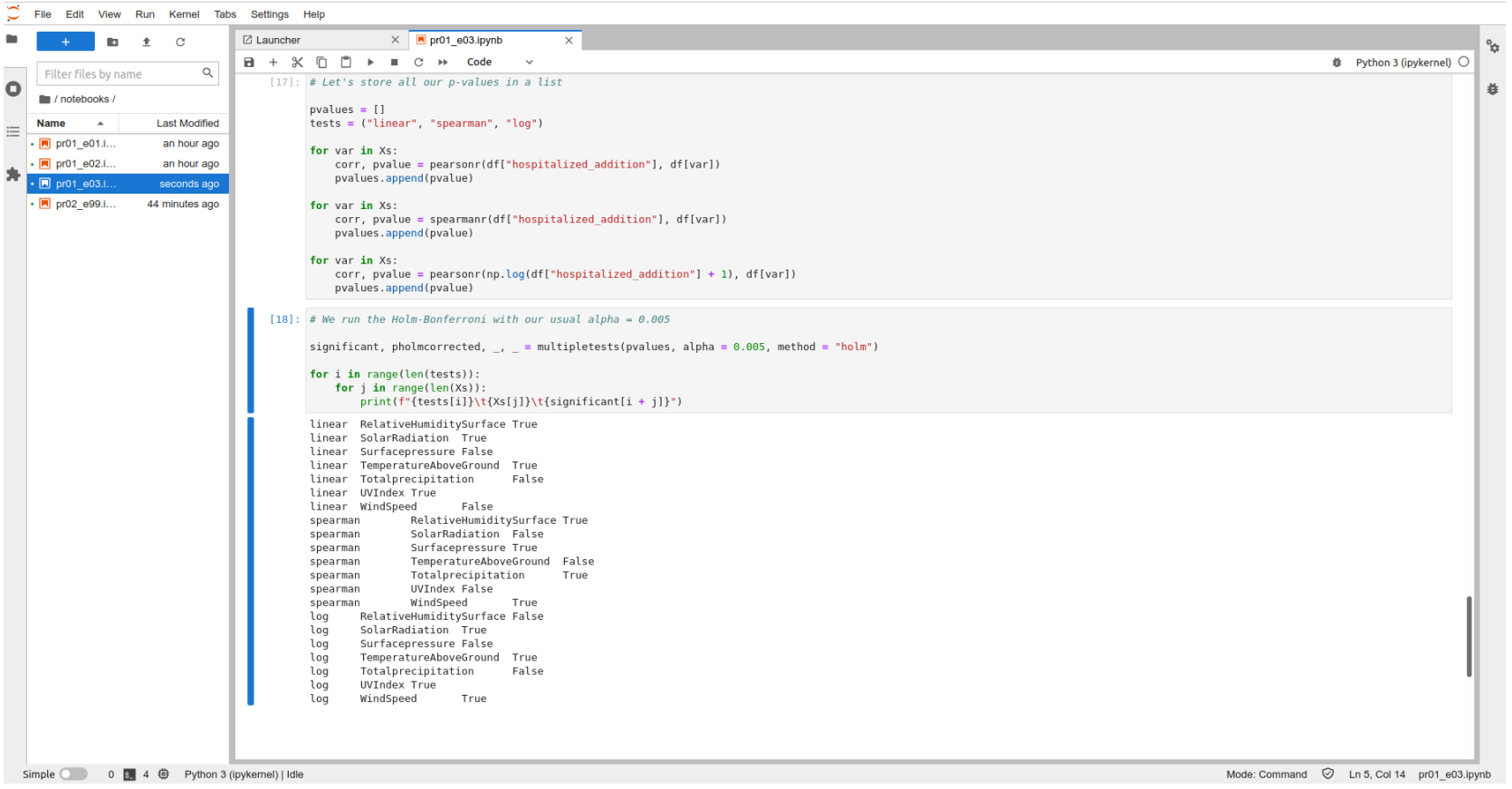
But!

- What I said applied if your questions are **independent**
- In your case they are **not**!
- So Bonferroni is too strict
- Alternatives?

Holm-Bonferroni

- Sort your p-values in ascending order
- First p-value → $p_1 < \alpha / N$
- Second p-value → $p_2 < \alpha / (N - 1)$
  - Because we already know p_1 passes!
- And so on...

Let’s get our hands dirty...



Q&A