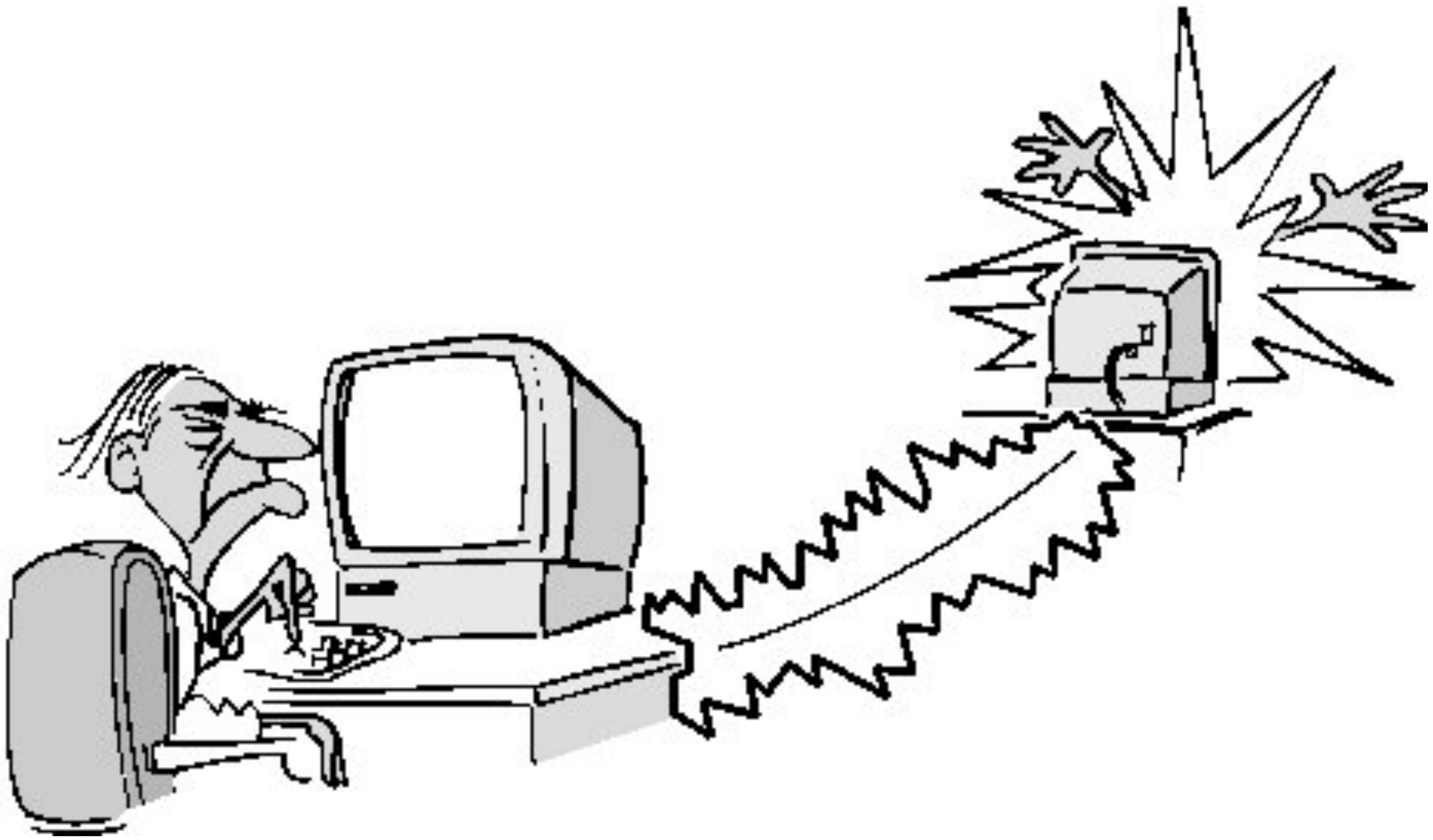
CORRELATION

10.29.2018

HOMEWORK 3

* due friday!



OFFICE HOURS TODAY MOVED TO TOMORROW

* i’m moving my office hour from monday (today) 1:30-3pm to **TUESDAY 4-5:30pm**

* my office hours wednesday (1:30-3pm) will stay the same

RECAP

* statistical **power**: how often a test says “significant” when there actually is an effect

* effect size

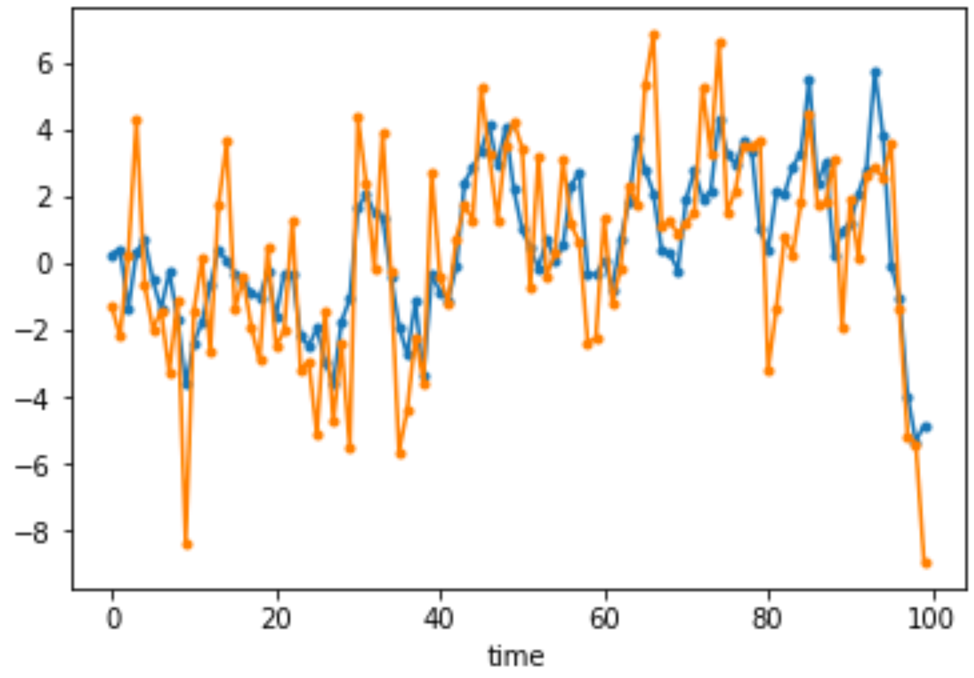
RECAP

* **permutation test**
  * “if these two samples were actually the same, it shouldn’t matter if we scramble them up and then re-divide them into two new samples…”

RELATIONSHIPS BETWEEN SAMPLES

* you record fMRI responses while someone listens to a podcast and plot the response of one voxel in auditory cortex (orange)

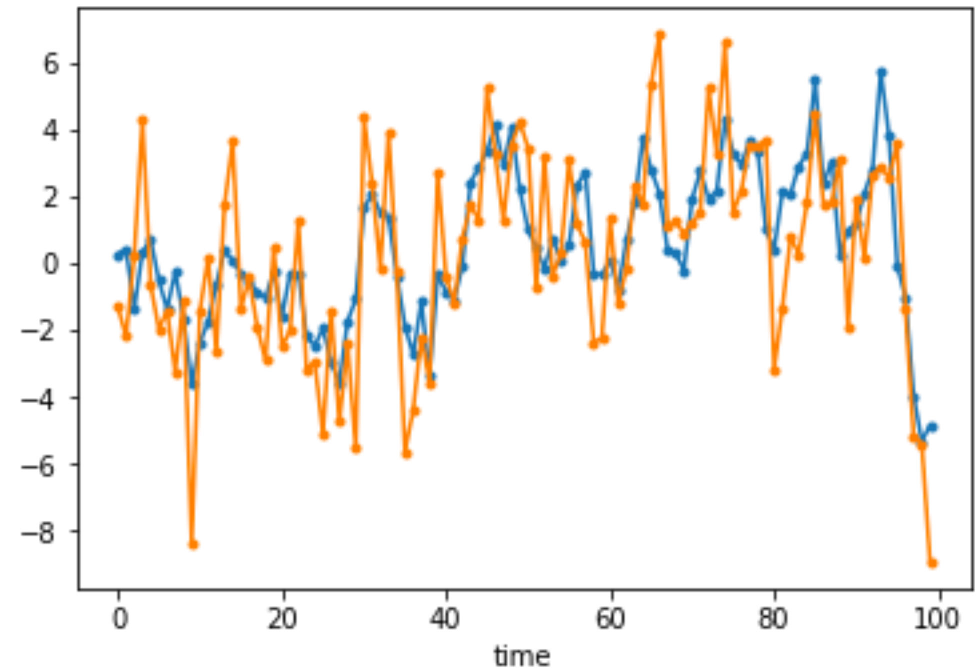
* you also measure how loud the sound is at every timepoint, and plot that (blue)

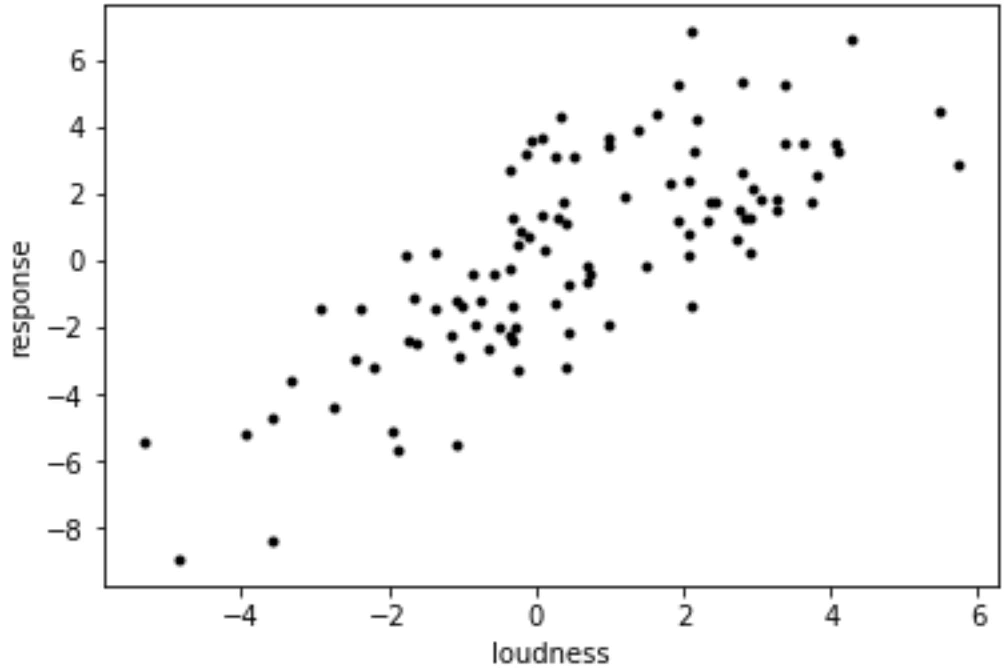


RELATIONSHIPS BETWEEN SAMPLES

* you can also plot loudness vs. fMRI response in a scatter plot (bottom)

* these two seem related. how related? how do we measure?



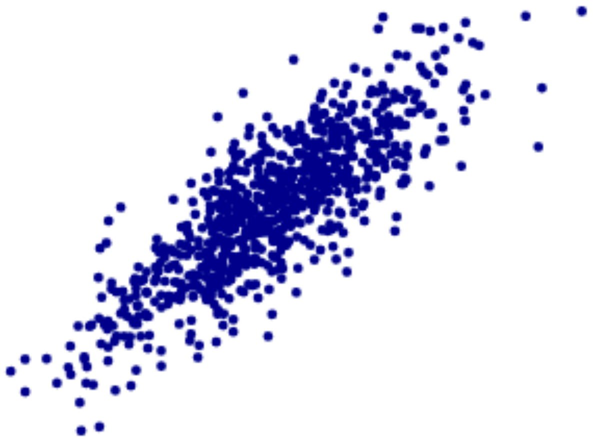


CORRELATION

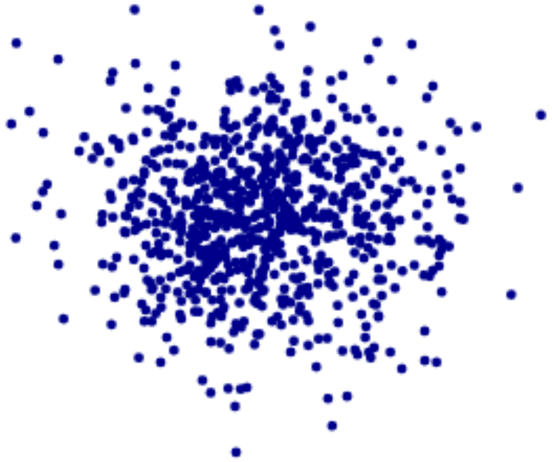
Karl Pearson

* “are these two sets of numbers (linearly) related?”

yes



no



CORRELATION

* the (Pearson) correlation between two variables is their covariance divided by the produce of their standard deviations

$$r_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

r_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}

WHAT THE HECK IS COVARIANCE

* recall:

$$var(X) = \sigma^2_X = \frac{1}{n}\sum_i^N (X_i-\bar{X})^2 = \frac{1}{n}\sum_i^N (X_i-\bar{X})(X_i-\bar{X})$$

var(X) = \sigma^2_X = \frac{1}{n} \sum_i^N (X_i-\bar{X})^2 = \frac{1}{n} \sum_i^N (X_i-\bar{X}) (X_i-\bar{X})

WHAT THE HECK IS COVARIANCE

* in covariance we replace one of the terms with Y:

$$cov(X,Y) = \frac{1}{n}\sum_i^N (X_i - \bar{X})(Y_i - \bar{Y})$$

cov(X,Y) = \frac{1}{n}\sum_i^N (X_i-\bar{X}) (Y_i-\bar{Y})

CORRELATION

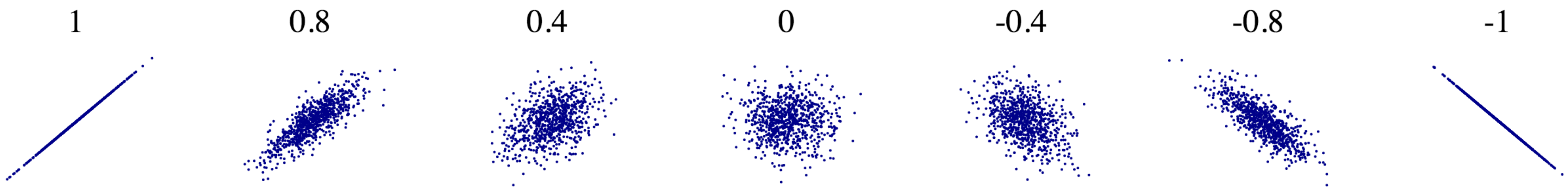
* is covariance, but normalized by the product of the standard deviations

* and thus is always in the range -1…1

  * which is nice

CORRELATION

* tells you how **linearly related** two variables are



CORRELATION

* incidentally, another way to think of correlation:

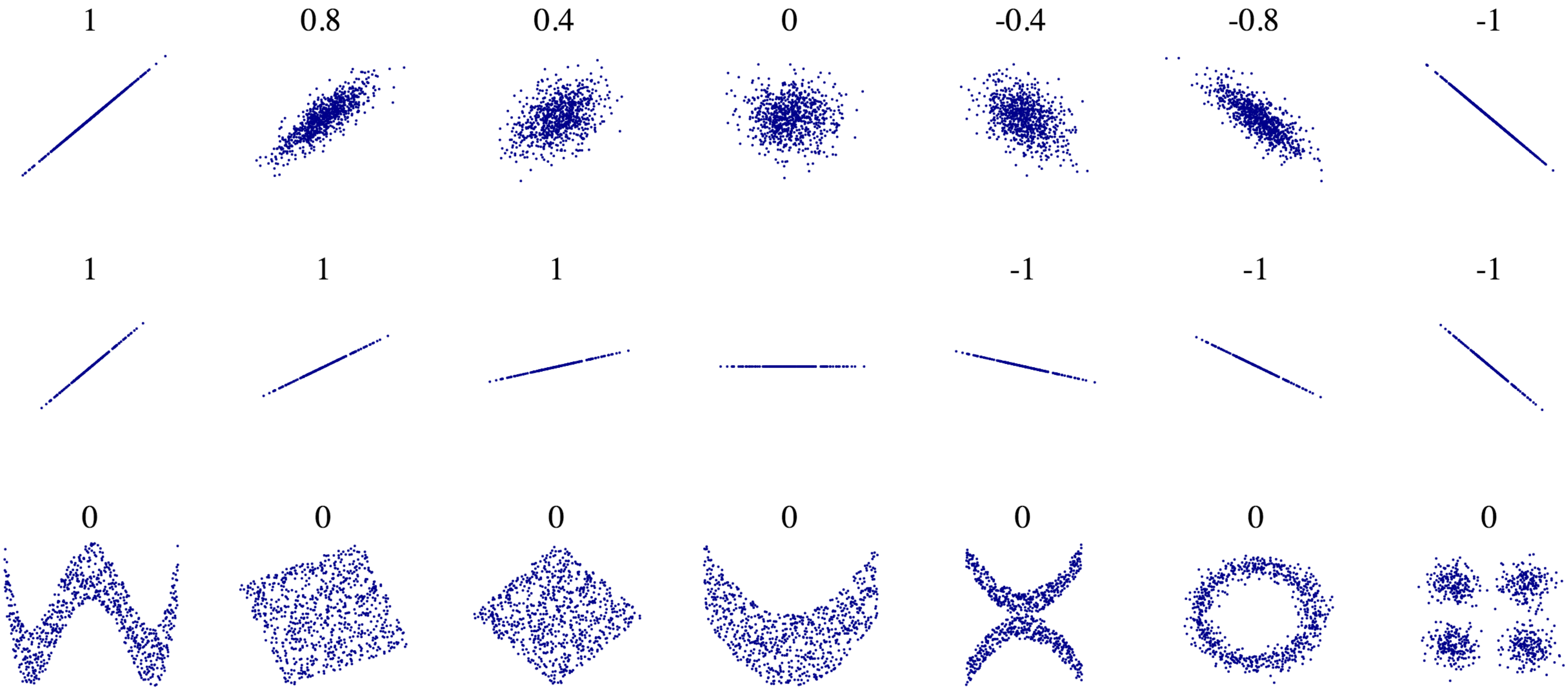
* z-score X and Y to get z(X) and z(Y), then fit a line: z(Y) = m*z(X) + b

* corr(X,Y) = m (the slope of the line)

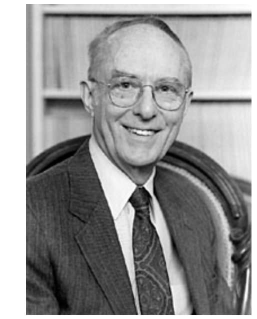
DANGERS OF CORRELATION

* just computing correlation can be dangerous when your variables are related in weird non-linear ways

DANGERS OF CORRELATION

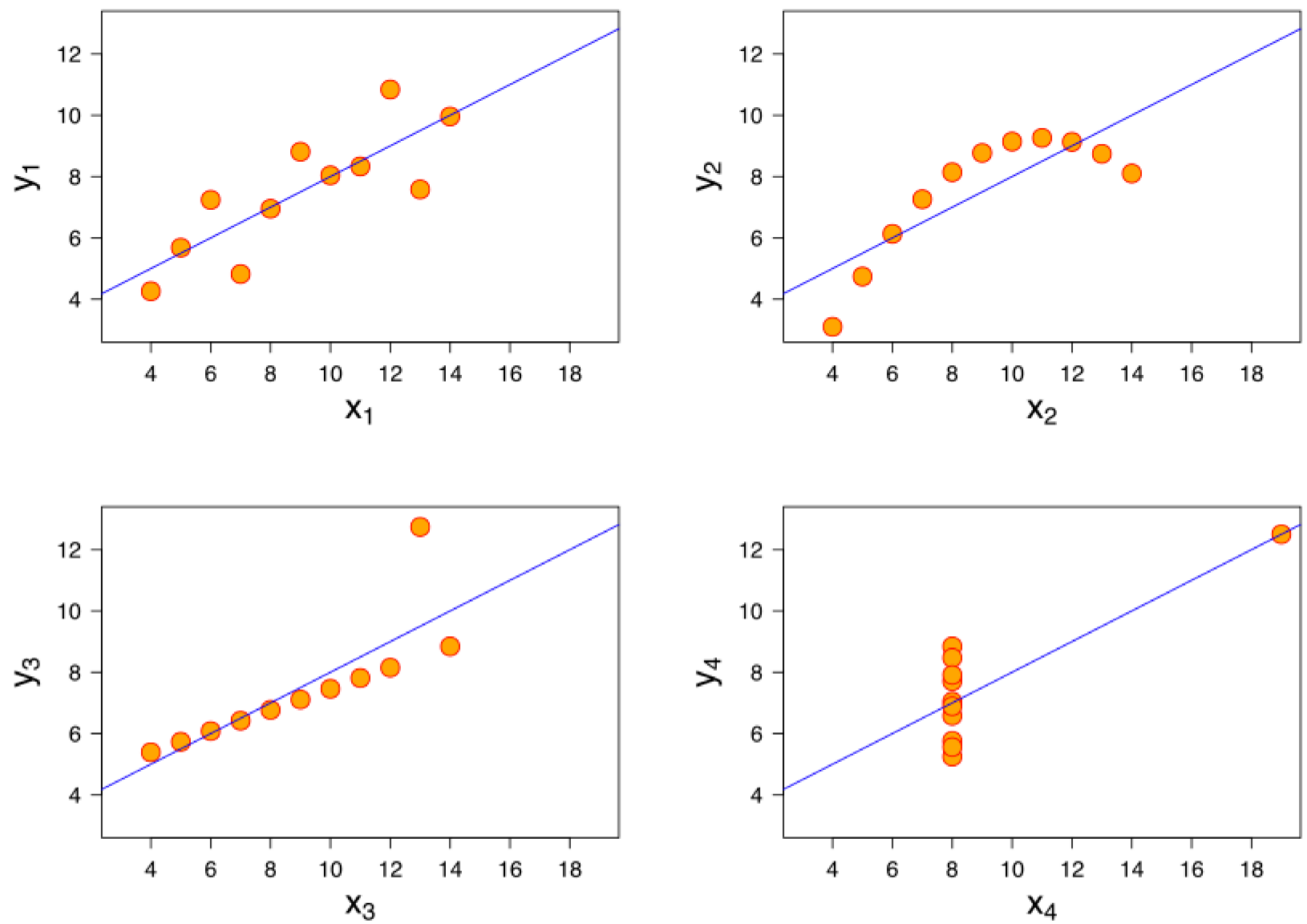


ANSCOMBE’S QUARTET



Frank

* all datasets have identical:
  * correlation
  * mean
  * variance
  * slope
  * R^2



COMPUTING CORRELATION

* **np.corrcoef(arr1, arr2)**
    * computes the correlation between two arrays
    * but weirdly, gives you a 2x2 array back, e.g.:
        * [[1., 0.76],
           [0.76, 1.]]

COMPUTING CORRELATION

* **np.corrcoef([arr1, arr2, arr3, …])**
  * computes the correlation between many arrays
  * for N arrays, gives you back an NxN matrix of correlations

CORRELATION SIGNIFICANCE

* suppose the correlation between X and Y is 0.15

* is this “real”, or is it something you’d see by chance?

* how do we figure this out?

CORRELATION SIGNIFICANCE

* permutation test:

  * correlation depends on X and Y being ordered the same way. but if they are actually uncorrelated, then it shouldn't matter if we re-order them randomly

CORRELATION SIGNIFICANCE

* bootstrap test:

  * bootstrap X and Y (simultaneously, to preserve ordering!) and compute correlation to find a confidence interval & standard error of the correlation measure

  * does the confidence interval include zero? no? then it’s significant!

CORRELATION SIGNIFICANCE

* exact test:

  * if we assume that X and Y are gaussian RVs, then there is an exact formula for what the distribution of correlations look like assuming they are unrelated

  * this can be used to find a p-value

  * implemented in **scipy.stats.pearsonr**

NEXT TIME

* we’re gonna start something totally new: timeseries!

END