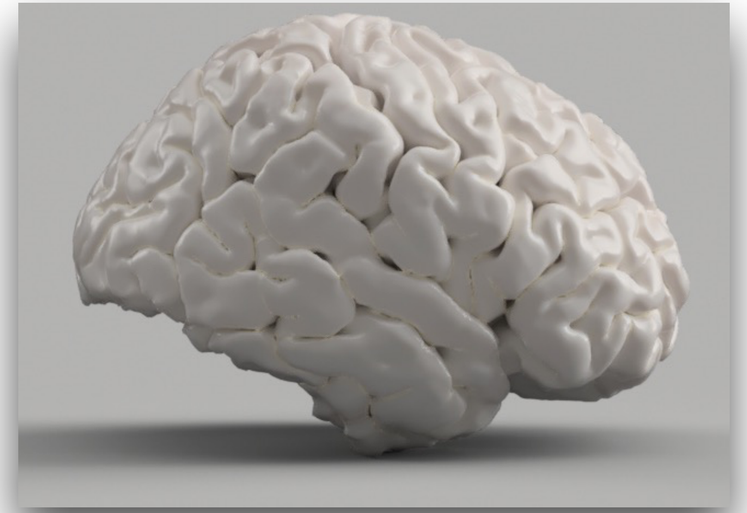
COGS109: Lecture 8



Data Analysis II: Variability

July 11, 2023

Modeling and Data Analysis

Summer Session 1, 2023

C. Alex Simpkins Jr., Ph.D.

RDPRobotics LLC | Dept. of CogSci, UCSD

- D3
- D4

A1 description

- file handout
- how to turn it in

Project description and components

- Paper review (week2 Friday)
- Proposal/checkpoint 1: Data (week3 Sunday)
- Checkpoint 2: EDA and Modeling (week4 Friday)
- Final report and video (group), group review (individual), video reviews (Extra credit) (week5 Friday)

Project proposal/data checkpoint

- **Project outline 109 SS1 23**
  - **Proposal, data checkpoint**
    - -question?
    - -hypothesis?
    - -data files
    - -wrangling cleaning importing
    - -what is the goal with the data?
    - -what variables
    - -format
    - -sample analysis
    - -sample timing
    - -transforms?
    - -what hypothesis do you want to test?
    - -what is the point I'd the model?
    - -confounds?
    - -significance?
  - **Analysis checkpoint**
  - **Final report**

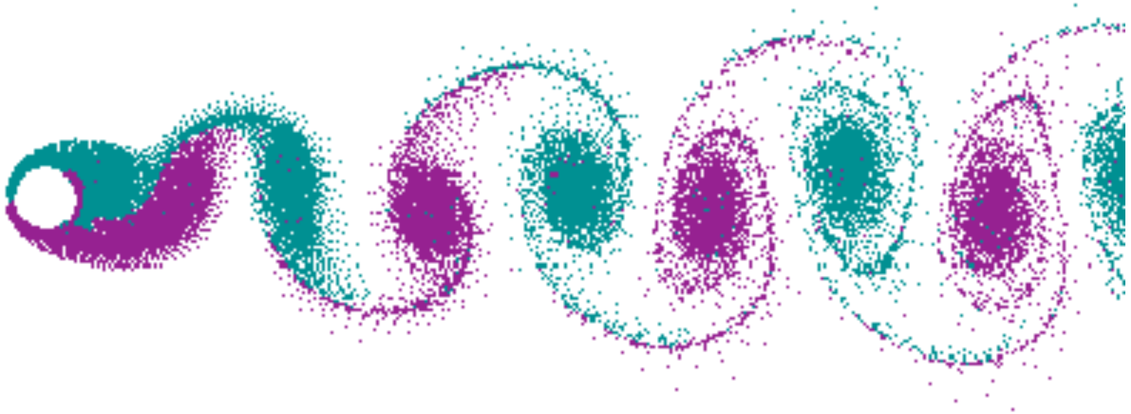
Underlying dynamics need to be exposed

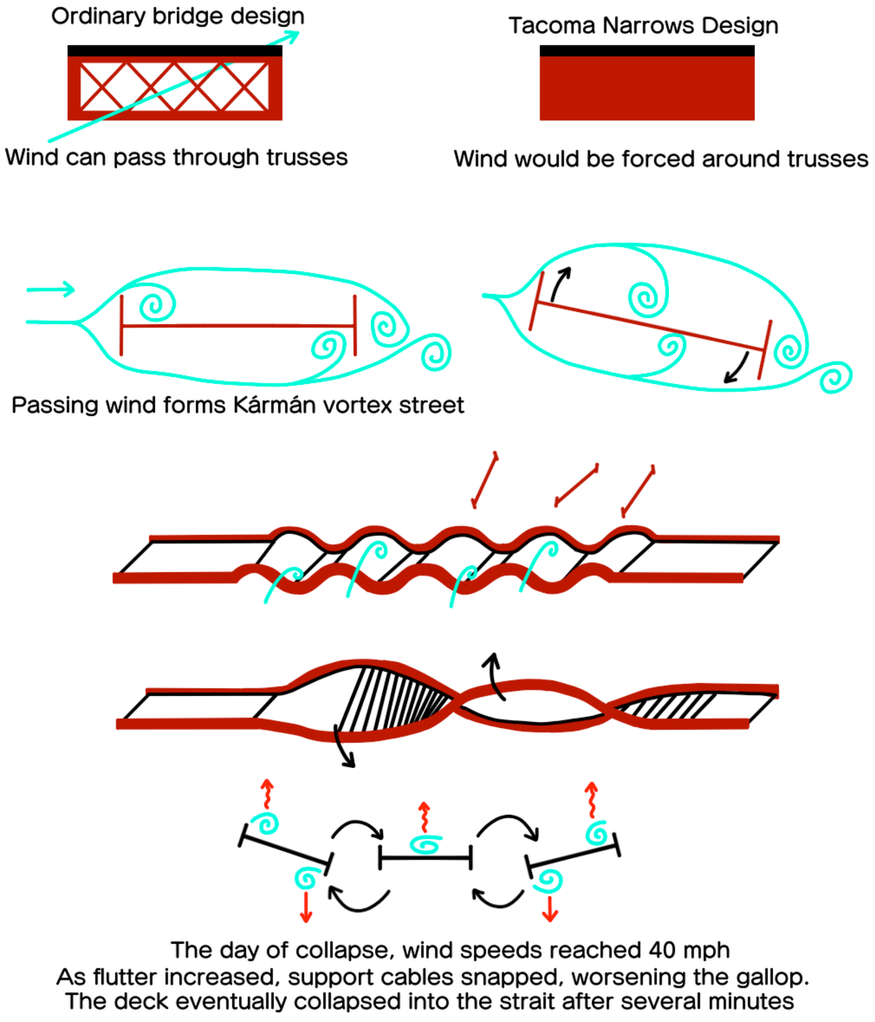
- Tacoma narrows bridge disaster
  - 1st order vs. higher order
  - https://en.wikipedia.org/wiki/Tacoma_Narrows_Bridge_(1940)

https://en.wikipedia.org/wiki/Tacoma_Narrows_Bridge_(1940)#/media/File:Opening_day_of_the_Tacoma_Narrows_Bridge,_Tacoma,_Washington.jpg

Underlying dynamics need to be exposed

- Designed to withstand first order forces but vibration was not considered
- Resonance
- Vortex shedding?
  - Forced oscillation close to resonance fre





https://en.wikipedia.org/wiki/Tacoma_Narrows_Bridge_(1940)

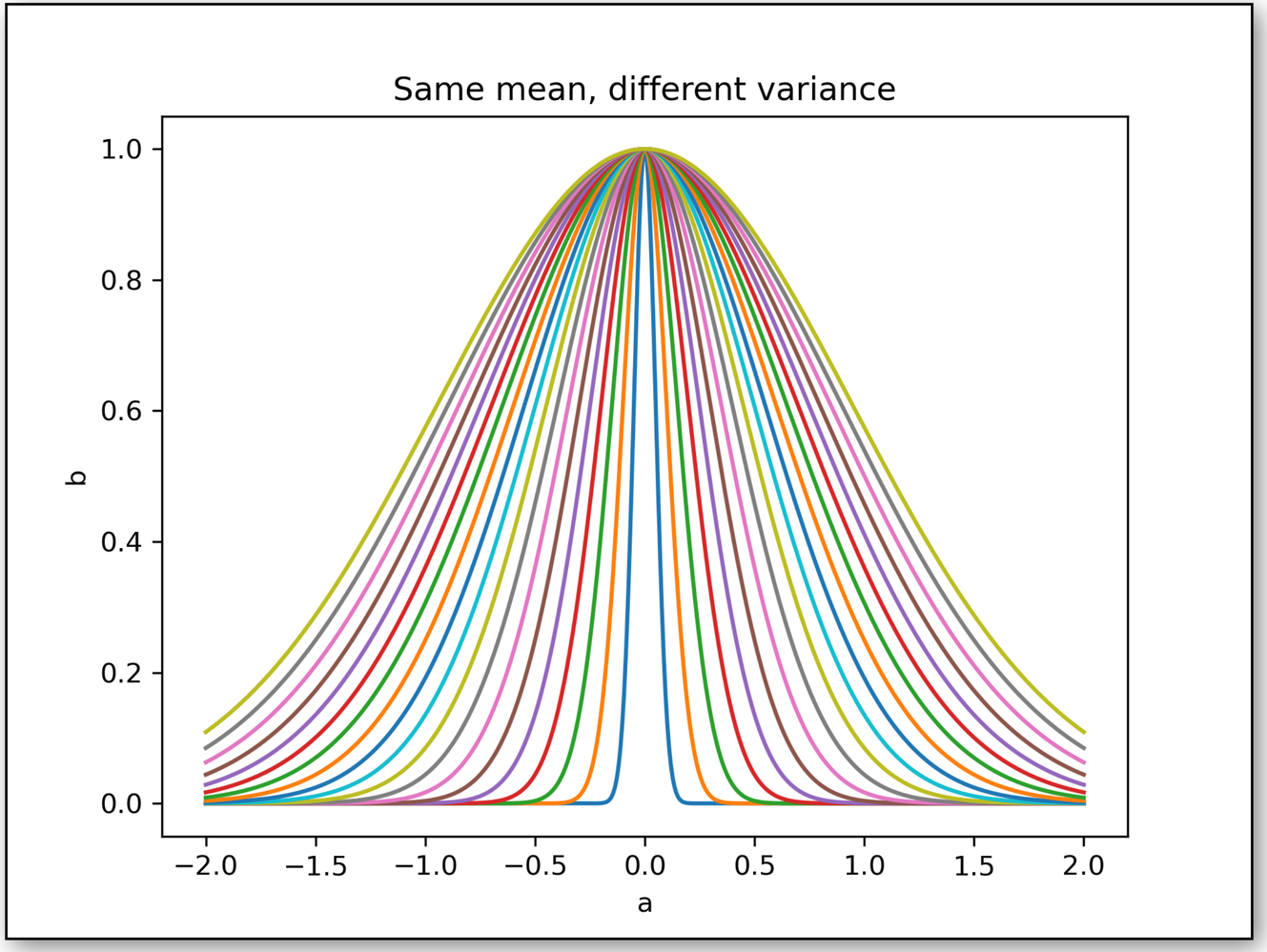
How do we then learn about unknown dynamics?

- Learn by **experience**, **experimentation**, **hypothesis generation**, **data science!**

  - “The Tacoma Narrows bridge failure has given us invaluable information ... It has shown [that] every new structure [that] projects into new fields of magnitude involves new problems for the solution of which neither theory nor practical experience furnish an adequate guide. It is then that we must rely largely on judgment and if, as a result, errors, or failures occur, we must accept them as a price for human progress.” [Othmar Ammann]

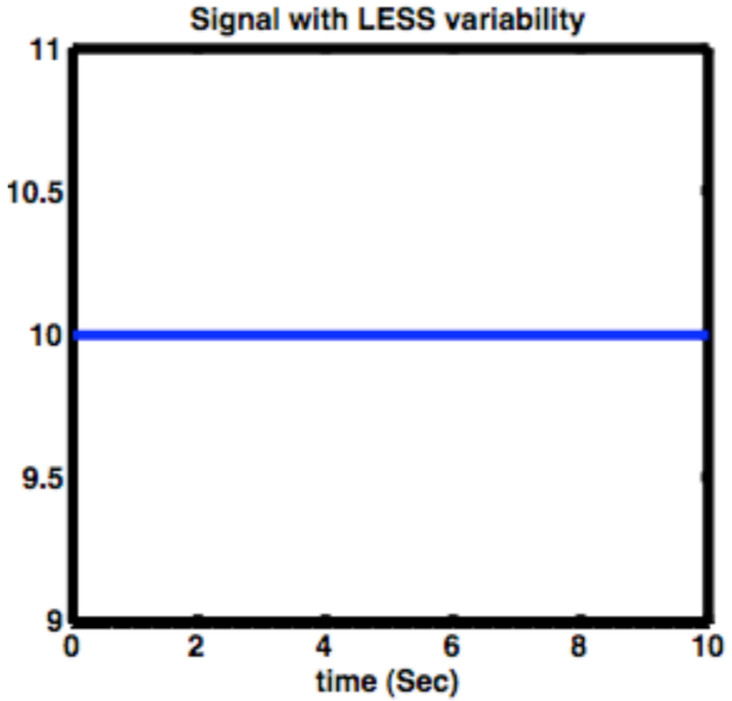
  - Following the incident, engineers took extra caution to incorporate aerodynamics into their designs, and wind tunnel testing of designs was eventually made mandatory.

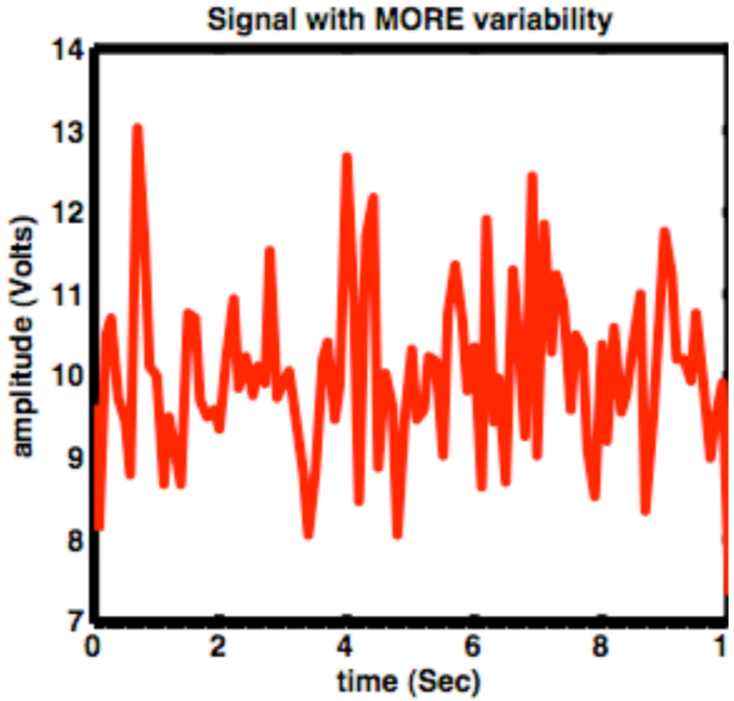
The mean isn’t everything, these have the same mean!



Why we need a measure of variability

Same means, different variability of the signal





We need a measure of Variability, here are a few…

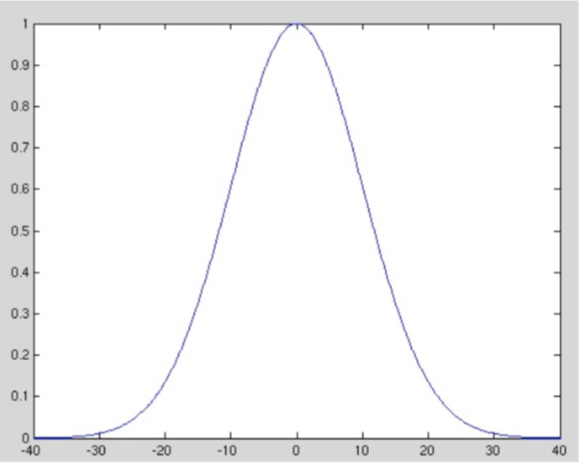
- Range
  - From math review, difference between max and min values of the data

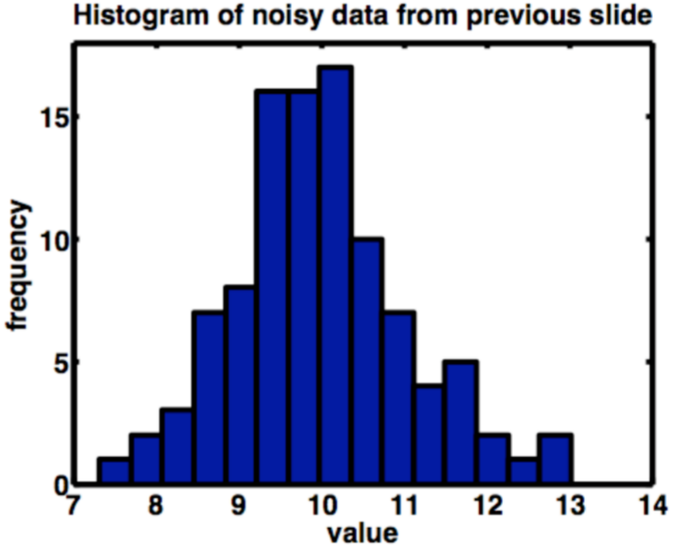
$$Range(x) = Max(x) - Min(x)$$

- Variance
  - Mean of squared deviations from the mean
  - In square units of the sample variable
- Standard deviation
  - Square root of variance
  - In units of the sample variable - sometimes easier to interpret

Returning to the normal distribution…and considering our data in terms of a histogram…

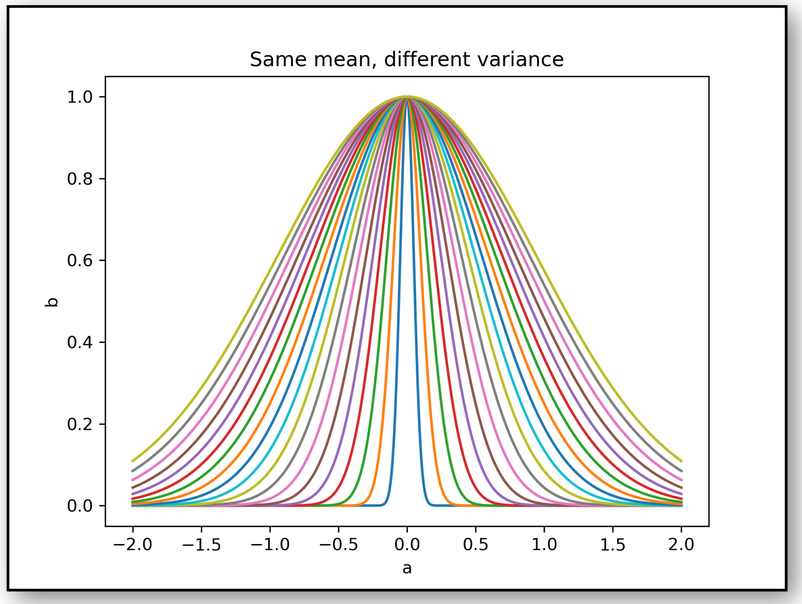
- The distribution of points about the mean can be considered in terms of probabilities
- How likely is a point to deviate from the mean?
- We call the normal distribution a *probability density function (PDF)* because it allows us to predict the likelihood that a sample will take on a particular value





Variance

- Whereas the mean defines a measure for the most likely point in state space (the center ‘location’ of a normal distribution)
- We can define the spread of the normal distribution about the mean by its *variance*



Variance (part II)

- Steps to compute the variance
  - **Compute the deviations from the mean for all the data**

$$d_i = \left(x_i - \bar{x}\right)$$

  - **Compute the square of each of the deviations**

$$sd_i = \left(d_i\right)^2$$

  - **Sum up all these squared deviations**

$$ssqd = \sum_{i=1}^{N} \left(sd_i\right)$$

  - **Divide the mean squared deviations by N, the number of observations**

$$Var = \frac{ssqd}{N}$$

Let’s find variance together

Standard Deviation

- Typical ‘deviation’ from the mean
- Ie how far on average scores depart on either side from the mean
- Easy to compute after the variance - just take the square root of the variance

$$SD = \sqrt{Var} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

$$\bar{x} = \frac{\sum x_i}{N}$$

Let’s look up standard deviation together

Z scores

- A Z score is simply a measure of how many standard deviations away from the mean a score is
- Units are standard deviations

$$Z_i = \frac{X_i - \mu}{SD}$$

Let’s look up z scores in python

- (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.zscore.html)

Covariance

- Covariance is very commonly used in statistical analysis as the basis for advanced statistics
- Gives a quantitative measure of the relationship between two variables

$$Cov(X,Y) = E\left[\left(X - \mu_x\right)\left(Y - \mu_y\right)^T\right]$$

$E = \text{exp}ectation$

$\mu = mean$

More Covariance

- If the two variables are independent, the covariance is 0
  - **(BUT IF COVARIANCE IS 0 THAT DOESN'T MEAN THE VARIABLES ARE INDEPENDENT!!!)**
- If they are totally dependent the covariance of data, can be arbitrarily large
  - **(AGAIN THE CONVERSE IS NOT NECESSARILY TRUE)**
- The diagonals are the variance of each variable
- If each row is an observation, and each column a variable…

$$\text{cov}(X) = \left(\frac{1}{N-1}\right)\left(X - mean(X)\right)\left(X - mean(X)\right)^T$$

Covariance in python

Correlation coefficient motivation

- We want to define a measure of how related our dependent and independent variables are
  - Variance, STD – variation of a single variable
  - Covariance – how two things vary in relation to each other
  - How do we compute the linear dependence of one variable to another?
- Correlation coefficient!

Intuitive arrival at the Correlation Coefficient

- Many kinds (we are going to discuss Pearson's product moment coefficient by Galton)
- A test for linear independence
- We want to measure how two things co-vary
  - We observe one thing varying (e.g. sunset)
  - We observe another thing varying (e.g. air temp. decrease)

Intuitive arrival at the correlation coefficient (II)

- **Positive Correlation** - When one thing's magnitude varies positively, and another thing's magnitude varies positively
  - **and if both vary negatively, also this is referred to as positive correlation**
- **Negative correlation** - When one thing's magnitude varies positively, and another thing's magnitude varies negatively
  - **And if one varies positively while the other varies negatively, this is also referred to as negative correlation**

Intuitive arrival at the correlation coefficient (III)

- We want our measure to be a single number
- In some way we'll need to scale the calculations so that the number is unitless
  - **The variables we're comparing may be in different units**
  - **We also don't care about bias - we're interested in variations, so we make our measures about zero, and normalize each**
  - **Remember when we presented z-scores as a normalized measure of how far from the mean a particular sample is in a dataset?**

$$Z_i = \frac{X_i - \mu}{SD}$$

Intuitive arrival at the correlation coefficient (IV)

- We arrive at the correlation coefficient by multiplying each z-score from one variable by the z-score from the other variable, then averaging all those results
  - **Thus if both tend to vary positively?**
    - Positive correlation
  - **If both tend to vary negatively?**
    - Positive correlation
  - **If one varies positively, and the other negatively?**
    - Negative correlation
  - **If sometimes they both vary positively or negatively, sometimes they vary oppositely?**
    - Small or near zero correlation

Correlation coefficient

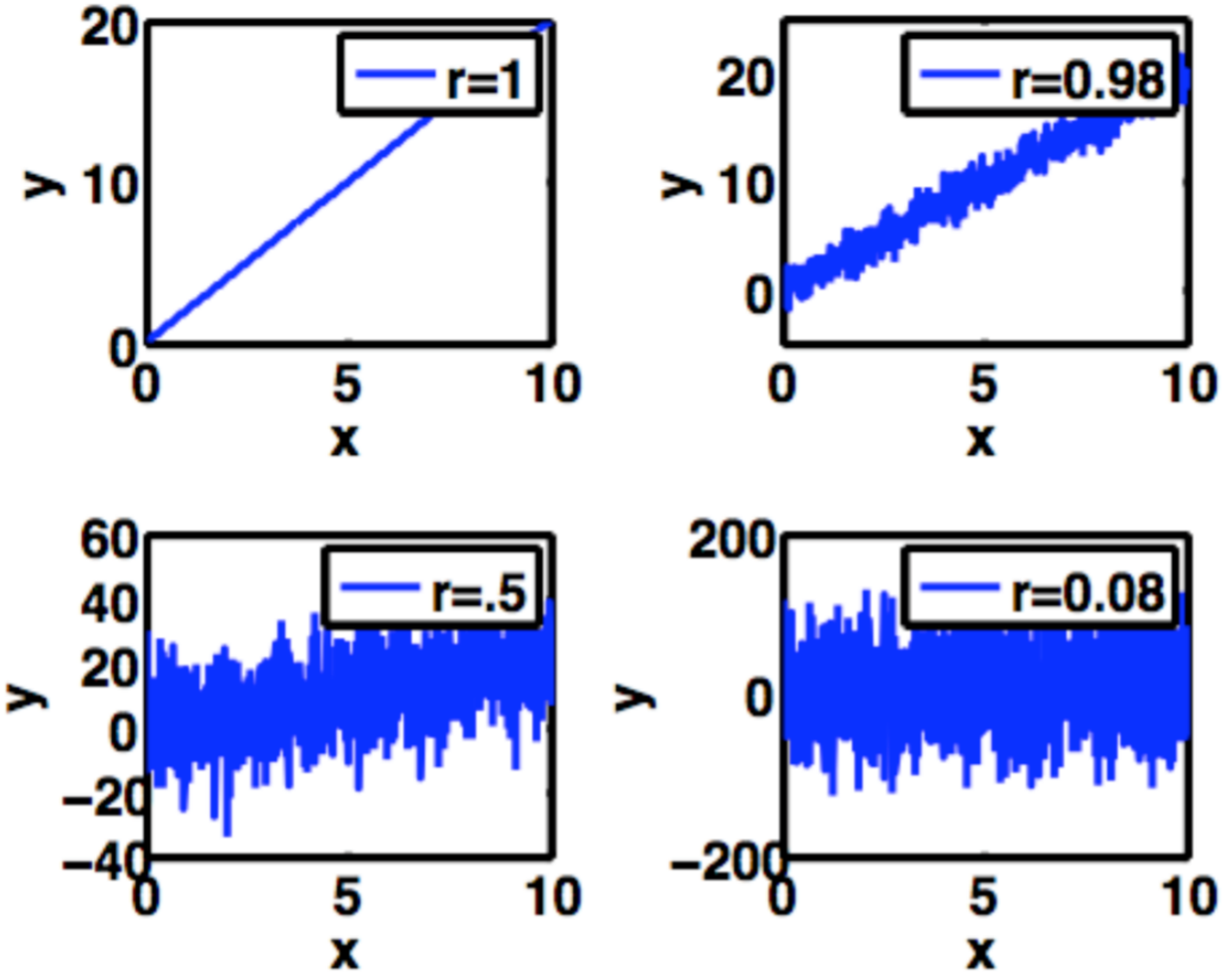
$$\rho(j,k) = \frac{\sum_{i=1}^{N} Z_{ij} Z_{ik}}{N}$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$\rho(X,Y) = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

Characteristics

- Range
  - **-1<=r<=1**
- Interpretation - independence
  - **Statistical independence**
    - The more distinct and unrelated the covariation, the closer to zero the correlation coefficient
    - Statistically independent if their correlation is zero
  - **Linear independence**
    - Two things varying perfectly together are linearly dependent, variables with less than perfect correlation are linearly independent


r=1
r=0.98
r=.5
r=0.08
x
y
0
5
10
20
60
40
−20
−40
200
−200

Correlation coefficient in python