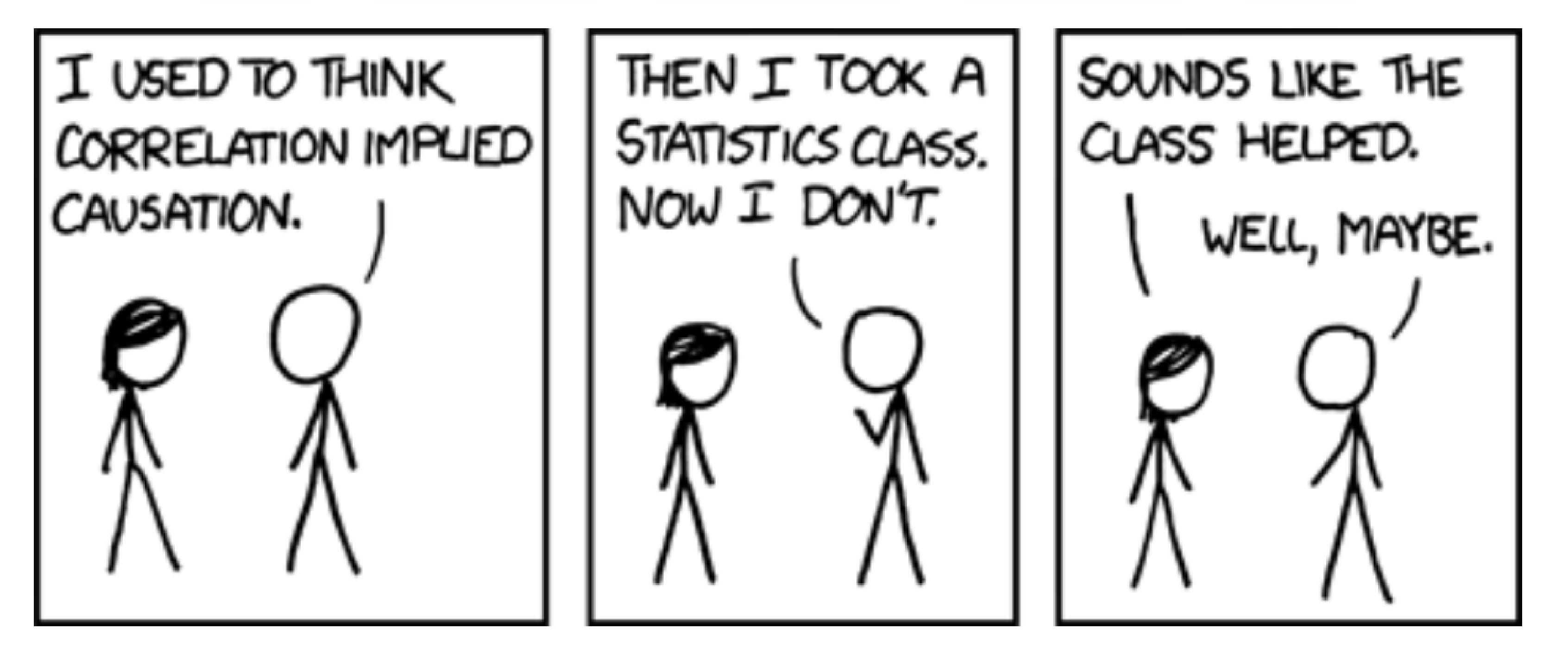
Session 5: Statistics and Exploratory Data Analysis

STARFISH SCHOOL 2022

Introduction to Basics in Statistics



https://xkcd.com/552/

Types of Data

Types of Data

Quantitative

- Continuous
  - Real or complex numbers

- Discrete
  - integers

Categorical

- Nominal
  - e.g., categories A, B, C, or I, II, III

- Ordinal
  - Ordering matters, e.g., a *Likert Scale* used in a survey: 1,2,3,4,5

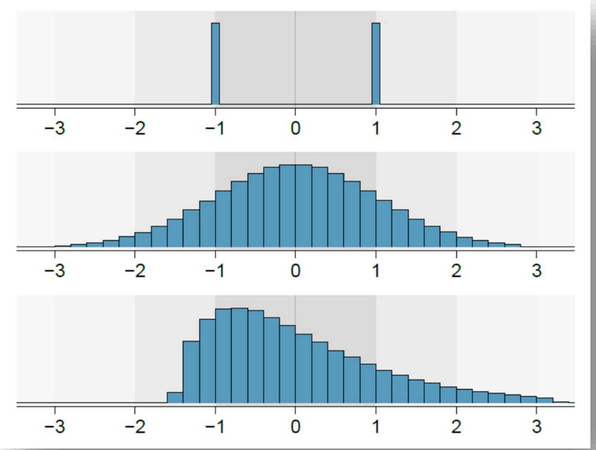
What astronomy examples can you think for each type?

Distributions

What exactly is a distribution?

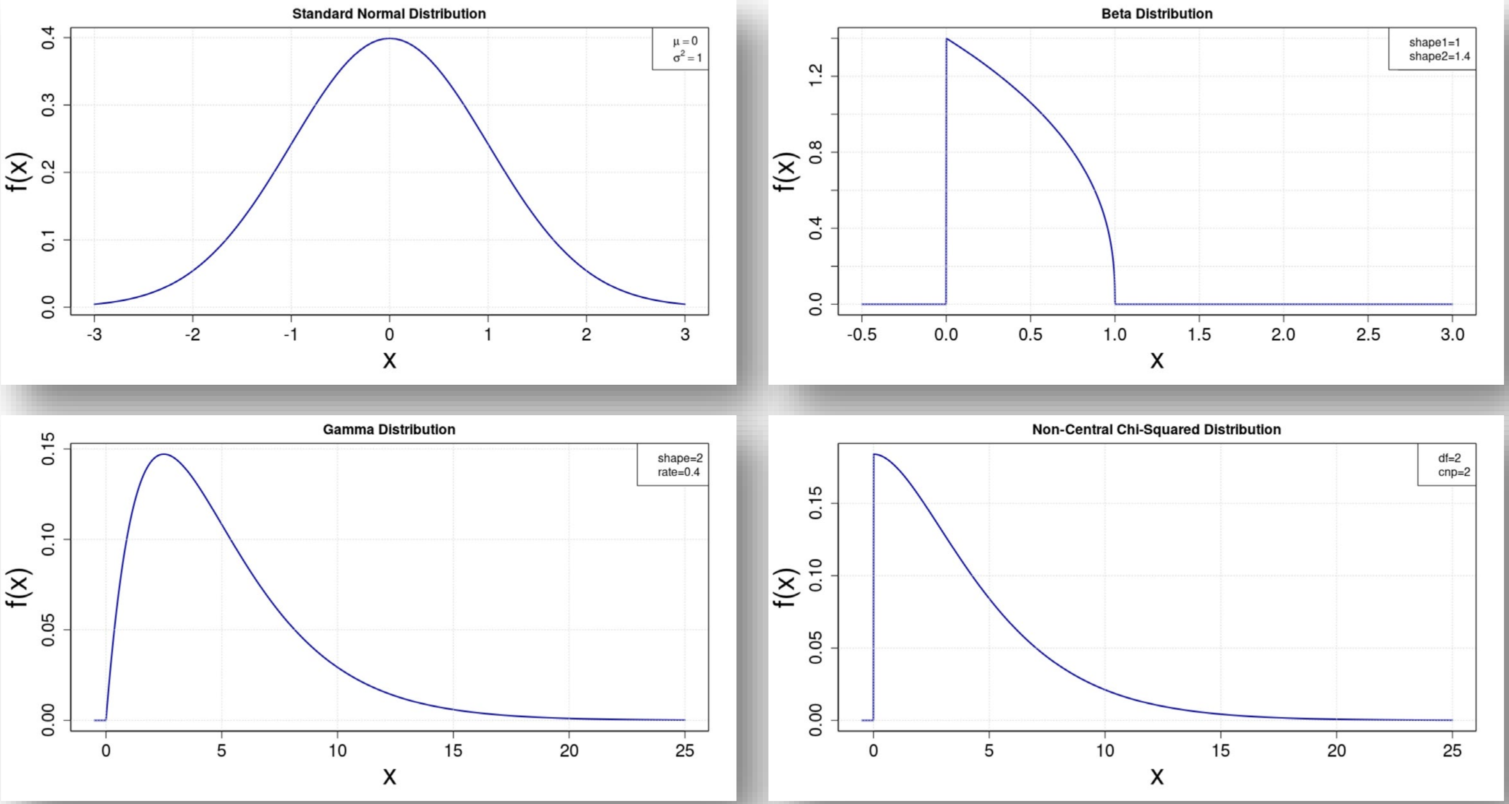
A distribution…

- Tells you the frequency or relative frequency of each possible value/event, or of some data that was collected
- **Could be empirical or analytic**
- Can be useful for modelling a population of objects
- Is often a foundation of statistical reasoning
- Can be continuous or discrete
- That is analytic has parameters that define its shape
- Can be univariate or multivariate



Example histograms (figure from Open Intro Statistics 4th ed.)

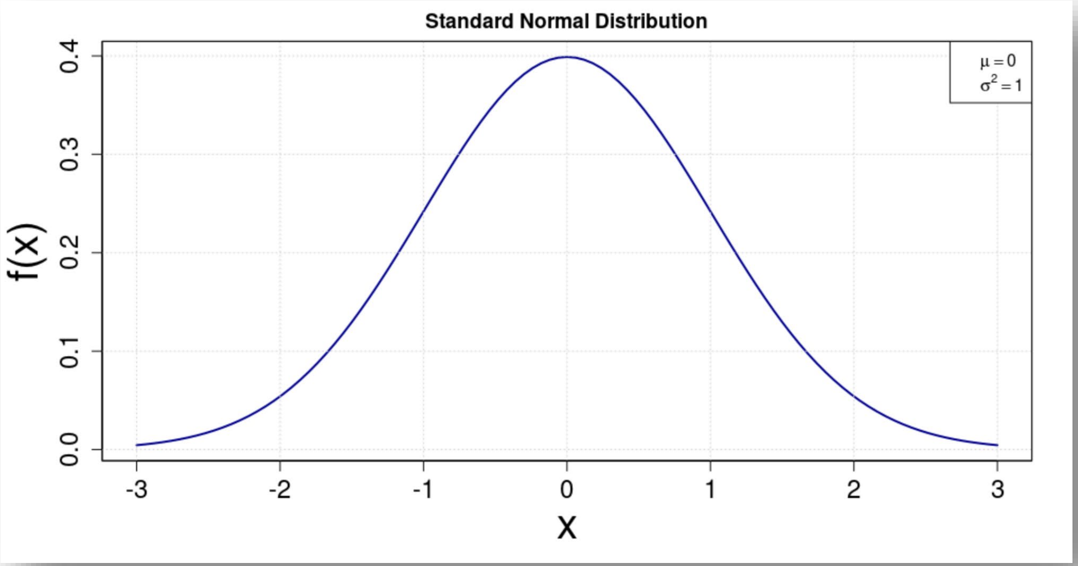
Some analytic probability distributions



Probability Distributions

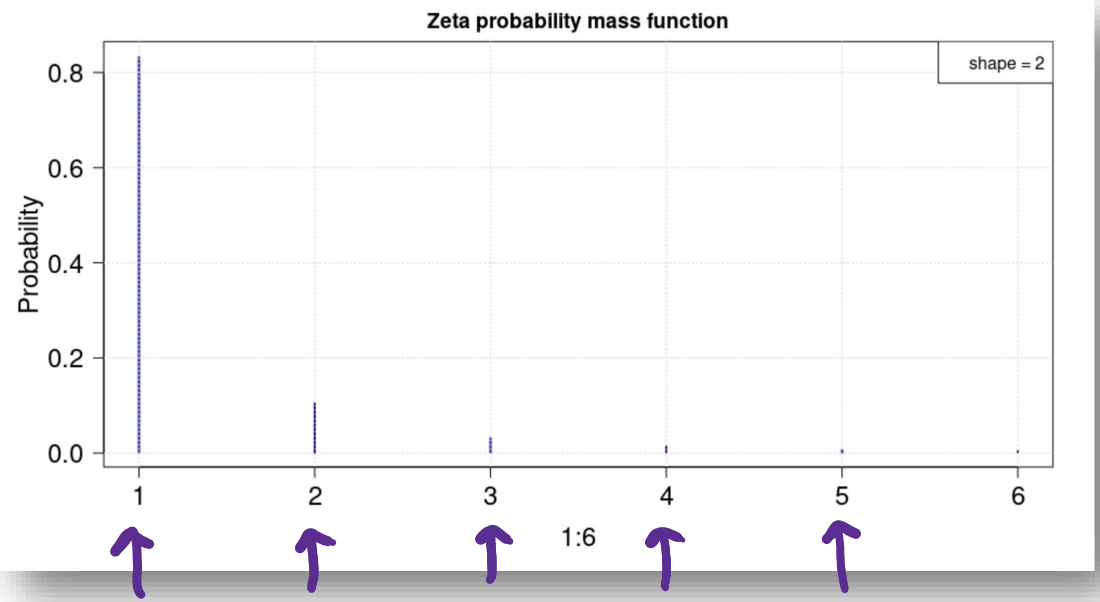
Continuous quantities

probability density function (pdf)

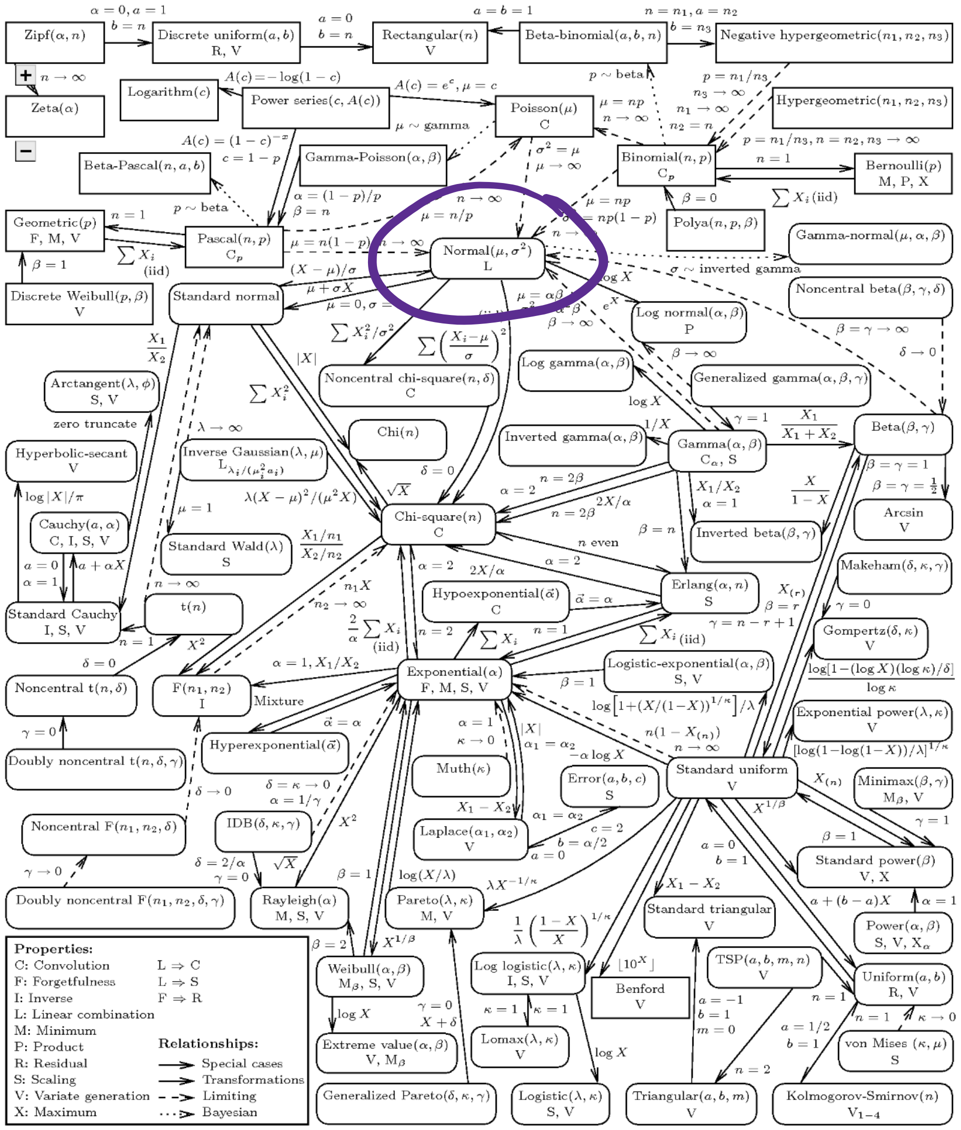


Discrete quantities

Probability mass function (pmf)

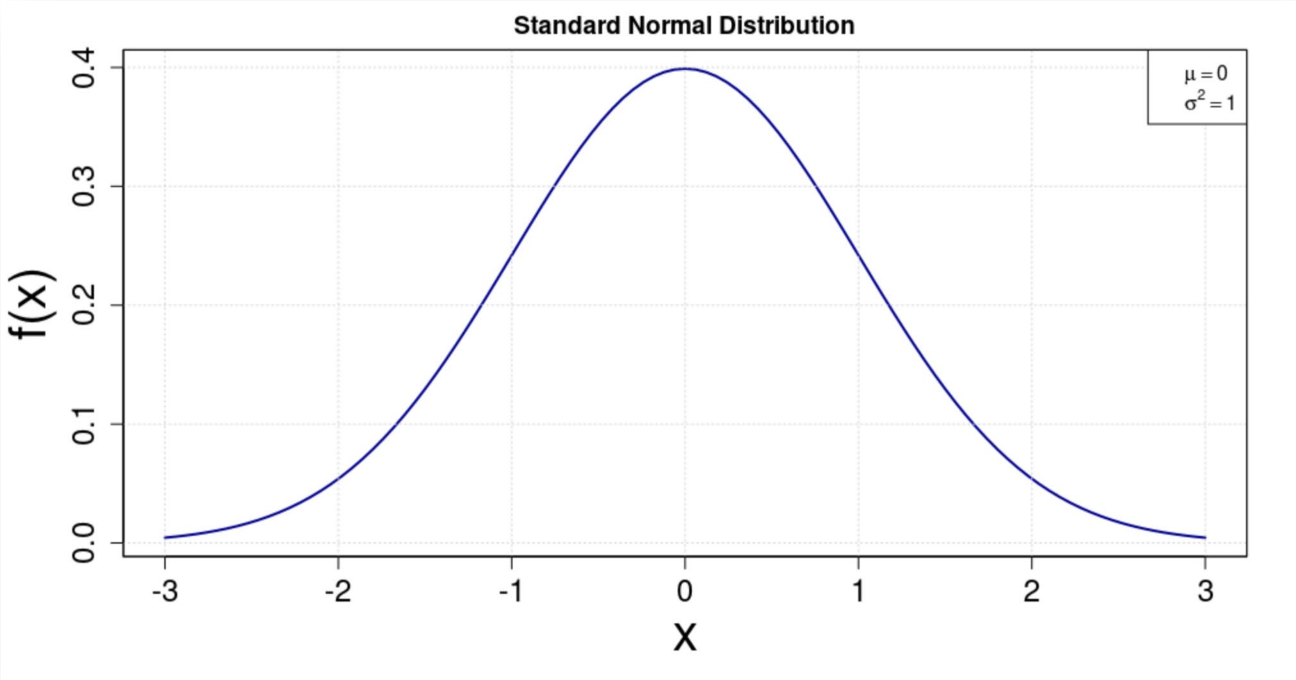


There are many univariate distributions!



http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

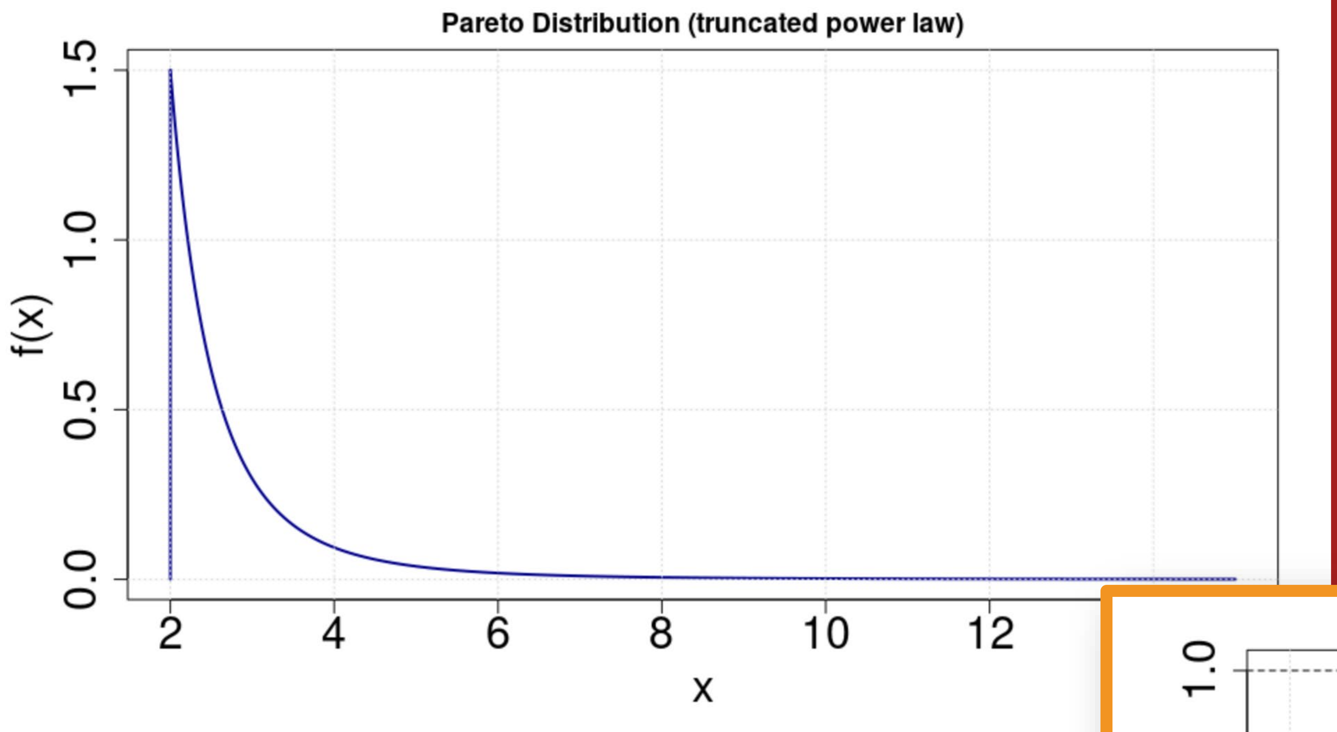
The Normal Distribution



$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Example:

Pareto distribution (truncated power-law)

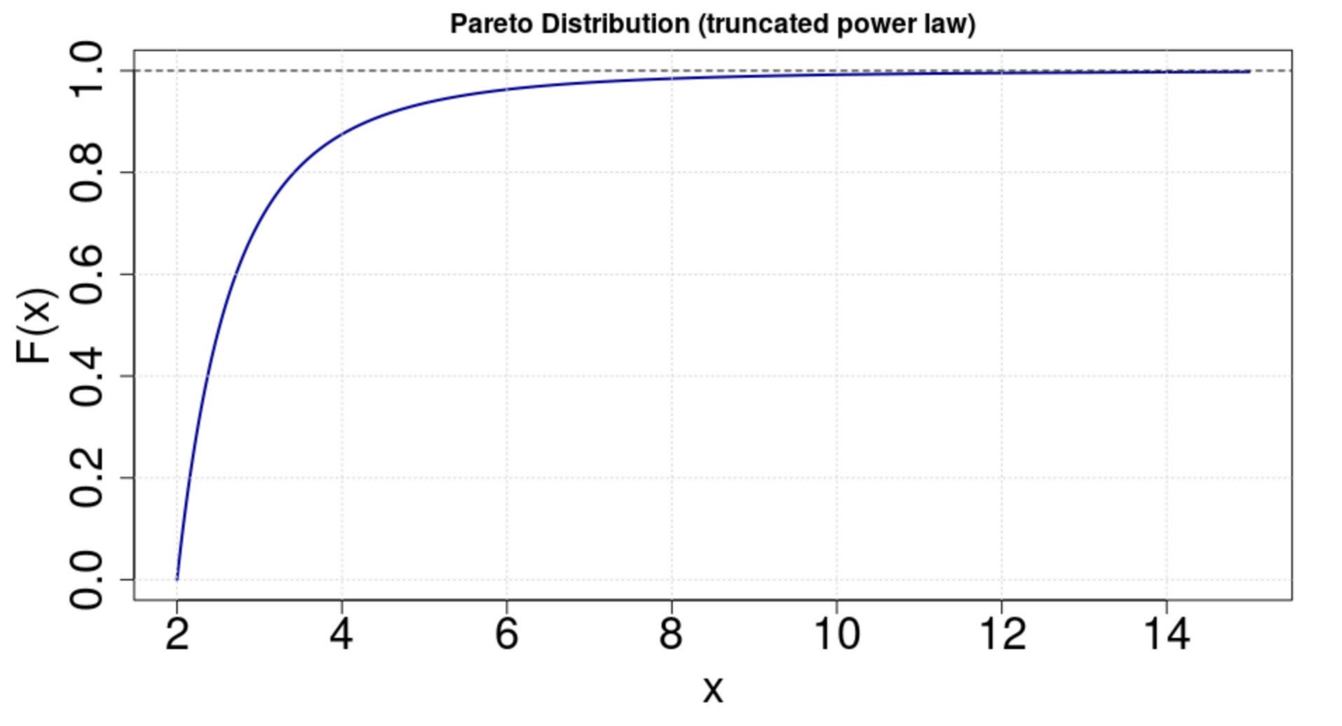


Probability distribution function (pdf)

$$f(x) = \frac{\alpha x_{\min}^{\alpha}}{x^{\alpha+1}}$$

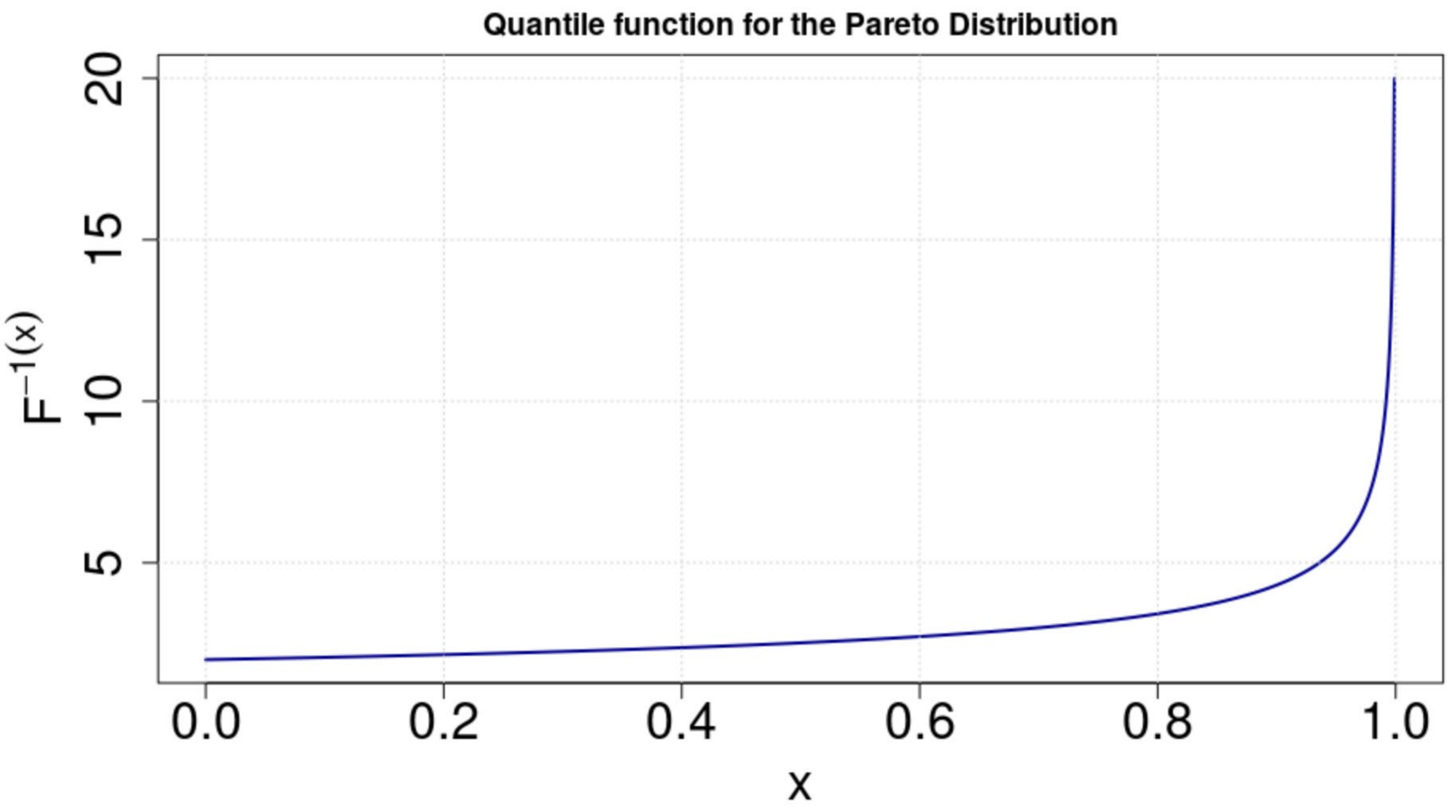
$$F(x) = P(X \leq x) = 1 - \left(\frac{x_{\min}}{x}\right)^{\alpha}$$

Cumulative distribution function (cdf)



Quantile Function

- This is the inverse of the cumulative distribution function (cdf)



Random Variables

Random Variable

- A random variable X is a *function* that maps an *outcome* to a *real number*
  - e.g., Let's say we decide to flip a coin repeatedly, and each time we flip it we record whether we get heads or tails with a 1 or a 0 respectively.

- In other words, X is a **function**. Little x represents the data --- *realizations* of that random variable.

A Random Variable follows a distribution

The standard statistics notation to show what distribution a random variable follows is:

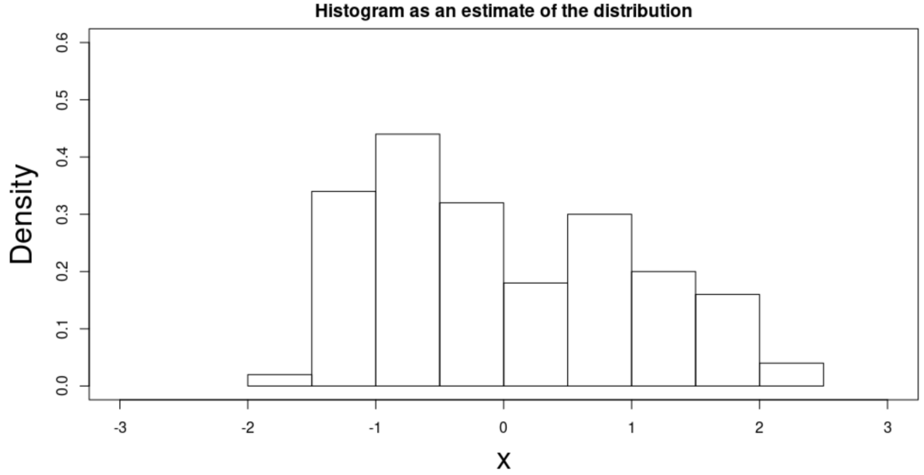
$$X \sim N(\mu, \sigma^2)$$

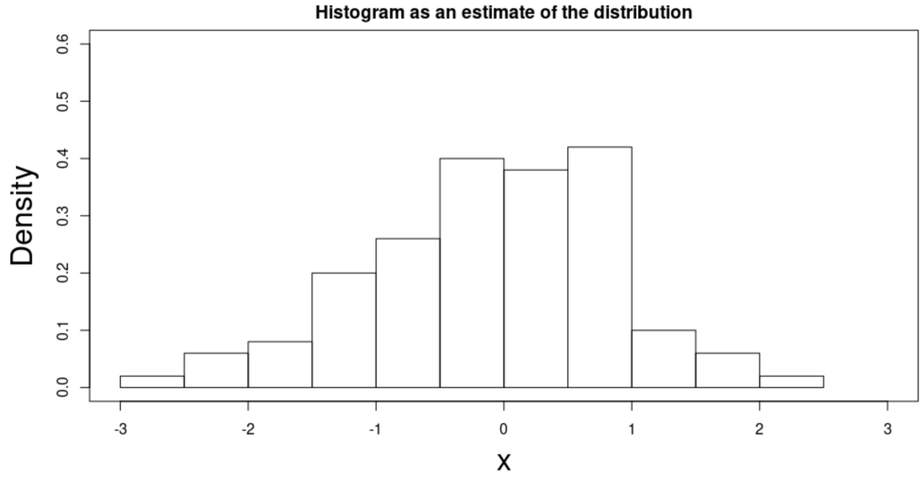
For example, we might assume that are data x (e.g. the photon counts from a star) follows a Poisson distribution

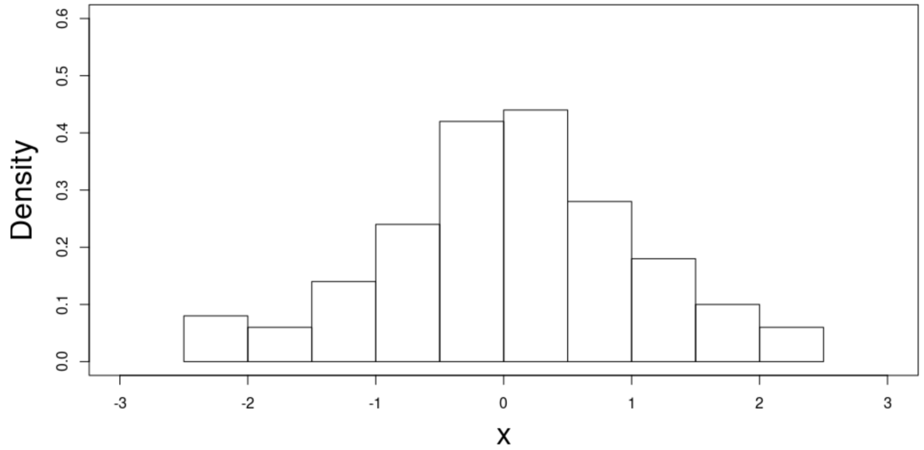
$$X \sim \mathrm{Pois}(\lambda)$$

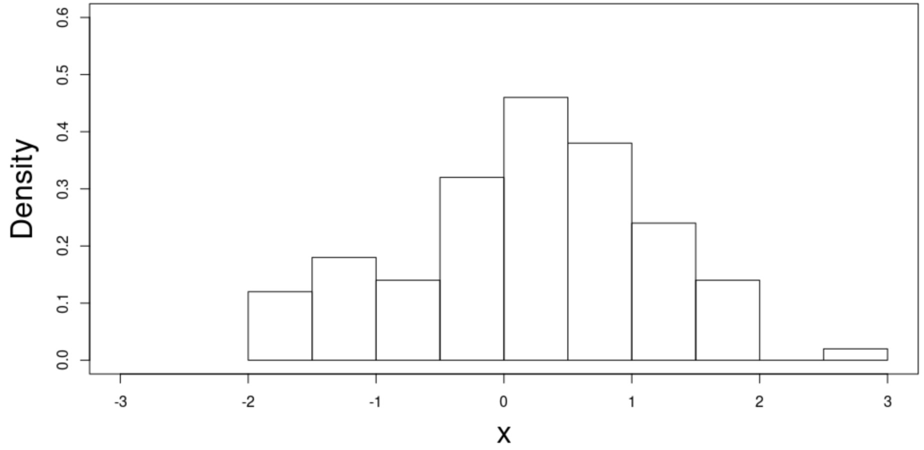
Randomness in Data

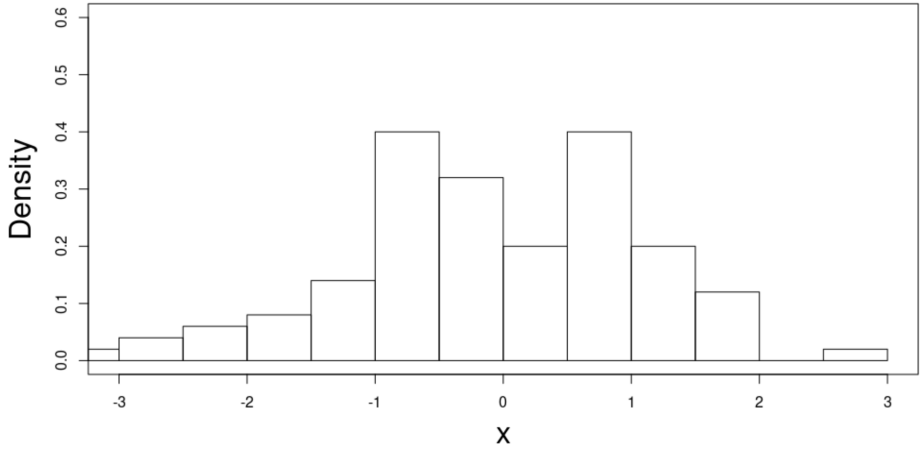
- All these histograms were generated from 100 draws from a standard normal











Sampling from a distribution (two basic approaches)

Inverse cdf Method

- First choice if the inverse cdf is tractable

Accept/Reject Algorithm

- Useful when you can't write down the inverse cdf

Exercise 1 -- Exercise_1.ipynb

1. Generate a random variable follow a **uniform distribution** between 0 and 50

2. Generate a random variable follow a **normal distribution** with mean = 100 and standard deviation of 50

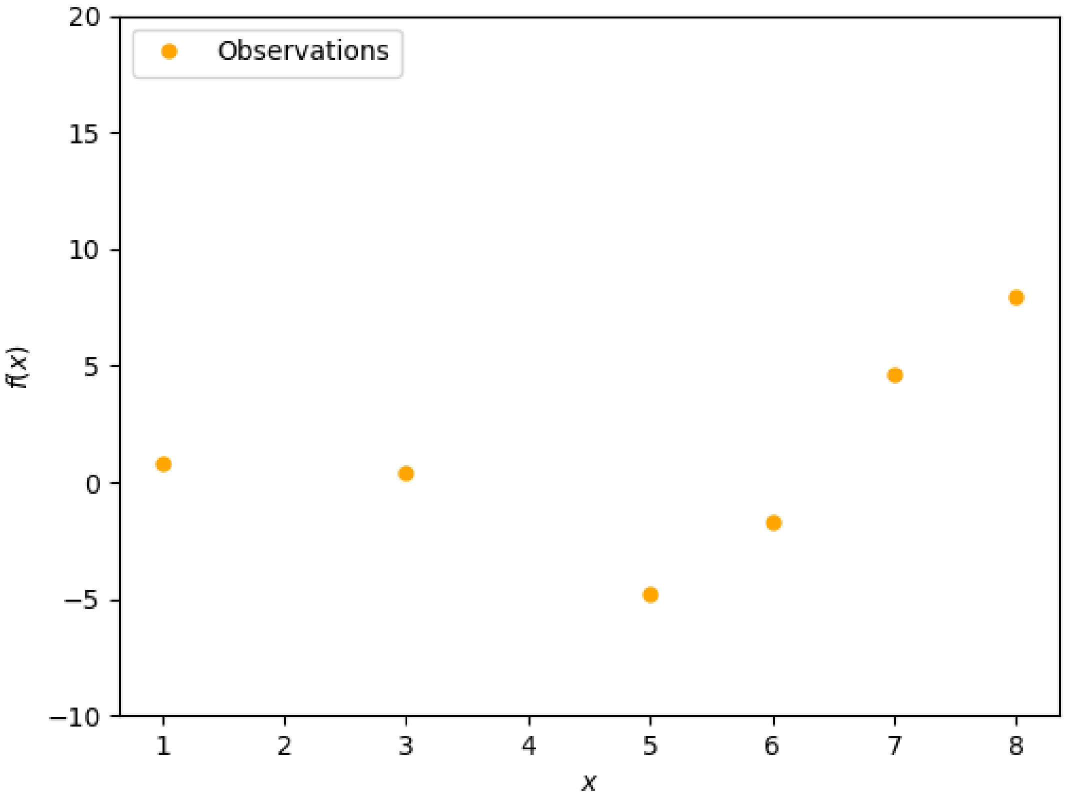
3. Use the ***accept-reject* approach** to transform numbers generated from a uniform distribution into those following the distribution: $\boldsymbol{P(x) = \left(\frac{1}{e-1}\right) e^{-x}}$ for $\mathbf{0 < x < 1}$ and **0** elsewhere.
   1. Draw a random samples x^* from the $U(0,1)$ distribution and a random sample y^* from the $U(0,c)$ distribution.
   2. If $y^* < f(x^*)$, keep x^*. If not, return to step 1.
   3. Continue until you have 100,000 samples.
   4. Plot a normalized histogram of the samples and then overplot the PDF.

4. Use ***cdf sampling*** to do the same thing above.
   1. To do this, compute the cdf $F(X)$ by integrating the PDF $P(x)$ from $-\infty$ to X.
   2. Then find the inverse $F^{-1}(X)$ of the CDF. *[HINT: Remember an inverse function $F^{-1}(x)$ is such that $F\left(F^{-1}(x)\right) = x$]*
   3. Draw a random sample u^* from the $U(0,1)$ distribution.
   4. Then the variable $y = F^{-1}(u^*)$ will have the probability distribution you seek.
   5. Continue until you have 100,000 samples.
   6. Plot a normalized histogram of the samples and then overplot the PDF.

Smoothing and Interpolation

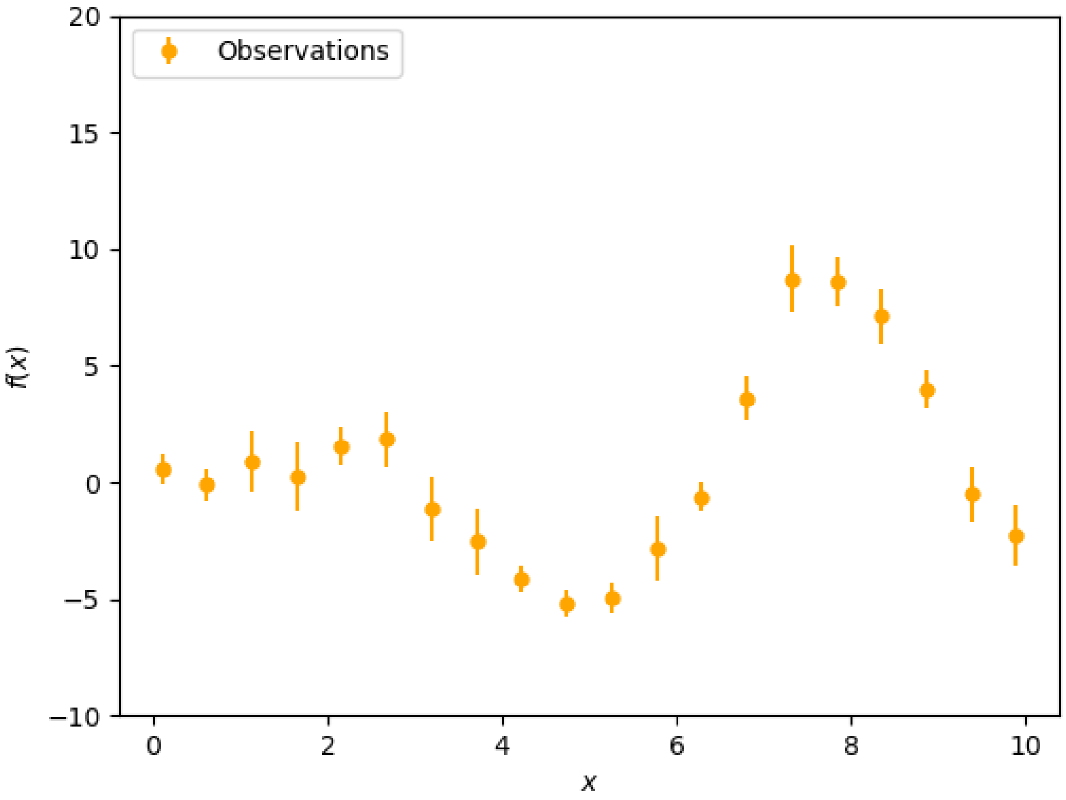
How to make data more pliable

- Sometimes you'll get data that looks like this:



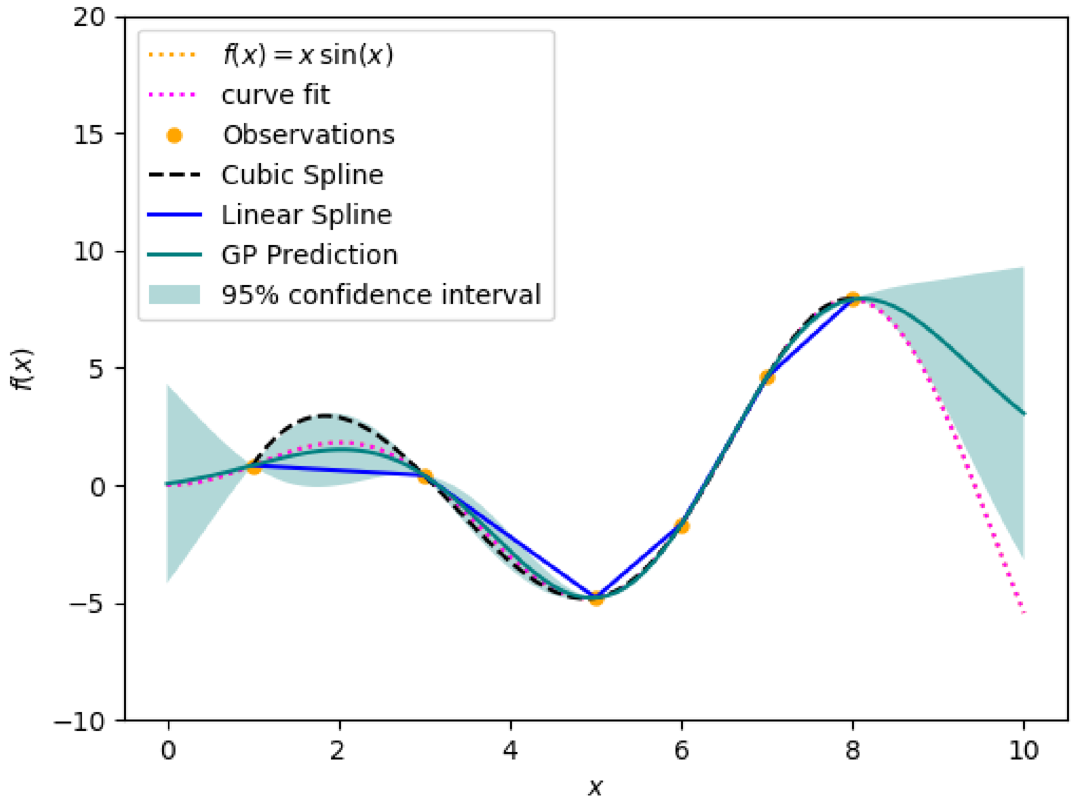
How to make data more pliable

- or this (with error bars)



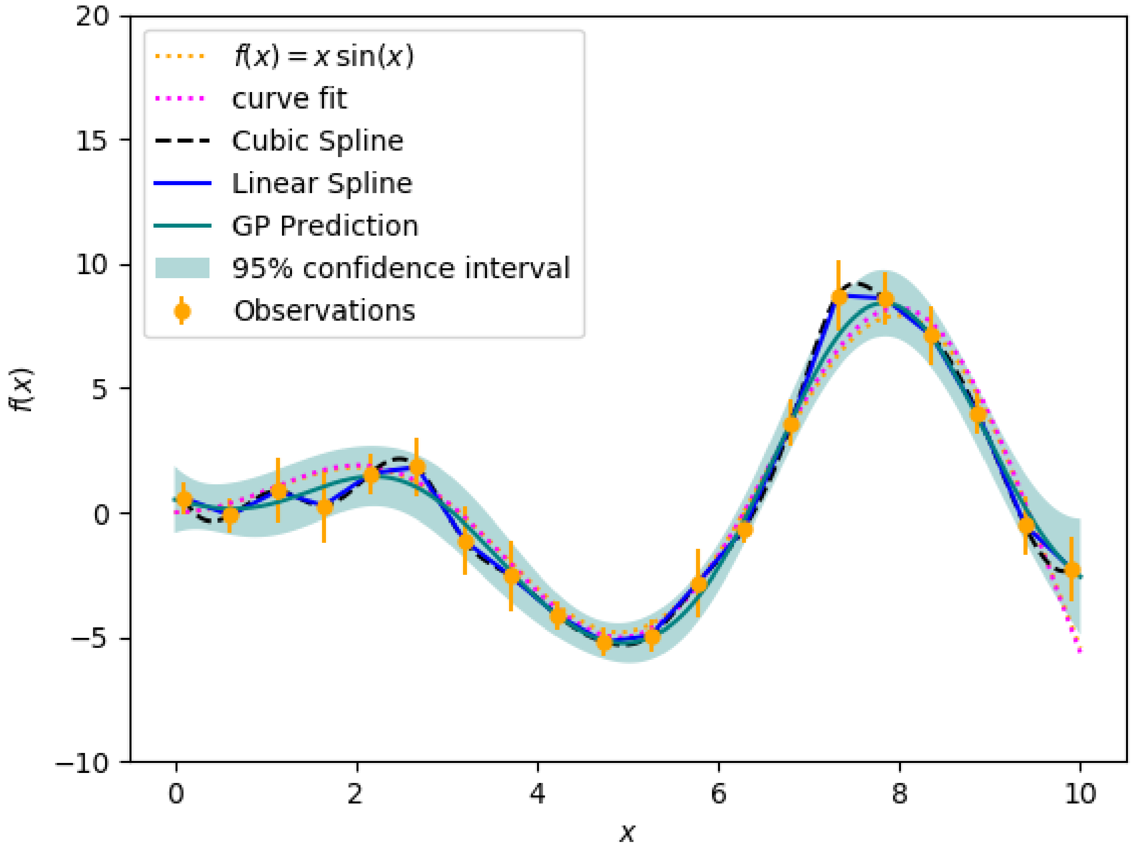
Examples

- Simple example from Vanderplas ++
- This code is provided in your exercise set – and shows a combination of methods
  - Spline_GP_demo.ipynb



Examples

- Simple example from Vanderplas ++
- This code is provided in your exercise set – and shows a combination of methods
    - Spline_GP_demo.ipynb



Fitting a Model to Data

How to fit a model to data?

Follow along with the Jupyter notebook:
Fit_your_data_demo.ipynb