Computational Methods for Linguists

Ling 471

Olga Zamaraeva (Instructor)
Yuanhe Tian (TA)
05/04/21

Thanks

- **Many thanks** for filling out evaluations
  - (and lecture surveys)
  - Feedback:
    - Term **definitions**
      - I will try; please type in the chat or even just interrupt outloud: “Blah: definition?” when I miss something
    - More **activities**
      - Let’s try today!
    - Ability to add **specific** (optional) comments in lecture survey
      - **Done**!
    - Canvas is bad (e.g. for discussion threads)
      - Fully agreed :) :) :)

Reminders

- Assignment 3 due Thursday
- Blog due today
  - Comments due by Thursday

Plan for today

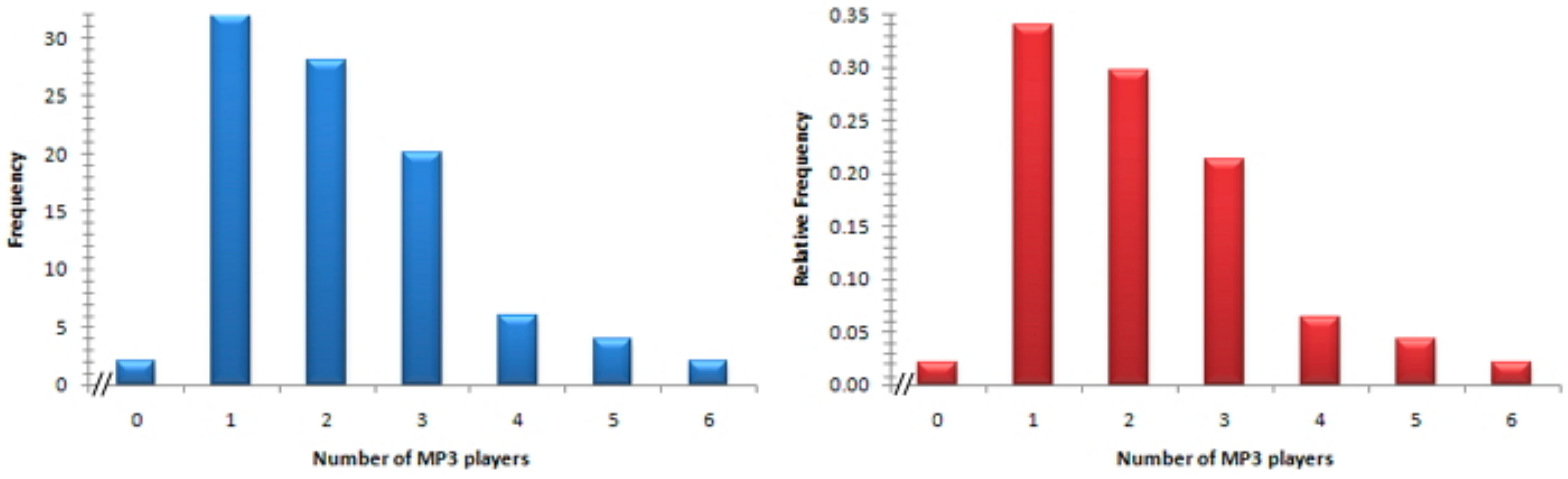
- Finish probability theory basics:
  - Probability mass and density
  - Distributions
  - Gaussian/Normal distribution
    - in Linguistics???
  - Group activity/exercise:
    - Implement the Gaussian formula in python and visualize the data and the distribution
    - Implementing formulas is scary
      - Goal: tackle some of that fear :)

Probability Mass, Density, and Distributions

(this is all a bit abstract)

(I will try to present the same info from different angles on different slides)

Relative frequencies
and histograms

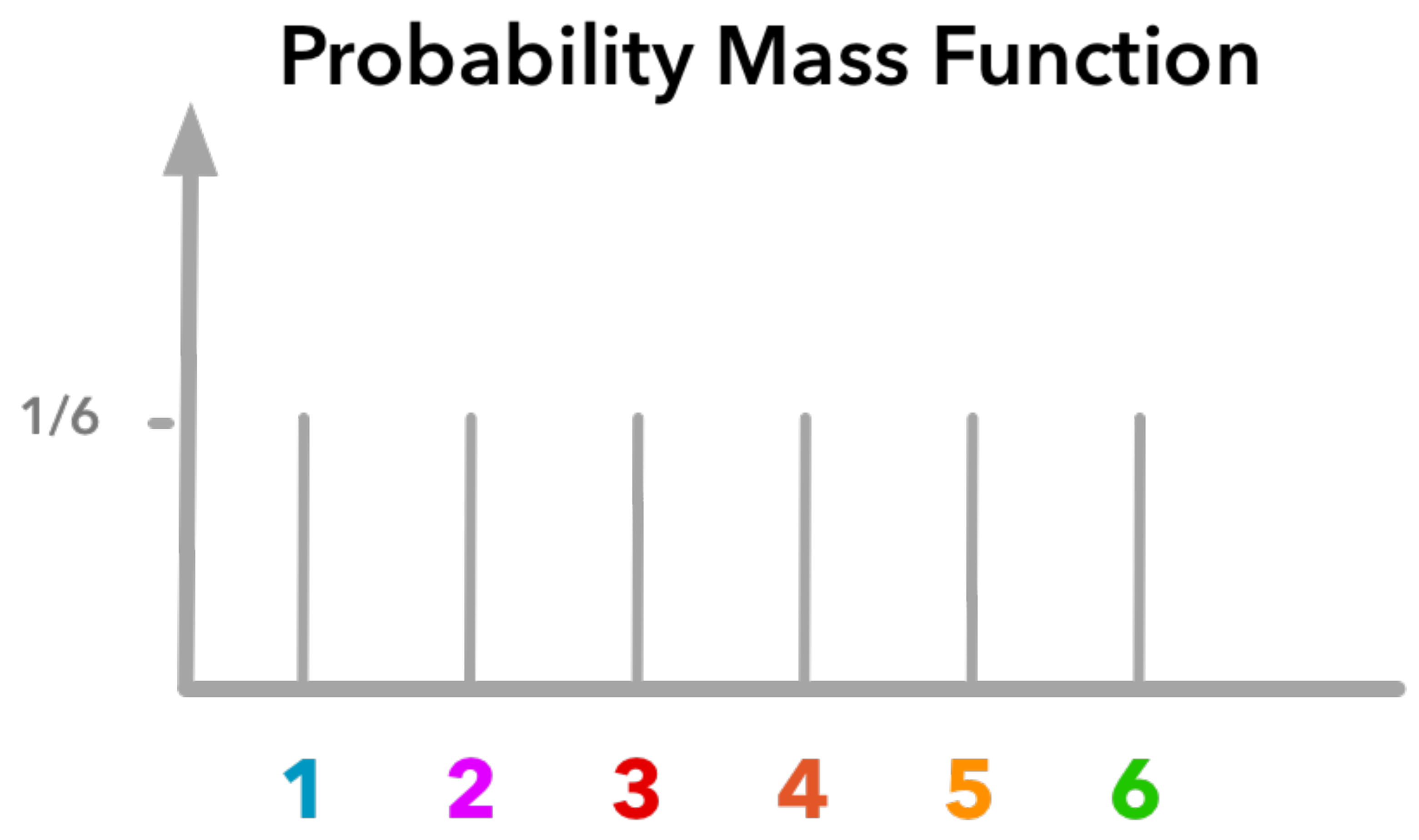


https://www.brainfuse.com/jsp/alc/resource.jsp?s=gre&c=37161&cc=108833

- Frequency:
  - How many times an **outcome** occurred
- Relative frequency:
  - What **percentage** of all outcomes does the outcome represent
  - Estimator for **probability**
- Histogram:
  - **x-axis**: outcomes
  - **y-axis**: frequencies or relative frequencies
    - relative frequencies will **sum to 1**
    - and then our histogram becomes a **Probability Mass Function**

Probability mass function

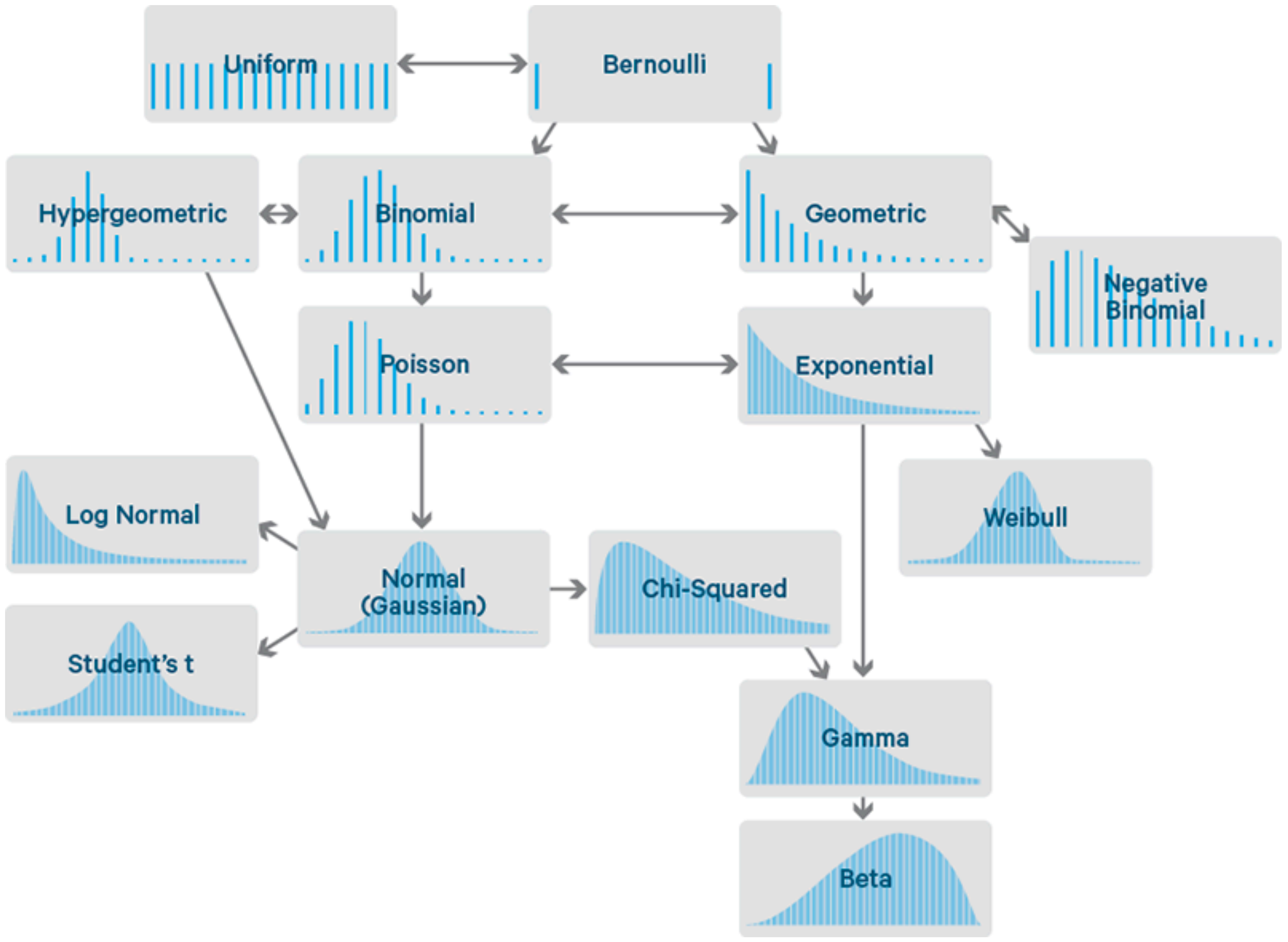
- the PMF is the **distribution** of probabilities for discreet RVs
  - given a possible value for a Random Variable
    - P(X = x)
  - returns the probability of that outcome
    - i.e. how are probabilities **distributed** between possible outsome values
  - e.g. for rolling a die:
    - P(X=x) = 1/6
    - X: {x=1,x=2,x=3,x=4,x=5,x=6}
    - can be visualized as a graph/histogram
  - the sum of all "bars" of PMF sums to 1



https://www.kdnuggets.com/2019/05/probability-mass-density-functions.html

Probability distributions
review

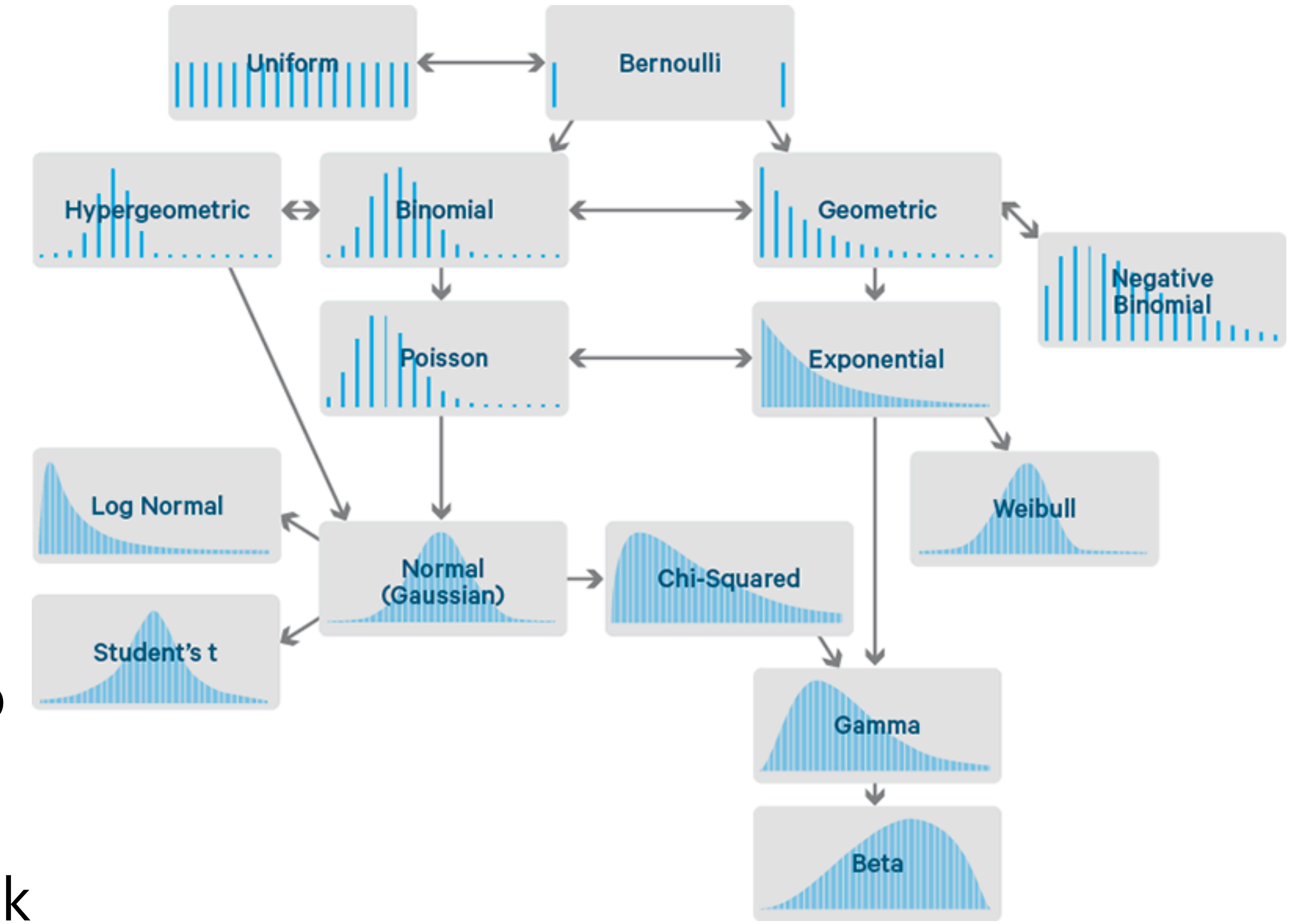
- Long time ago (18th century or earlier):
  - Mathematicians collecting/analyzing data noted
    - the same **shapes** kept reappiaring
- Distribution:
  - Represent how the outcome **relative** frequencies are distributed
  - as a function/formula (curve)
  - use the curve to predict future outcomes
- Distribution curves/shapes appropximate the truth
  - based on many empirical observations
- **Known**/defined functions have parameters
  - which can be estimated based on the observed data
  - which values for parameters make the observed data the likeliest?
- Not all distributions resemble **known** functions
  - the ones which are known were simply observed **more**, to eventually get names
- When approximating, we are restricted to a set of functions for which the area under the curve sums up to 1
  - Otherwise, would not be able to interpret the function as probability disribution



https://www.datasciencecentral.com/profiles/blogs/common-probability-distributions-the-data-scientist-s-crib-sheet

Probability distributions
informal summary

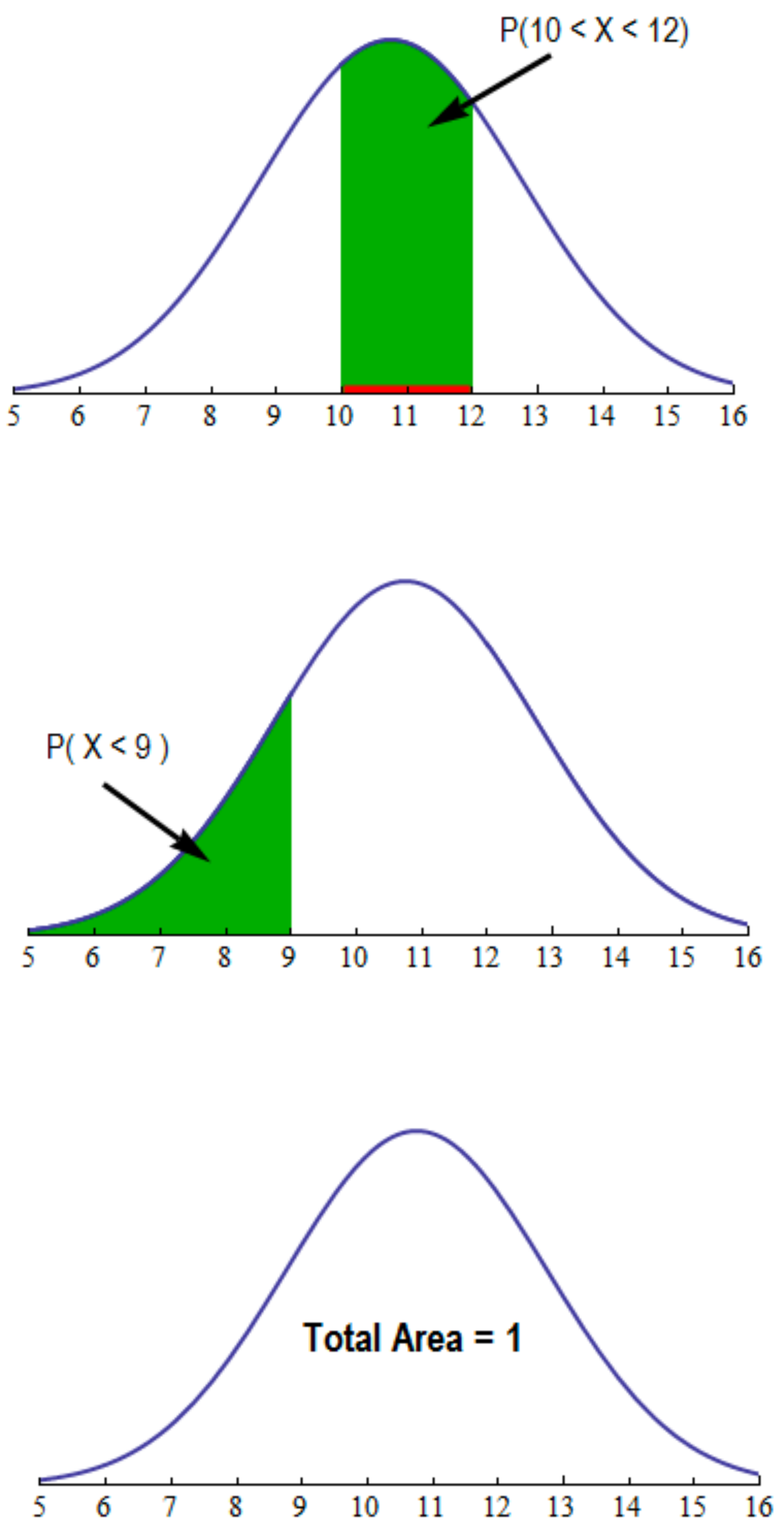
- We want to be able to **predict** a phenomenon
  - e.g. bus arrivals
- **How** do we know which probability function to use?
  - **Option 1:** Make tons of experiments, plot them, look the shape of **best-fitting** curve and see which known function it resembles the most, or invent a new name for the function
  - **Option 2:** Maybe "waiting time" is a common phenomenon and there is a well-known distribution already (i.e. somebody has already done Option 1)



https://www.datasciencecentral.com/profiles/blogs/common-probability-distributions-the-data-scientist-s-crib-sheet

Probability Density
Continuous Random Variables

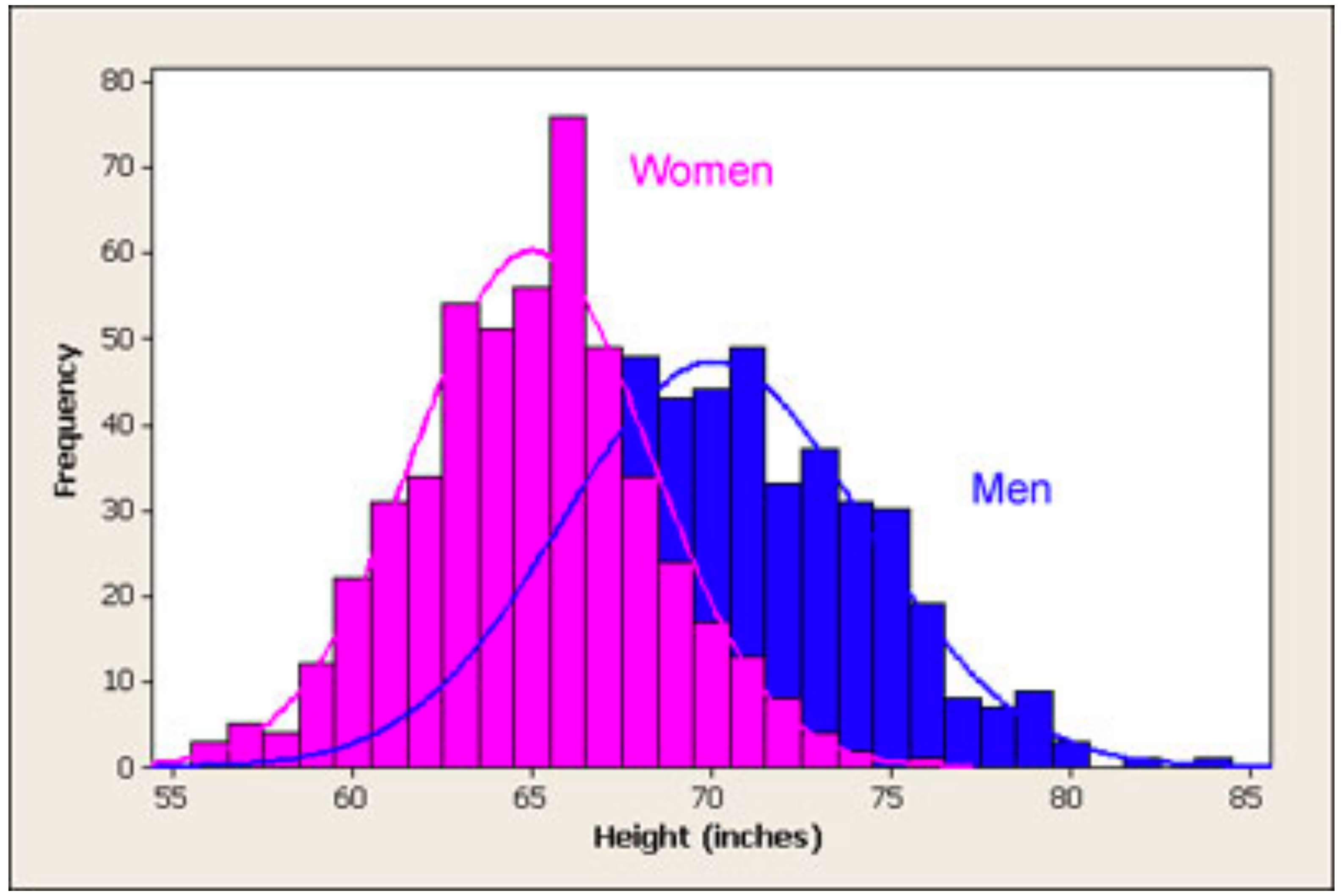
- Some random variables are continuous
  - set of values is a range, e.g. from 0 to 1, or from 0.99 to 99.99...
    - Age: 25 years, 10 months, 2 days, 5 hours, 4 seconds, 4 milliseconds, 8 nanoseconds, 99 picosends...**and so on**.
  - as opposed to discreet (coin toss, die roll...)
    - set of countable values: {H,T}; {1,2,3,4,5,6}
- Continuous random variables's distributions are defined by their probability **density** functions
  - The area under the curve sums to 1



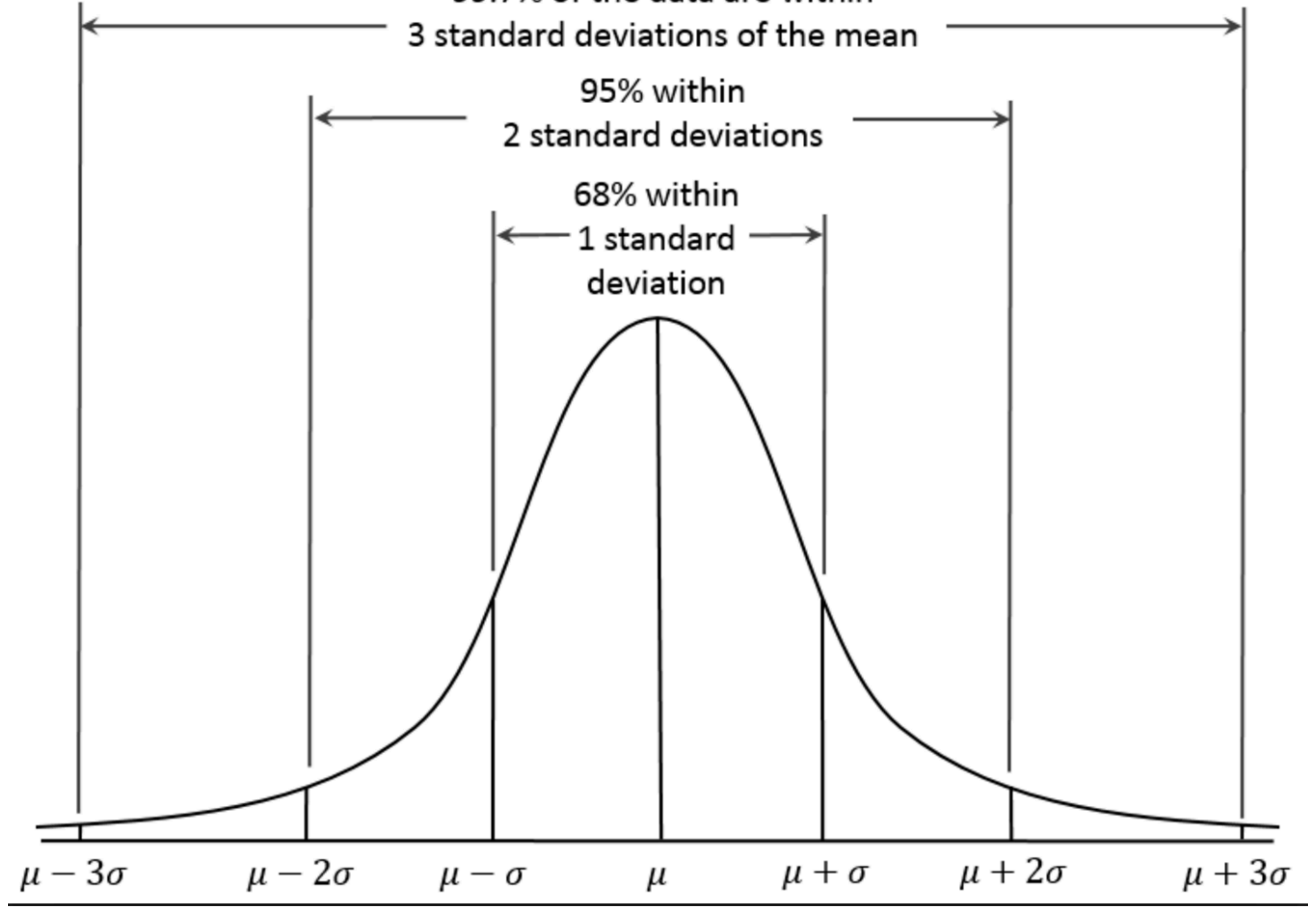
https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/continuous-probability-distribution-2-of-2/

Normal distribution
aka Gaussian

- The most famous probability distribution
- e.g. "people's heights are normally distributed"
- ...as for linguistics:
  - Sociolinguistic phenomena may be associated with normal distributions
  - but textual phenomena not so much!
  - words in text are not independent and are not continuous
- Still, knowing basic properties of the Gaussian is important
  - ...often, it is useful to assume a Gaussian even where there isn't any!



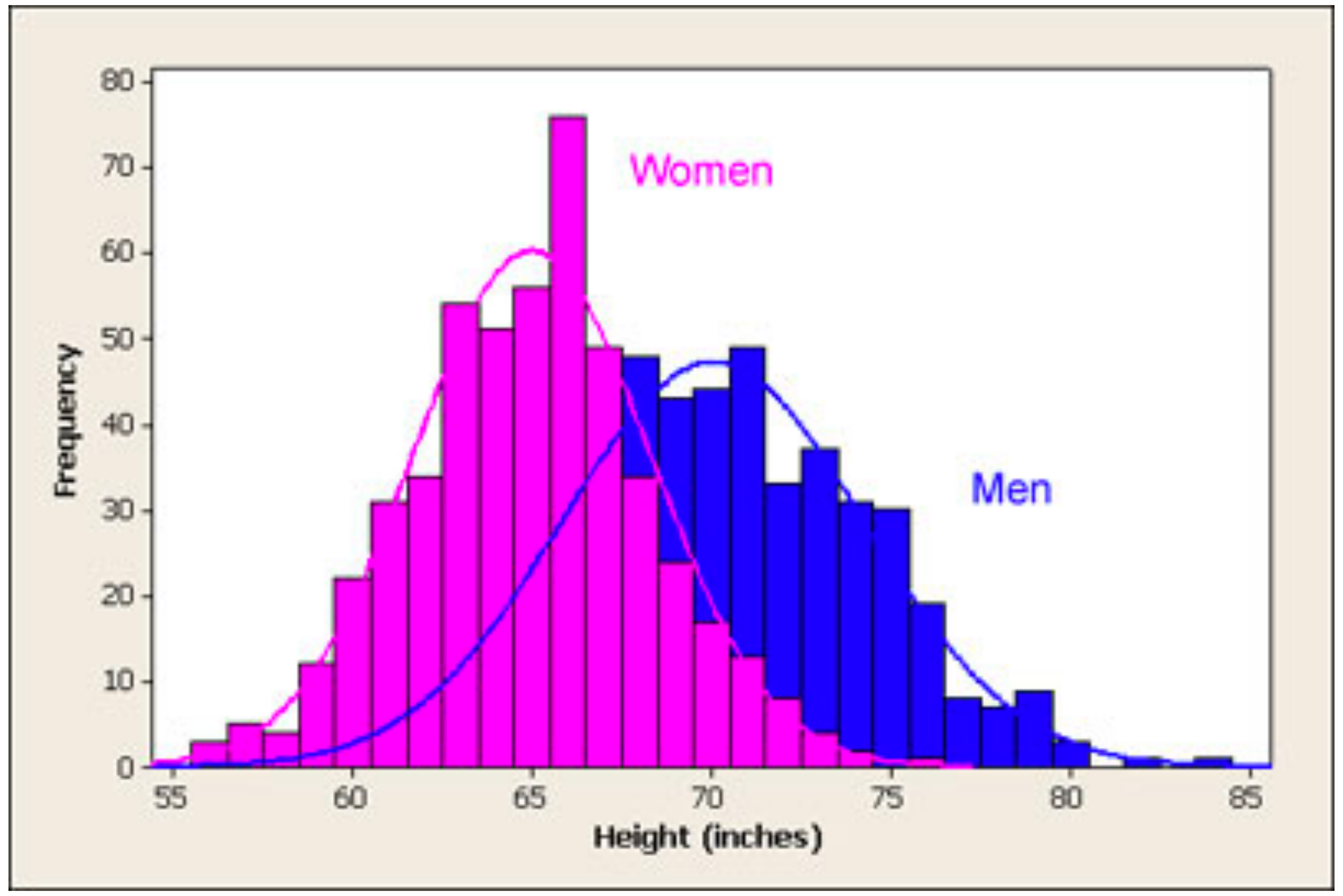
https://www.usablestats.com/lessons/normal



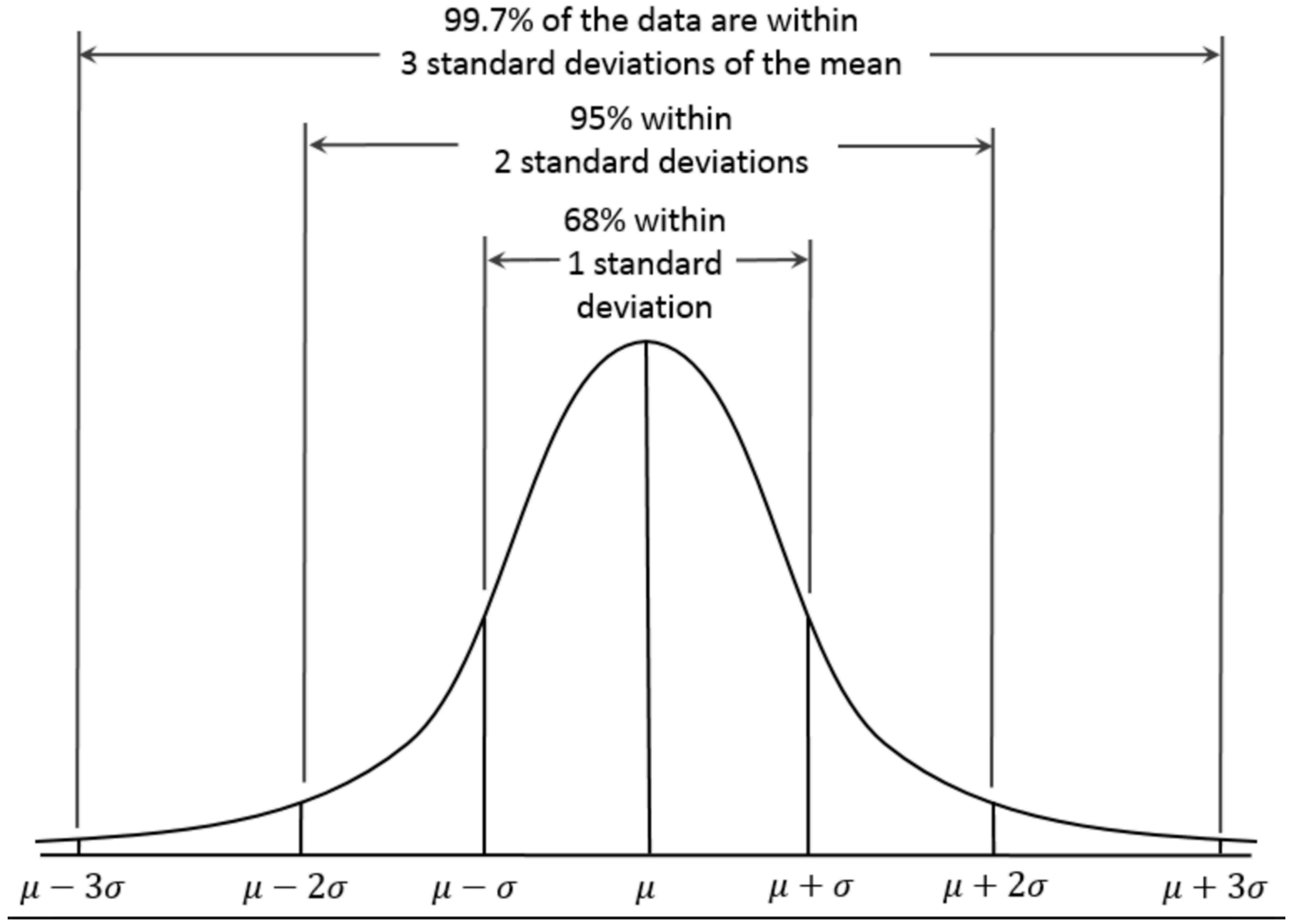
https://medium.com/@ar3441/the-central-limit-theorem-9ede4ebfafa5

Normal distribution
aka Gaussian

- Much of data out there is normally distributed
  - If data is normal, can use “parametric methods”
    - a range of powerful methods that assume a distr.
  - if not, should use other, less powerful methods!
    - except, can sometimes assume data is “normal enough” :)
- Is language data normally distributed?
  - depends on what kind, of course



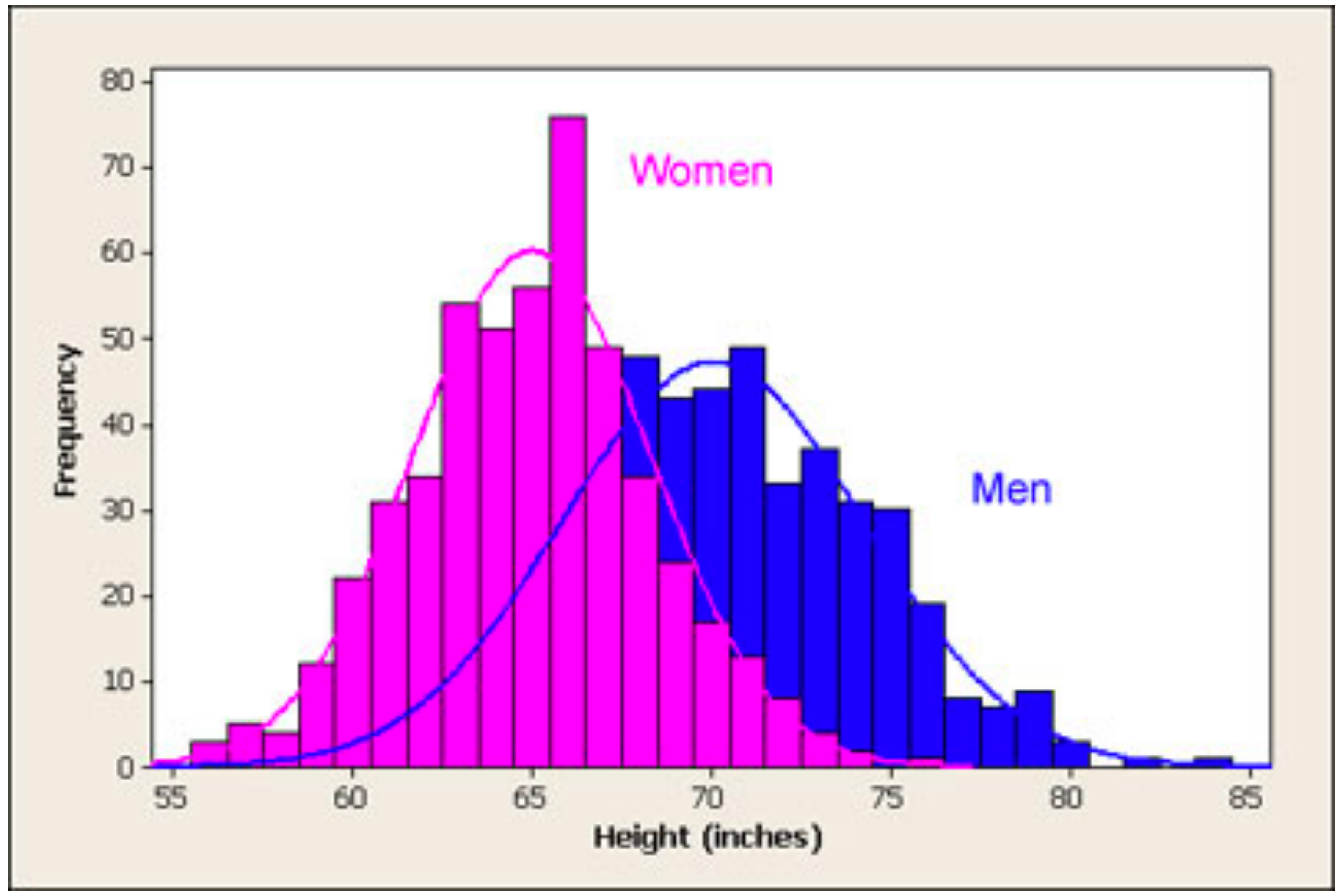
https://www.usablestats.com/lessons/normal



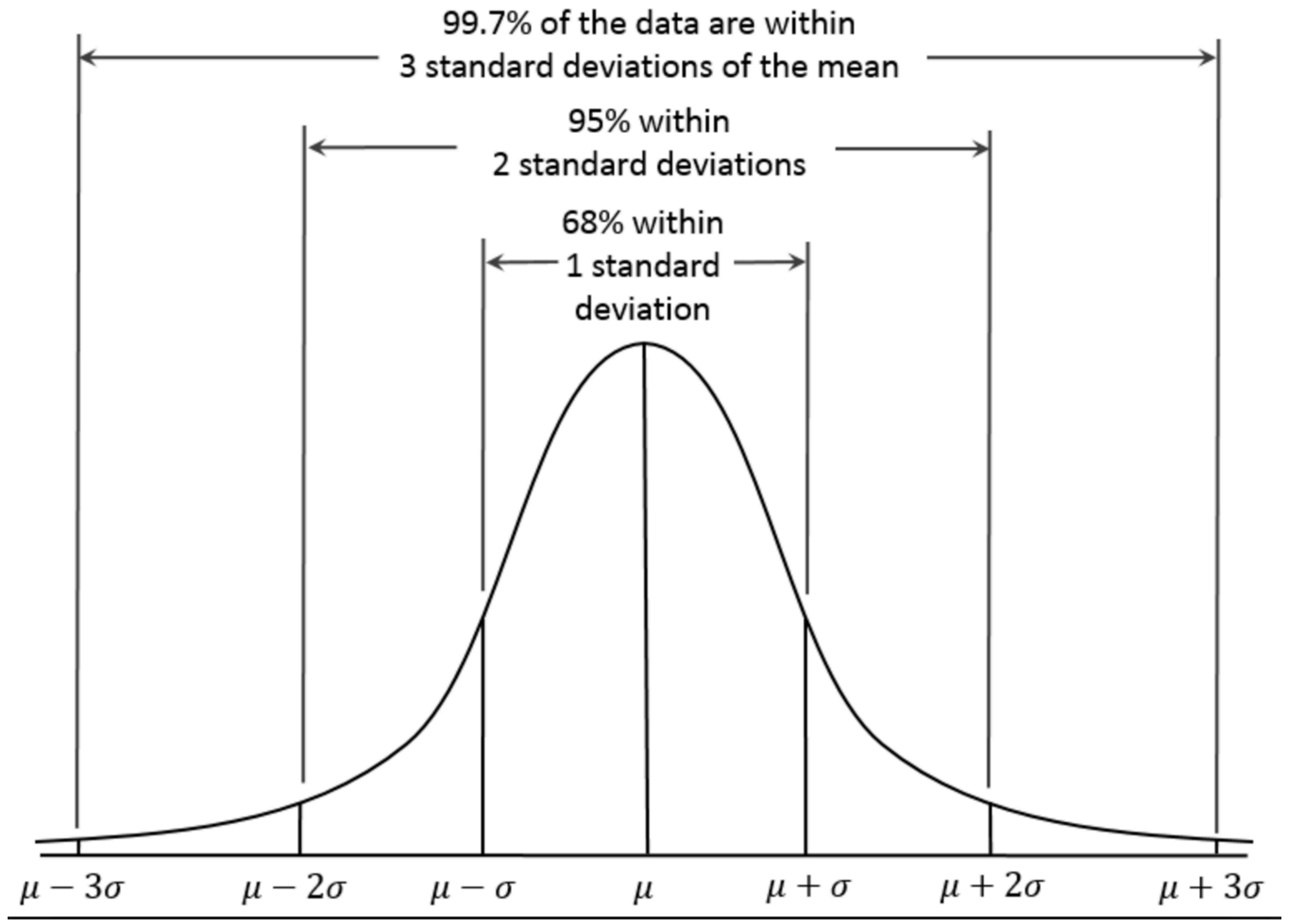
https://medium.com/@ar3441/the-central-limit-theorem-9ede4ebfafa5

Normal distribution
aka Gaussian

- aka the Bell-shaped curve
  - One of the most important distributions in the world :)
- Center: the average value (the mean)
  - e.g. the average height in a population
- Tails: the outliers
- Standard Deviation:
  - a value such that 2/3 of observations fall within 1 std. dev. from the mean
  - the smaller the std. dev. the better
    - the data then is less widely variable and easier to reason about
- What's the Y-axis here?
  - "frequency", as in how many people have that height
- What's up with the bars and the curve?
  - the bars are actual observations
  - the curve is a "fit"
- What would we use the curve for?
  - if standardized to range from 0 to 1, it's the probability distrubution!



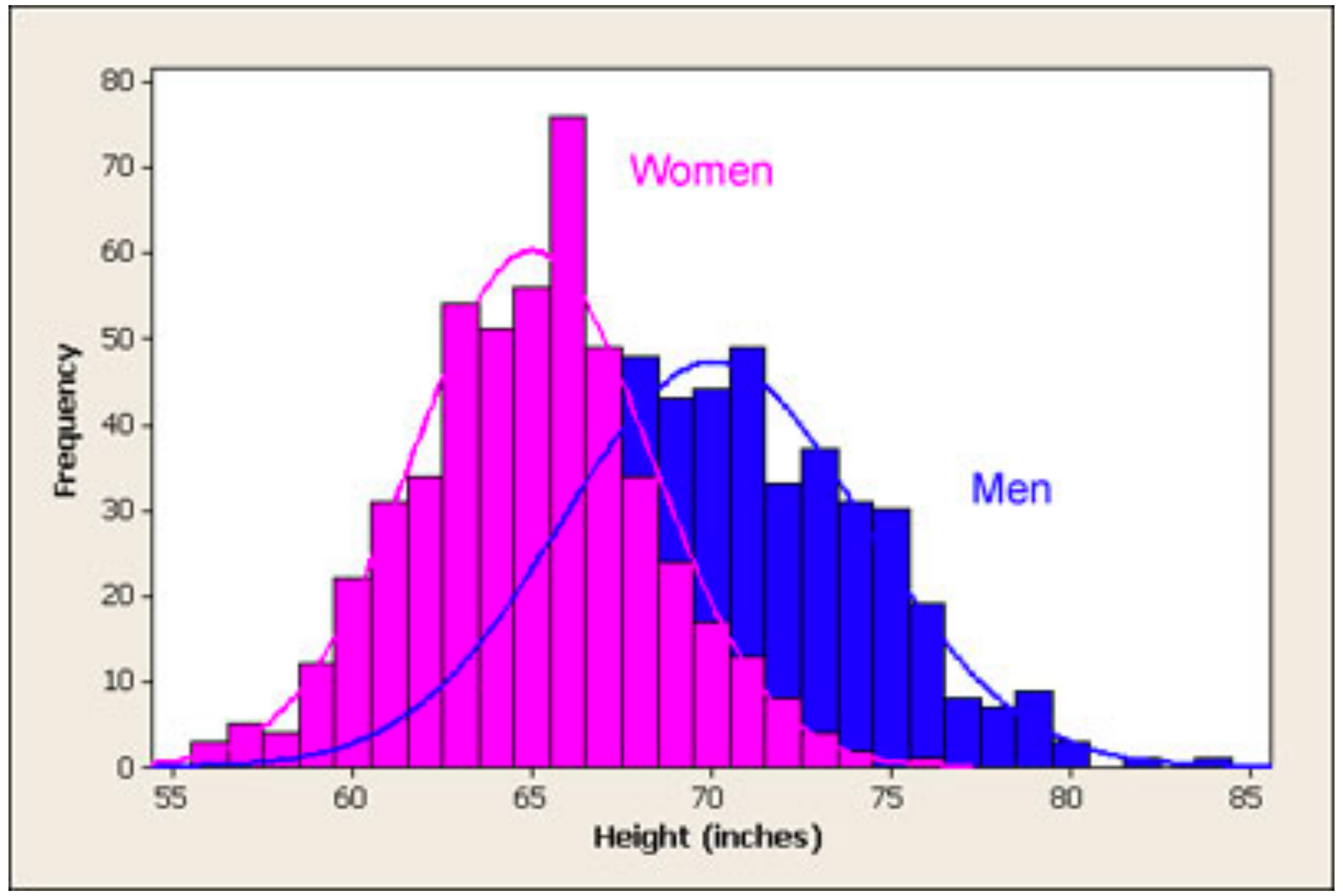
https://www.usablestats.com/lessons/normal



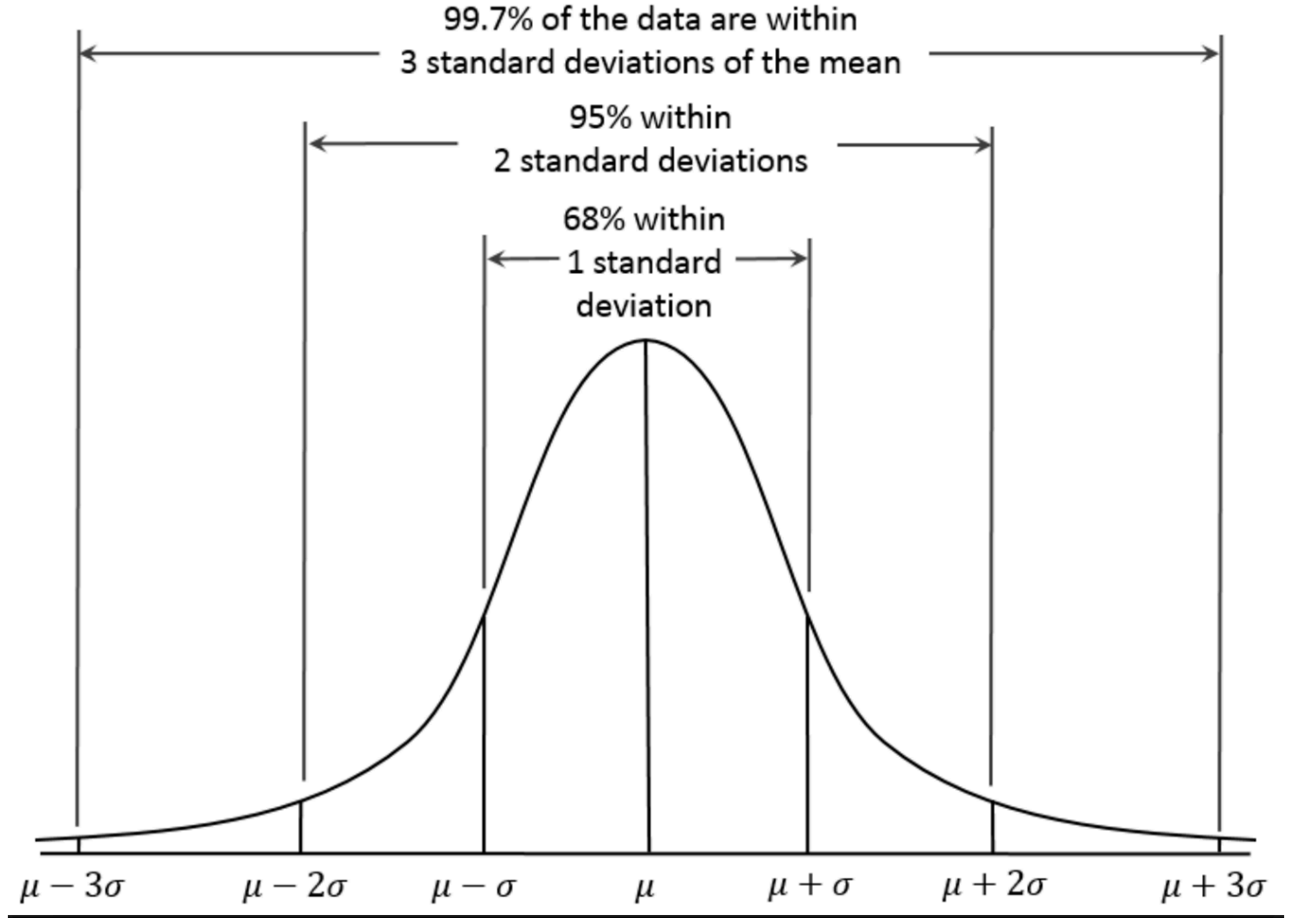
https://medium.com/@ar3441/the-central-limit-theorem-9ede4ebfafa5

Normal distribution
aka Gaussian

- How are Gaussians used?
  - the curves are defined by two parameters:
    - mean
    - standard deviation
  - Estimate the parameters from **samples**
    - will never know about the **actual** population!
  - given these parameters (i.e. the curve!):
    - can predict how many outcomes to expect in which range
      - ...for the entire population!
    - e.g. how many shoes to produce of which size



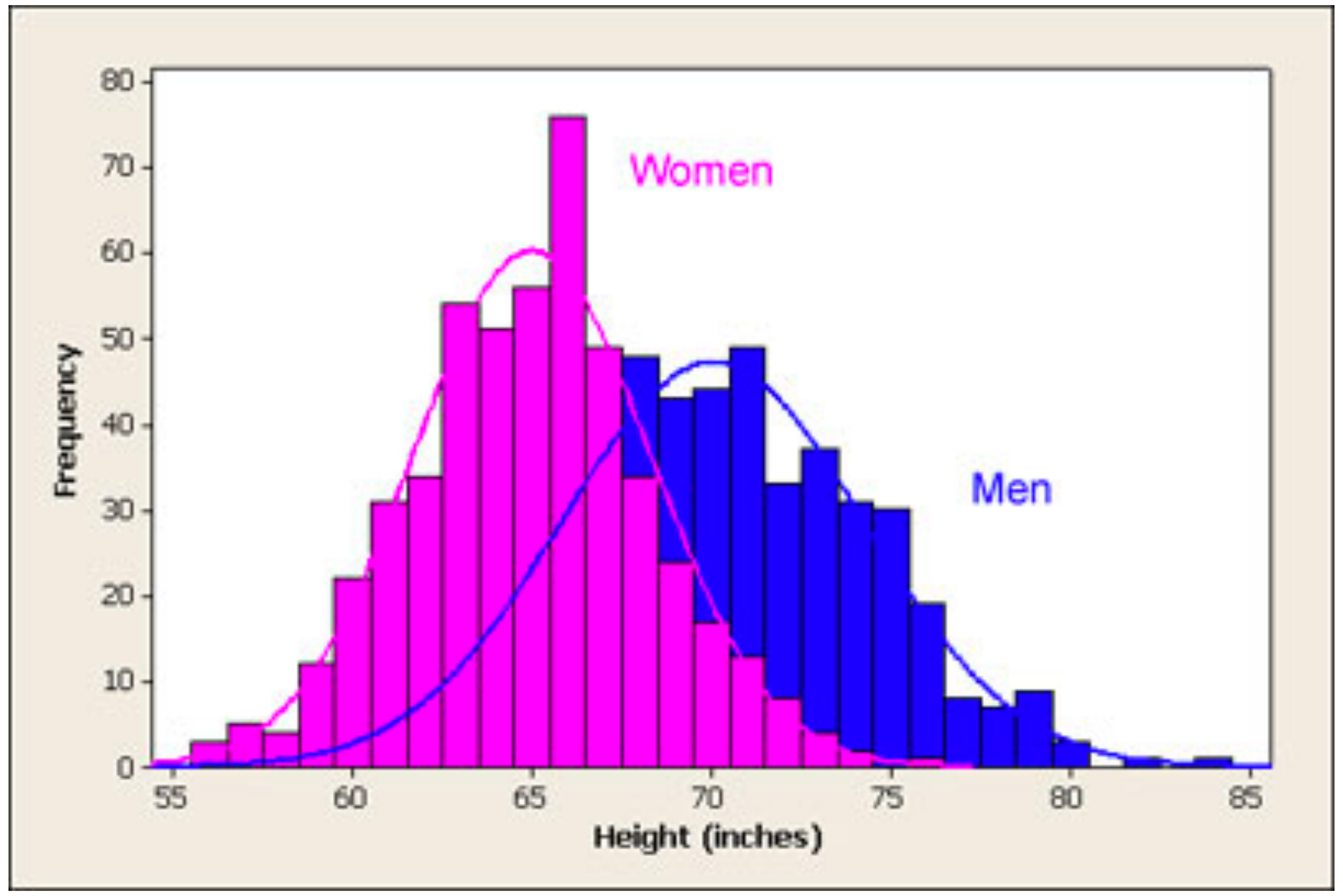
https://www.usablestats.com/lessons/normal



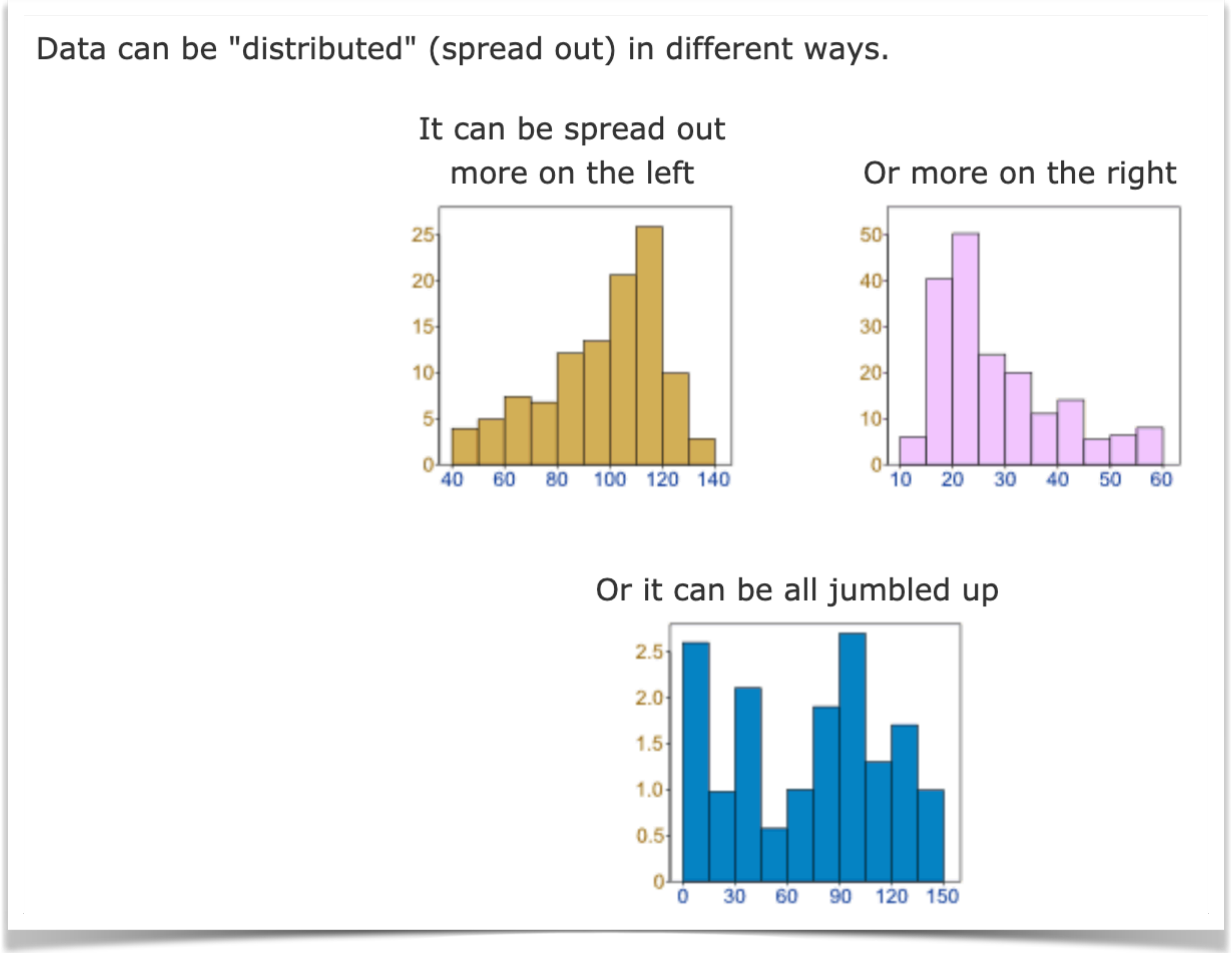
https://medium.com/@ar3441/the-central-limit-theorem-9ede4ebfafa5

Normal distribution
aka Gaussian

- When is mean (average) not enough?
  - Knowing the mean height is useful:
    - e.g. how many shoes of which size to produce
- What about knowing the mean wealth?
  - if the data is not actually normally distributed:
    - knowing the **mean** doesn't give you as much
    - because fitting a bell curve onto the data would be inaccurate
    - e.g. the median wealth in Seattle is different from the mean wealth
- What about language?
  - Sociolinguistic variables may be normally distributed
  - Syntactic phenomena?..
    - Maybe!



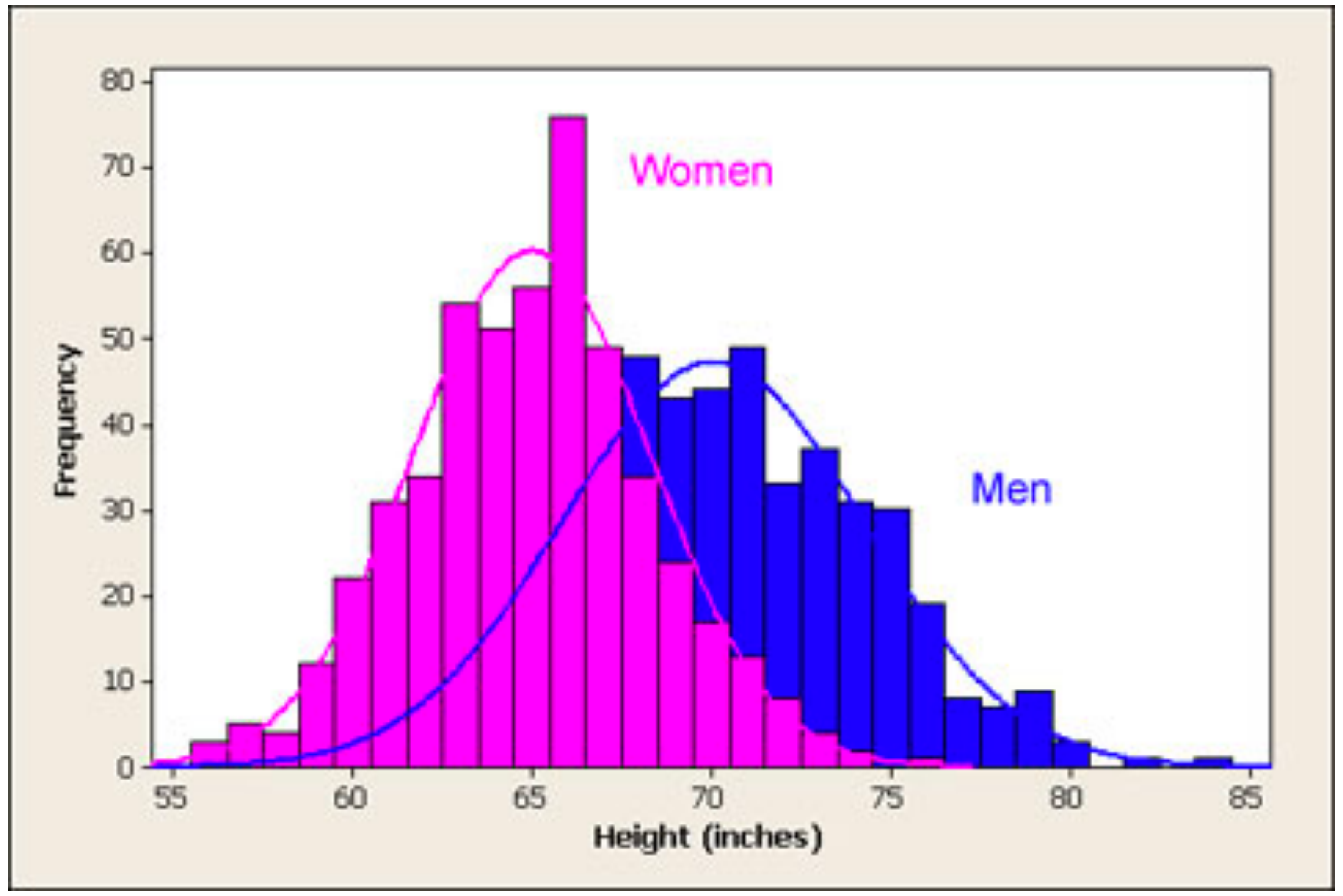
https://www.usablestats.com/lessons/normal



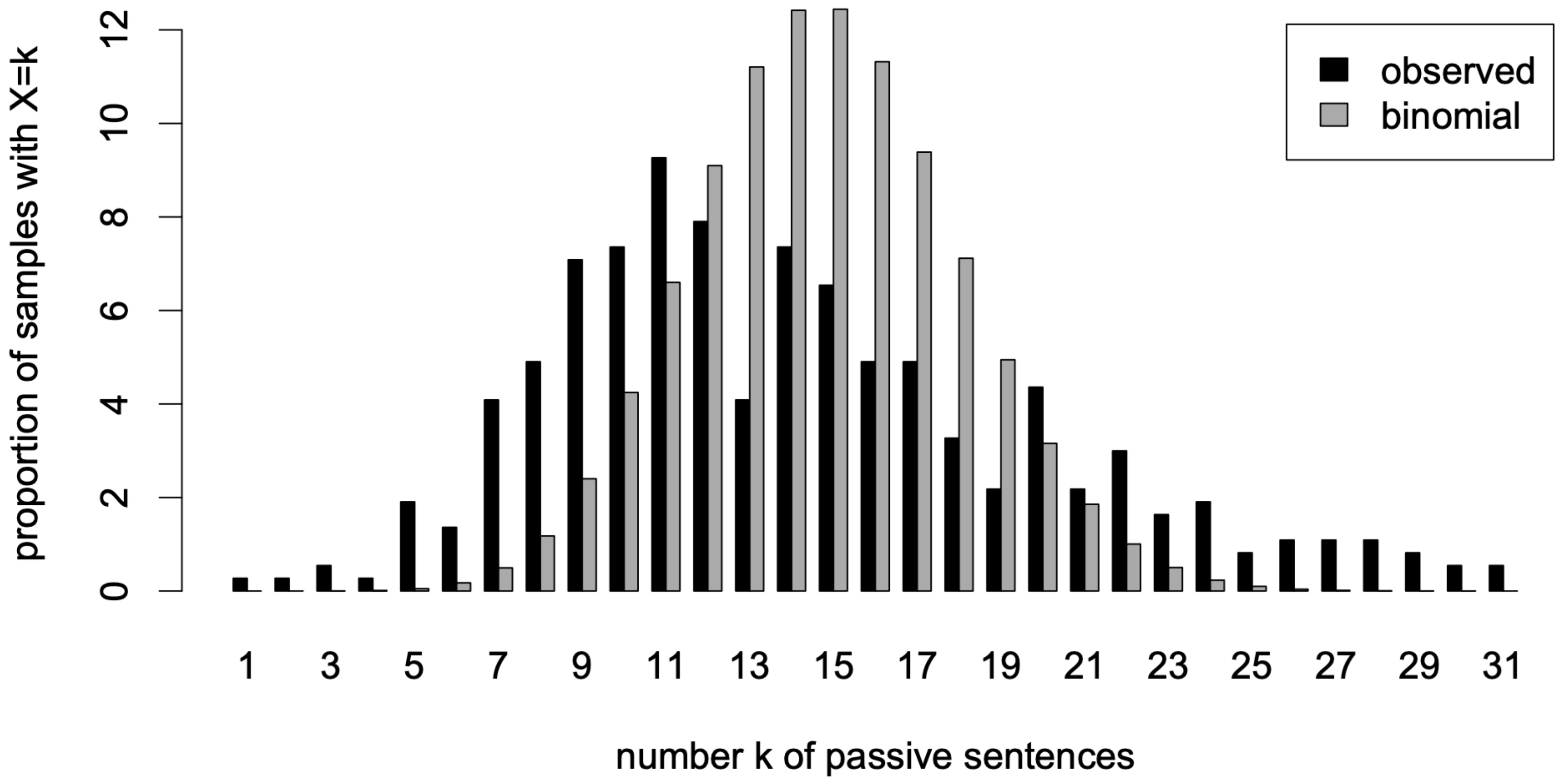
https://www.mathsisfun.com/data/standard-normal-distribution.html

Normal distribution
in language

- Sociolinguistic variables may be normally distributed
- Syntactic phenomena?..
  - Maybe!
  - E.g. passive sentences in samples from the Brown corpus
    - “Binomial” distribution:
      - Two possible outcomes (like coin toss)
      - passive/not passive
      - looks like Gaussian in shape, but is discreet



https://www.usablestats.com/lessons/normal

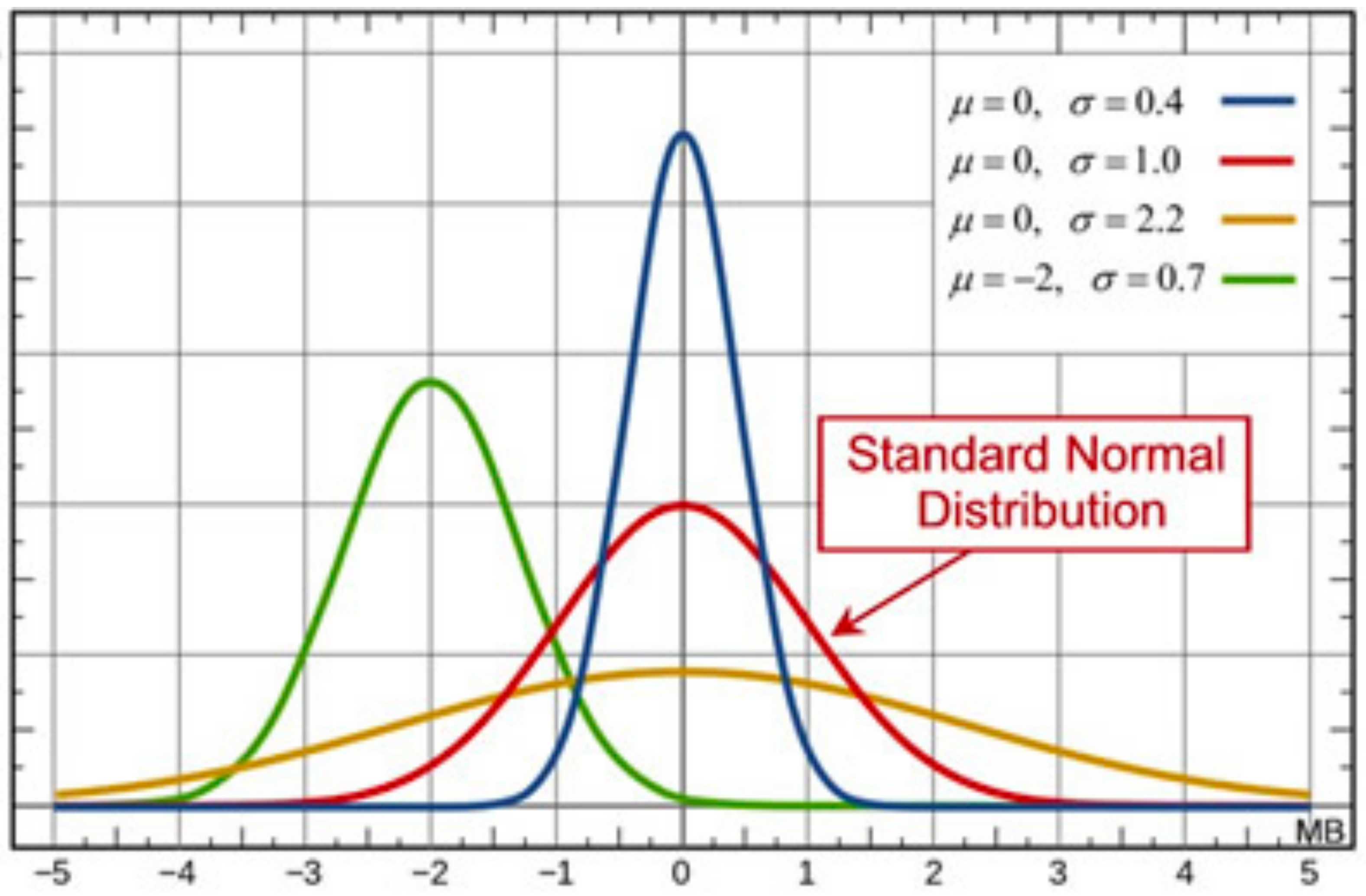


https://marcobaroni.org/publications/chapter38.pdf

Normal distribution

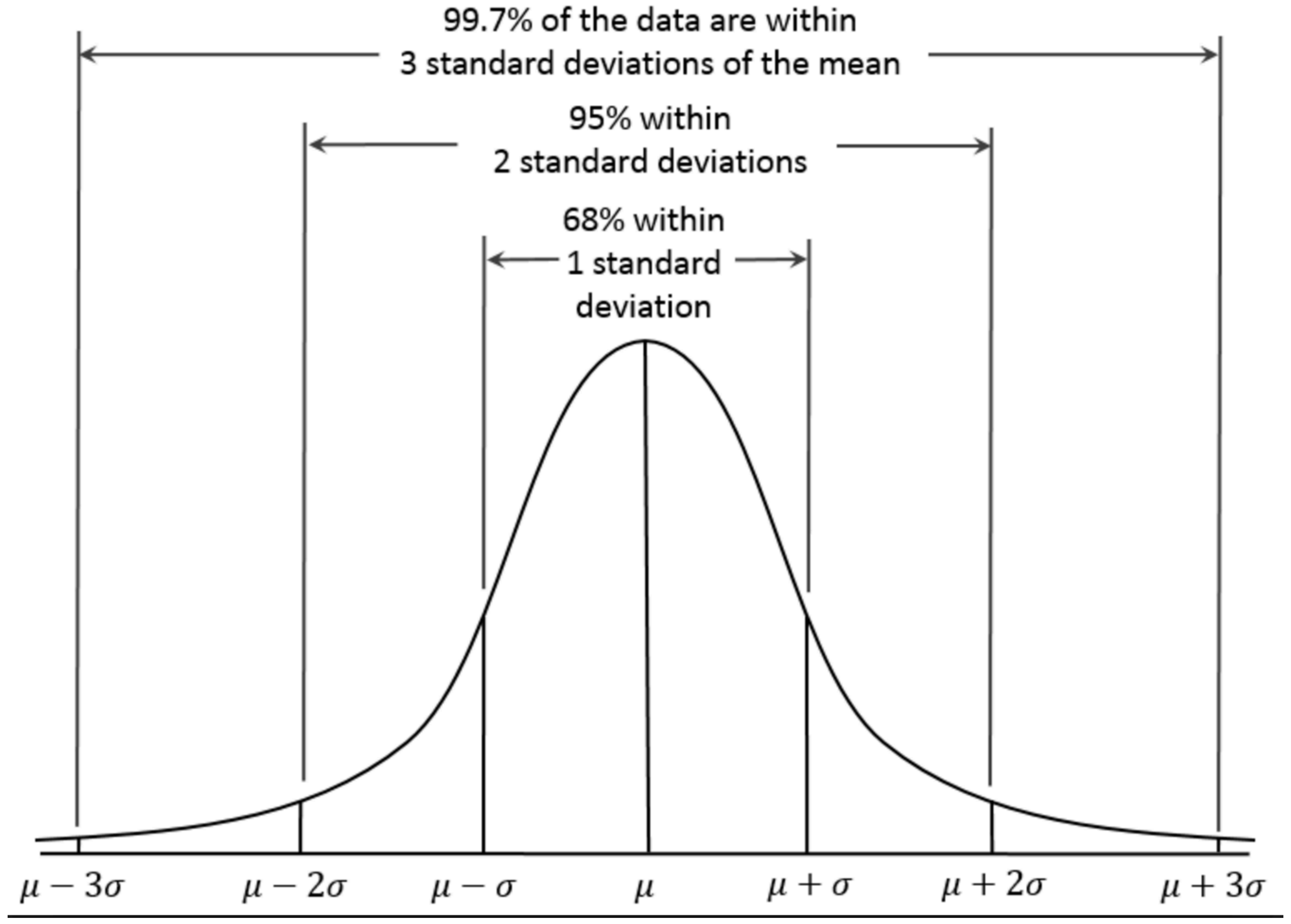
aka Gaussian

- The formula:
  - $$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
  - **parameters**:
    - μ: **mean**, average; where bell curve is centred
      - most often, you will get close to this x value when you sample
    - σ: **standard deviation**; how widely the curve is spread
    - both parameters can be **estimated** based on observation
      - the process is more involved than with coin toss
    - (π, e: known mathematical constants, numbers which turn out useful in the natural world)



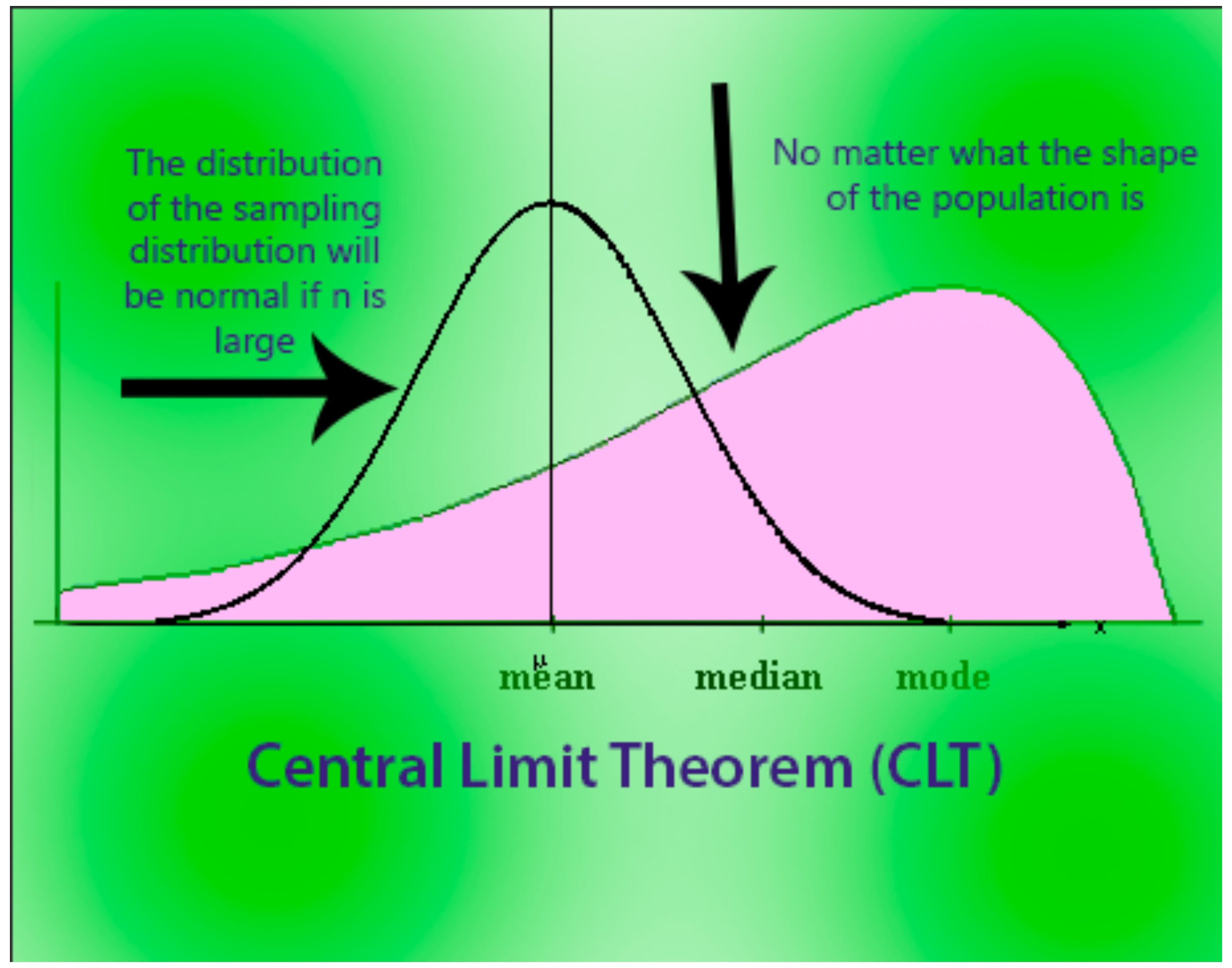
https://mathbitsnotebook.com/Algebra2/Statistics/STstandardNormalDistribution.html

https://www.usablestats.com/lessons/normal

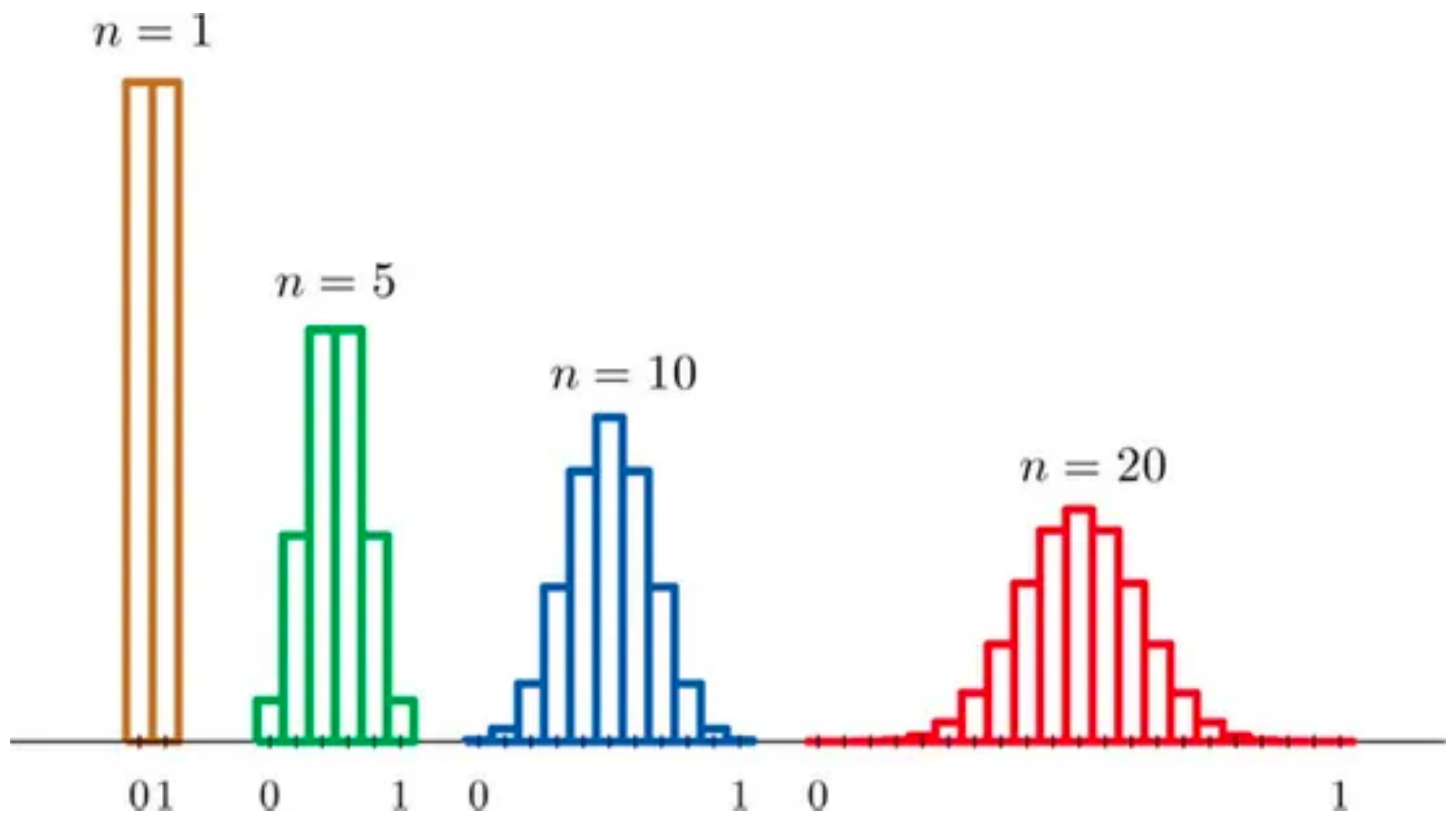


The Central Limit Theorem
for independent variables

- In real life:
  - impossible to observe the entire population
- Also in real life:
  - The distribution of sample **means** will be approximately Normal if the sample is large
  - ...and **equal** to the population mean...
  - ...**even** if the distribution of the entire population is **not** Normal
- So, take many samples, get the mean of each
  - the means will be distributed normally
  - ...now you can reason **pretty well** where the **actual** mean is



https://medium.com/@seema.singh/central-limit-theorem-simplified-46ddefeb13f3



https://www.simplypsychology.org/central-limit-theorem.html

Demo: The python numpy package
and
matplotlib.pyplot package

Activity:
Generating Gaussian data and visualizing it:

https://olzama.github.io/Ling471/assignments/ex-gauss.html

- **Goals:**
  - Strengthen built-in method calls
    - Passing the right arguments in the right order
    - Storing return values in variables
  - Overcome the fear of Greek letters, fractions, and exponents
    - …to some extent
  - Practice writing code which doesn't really work until you finish all of it (you need to build a bigger picture of what's happening in your head)

Lecture survey:
in the chat