Rensselaer

why not change the world?®

Stats Review, Distributions, Hypothesis Testing

Ahmed Eleish

Data Analytics

ITWS-4600/ITWS-6600/CSCI-4600 MGMT-4600/6600/BCBP 4600

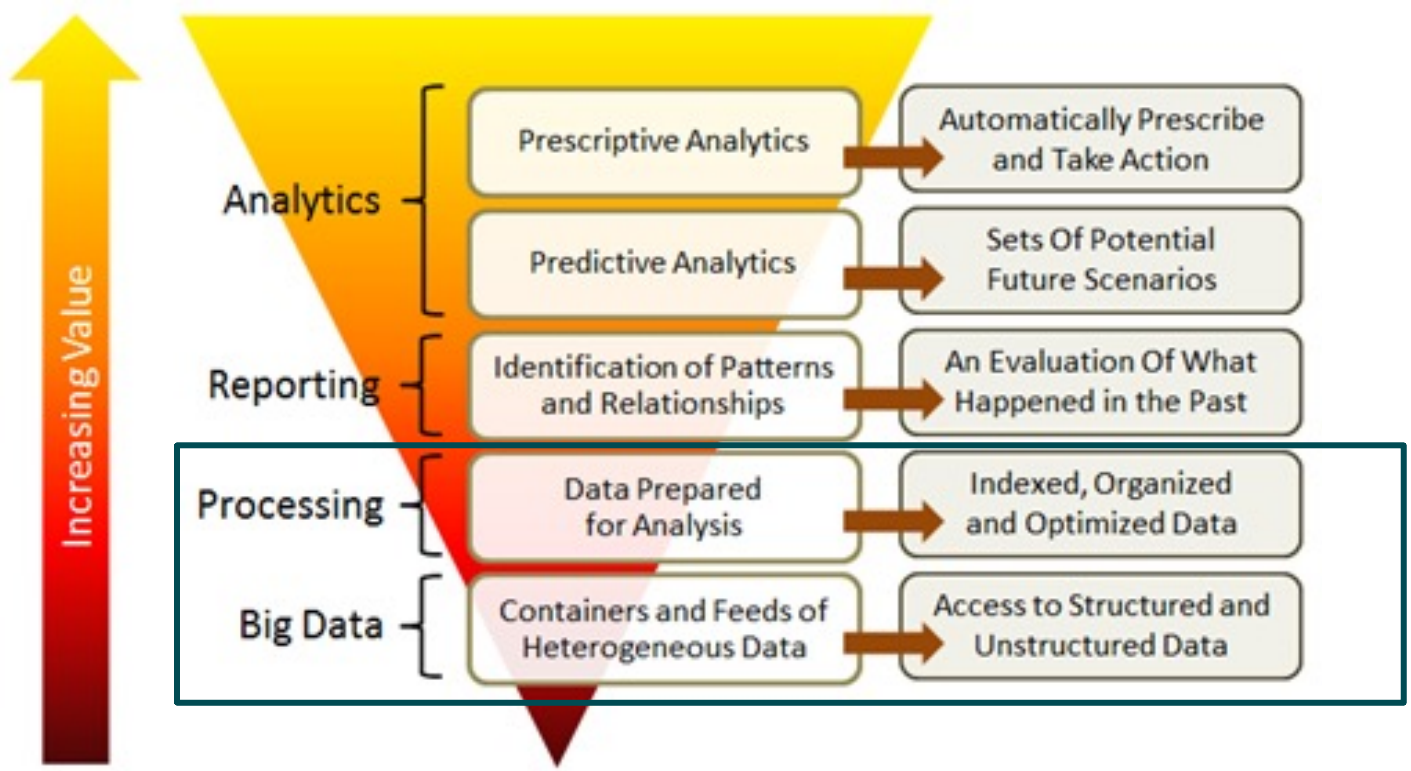
Group 1 Module 3, September 9th, 2025

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

- Stats review cont'd
- Exploring
  - Distributions
  - Visualization
- Testing and evaluating the results (beginning)

Lower layers in the Analytics Stack


Increasing Value
Analytics
Prescriptive Analytics
Automatically Prescribe and Take Action
Predictive Analytics
Sets Of Potential Future Scenarios
Reporting
Identification of Patterns and Relationships
An Evaluation Of What Happened in the Past
Processing
Data Prepared for Analysis
Indexed, Organized and Optimized Data
Big Data
Containers and Feeds of Heterogeneous Data
Access to Structured and Unstructured Data

Stats review – cont’d

Definitions/ topics

- Statistic
- Statistics
- Population and Samples
- Sampling
- Distributions and parameters
- Central Tendencies
- Frequency

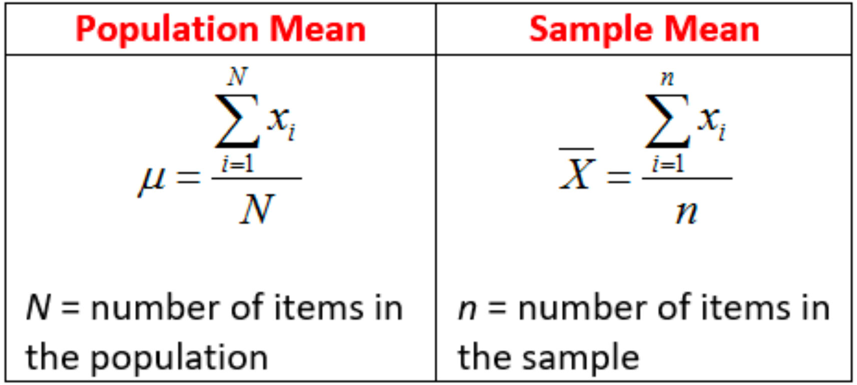
Previous class

- Frequency distributions
- Probability
- Hypothesis (null and alternate)
- Significance tests
- P-value

Today’s class

Measure of Central Tendency

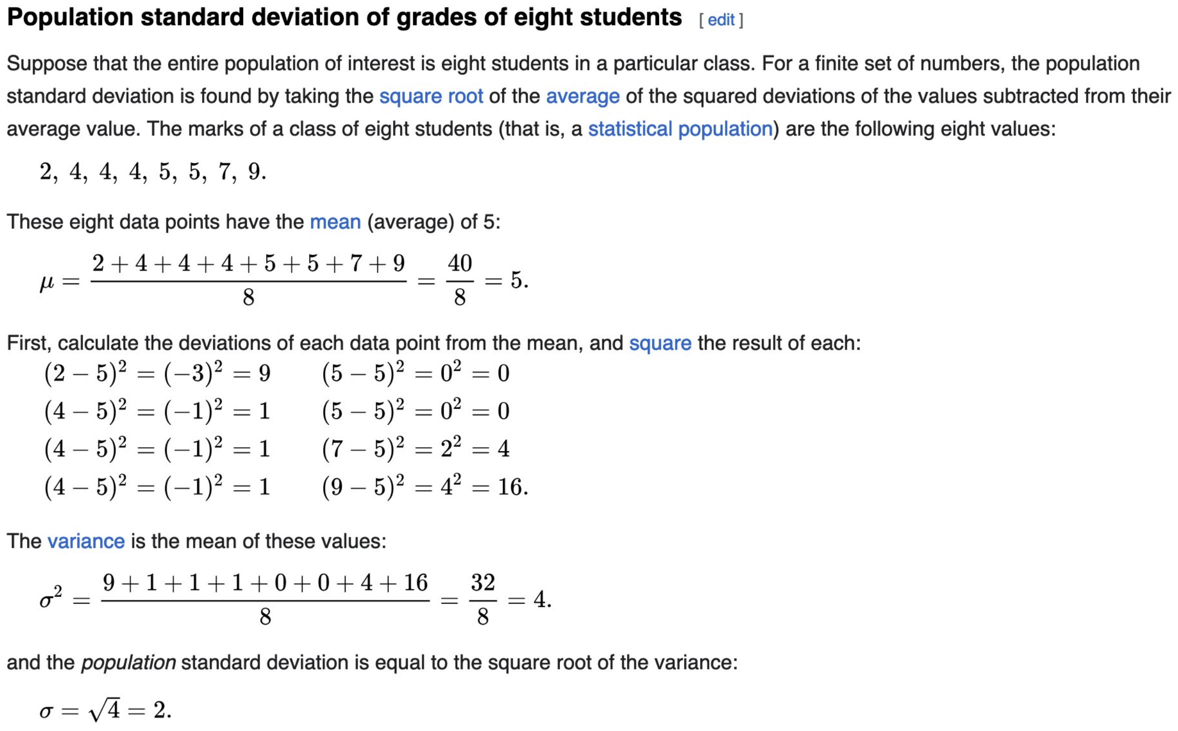
- Mean: The most commonly used measure of central tendency, commonly referred to as "Average", sensitive to extreme values (sensitive to outliers)
  - Population Mean
  - Sample Mean



Population Mean	Sample Mean
$\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	*n* = number of items in the sample

Image Resource: https://www.onlinemathlearning.com/population-mean.html Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Standard Deviation



Population standard deviation of grades of eight students [edit]

Suppose that the entire population of interest is eight students in a particular class. For a finite set of numbers, the population standard deviation is found by taking the square root of the average of the squared deviations of the values subtracted from their average value. The marks of a class of eight students (that is, a statistical population) are the following eight values:

$$2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9.$$

These eight data points have the mean (average) of 5:

$$\mu = \frac{2+4+4+4+5+5+7+9}{8} = \frac{40}{8} = 5.$$

First, calculate the deviations of each data point from the mean, and square the result of each:

$$\begin{array}{ll} (2-5)^2 = (-3)^2 = 9 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (7-5)^2 = 2^2 = 4 \\ (4-5)^2 = (-1)^2 = 1 & (9-5)^2 = 4^2 = 16. \end{array}$$

The variance is the mean of these values:

$$\sigma^2 = \frac{9+1+1+1+0+0+4+16}{8} = \frac{32}{8} = 4.$$

and the *population* standard deviation is equal to the square root of the variance:

$$\sigma = \sqrt{4} = 2.$$

https://en.wikipedia.org/wiki/Standard_deviation

Sample vs. Population

- Population
  - *All possible data points*
  - *May be of finite size (N) or infinite*
  - *Greek letters for parameters (μ, σ)*
  - *Parameters are estimated*

- *Sample*
  - *Finite subset of the population*
  - *Of finite size (n)*
  - *Latin letters for statistics (m, s)*
  - *Statistics are computed*

Sample vs. Population

- If sample *is* population, then $\mu = m$
- Realistically while they are not equal, m is a good estimator for μ

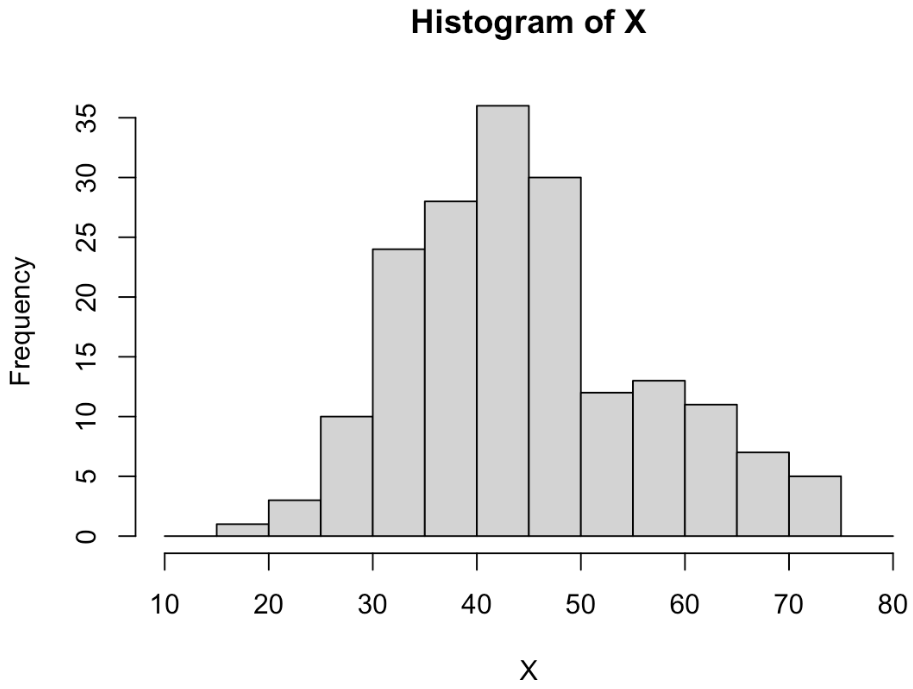
Law of large numbers

- *“the average of the results obtained from a large number of independent random samples converges to the true value, if it exists”*
- *“given a sample of independent and identically distributed values, the sample mean converges to the true mean."*

For this course

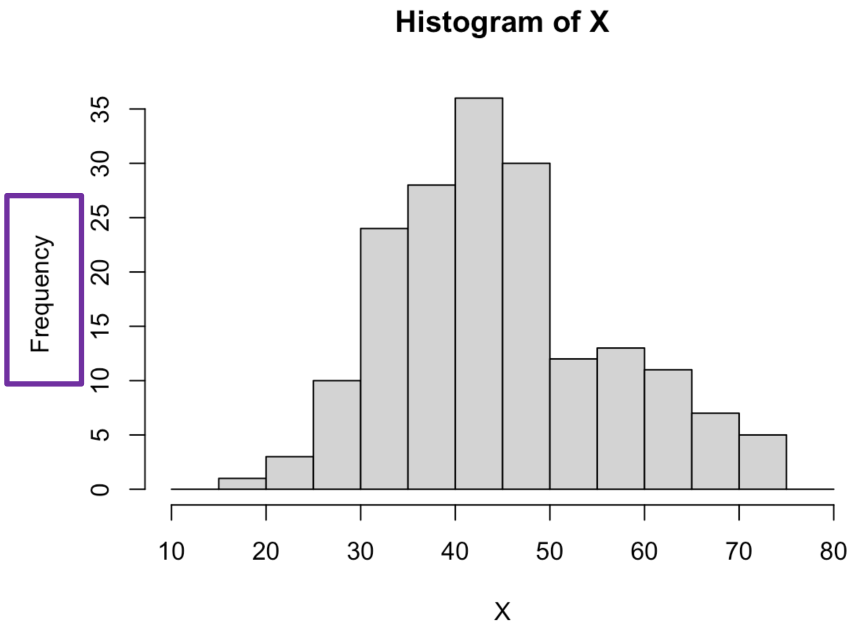
- Consider the observations in the given/acquired dataset the **entire** population.

Grouped Frequency Distribution aka binning

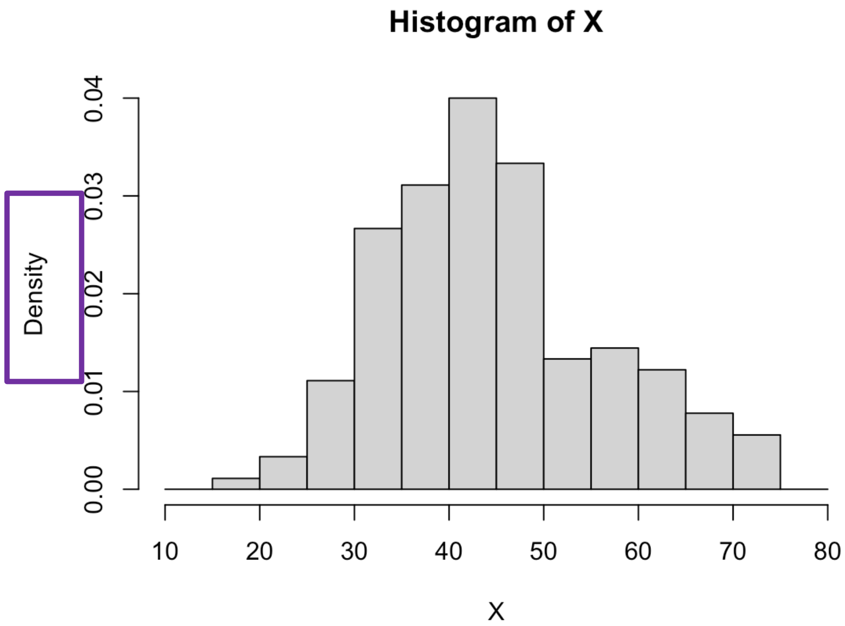


Bin size = 5

Frequency vs. Density

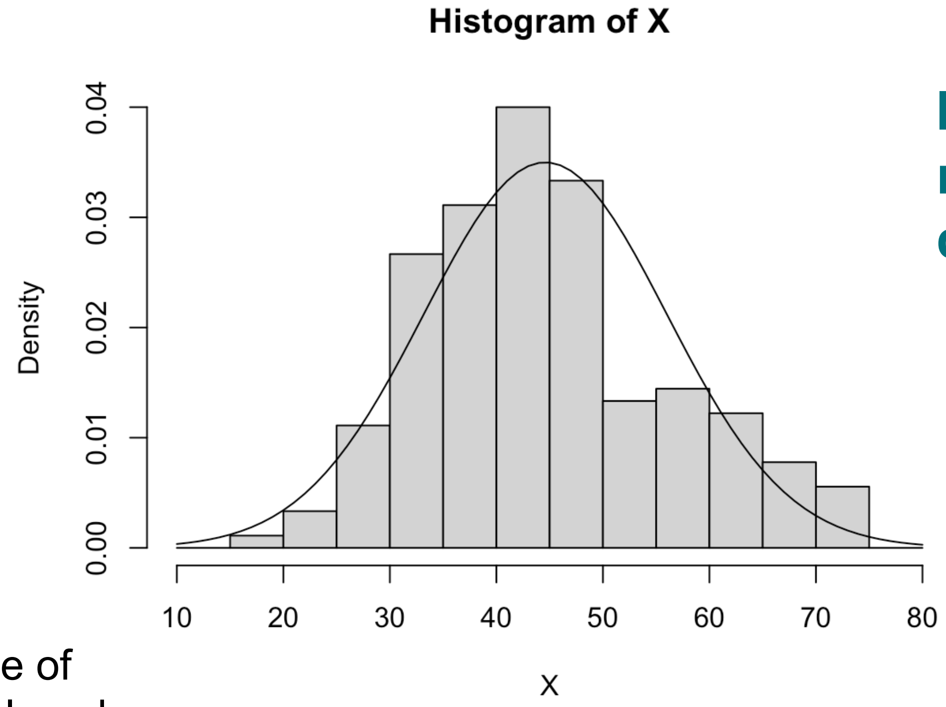


- 36 observations where 40 < x < 45
- 36/180 (total) = 0.2 or 20%



- density at bar = 0.04 where 40 < x < 45
- area of bar = 0.04 * 5 (width of bar) = 0.2 or 20%

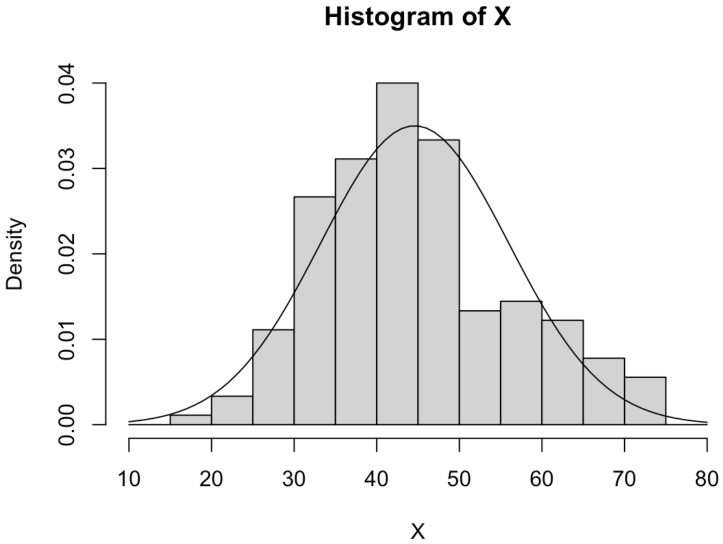
Empirical vs. Theoretical



Is variable X normally distributed??

- Probability density curve of normal distribution overlayed

Empirical vs. Theoretical



Is variable X normally distributed??

No!!

Shapiro-Wilk normality test

data: X
W = 0.96964, p-value = $\boxed{0.0005824}$

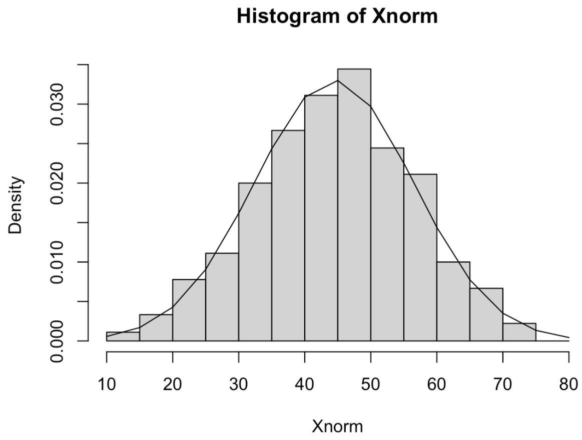
Less than 0.05!

Empirical vs. Theoretical

Let’s generate 180 numbers drawn at random from a normal distribution with the same mean and sd as X

Is variable X normally distributed??

Yes!!



Shapiro-Wilk normality test

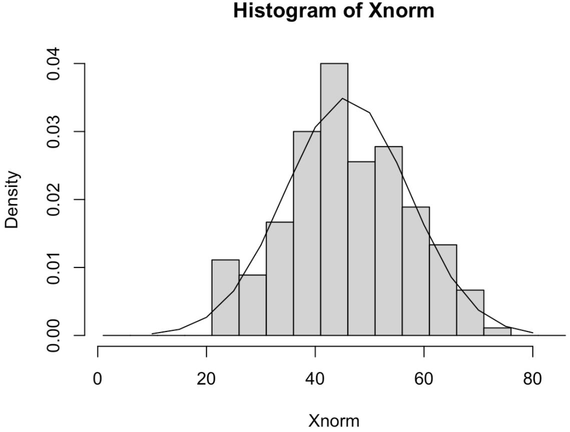
data: Xnorm
W = 0.99517, p-value = 0.8308

greater than 0.05!

Empirical vs. Theoretical

Let's generate 180 **NEW** numbers drawn at random from a normal distribution with the same mean and sd as X

Is variable X normally distributed??



Yes!!

```
Shapiro-Wilk normality test

data:  Xnorm
W = 0.98952, p-value = 0.2079
```

greater than 0.05!

Frequencies v. Probabilities

- Actual rate of occurrence in a sample or population – frequency
- Expected or estimate likelihood of a value or outcome – probability

- Coin toss – two outcomes (binomial) p= 0.5 (of “heads”)
- Major of study
- School year
- Which US State you live in

Distributions

- Shape
- Parameter(s)
  - Mean
  - Standard deviation
  - Skewness

…

- Which one fits?

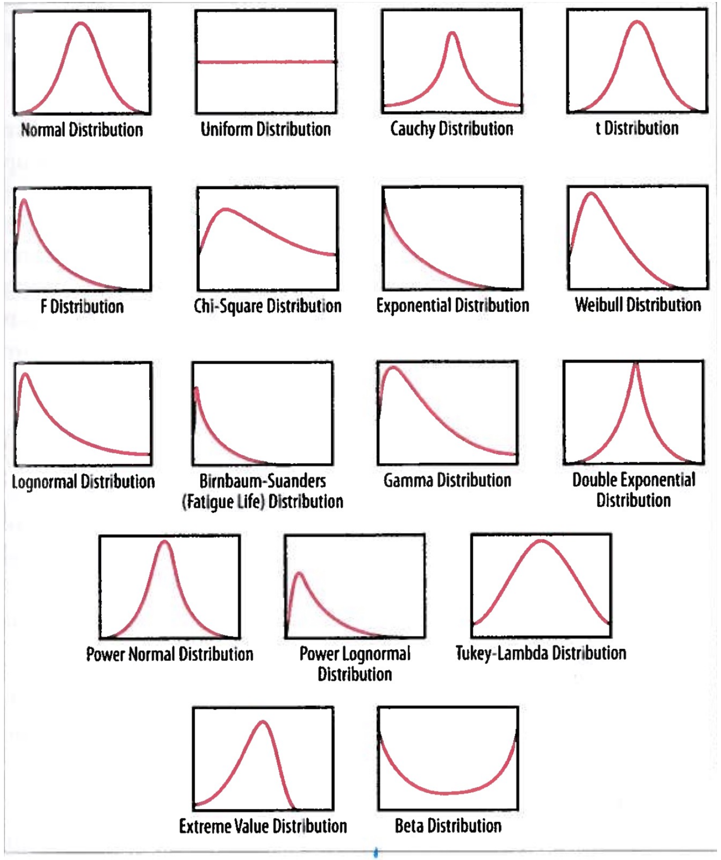


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

Binomial Distribution

- Describes the outcome of coin toss experiments.

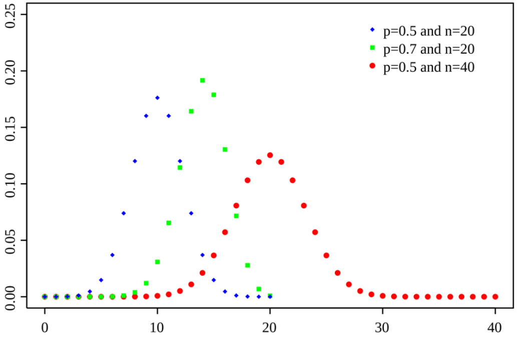
- Binomial distributions are discrete and are defined by 2 parameters: *p* (probability of success) and *n* (number of trials)

- Probability Mass Function

$$f(k,n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, ..., n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$



Probability Mass Function of Binomial Distribution

e.g. Probability of obtaining exactly 1 head in 2 coin tosses:

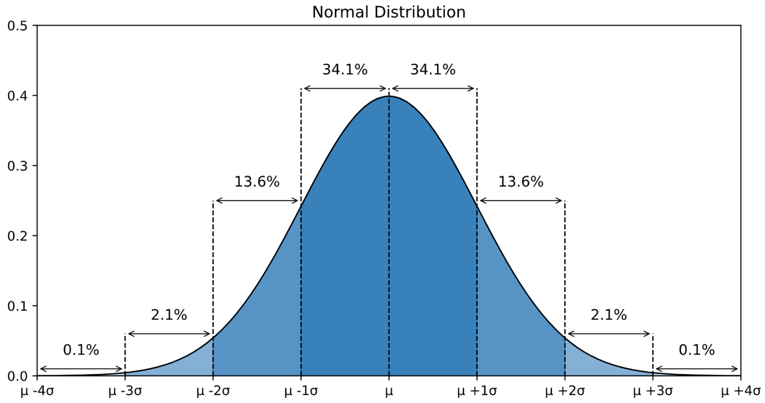
$$C_1^2 \ * 0.5^1 * 0.5^{2-1} = 0.5$$

https://en.wikipedia.org/wiki/Binomial_distribution

Image credit: Tayste

Normal Distribution

- The normal distribution implies tight bounds on the probability of lying far from the mean. 68% of the values must lie within one sigma (standard deviation) of the mean, and 95% within two times the sigma (standard deviation) and 99.7% lie within the three the sigma (standard deviation)



- Roughly 68.3% of the data is within 1 standard deviation of the average (from μ-1σ to μ+1σ)
- Roughly 95.5% of the data is within 2 standard deviations of the average (from μ-2σ to μ+2σ)
- Roughly 99.7% of the data is within 3 standard deviations of the average (from μ-3σ to μ+3σ)

Image Credit: W3C school:
https://www.w3schools.com/statistics/statistics_normal_distribution.php

Normal Distribution

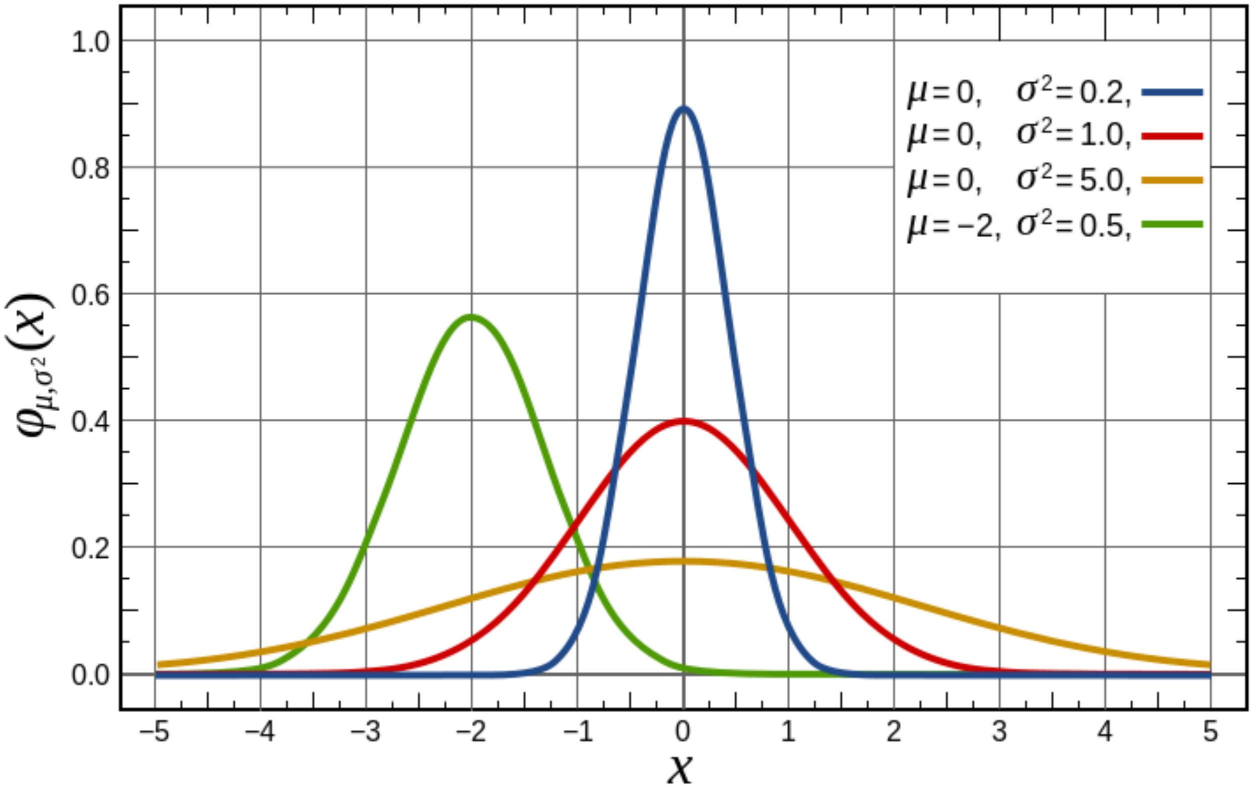
- Many naturally occurring phenomena are modeled by the normal distribution.

- Normal distributions are continuous: generalization of the binomial distribution, where n->infinity and the degree of concentration around the mean is specified by the parameter sigma.

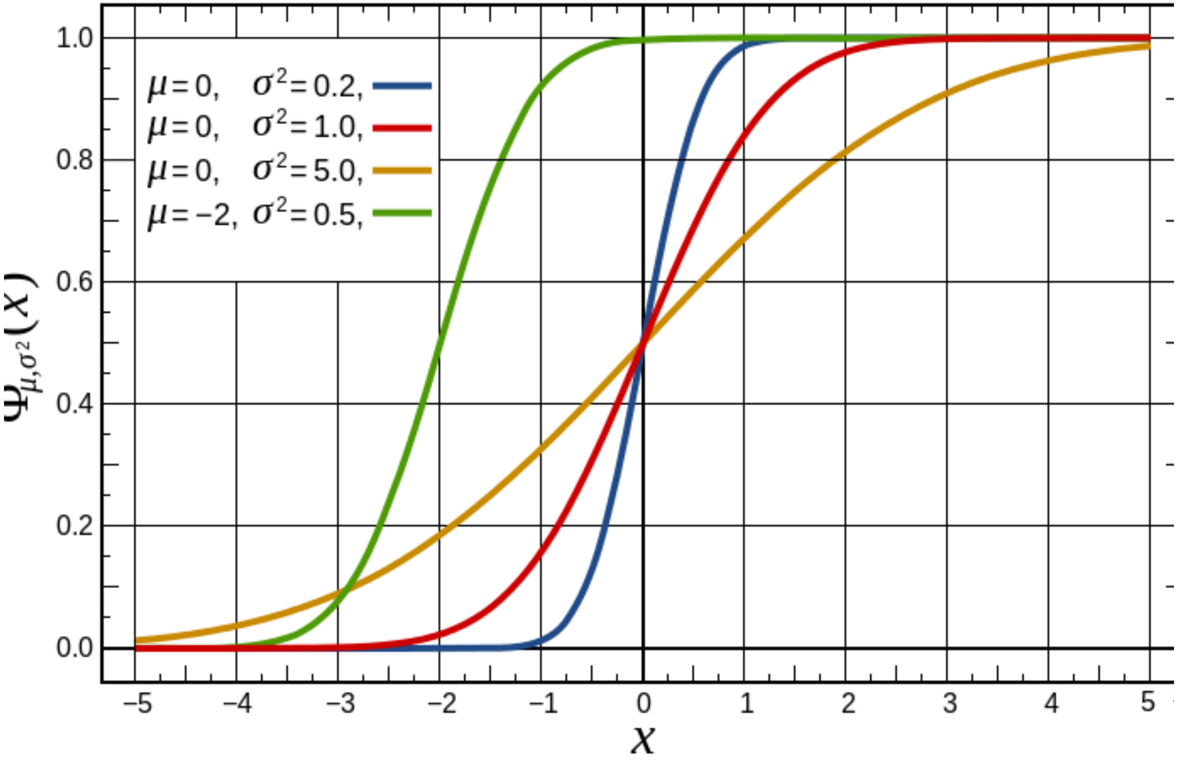
- Bell-shaped curve or Gaussian distribution, which is parameterized by its mean and standard deviation.

https://www.youtube.com/watch?v=4HpvBZnHOVI

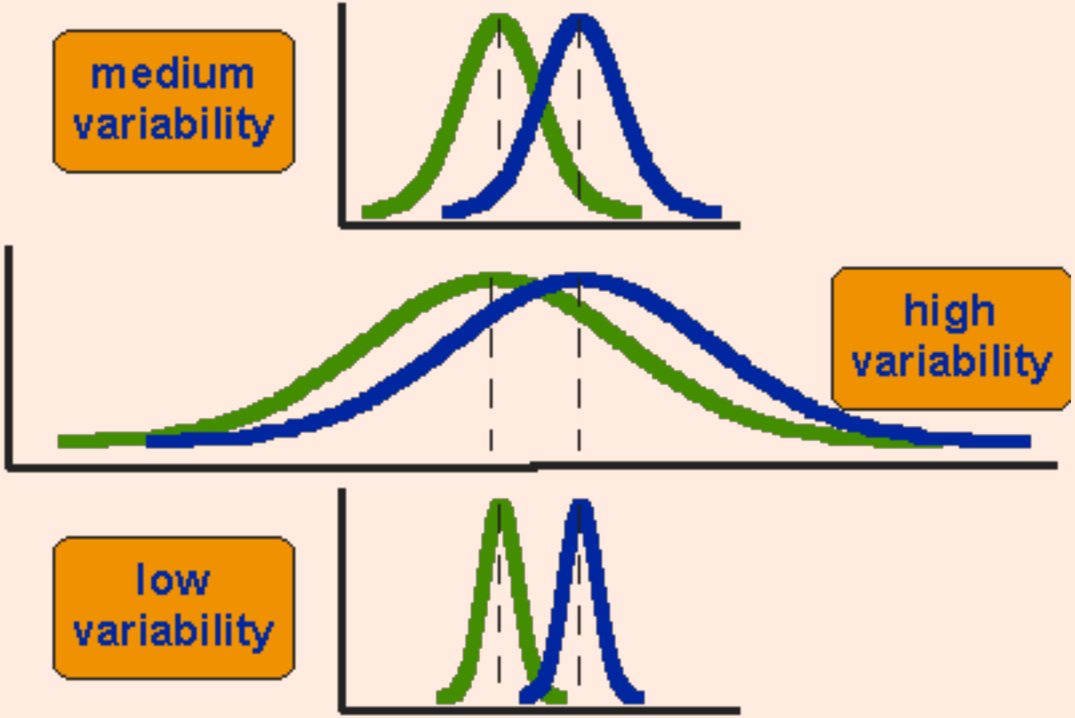
Probability Density



Cumulative


$\mu=0, \quad \sigma^2=0.2,$ 🟦
$\mu=0, \quad \sigma^2=1.0,$ 🟥
$\mu=0, \quad \sigma^2=5.0,$ 🟨
$\mu=-2, \quad \sigma^2=0.5,$ 🟩
$\Psi_{\mu,\sigma^2}(x)$
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5
x

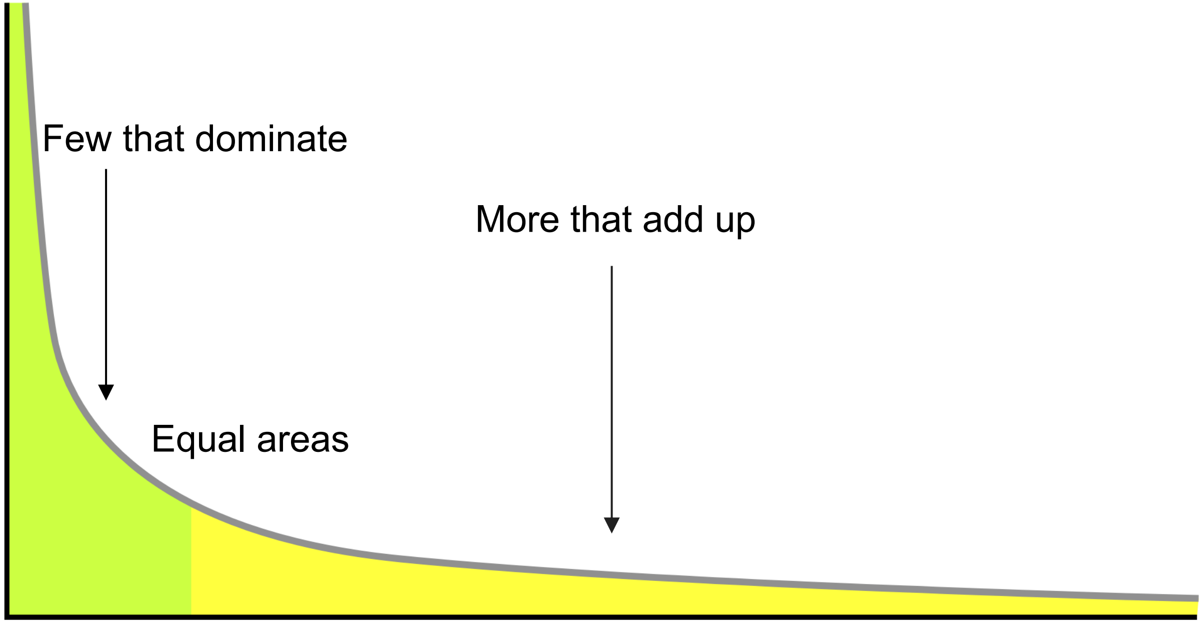
Variability in normal distributions



Heavy-tail distributions

- Probability distributions whose tails are not exponentially bounded

e.g. long-tail distributions - common in business, marketing, social media mechanisms



http://en.wikipedia.org/wiki/Heavy-tailed_distribution

Plotting these distributions

- Histograms and binning
- Getting used to log scales
- Going beyond 2-D

Distribution tests

Most distributions have tests:

- Wilcoxon-Mann-Whitney test
  - Comparing populations

  Two data samples are independent if they come from distinct populations and the samples do not affect each other. Using the Mann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.
http://www.r-tutor.com/elementary-statistics/non-parametric-methods/mann-whitney-wilcoxon-test

- Kolmogorov-Smirnov
- Shapiro–Wilk
- Anderson–Darling

- It got out of control when people realized they can name the test after themselves, v. someone else...

Probability

Before dive into Naïve Bayes in upcoming classes, lets go over some definitions in probability.

- Probability is the measure of the likelihood that an event will occur.
- In other words, probability is a measurement of how likely an event occurs.
- • *Probability of event **A***:

$$P(A) = \frac{Number\ of\ ways\ \boldsymbol{A}\ can\ occur}{Number\ of\ possible\ outcomes}$$

Reference: https://en.wikipedia.org/wiki/Probability

Probability

- You should know/understand the two probability concepts:

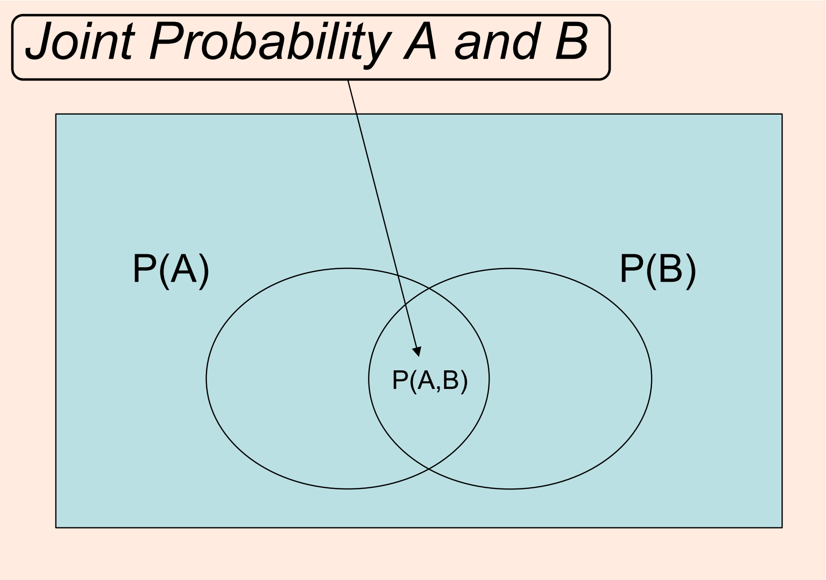
1) Joint Probability

2) Conditional Probability

Reference: https://en.wikipedia.org/wiki/Probability

Probability

Joint Probability: *specifies the probability of event A and event B occurring together.*



https://en.wikipedia.org/wiki/Joint_probability_distribution

Probability

Joint Probability: *specifies the probability of event A and event B occurring together*.

If the two events are independent,

What is the probability of getting two 6’s when you roll two dice?

The probability of rolling(getting) two 6’s:

$$P(A,B) = P(A) * P(B) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$$

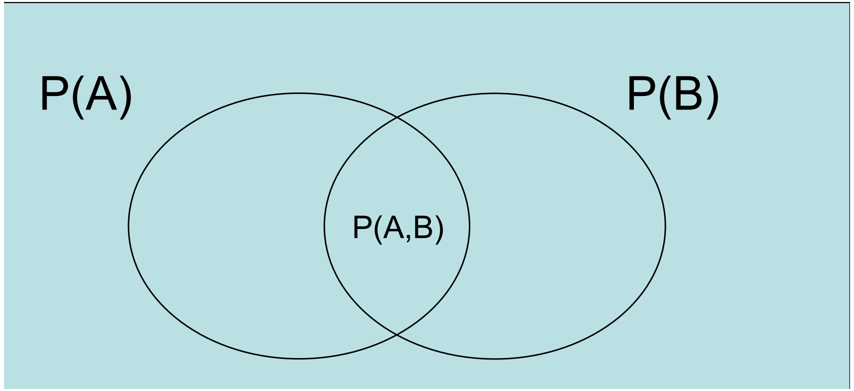
https://en.wikipedia.org/wiki/Joint_probability_distribution

Image/Photo Credit: https://pngtree.com/freepng/two-dice_1504759.html

Probability

<u>Conditional Probability:</u> *probability of event A occurring, given that event B occurred.*

$P(A|B) = \frac{P(A,B)}{P(B)}$ = *Probability of A, given B ; P(B)>0*



https://en.wikipedia.org/wiki/Conditional_probability

The Complement Rule

$$P(\text{not } A) = 1 - P(A)$$

$$P(B|\text{not } A) = 1 - P(\text{not } B|\text{not } A)$$

Hypothesis

1. Write the original claim and identify whether it is the null hypothesis or the alternative hypothesis.
2. Write the null and alternative hypotheses. Use the alternative hypothesis to identify the type of test.
3. Write down all information from the problem.
4. Find the critical value using the tables
5. Compute the test statistic
6. Make a decision to reject or fail to reject the null hypothesis. A figure showing the critical value and test statistic may be useful.
7. Write the conclusion.

Hypothesis

- What are you exploring?

- “Traditional” data analytics features ~ well defined hypotheses

– Big Data messes that up

- e.g. Stock market performance / trends versus unusual events (crash/ boom):

– Populations versus samples
– which is which? Why?

- e.g. Election results are predictable from exit polls

Null and Alternate Hypotheses

- H0 - null
- H1 – alternate

- If a given claim contains equality, or a statement of no change from the given or accepted condition, then it is the null hypothesis, otherwise, if it represents change, it is the alternative hypothesis.

- **It never snows in Troy in January**
- **Students will attend their scheduled classes**

Accept or Reject?

- **Reject the null hypothesis if the p-value is less than the level of significance.**

- **You will fail to reject the null hypothesis if the p-value is greater than or equal to the level of significance.**

- **Typical significance 0.05 (!)**

Random Numbers

- Can a computer generate a random number?
- Can you?
- Why? – to reduce selection bias!
- In R – many ways – see help on Random {base} and get familiar with set.seed()

Preliminary (Exploratory) Analysis

- Determining if there is one or more common distributions involved – i.e. parametric statistics (assumes or asserts a probability distribution)
- Fitting that distribution -> provides a model!
- Or NOT

– A hybrid model or

– Non-parametric (statistics) approaches are needed – more on this to come

Considerations

- Quality, uncertainty and bias – you will often spend a lot of time with the data
- Distributions – the common and not-so common ones and how cyber vs. natural data can have distinct distributions
- How simple statistical distributions can mislead us
- Populations and samples and how inferential statistics will lead us to model choices (no we have not actually done that yet in detail)
- Preparing data for exploratory analysis

Reminder:

- Make sure you have R & RStudio installed for Lab 1

- Experiment with R

- Create the Github repository for this class if you have not created it yet and email the repo URL to me (eleisa2@rpi.edu)

Thanks!
(See you Friday)

*** Experiment with R!