Rensselaer

why not change the world?®

Data and Information Resources, Role of Hypothesis, Exploration and Distributions

Ahmed Eleish

Data Analytics

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960 MGMT- 4962/6962 BCBP 4960

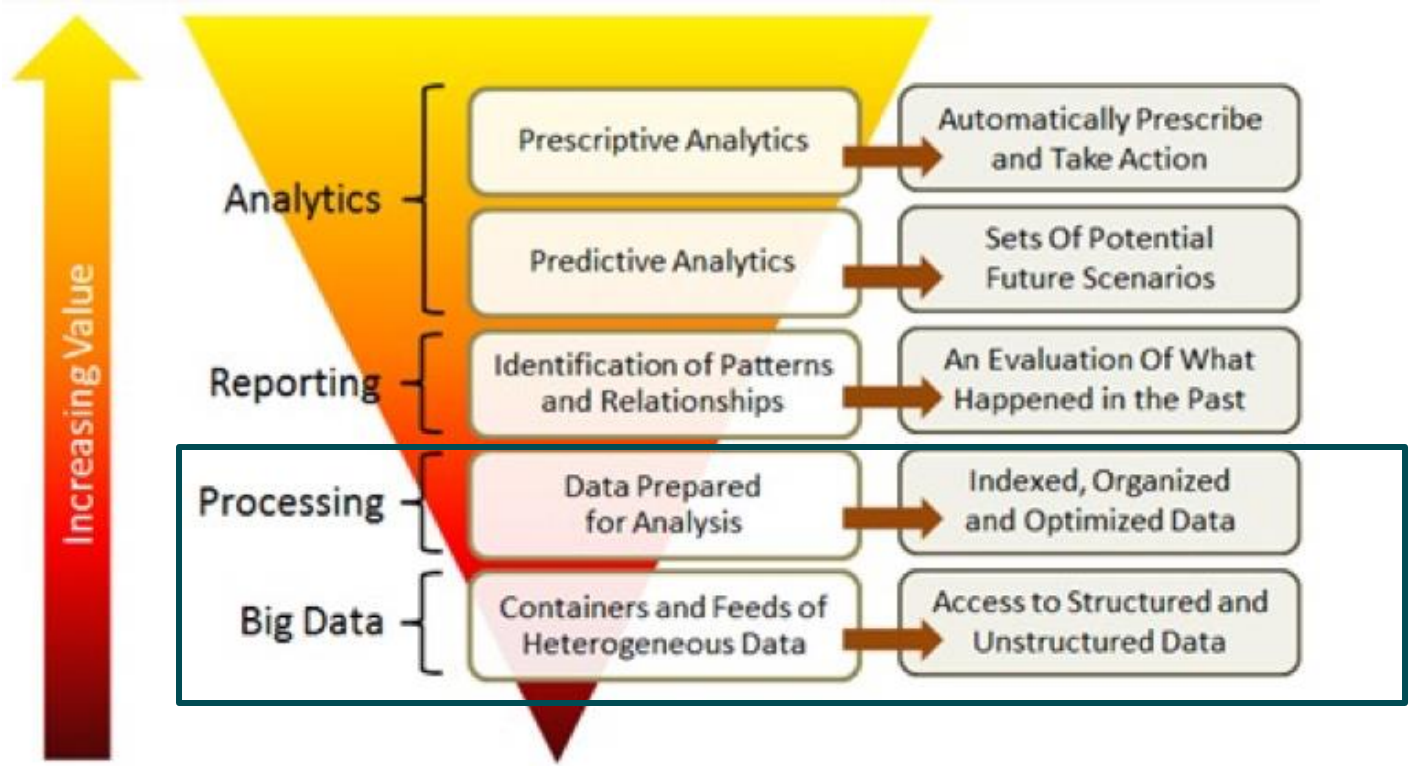
Group 1 Module 3(a), September 10th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

- Data sources
  - Cyber
  - Human
- “Munging”, “wrangling”, … preparation!
- Exploring
  - Distributions...
  - Summaries
  - Visualization
- Testing and evaluating the results (beginning)

Lower layers in the Analytics Stack




"Cyber Data"
...
iStock.
by Getty Images™

“Human Data” ...

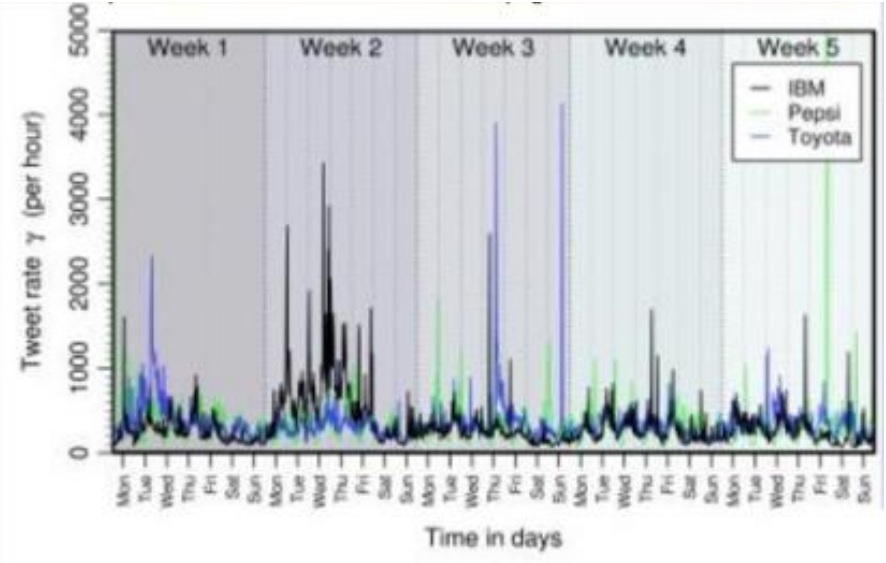


Image Credit: http://images.sciencedaily.com/2013/10/131007151731-large.jpg http://www.downsidehedge.com/wp-content/uploads/2013/09/130915StockTwitsSPX.png

Stats review – cont'd

Definitions/ topics

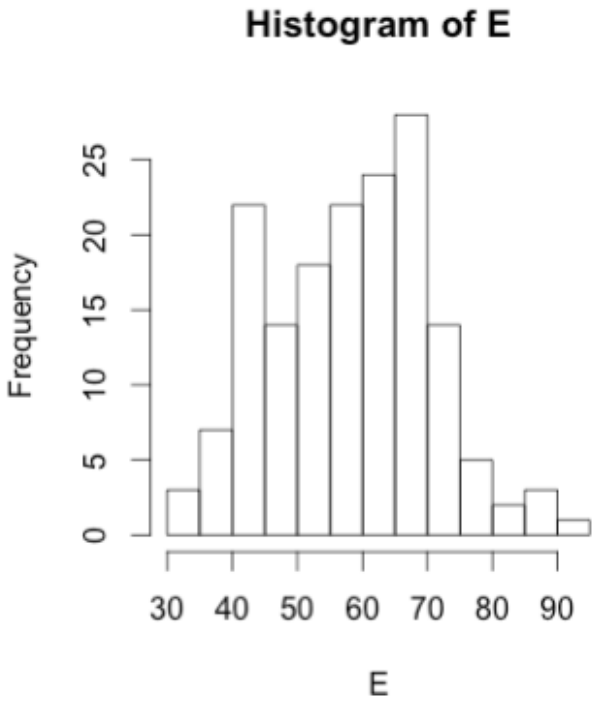
- Statistic
- Statistics
- Population and Samples
- Sampling
- Distributions and parameters
- Central Tendencies
- Frequency

Previous class

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

Today's class

Grouped Frequency Distribution aka binning



Distributions

- Shape
- Parameter(s)
- Which one fits?

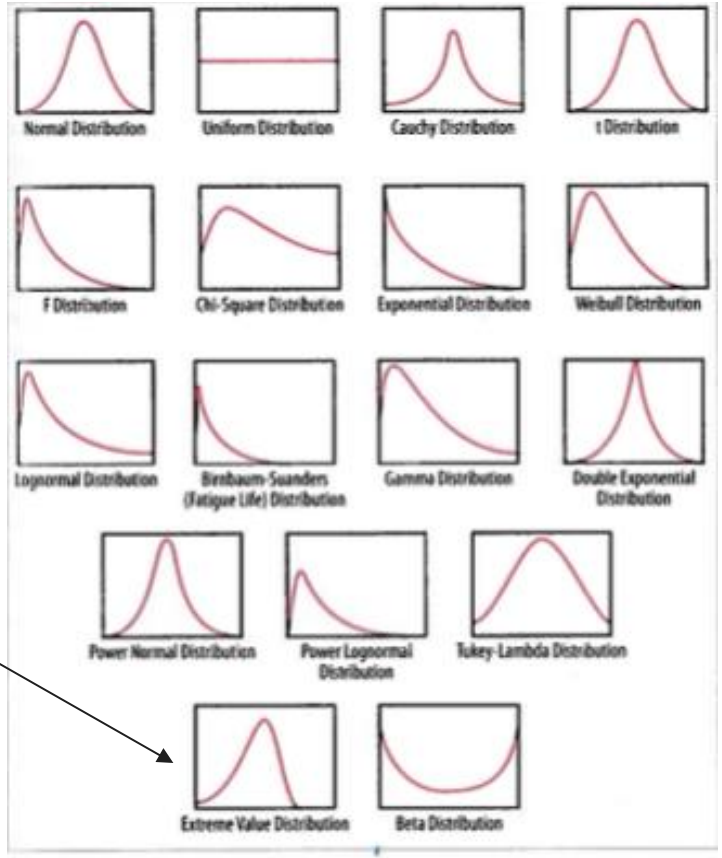


Figure 2-1. A bunch of continuous density functions (aka probability distributions)

Distributions

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

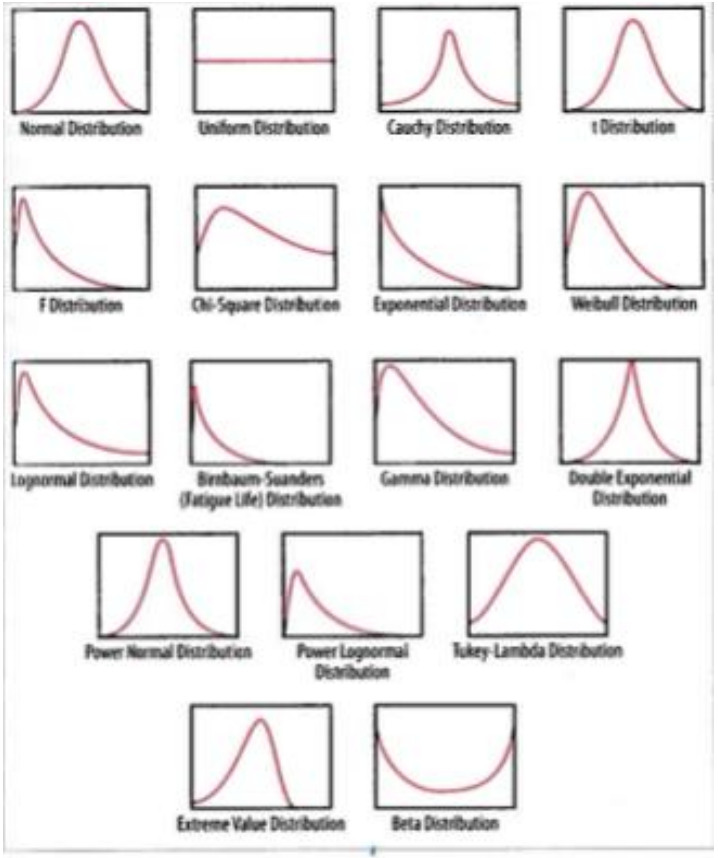


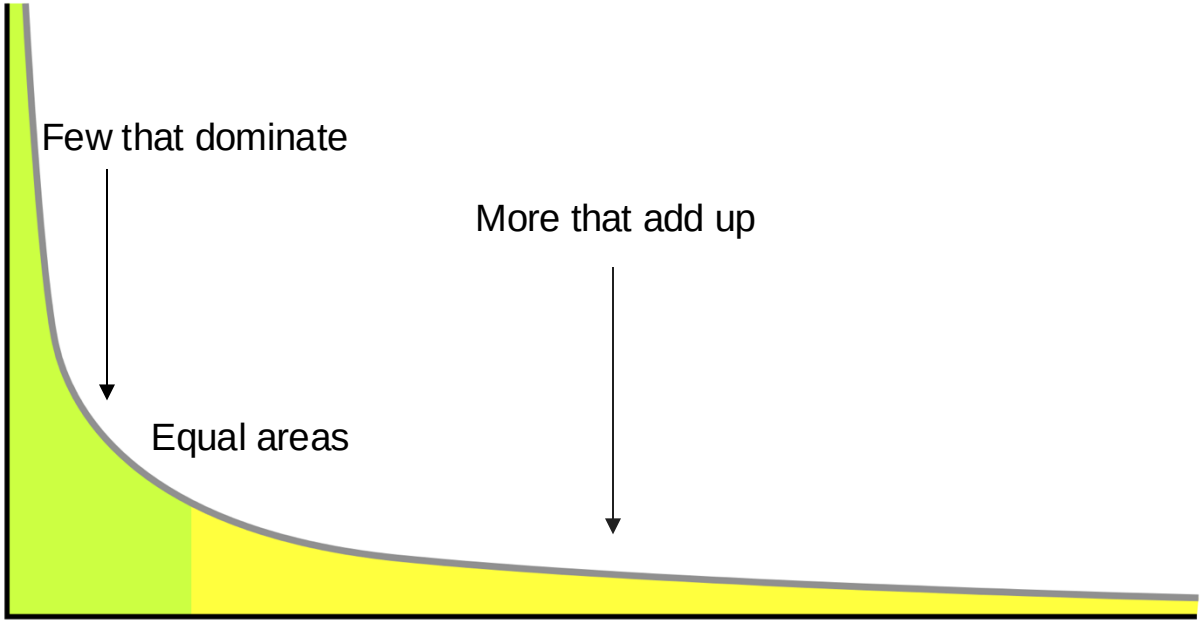
Figure 2-1. A bunch of continuous density functions (aka probability distributions)

Plotting these distributions

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

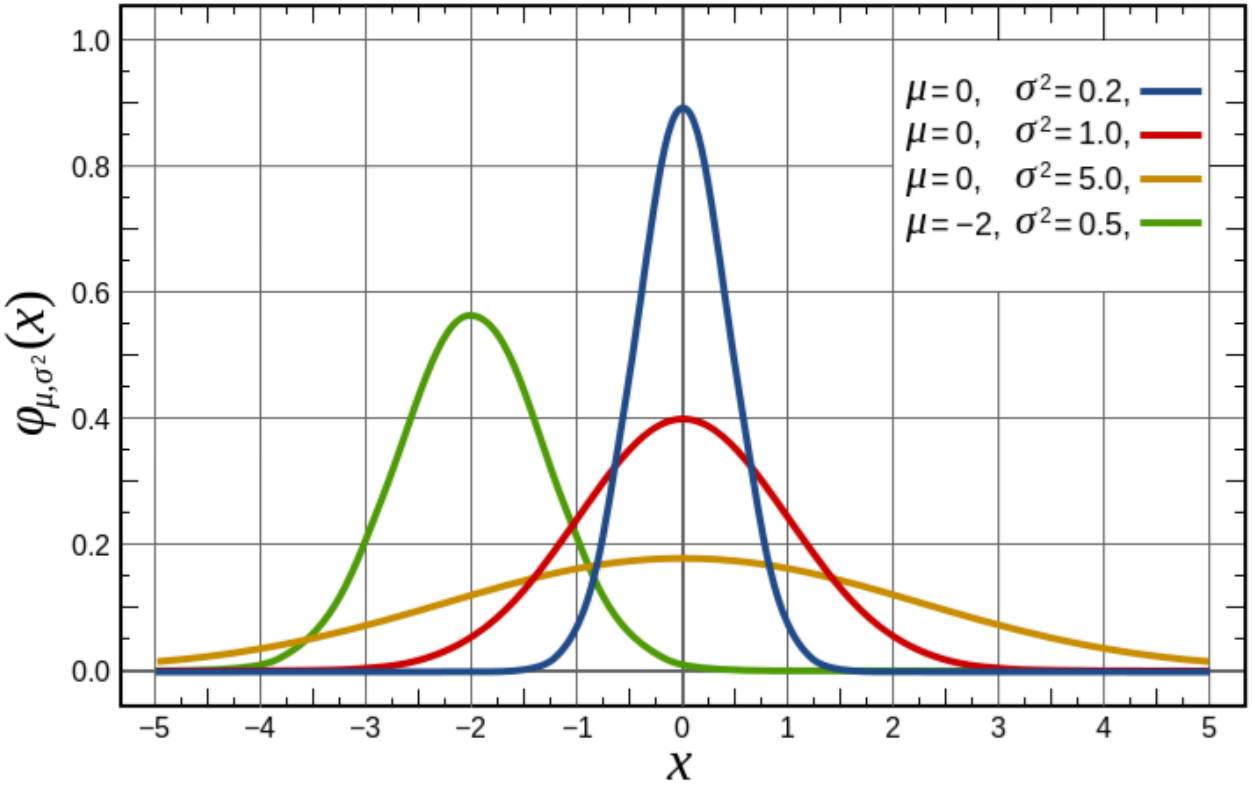
Heavy-tail distributions

- Probability distributions whose tails are not exponentially bounded
  e.g. long-tail distributions - common in nature …

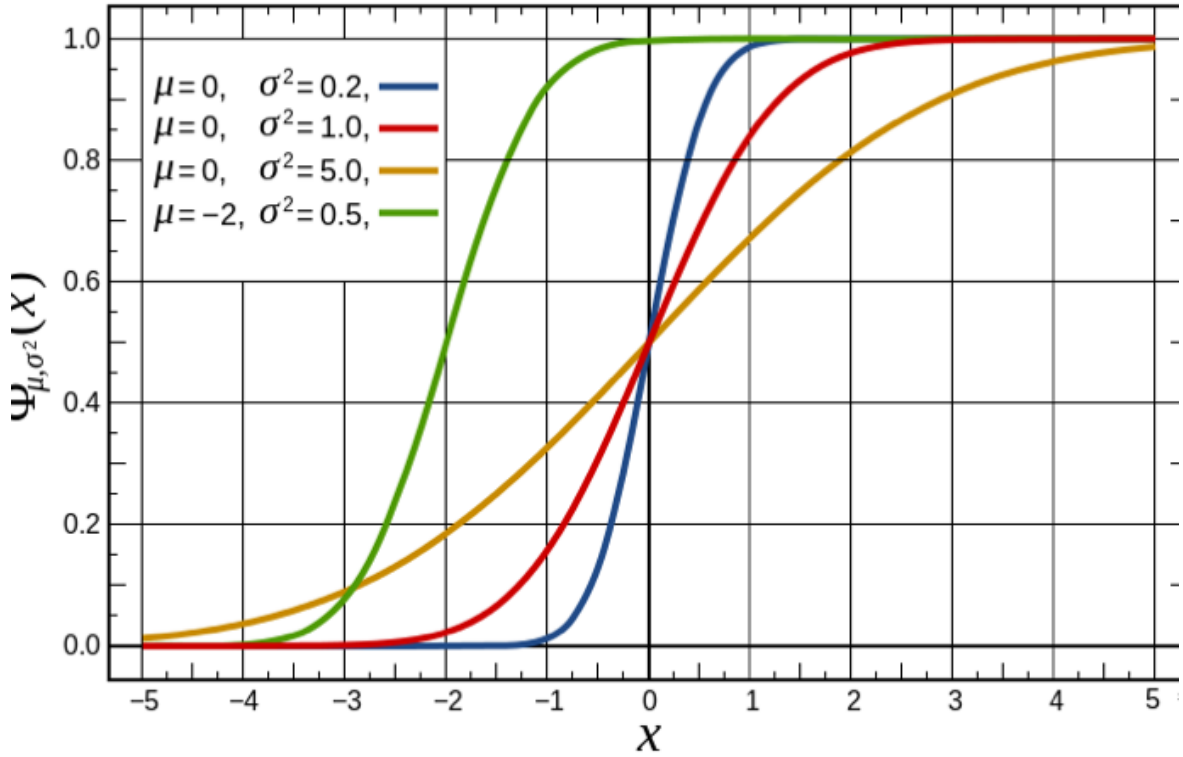


http://en.wikipedia.org/wiki/Heavy-tailed_distribution

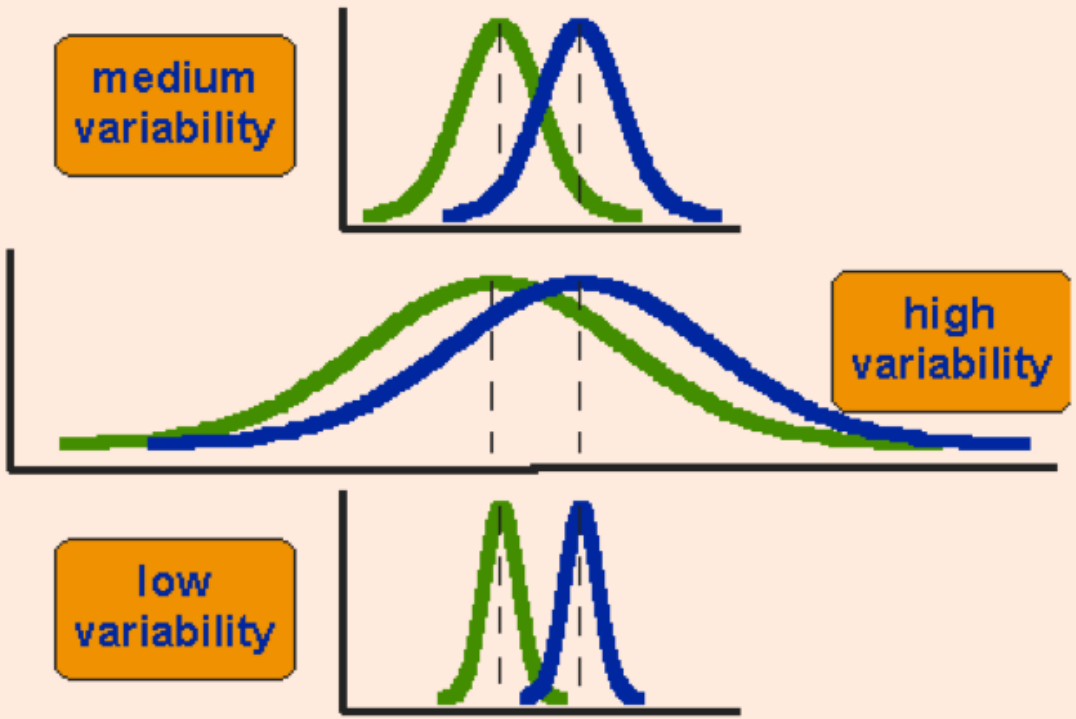
Probability Density



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

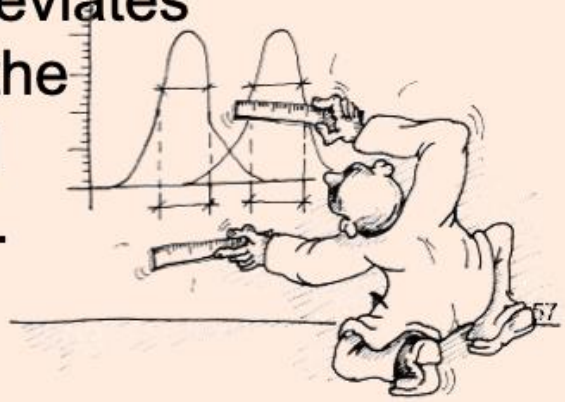
Variability in normal distributions



F-test

$F = S_1^2 / S_2^2$
where S_1 and S_2 are the sample variances.

The more this ratio deviates from 1, the stronger the evidence for unequal population variances.

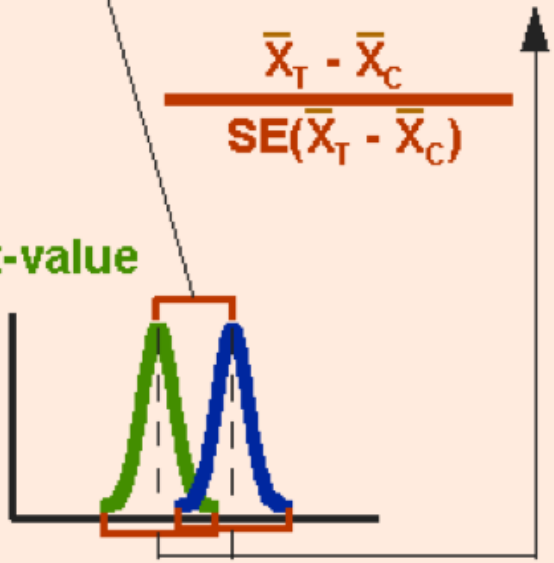


T-test

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

$$= \text{t-value}$$



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:

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

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

Standard Error

- Versus standard deviation = SD (i.e. from the mean)
- SE ~ SD/sample size
- So, as size increases SE << SD !! Big data

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
- University year
- Which US State you live in

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.

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)

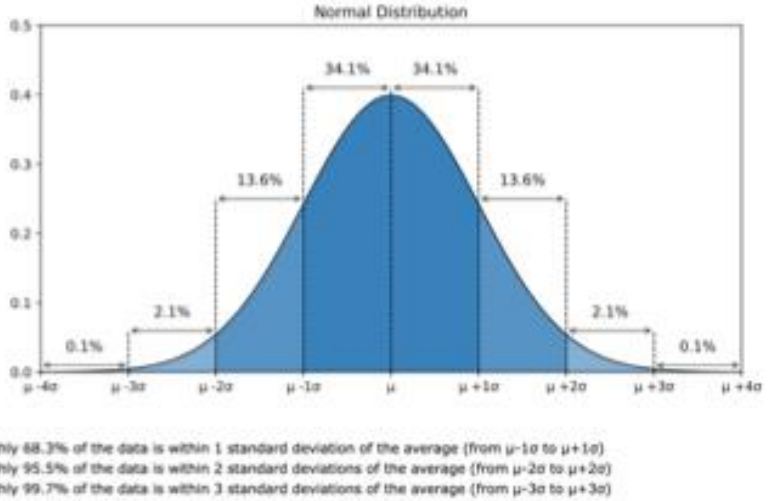


Image Credit: W3C school:
https://www.w3schools.com/statistics/statistics_normal_distribution.php

Distribution tests

most distributions have tests

•Wilcoxon (Mann-Whitney)

– 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 (KS)

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

Distributions

• Go over the distributions listed here (make yourself familiar with these distributions):

http://www.quantitativeskills.com/sisa/rojo/alldist.zip

Probability

• Before dive into the Naïve Bayes lecture 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

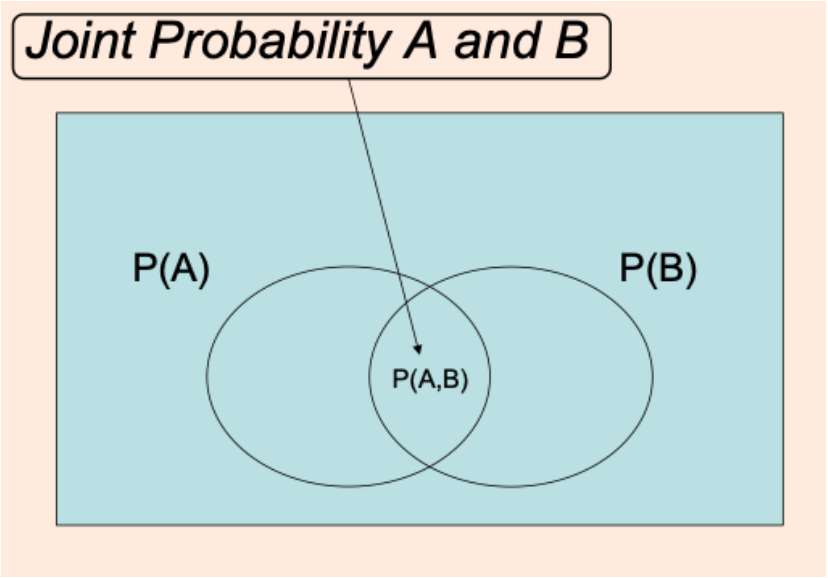
• You should know/understand the two probability concepts:

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:

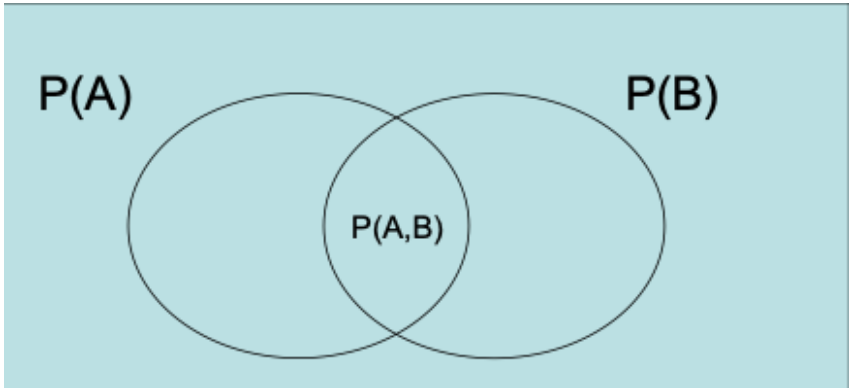
$$P(A,B) = P(A) * P(B) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$$

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

Bayes Theorem

- The relationship between conditional probabilities, P(B|A) and P(A|B) can be expressed using the Bayes Theorem.

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

Reference: https://en.wikipedia.org/wiki/Bayes%27_theorem

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?
- Regular 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

P-value

• One common way to evaluate significance, especially in R output

– approaches hypothesis testing from a different manner. Instead of comparing z-scores or t-scores as in the classical approach, you're comparing probabilities, or areas.

• The level of significance (alpha) is the area in the critical region. That is, the area in the tails to the right or left of the critical values.

P-value

• The p-value is the area to the right or left of the test statistic.
– If it is a two tail test, then look up the probability in one tail and double it.

• If the test statistic is in the critical region, then the p-value will be less than the level of significance.
– It does not matter whether it is a left tail, right tail, or two tail test. This rule always holds.

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 (!)**

E.g. Election prediction

- Exit polls versus election results
- How is the “population” defined here?

- What is the sample, how is it chosen?
  – What is described and how is that used to predict?
  – Are results categorized? (where from, specialty, age)

- What is the uncertainty?
  – It is reflected in the “sample distribution”
  – And controlled/ constraints by “sampling theory”

Random Numbers

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

Summary: exploration

- Going from preliminary to initial 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

Summary - considerations

- Cyber and Human data; 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 and human 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)
- Munging/wrangling toward exploratory analysis
- Toward models!

Reminder: finish Lab 0

- Reminder to finish the last week intro to Lab (Lab 0 – installing R)
- R! (how is your learning/coding in R going?) keep learning/coding...
- 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!