ALGONQUIN COLLEGE

CST8390 BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 1

Introduction to Data Analytics

Professor: Dr. Anu Thomas
Email: thomasa@algonquincollege.com
Office: T314

Data Analytics

- Definition: Data Analytics is the process of aggregating large data sets in order to detect underlying patterns that might not be visible by just looking at raw data.
- These patterns give insight to maximize profits, improve health, lower electricity usage, etc.

Mountains of Data

- We now have more data being gathered/collected. Governments are starting to adopt openness policies of making public data freely available on the internet.
- Canada Open Government: http://open.canada.ca/en
- Seattle Open Data https://data.seattle.gov/
- Ontario Open Data https://www.ontario.ca/search/data-catalogue
- Ottawa Open Data: http://data.ottawa.ca/

Seattle Bicycle Traffic

- https://data.seattle.gov/Transportation/daily-bike-traffic/d4dx-u56x
- A traffic counter counts number of bicycles on the East and West sidewalks.
- There are traffic spikes from 7-9 am on the West side, and from 5-6pm, but only 5 days a week.

Profit

- As an entrepreneur, where would you sell hot dogs, or advertise?

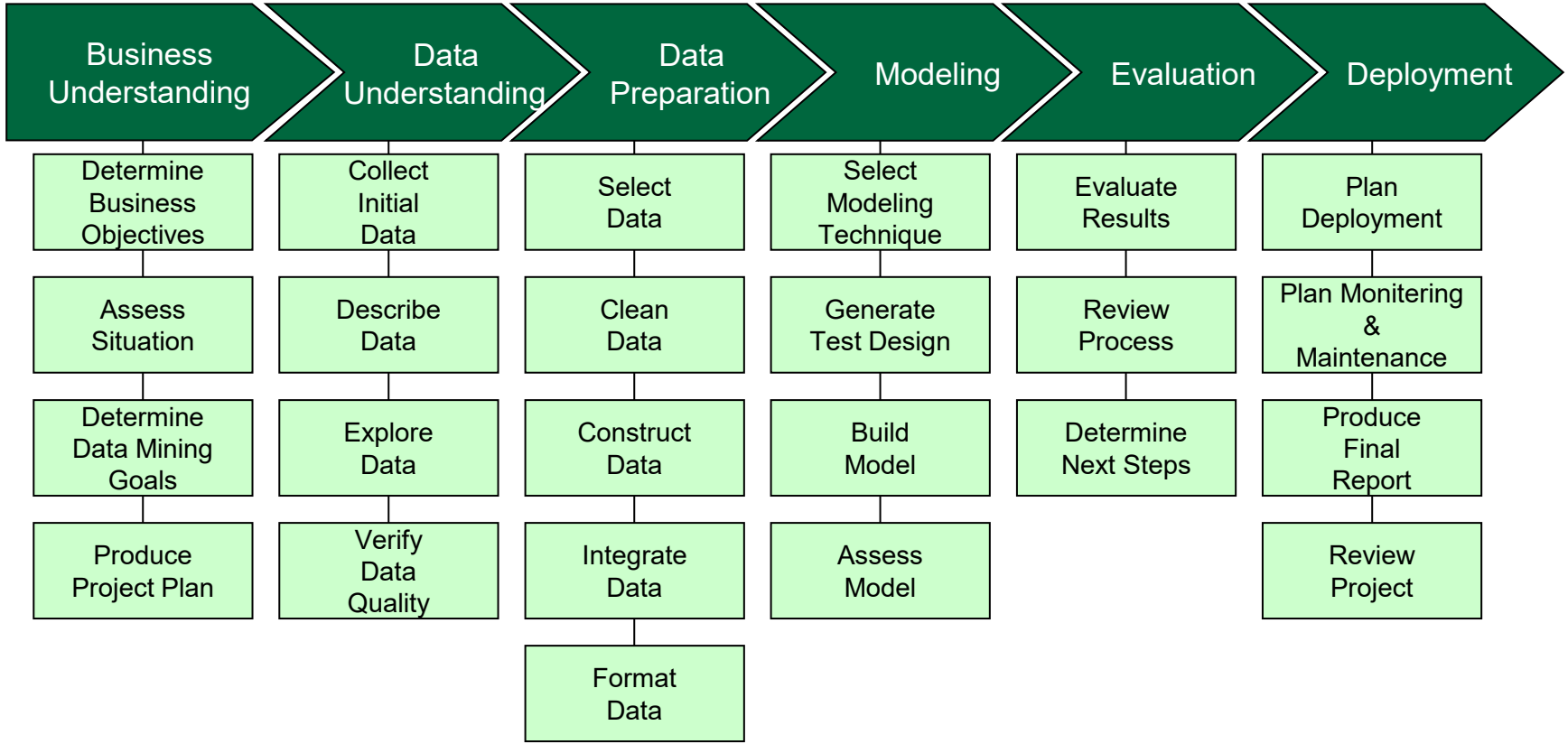
http://www.blogto.com/eat_drink/2015/07/everything_to_know_about_hot_dog_stands_in_toronto/

Healthcare

- The Real-World Benefits of Machine Learning in Healthcare
- Machine Learning Healthcare Applications – 2018 and Beyond

CRISP-DM

- Cross-industry Standard Process for Data Mining
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modeling
  - Evaluation
  - Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Determine Business Objectives
Assess Situation
Determine Data Mining Goals
Produce Project Plan
Collect Initial Data
Describe Data
Explore Data
Verify Data Quality
Select Data
Clean Data
Construct Data
Integrate Data
Format Data
Select Modeling Technique
Generate Test Design
Build Model
Assess Model
Evaluate Results
Review Process
Determine Next Steps
Plan Deployment
Plan Monitering & Maintenance
Produce Final Report
Review Project

Understanding the business

- Identify what you want to accomplish from a business perspective
- Assess current situation
- Determine goals
- Produce project plan

Understanding data

- Describe data
- Explore data
- Verify data quality

Preparing data

- Select your data
- Clean your data
- Construct required data
- Integrate data

Types of Data

- Typically data fall in to one of 3 categories:
- Structured – The data are highly structured, where every element has the same fields: age, first name, last name, address, etc. Databases, objects are good examples
- Unstructured – The data have no common structure. News articles, websites, video, audio and photographs.
- Semi-structured – The data use some structure, but it is not common. This includes tree-type data like XML and JSON.

Data Preparation

- When getting data from different sources, some work is needed when putting it together:
  - Cleaning and filtering: Remove duplicate data, missing data, resolve incomplete data. Something like: ***Woodroffe Ave, Woodroffe, Woodroffe Avenue*** should all be the same.
  - Remove outliers: (data that is far outside the average). Every semester, some students register for a course but don't drop it. This means they get 0 for everything and lowers the class average. Another example is that sales for a store are $0 for regional some regional holidays.
  - Variable transformations. Changing how variables are represented (metric / imperial)

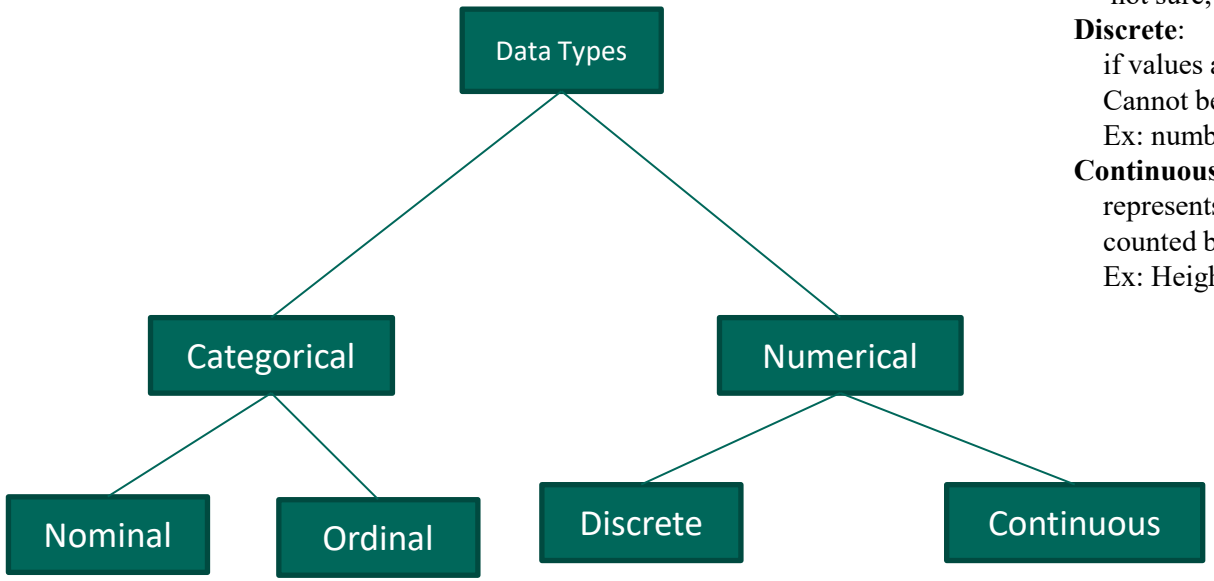
Data Preparation

- You get a data matrix with variables, attributes or features (columns)
- Instances are the rows (N).

Instance	Date	Price	Quantity	Label
1	May 2, 2016	5.50	1	Regular
2	May 28, 2016	3.79	2	Sale

- The data are typically brought together from various parts of the organization. They must be transformed into a single data format, for instance, price must all be $USD, or temperatures must all be Celsius instead of Fahrenheit.

Data Types



Nominal:
Gender: male, female

Ordinal:
survey questions: Strongly agree, agree, not sure, disagree, strongly disagree

Discrete:
if values are distinct and separate.
Cannot be measured but can be counted
Ex: number of heads in 100 coin flips

Continuous:
represents measurements. Values cannot be counted but can be measured
Ex: Height, salary

Data cleansing

- How do you detect outliers?
  - One method is to sort the data. The outliers will be at either end of the sorted sequence.
- For tagged data, make all similar tags the same: Woodroffe Ave.
- What about missing data? Replace with random numbers from average and standard deviation?
- Replace with “Missing” or “Unknown” tag.

Real Data from UCI Machine Learning Repository

Wine Dataset

```
1,13.9,1.68,2.12,16,101,3.1,3.39,.21,2.14,6.1,.91,3.33,985
1,14.1,2.02,2.4,18.8,103,2.75,2.92,.32,2.38,6.2,1.07,2.75,1060
1,13.94,1.73,2.27,17.4,108,2.88,3.54,.32,2.08,8.90,1.12,3.1,1260
1,13.05,1.73,2.04,12.4,92,2.72,3.27,.17,2.91,7.2,1.12,2.91,1150
1,13.83,1.65,2.6,17.2,94,2.45,2.99,.22,2.29,5.6,1.24,3.37,1265
1,13.82,1.75,2.42,14,111,3.88,3.74,.32,1.87,7.05,1.01,3.26,1190
1,13.77,1.9,2.68,17.1,115,3,2.79,.39,1.68,6.3,1.13,2.93,1375
1,13.74,1.67,2.25,16.4,118,2.6,2.9,.21,1.62,5.85,.92,3.2,1060
1,13.56,1.73,2.46,20.5,116,2.96,2.78,.2,2.45,6.25,.98,3.03,1120
1,14.22,1.7,2.3,16.3,118,3.2,3,.26,2.03,6.38,.94,3.31,970
1,13.29,1.97,2.68,16.8,102,3,3.23,.31,1.66,6,1.07,2.84,1270
1,13.72,1.43,2.5,16.7,108,3.4,3.67,.19,2.04,6.8,.89,2.87,1285
2,12.37,.94,1.36,10.6,88,1.98,.57,.28,.42,1.95,1.05,1.82,520
2,12.33,1.1,2.28,16,101,2.05,1.09,.63,.41,3.27,1.25,1.67,680
2,12.64,1.36,2.02,16.8,100,2.02,1.41,.53,.62,5.75,.98,1.59,450
```

Breast Cancer Dataset

```
842302,M,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,153.4,0.006399,0.04904,0.05373,0.01587,0.03003,0.006193,25.38,17.33,184.6,2019,0.1622,0.6656,0.7119,0.2654,0.4601,0.1189
842517,M,20.57,17.77,132.9,1326,0.08474,0.07864,0.0869,0.07017,0.1812,0.05667,0.5435,0.7339,3.398,74.08,0.005225,0.01308,0.0186,0.0134,0.01389,0.003532,24.99,23.41,158.8,1956,0.1238,0.1866,0.2416,0.186,0.275,0.08902
84300903,M,19.69,21.25,130,1203,0.1096,0.1599,0.1974,0.1279,0.2069,0.05999,0.7456,0.7869,4.585,94.03,0.00615,0.04006,0.03832,0.02058,0.0225,0.004571,23.57,25.53,152.5,1709,0.1444,0.4245,0.4504,0.243,0.3613,0.08758
84348301,M,11.42,20.38,77.58,386.1,0.1425,0.2839,0.2414,0.1052,0.2597,0.09744,0.4956,1.156,3.445,27.23,0.00911,0.07458,0.05661,0.01867,0.05963,0.009208,14.91,26.5,98.87,567.7,0.2098,0.8663,0.6869,0.2575,0.6638,0.173
84358402,M,20.29,14.34,135.1,1297,0.1003,0.1328,0.198,0.1043,0.1809,0.05883,0.7572,0.7813,5.438,94.44,0.01149,0.02461,0.05688,0.01885,0.01756,0.005115,22.54,16.67,152.2,1575,0.1374,0.205,0.4,0.1625,0.2364,0.07678
843786,M,12.45,15.7,82.57,477.1,0.1278,0.17,0.1578,0.08089,0.2087,0.07613,0.3345,0.8902,2.217,27.19,0.00751,0.03345,0.03672,0.01137,0.02165,0.005082,15.47,23.75,103.4,741.6,0.1791,0.5249,0.5355,0.1741,0.3985,0.1244
844359,M,18.25,19.98,119.6,1040,0.09463,0.109,0.1127,0.074,0.1794,0.05742,0.4467,0.7732,3.18,53.91,0.004314,0.01382,0.02254,0.01039,0.01369,0.002179,22.88,27.66,153.2,1606,0.1442,0.2576,0.3784,0.1932,0.3063,0.08368
84458202,M,13.71,20.83,90.2,577.9,0.1189,0.1645,0.09366,0.05985,0.2196,0.07451,0.5835,1.377,3.856,50.96,0.008805,0.03029,0.02488,0.01448,0.01486,0.005412,17.06,28.14,110.6,897,0.1654,0.3682,0.2678,0.1556,0.3196,0.1151
844981,M,13,21.82,87.5,519.8,0.1273,0.1932,0.1859,0.09353,0.235,0.07389,0.3063,1.002,2.406,24.32,0.005731,0.03502,0.03553,0.01226,0.02143,0.003749,15.49,30.73,106.2,739.3,0.1703,0.5401,0.539,0.206,0.4378,0.1072
84501001,M,12.46,24.04,83.97,475.9,0.1186,0.2396,0.2273,0.08543,0.203,0.08243,0.2976,1.599,2.039,23.94,0.007149,0.07217,0.07743,0.01432,0.01789,0.01008,15.09,40.68,97.65,711.4,0.1853,1.058,1.105,0.221,0.4366,0.2075
845636,M,16.02,23.24,102.7,797.8,0.08206,0.06669,0.03299,0.03323,0.1528,0.05697,0.3795,1.187,2.466,40.51,0.004029,0.009269,0.01101,0.007591,0.0146,0.003042,19.19,33.88,123.8,1150,0.1181,0.1551,0.1459,0.09975,0.2948,0.08452
84610002,M,15.78,17.89,103.6,781,0.0971,0.1292,0.09954,0.06606,0.1842,0.06082,0.5058,0.9849,3.564,54.16,0.005771,0.04061,0.02791,0.01282,0.02008,0.004144,20.42,27.28,136.5,1299,0.1396,0.5609,0.3965,0.181,0.3792,0.1048
846226,M,19.17,24.8,132.4,1123,0.0974,0.2458,0.2065,0.1118,0.2397,0.078,0.9555,3.568,11.07,116.2,0.003139,0.08297,0.0889,0.0409,0.04484,0.01284,20.96,29.94,151.7,1332,0.1037,0.3903,0.3639,0.1767,0.3176,0.1023
846381,M,15.85,23.95,103.7,782.7,0.08401,0.1002,0.09938,0.05364,0.1847,0.05338,0.4033,1.078,2.903,36.58,0.009769,0.03126,0.05051,0.01992,0.02981,0.003002,16.84,27.66,112,876.5,0.1131,0.1924,0.2322,0.1119,0.2809,0.06287
```

Statistics

Random Number Generators

- The numbers seem random but they're not. They're pseudo-random
- The sequence of numbers generated depends on the starting "seed"
- ***Two number generators will produce the same numbers if they have the same seed.***
- Create two Random Objects with the constructor: Random(int seed)
- call nextInt() on both. They will produce the same sequence.
- To create a new sequence every time, use: System.currentTimeMillis() as the seed

Statistics

- Given a set of data (Numbers), there are several things we can compute:
- Mean: What is the average? $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$
- Median: What is the middle item? Array[size/2]
- Mode: What item appears the most often:
- 1, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5 ?
- Order Statistics: What is the 3rd largest number? What is (n-2)nd largest number?

Computing Mean

- Computing the mean tells the average value, but sometimes you want to use different formulae:
- Moving average: What is the average of values in the last 5 days?
- If N is the total sum, and the day changes, you don't need to recompute the mean. Instead, add the new values, subtract the expired values, and divide by the new number of items, N.

Computing Rolling Mean

- If the mean, N, represents the average of 5 days:

  [1, 2],[3, 4], [4, 6], [5, 7], [9, 11]

- $N = (1 + 2 + 3 + 4 + 4 + 6 + 5 + 7 + 9 + 11)/10 = 5.2$
- Now new data are ready to be added: 10, 12
- Imagine that [1, 2] are now out of date and no longer part of your computation. Just add the new data, subtract the old data, and divide by the number of data:
- $N' = N + [(10+12) - (1+2)]/10$

  $= 5.2 + 1.9$

  $= 7.1$

- Verify: $(3 + 4 + 4 + 6 + 5 + 7 + 9 + 11 + 10 + 12)/10 = 7.1$

Weighted Average

- Suppose you want more recent data to be worth more than older data, like for predicting gas prices.
- Decide the comparative weights of the components:

$$5*(\text{Price}_{\text{days-1}}) + 3*(\text{Price}_{\text{days-2}})+2*(\text{Price}_{\text{days-3}})+1*(\text{Price}_{\text{days-4}})$$

- Now divide by the weighted number of elements:

$$5*(\text{Price}_{\text{days-1}}) + 3*(\text{Price}_{\text{days-2}})+2*(\text{Price}_{\text{days-3}})+1*(\text{Price}_{\text{days-4}}) \\ / (5 + 3 + 2 + 1)$$

Sample vs Population

- A population data set contains all members of a specified group (the entire list of possible data values)
  - Example: all people in Ottawa

- A sample data set contains a part, or a subset, of a population. The size of a sample is always less than the size of the population from which it is taken.
  - Example: some people in Ottawa

Standard Deviation

- Let X be an array: X = {21, 37, 13, 25, 32, 8 }
- What is the average? $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$ = 22.6667
- **Standard deviation**: N is the number of elements
- $\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2/N}$. This formula is used when you have measured the entire population. In Excel, this is stdev.p() = 10.07748
- $\sqrt{\sum_{i=1}^{N}(x_i - \mu)^2/(N-1)}$. This formula is used when you have only part of the data. In Excel, this is stdev.s() = 11.03932

Standard Deviation

	x	x-Mean	$(x\text{-Mean})^2$
	21	-1.6667	2.7778
	37	14.3333	205.4444
	13	-9.6667	93.4444
	25	2.3333	5.4444
	32	9.3333	87.1111
	8	-14.6667	215.1111
Mean	22.6667		

Sum of $(x\text{-Mean})^2$	609.3333
Sum / 6	101.5556
sqrt (sum/6)	10.0775

Mean and Variance

- Difference of means is the difference between the averages of two samples
- Variance is just standard deviation squared.

- https://www.khanacademy.org/math/probability/data-distributions-a1/summarizing-spread-distributions/v/range-variance-and-standard-deviation-as-measures-of-dispersion

- https://www.khanacademy.org/math/statistics-probability/displaying-describing-data/sample-standard-deviation/v/statistics-sample-variance

Distributions

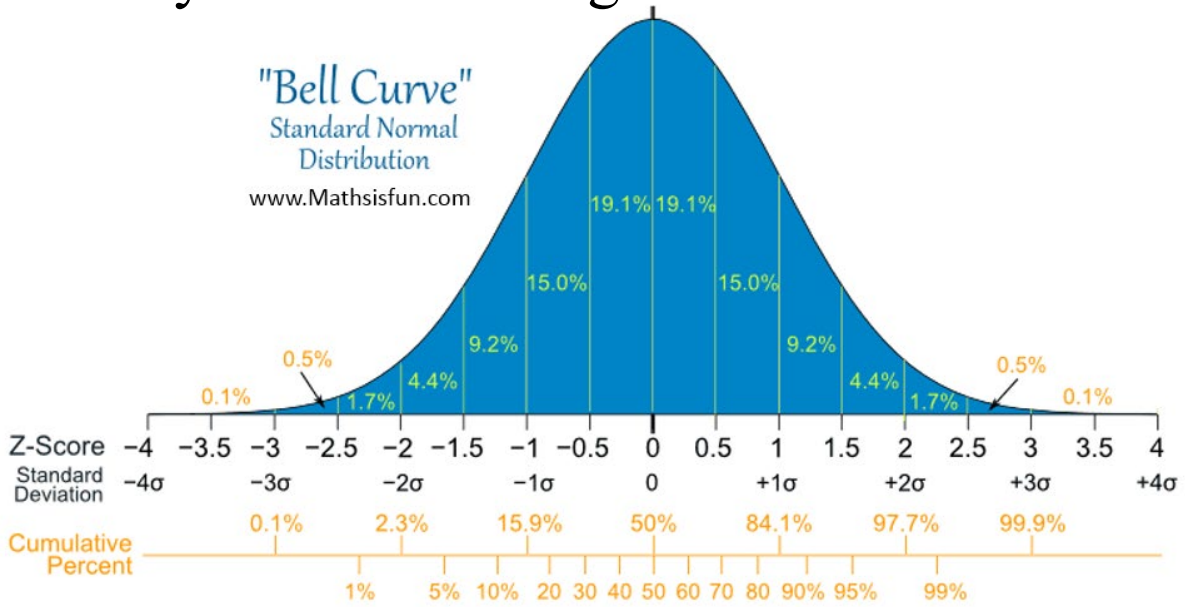
- Uniform – The probability of an event is equal (uniform). The probability of getting tails for flipping a coin, or rolling a 1 with a die.
- Gaussian (Normal) – The values are centered around a midpoint (mean), but decrease as you get farther from the mean: grades on a test.
- Geometric, Poisson, Exponential – These are other distributions that exist, but we don't have time to cover.

Normal / Gaussian

- Also called a bell curve, with midpoint of μ (pronounced me-you), and standard deviation of σ (pronounced sigma).
- The density of an event $(x \mid \mu, \sigma^2)$ is: $\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\sigma^2\pi}}$
- The variance is σ^2
- If $\mu = 0$ and $\sigma^2 = 1$ then this is called Standard Normal Distribution

Normal / Gaussian

- The area under the curve must add up to 1. Probabilities are calculated by a number being less than a number.



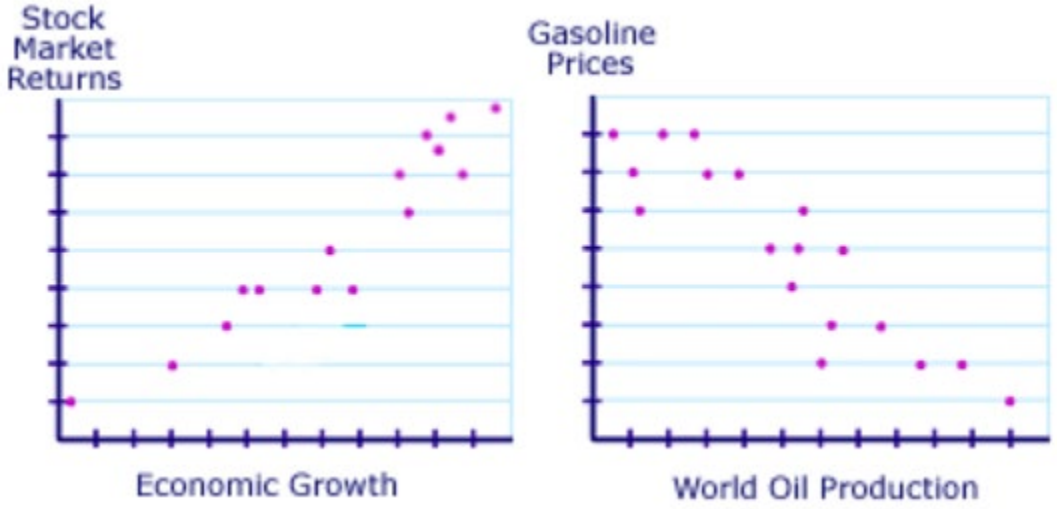
Rank Statistics

- Rank statistics compute where a number compares to the rest of the data, for instance to 5%, bottom 15%, etc.
- They are described in percentiles, meaning how much of the data is less than the number. 95% percentile means that 95% of the numbers are less. The median is the 50% percentile.

Covariance and correlation

- The covariance computes the strength and direction of the relationship of two sets of values. Do they both get bigger together? Does one get bigger as the other gets smaller? Is there no relationship?
- Covariance is calculated by: $$S_{xy} = \sum_{i=1}^{n} \frac{(X_i - \mu_x) * (Y_i - \mu_y)}{(n-1)}$$
- If S_{xy} is positive (large) then X and Y increase together. If it is negative, then X and Y are inversely related. If it is 0, then there is no relationship.

Covariance



http://ci.columbia.edu/ci/premba_test/c0331/s7/s7_5.html

Covariance and correlation

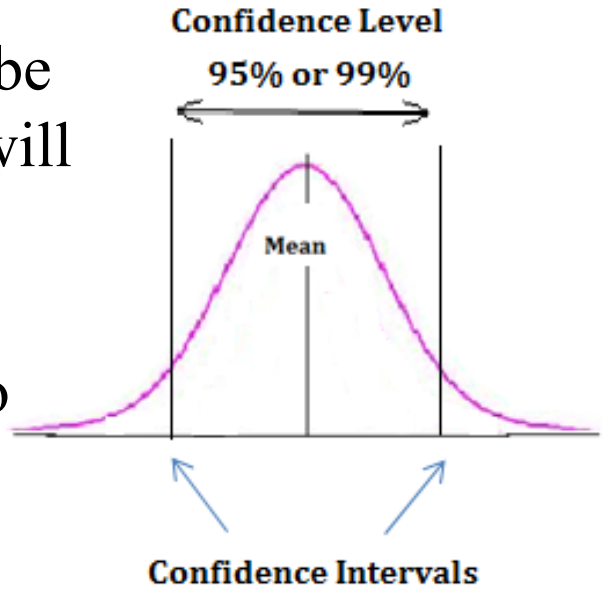
- Correlation also tells you the degree to which the variables tend to move together
- Strength of the relationship
- Correlation R_{xy} is calculated by:

$$R_{xy} = \frac{S_{xy}}{\sigma_x \sigma_y}$$

- It turns the Correlation into a number between -1 and +1

Confidence Intervals

- Confidence intervals describe the uncertainty of a parameter. If you repeatedly take a small sample to measure, your values will always be different. The mean you repeatedly sample will also follow a normal distribution.
- The confidence intervals calculates the probability that the actual mean falls close to your measured number.



Confidence Intervals

- The formula for calculating the confidence interval is: $\mu \pm z_c \left(\frac{\sigma}{\sqrt{n}}\right)$
- Z_c is the "critical value" for difference confidence levels: 90% is 1.645, 95% is 1.96, 99% is 2.575.
- This computes the limits for 90, 95 or 99% of the data.

- The 90% confidence interval says there is a 90% chance that the true mean falls within the range you have measured +/- some error
- The 95% confidence interval says there is 95% chance that it falls within a larger range.

Review

- Don’t memorize the formulas for probabilities! Just be familiar with the names: Uniform, Gaussian/Normal.
- What are rank statistics and percentiles?
- Know the terms: mean median, mode and variance
- Learn the formula for calculating mean and standard deviation for a data set.