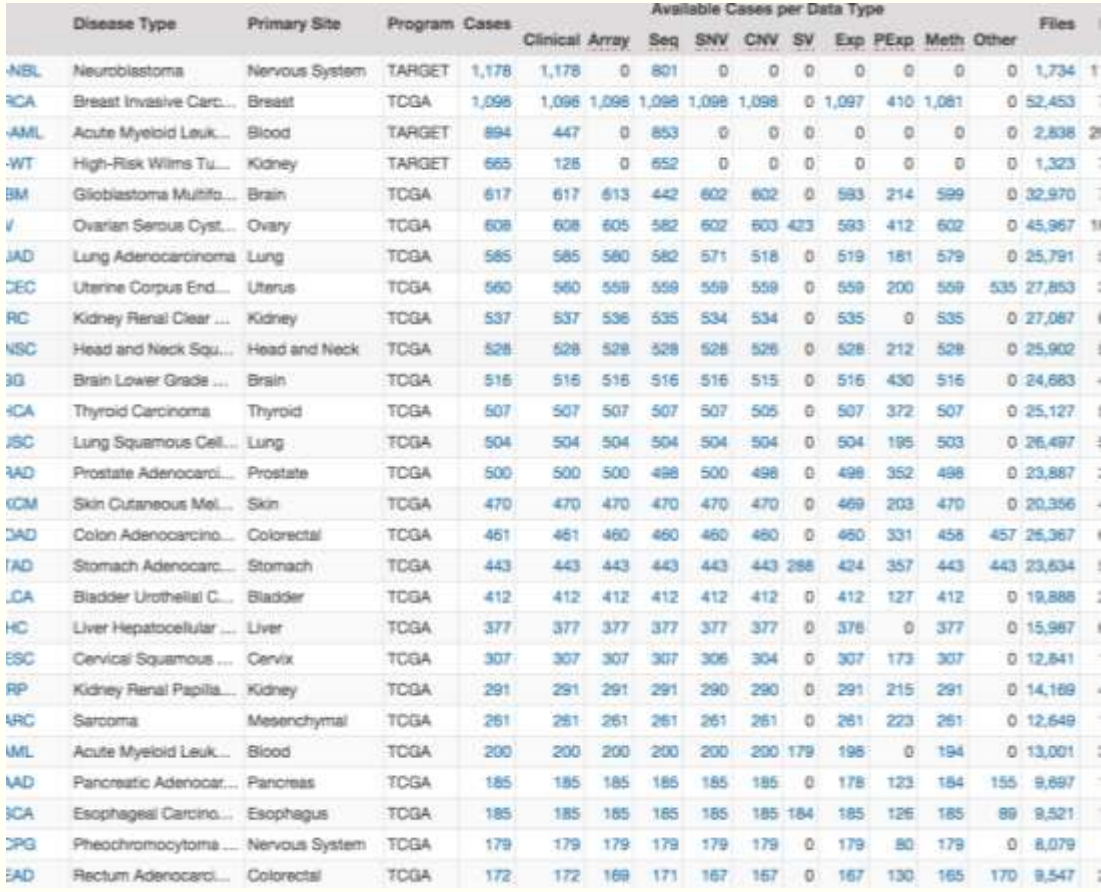
UNDERSTANDING DATA & PROBABILITY

Lecture 1A



	Disease Type	Primary Site	Program	Cases	Available Cases per Data Type										Files	
					Clinical	Array	Seq	SNV	CNV	SV	Exp	PExp	Meth	Other		
NBL	Neuroblastoma	Nervous System	TARGET	1,178	1,178	0	801	0	0	0	0	0	0	0	1,734	1
BCA	Breast Invasive Carc...	Breast	TCGA	1,098	1,098	1,098	1,098	1,098	1,098	0	1,097	410	1,081	0	52,453	1
AML	Acute Myeloid Leuk...	Blood	TARGET	894	447	0	853	0	0	0	0	0	0	0	2,838	2
WT	High-Risk Wilms Tu...	Kidney	TARGET	665	126	0	652	0	0	0	0	0	0	0	1,323	1
BM	Glioblastoma Multifo...	Brain	TCGA	617	617	613	442	602	602	0	593	214	599	0	32,970	1
V	Ovarian Serous Cyst...	Ovary	TCGA	608	608	605	582	602	603	423	593	412	602	0	45,967	1
LAD	Lung Adenocarcinoma	Lung	TCGA	585	585	580	582	571	518	0	519	181	579	0	25,791	1
UEC	Uterine Corpus End...	Uterus	TCGA	560	560	559	559	559	559	0	559	200	559	535	27,853	3
KRC	Kidney Renal Clear ...	Kidney	TCGA	537	537	536	535	534	534	0	535	0	535	0	27,087	1
HNSC	Head and Neck Squ...	Head and Neck	TCGA	528	528	528	528	528	526	0	528	212	528	0	25,902	1
LGG	Brain Lower Grade ...	Brain	TCGA	516	516	516	516	516	515	0	516	430	516	0	24,683	1
THCA	Thyroid Carcinoma	Thyroid	TCGA	507	507	507	507	507	505	0	507	372	507	0	25,127	1
LUSC	Lung Squamous Cell...	Lung	TCGA	504	504	504	504	504	504	0	504	195	503	0	26,497	1
PRAD	Prostate Adenocarci...	Prostate	TCGA	500	500	500	498	500	498	0	498	352	498	0	23,887	2
SKCM	Skin Cutaneous Mel...	Skin	TCGA	470	470	470	470	470	470	0	469	203	470	0	20,356	1
COAD	Colon Adenocarcino...	Colorectal	TCGA	461	461	460	460	460	460	0	460	331	458	457	26,367	1
STAD	Stomach Adenocarc...	Stomach	TCGA	443	443	443	443	443	443	288	424	357	443	443	23,634	1
BLCA	Bladder Urothelial C...	Bladder	TCGA	412	412	412	412	412	412	0	412	127	412	0	19,888	2
LIHC	Liver Hepatocellular ...	Liver	TCGA	377	377	377	377	377	377	0	376	0	377	0	15,987	1
CESC	Cervical Squamous ...	Cervix	TCGA	307	307	307	307	306	304	0	307	173	307	0	12,841	1
KIRP	Kidney Renal Papilla...	Kidney	TCGA	291	291	291	291	290	290	0	291	215	291	0	14,169	1
SARC	Sarcoma	Mesenchymal	TCGA	261	261	261	261	261	261	0	261	223	261	0	12,649	1
AML	Acute Myeloid Leuk...	Blood	TCGA	200	200	200	200	200	200	179	198	0	194	0	13,001	2
PAAD	Pancreatic Adenocar...	Pancreas	TCGA	185	185	185	185	185	185	0	178	123	184	155	9,697	1
ESCA	Esophageal Carcino...	Esophagus	TCGA	185	185	185	185	185	185	184	185	126	185	89	9,521	1
PCPG	Pheochromocytoma ...	Nervous System	TCGA	179	179	179	179	179	179	0	179	80	179	0	8,079	
READ	Rectum Adenocarci...	Colorectal	TCGA	172	172	169	171	167	167	0	167	130	165	170	9,547	2

Data

- Result of measurements on a set of objects
- Can be of two types – quantitative (numerical) or qualitative
- The *level of measurement* dictates the types of analysis that can be performed on the data (or scope for analysis) and consequently the analytical value of (or from) the data
  - Analytical value can take many forms but we will use it to mean the amount of information we can extract from the data through analysis
- Level of measurement (lowest to highest): Nominal, Ordinal, Interval, Ratio (NOIR)

Levels of Measurement / Variable Types

- **Nominal** – Can be assigned to categories – *Gender*
- **Ordinal** – Can be categorized and ordered – *Education*
- **Interval** – Can be categorized, ordered and allows for meaningful interpretation of intervals but not ratios of values. True interval scales are rare.
  - *Example:* Fahrenheit scale for temperature as a measure of warmth. Equal differences on this scale represent equal differences in temperature (15 degrees is 5 degrees warmer than 10 degrees, and is the same as the difference between 25 degrees and 20 degrees), but a temperature of 30 degrees is not twice as warm as one of 15 degrees.
  - This is because zero degrees Fahrenheit is not the same as no warmth. That is we do not have a true zero and zero temperature does not mean no temperature (Dates are also measured on an interval scale; similarly *Credit Score* does not have a zero; the values range from 200 to 800)
- **Ratio** – Can be categorized, ordered and allows for meaningful interpretation of intervals, as well as ratios of values – *Age, Income, Net Worth, Sales*

Table 1

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Levels of Measurement – Scope for Analysis

- Ratio data have the greatest scope for analysis and nominal data the least. Table 2 below summarizes possible types of operations on the data.
  - Higher the level, greater the scope for analysis, and higher the analytical value
  - Remember: If you convert data from a higher level to a lower level of measurement (say convert Age to categories) YOU WILL LOSE ANALYTICAL VALUE

Table 2

Data Type	Operators	Operations
Nominal	=, !=	Grouping, Mode
Ordinal	Also >, <	Sorting, Median
Interval	Also +, -	Median, sometimes Mean, standard deviation etc.
Ratio	Also *, /	Many types of operations

Table 1

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Probabilistic/Statistical View of Data

- We can view the data in Table 1 as a collection of column variables using the following mapping:
  - {Age. Gender, Education, Credit Score, Income, Net Worth, Sales} → $\{X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$.
- Further, the values taken by each of the variables has an associated *probability.*
- Such variables are called *random variables* and form the basis of a probabilistic/statistical view of data.
- The collection of values and the associated probabilities comprise a *probability distribution of a random variable.*

Table 1

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Type of Data and Statistical Method

- We can categorize the major methods in this course depending on the type of data
- Most of the models will use a one or more independent (predictor) variables to explain and predict dependent variables

Dependent Variables	Outcome	Predictor Variables	Statistical Method of Analysis	Estimation Method
Ratio (Continuous)	Value Prediction	Categorical (Nominal or Ordinal) and/or Continuous	Regression	Least Squares
Ratio (Continuous)	Value Prediction	Categorical (Nominal)	ANOVA (Regression)	Least Squares
Nominal or Ordinal (Categorical)	Association (Dependence)	Categorical (Nominal)	Contingency Table	Chi-Square
Nominal or Ordinal (Categorical)	Category Value Probability, Classification	Categorical and/or Continuous	Logistic Regression	Maximum Likelihood

Type of Data and Statistical Method - Example

- Predict Income (continuous) based on Net Worth (Continuous) & Gender (Categorical) –Regression
- Predict Income Differences (continuous) based on Gender (categorical) – ANOVA or Regression
- Predict/Classify Probability of Gender (categorical) given Education (categorical) – Contingency Table
- Predict/Classify Probability of Gender (categorical) given Income (continuous) – Logistic Regression

Table 1

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Other Questions you can Answer knowing Probability Distributions

- If you choose 10 customers, what is the probability that there are no more than 3 females?
- If you choose 10 customers, what is the probability that there will be 2 PhDs, 3 Masters, 3 Bachelors and 2 HS?
- What is the probability that you will see 9 non-PhD customers, before you see the first PhD?
- Many such questions…

Table 1

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Differences between this Course and a Traditional Statistics Course

	Statistics for Data Science	Traditional Statistics Course
Data Source	Secondary Data Sets	Primary Data
Data Selection and Collection	Transactions from Processes (Business, Health, Financial)	Random Sampling & Variations; Surveys (such as polls)
Data Set Size	Medium to Large	Small
Study (Research) Design	Observational/Non-experimental	Experimental
Focus	Models for Prediction/Classification/Patterns	Models for Statistically Significant Effects
Role of Statistical Significance	Supportive, but secondary	Primary
Nature of Study	Often exploratory	Often confirmatory
Goal	Understanding what model is telling us	Obtaining valid cause-effect conclusions

Commonalities with Traditional Statistics

- Concepts of Probability and Random Variables – Probabilistic View of Data
- Probability Distributions of Random Variables
- Concepts of Sample Statistics and Population Parameters
- Statistical Relationships (such as correlations)
- Choice of Method based on nature of variables
- Statistical Models such as Regression, ANOVA, Logistic Regression, Contingency Tables
- Model Selection and measures of Model Effectiveness
- Sampling distributions, confidence intervals and tests of significance