**FUNDAMENTALS of DATA SCIENCE**

(From Booz Allen Fundamentals of Deep Learning)

Identifying the unseen

Advanced computing enables us to make connections like never before. We can

teach machines to detect cancer, identify cybersecurity threats before they happen,

and optimize business operations – the possibilities are endless. So how can

machines continue to learn, without the dedication of abundant resources?

Deep learning, a sub-mechanism of artificial intelligence (AI), is a computing process that enables

machines to find patterns in data. Deep learning sifts through data and makes connections, identifying

odd trends that would be unfindable by the human senses. Imagine a world in which we could identify

risks before they happen, capitalize on trends that could save lives, and rely on machines to make

conclusions based off of data and fact, instead of emotion and perception – that’s the world of deep

learning.

**INTERMEDIATE STATISTICS**

What happens when you’re given sample data and you’re asked to infer from it answers to specific questions?

**SAMPLE**

Under the right circumstance, the sample can act as a representative of the entire population.

Considerations regarding quality of the sample: Sample size, How the sample was chosen (random),

**Simple Random Sample**: each individual has the same probability of being chosen at any stage. Any subset of k individuals has the same probability of being chosen as any other subset containing k individuals.

Must be **Unbiased** and the data points must be **Independent** (the selection of one member must not influence the selection of other members).

Is the sample representative of the population?

The greater the sample size, the smaller the range of the confidence interval becomes. We become more confident with larger sample sizes.

**Central Limit Theorem**

The more samples we take, the closer the means of our sample means will get to the population mean.

The sampling distribution of our sample means starts to look like a normal distribution. As sample size increases, the curve becomes more normal (taller and narrower, and the standard deviation gets smaller).

The mean of the means approaches the mean of the population.

With enough random samples, we get an excellent approximation of the population mean.

(The Central Limit Theorem suggests that) **The larger the sample size, the more confidence we have in our results.**

The average of the sample means comes with a standard error.

The **Standard Error** is the standard deviation of our proportion distribution. (The standard deviation of the sample means is our estimated standard error). With larger sample sizes, our standard error gets smaller.

The Central Limit Theorem is important as we use sample means to approximate the population mean.

**CONFIDENCE INTERVALS**

How confident are we that a single sample mean is near our actual population mean?

One random sample will allow us to calculate a range and attach to it a level of confidence.

The **Confidence Interval** indicates a level of confidence that the population’s actual results will fall within a range of values relative to the sample’s results. They provide a level of confidence for a given interval.

z-scores to us how many standard deviations away from the mean we would need to eb to capture a certain percentage of the total distribution.

Sample…. [10, 15, 9, 18, 12, 8, 12]

Sample mean

Sample standard deviation

**Sampling Error** = sample stdev/sqrt(n) [for proportions: sample proportion/sqrt(n)]

Confidence Interval (95%) for population mean:

**Upper limit** = sample mean (or proportion) + 1.96 \* (sampling error)

**Lower limit** = sample mean (or proportion) - 1.96 \* (sampling error)

**HYPOTHESIS TESTING**

50% men, 50% women in population

50 individuals are chosen for jury duty

The chosen sample/pool contained 14 men and 36 women

Was this randomly by chance?

Binomial Random Variable

n = number of trials

p = probability of outcome

4-step Process:

1a.) Develop Hypotheses (typically there are 2)

**H0** = Null Hypothesis (status quo, no significance to results)

Jury numbers happened by chance

Ho is p <= .5 for women to be chosen for jury duty

**Ha** = Alternative Hypothesis

Jury numbers are not by chance

Odds of women being chosen for jury duty is greater than 50%

Ha is p > .5 for women to be chosen for jury duty

1b.) State Significance Level

Significance level = 5%

If 36 or more women ending up on a jury have less than a 5% chance of occurring at random,

then we will reject our null hypothesis.

2.) Identify a Statistic That Will Assess the Validity of the Null Hypotheses

In this case, Binomial Probabilities

p=.50 (probability of a woman being chosen for the jury panel) (aka, Binomial Random Variable)

50 trials

Number of successful trials is 36

3.) Determine P value (for the test statistic)

This p-value is the probability that 36 women or more would be chosen by chance out of 50 trials.

p = .13%

4.) Compare P-value to significance level

There was only a .13% chance that, at random, 36 or more women would be chosen for a panel of 50 potential jurors.

Our chosen Alpha (significance level) was .05

p-value < significance level so we would reject the Null Hypothesis

Therefore, we can conclude that it is much more likely for a woman to be chosen versus a man.

**T-Distribution (small sample sizes, z-score is not valid))**

Sample size < 30

**z-test**: compares mean of sample to the population mean, sample comes from the population

**t-test**: compares 2 independent samples (they don’t have to come from the same population)

Not 1 curve, but multiple curves for different sample sizes

The critical scores for t-distributions are higher than for z-distributions

-----------------------------------------------------------------------------------------------------------------------------------------

**HYPOTHESIS TESTING**

Basic principle:

What is the probability of X happening by chance (if randomness is the only explanation)?

If the probability is low, you reject randomness as a likely explanation.

H0: Null Hypothesis

No systematic effect and random sampling error is the only explanation.

Ha/H1: Alternative Hypothesis

There is a systematic effect between the group means, association between the variables

**EXPLORATORY DATA ANALYSIS**

Sets the stage for modeling

**Visual Exploratoration**

Exploration is always a critical first step in any good data analysis

Precursor to numerical exploration

Review Data, Check Assumptions, Check Anomalies, Data Suggestions

Determine if your data appropriate to reach your conclusions?

Exploratory Graphs/Charts

Information Dense: they communicate so much more so much faster

Check for Shape, Gaps, Outliers

Start with single/univariate distributions (histogram)

Then evaluate joint/bivariate distributions (scatterplot)

Look for **unusual** cases (exceptional values)

Look for **errors** in the data (check for patterns)

Look for **missing** data (check for patterns, impute)

Bar Plot

Histograms

For quantitative variables

Shows the shape of then distribution

Box Plots

For quantitative variables

Good for identifying outliers ( a number which is less than Q1 or greater than Q3 by more than 1.5 times the interquartile range ( IQR=Q3−Q1 ).

Scatter Plot (Pair Plot)

Shows associations between quantitative variables

Multicollinearity: try using fewer variables/features or combining features (based on domain knowledge)

Conclusion:

Use a method that you find quick and easy

When you’re going through your charts you are determining:

Do you have what you need to answer your questions?

Are there gaps in the data?

Are there exceptional cases (outliers)?

Are there error in the data?

**Numerical (Statistical) Exploration**

First Graphics, then numbers

You’re still exploring, you’re not modeling the data

It’s good to get multiple perspectives on the data