# Data Science

Naga Nitish ◇ May 2020

## 1 Probability

first cue?

second cue?

### Introduction

- Probability is the liklihood of event occuring
- Trail – Observing an event occur and note the outcome
- Experiment – Collection of trails
- Expected value – The outcome we expect from an experiment
- Probability frequency distribution – collection of probabilities of each possible outcome of an event
- Permutations – represents the number of different possible ways we can arrange a set of elements – $n!$
- Variations – represents the number of different possible ways we can pick and arrange a number of elements
  - With repetition – $n^p$
  - Without repetition – ${}^nP_p = \frac{n!}{(n-p)!}$
- Combinations – represents number of different possible ways we can pick elements
- Baye's theorem – $P(A|B) = P(B|A) * P(A)/P(B)$

### Distributions

There are two types of distributions: They are Discrete and Continuous

**Discrete**

1. Uniform
2. Bernoulli
   - one trail – two possibilities
   - $E(Y) = p$
   - $Var(Y) = p(1-p)$
3. Binomial
   - measures the frequencies of occurrence of one of the possible outcomes over n trails
   - $P(Y = y) = C(y, n) \times p^y \times (1-p)^{n-y}$
   - $E(Y) = n \times p$

- $Var(Y) = n \times p \times (1-p)$

4. Poisson
   - measures the frequency over an interval of time or distance
   - only non-negative values
   - $P(Y = y) = \frac{\lambda^y}{y!e^{-\lambda}}$
   - $E(Y) = Var(Y) = \lambda$

**Continuous**

1. Normal
   - bell shaped, symmetric, thin tails
   - $E(Y) = \mu$
   - $Var(Y) = \sigma^2$
2. Students' T
   - a small sample size approximation of normal distribution
   - bell shaped, symmetric, flat tails
   - accounts for extreme values better than normal distribution
   - $Var(Y) = s^2 \times \frac{k}{k-2}$
3. Chi squared
   - asymmetric, skewed to right
   - it is square of T distribution
   - $E(Y) = k$
   - $Var(Y) = 2k$
4. Exponential
   - Both PDF and CDF plateau after certain point
   - $E(Y) = \frac{1}{\lambda}$
   - $Var(Y) = \frac{1}{\lambda^2}$
5. Logistic
   - The smaller the scale parameter, the quicker it reaches 1.0
   - $E(Y) = \mu$
   - $Var(Y) = \frac{s^2 \times \pi^2}{3}$

third cue?

**Summary**

1. something random
2. something random
3. something random

# 2 Statistics

### Types of Data

**Qualitative data or categorical data**

- Nominal - values not order
- Ordinal - there is order or ranking

**Quantitative data**

- Discrete
- Continuous

### Types of statistics

**Descriptive**

- To describe data
- Measure of central tendencies
  - **Mean or average** – sum of all values divided by number of values
  - **Median** – middle term in the sorted list
  - **Mode** – value with highest frequency
  - **Mid-range** – average of largest and smallest value
- Measure of dispersion
  - **Range** – largest minus smallest value
  - **Standard deviation** – square root of variance
  - **Variance** – average of squared differences of the mean
- Frequency distributions
- Histograms
  - It's a bar graph with equal width
  - Properties – symmetric, skewed and uniform or rectangular

**Inferential**

- To make inferences from data
- Hypothesis testing
- ANOVA
- Chi-squared tests
- Regression

**Some important points**

- **Skewness** – Left (negative) skewness means that the outliers are to the left
- **Covariance** – It is joint variability of two variables

$$\sigma_{xy} = \frac{\Sigma_{i=1}^{N}(x_i - \mu_x) \times (y_i - \mu_y)}{n-1}$$

- **Correlation**

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

## Central Limit Theorem

The Central Limit Theorem (CLT) is one of the greatest statistical insights. It states that no matter the underlying distribution of the dataset, the sampling distribution of the means would approximate a normal distribution. Moreover, the mean of the sampling distribution would be equal to the mean of the original distribution and the variance would be n times smaller, where n is the size of the samples. The CLT applies whenever we have a sum or an average of many variables (e.g. the sum of rolled numbers when rolling dice)

- **Estimator** is a mathematical function that approximates a population parameter depending only on sample information
- **Estimate** is the output that we get from estimator. Point estimate and confidence interval estimate

**Confidance interval estimate**

With population variance

$$\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Without population variance

$$\bar{x} \pm t_{n-1,\alpha/2} \times \frac{s}{\sqrt{n}}$$

where standard error is $s/\sqrt{n}$

### Summary

1. something random
2. something random
3. something random
4. something random
5. something random