**Statistics: *SciPy***

* Analyze and explore large data
* How to find Random difference?
* Test data, find facts

1. **Descriptive statistics**: summarize data (entire data)
2. **Inferential statistics**: measured conclusion (sample data)

* **Sampling:** it’s **subset of entire dataset**, select some records from population.
  + **Random:** shuffle and random pick-up record (exit poll population)
  + **Stratified:** solve class imbalance problem (male-female)
  + **Systematic:** kth element pick-up record
  + **Convenient:** condition / particular field based
* **Central Tendencies:** Indicate where does the **Center of the distribution** of our data.
  + **Mean, Median, Mode** (mean ~ median means it normally distributed)



* **Spread / Dispersion & Variability:** Heavily **spread out or not** it’s described.
  + **Range:** max – min
  + **Percentiles**: 100 records in x’s 90PR means 90 records have below x
    - **Find value for particular PR:** val\_idx = PR/100\*(n+1) 🡺 find value
  + **Quartile:** find outliers
    - **IQR**: Q3 - Q1 inter quartile range



* + - **Lower** **Fences**: (Q1-1.5\*IQR) / **Upper Fences:** (Q3+1.5\*IQR)



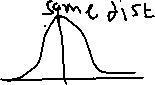
* + **Variance:** (**E** (xi - X\_mean) \*\*2) / N



* + - **Sample var:** (**E** (xi - x\_mean) \*\*2) / n-1 :. N-1 means degree of freedom
  + **Standard Deviance:** sqrt(var)



* **Gaussian/Normal Distribution:**



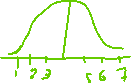
* + **Bell curve**, mean=median=mode
  + **Emperical formula**:



* + - Mean +- 1std (68.3%)
    - Mean +- 2std (95.4%)



* + - Mean +- 3std (99.7%)
  + What std fall of particular x value: **Z-score**=



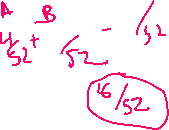
* **Standard Normal Distribution:** mean=0, std=1
  + It’s find using z-score
  + If dataset have many units so using z-score make SND it’s called **Standardization**
  + #**Normalization:** it’s gives to option by make upper and lower bound of data (lower range to upper range) ***$$MinMaxScaler*** It uses in CNN DL if pixel is 280\*280 means (0-280) 🡺 (0-1) every value divide by max range
* **Probability:** it’s measure of the likelihood of an event.



* + **Additional rule:** P(A|B) = P(A) + P(B) – P(A^B)



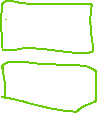
* + **Multiplication rule:**



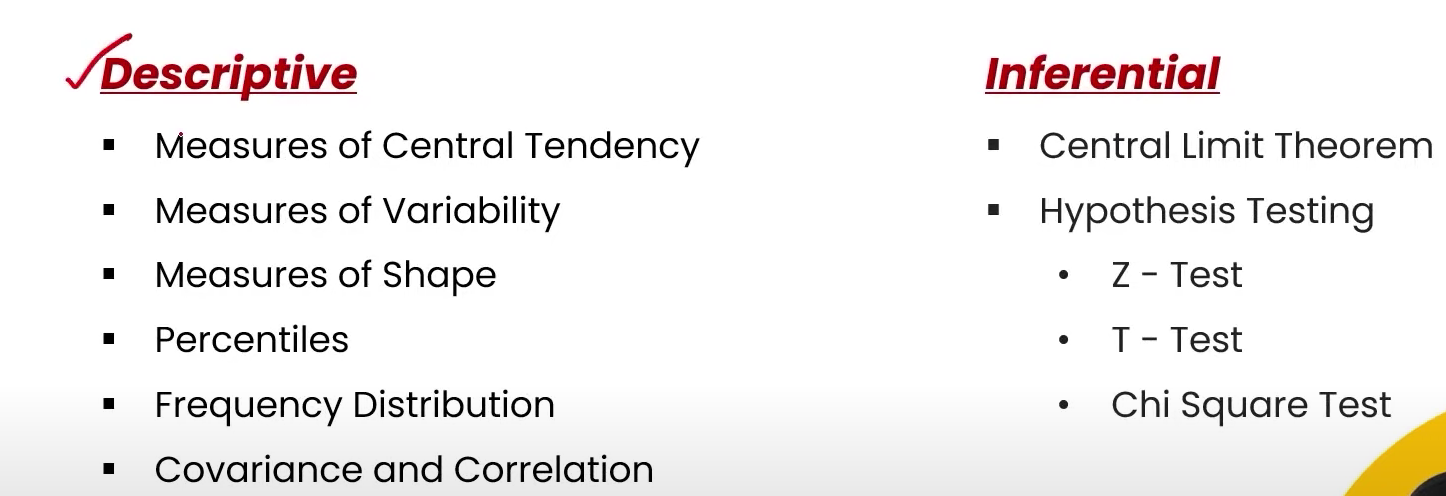
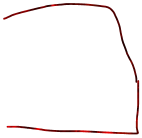
* + - **Independent event:** rolling dice 2 times P(A&B) = P(A)\*P(B)



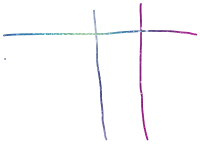
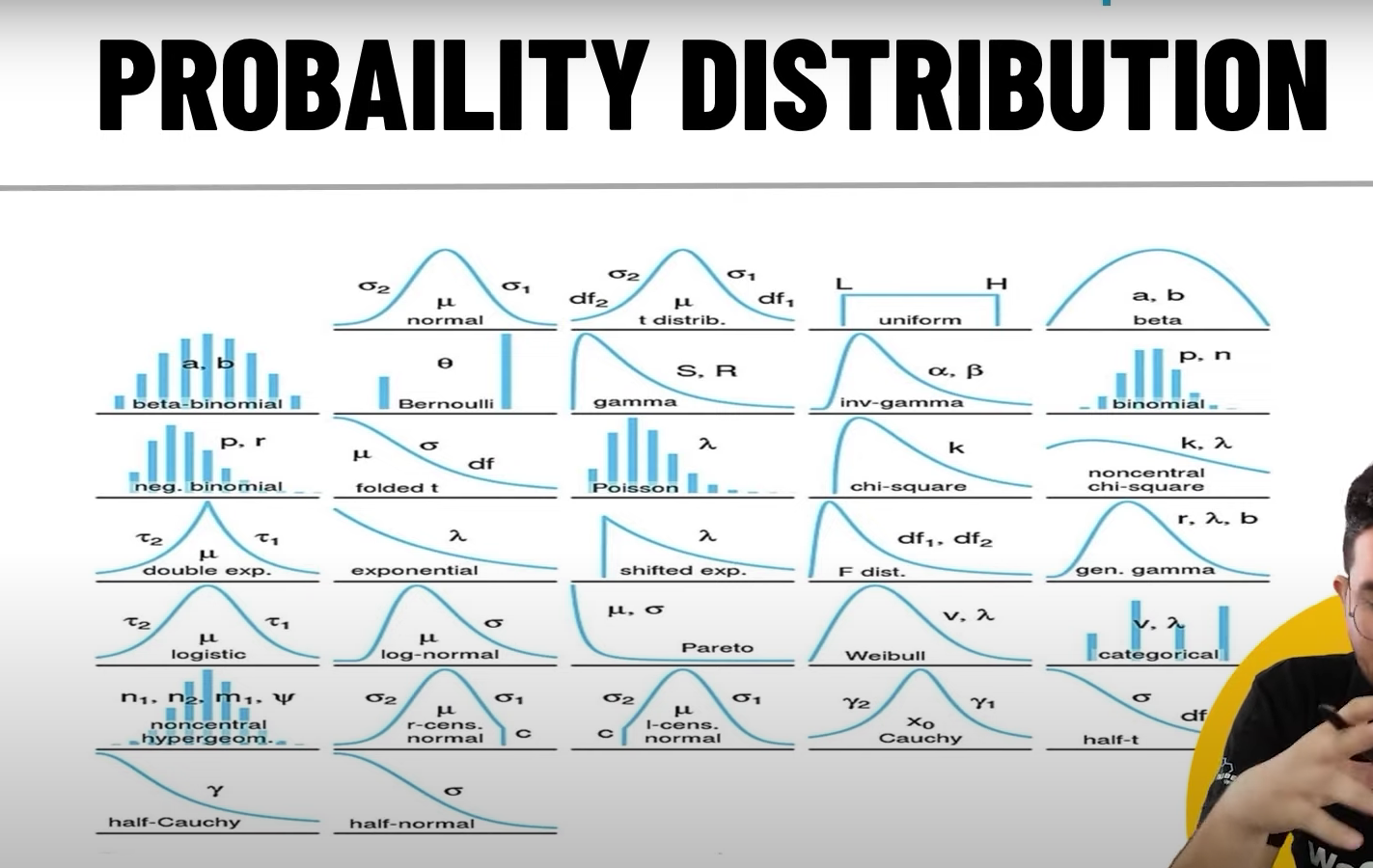
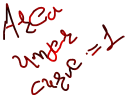
* + - **Dependent event:** first collect one black ball and then Next select red ball **$$Naïve bayes:** P(B|A) = P(A|B) \* P (B) / P(A)



* + **Permutation:**
    - nPr = n! / (n-r)! **:.** N=total r=how many types of combination make
  + **Combination:** reshape unique combination
    - **nCr = n! / r! (n-r)!**
* **P value:** probability of this point

****



* + **Mean Absolute Division:** MAD=E(xi-x\_mean)/n low spread value very imp in best performance
  + **Measures of Shape:** 
    - **Skewness:** measureasymmetry of the distribution :. **sk = df[col].skew()**
      * **0 means** perfectly symmetric dist**. sk=0**
      * ******Positive means** dist. is skewed to the right. Tail longer on right, mean > median > mode **sk>0**
      * **Negative means** dist. is skewed to the left. Tail longer on left, mean < median < mode **sk<0**
  + **Frequency Distribution & Cumulative Distribution:**
    - Cumulative means sum of frequency one by one
  + **Probability Distribution Function:**
    - **Prob. Density func. (PDF):** use for cont. data. For particular range how many % of chances comes.
    - **Prob. Mass func. (PMF):** use for discrete data. Freq of chance

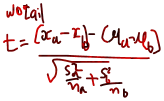


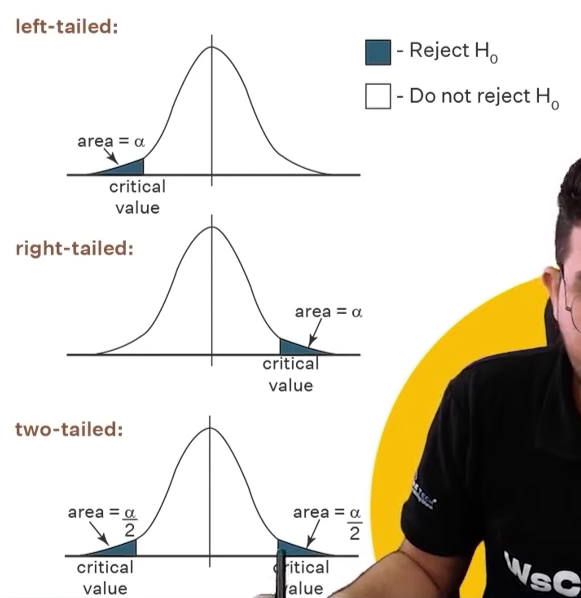
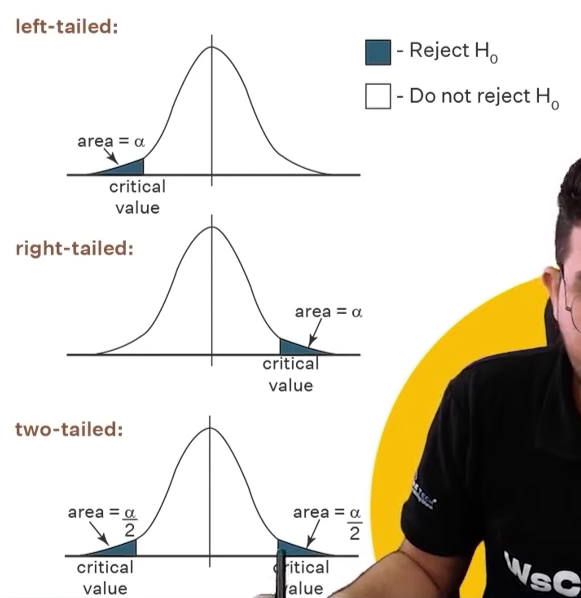
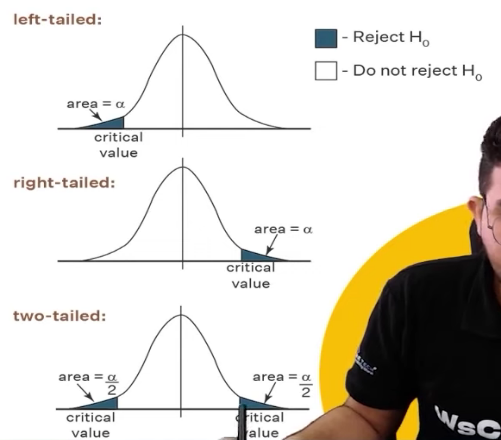
* + - **Cumulative Density func. (CDF):**



* **Central Limit Theorem:**



* + Population data if it’s normal / not-normal distribution dataset.
  + Then make samples and find means always all have normal distribution.
  + all sample means to make mean that main mean of population data.
  + **Assumption**:
    - Random sampling and sample size also < 10% population and n>30
    - Multicollinearity not allowing
* **Hypothesis Testing:**
  + Prev and new product/data test and check new is best or not.
  + **Null hypothesis (H0):** suggest no statistically significant exits b/w population.
    - Old is better than new or equals (=)
  + **Alternative hypothesis (Ha):** suggest there is a significant diff b/w population.
    - New is better than old or not equals (!=, >, <)
  + **Testing:**
    - **Z-Test:** Work on avg.,
      * known for population’s mean & std
      * sample data n > 30, sample’s mean
    - **T-Test:** Work on avg.,
      * known for population’s mean
      * sample data n < 30, sample’s mean & std
      * df=n-1, one/two tail
    - **Chi-Square Test:** Suggest goodness | Find relation b/w two data
  + **Practical steps:**
    - State find H0 and Ha:
    - Choose level of significance: (a/alpha) (100 - confidence interval)
      * Based on domain expert 1%-5% (mostly used 0.05)
      * Always assume situation has normal distribution



* + - Find test statistic: Z-test / T-test / Chi-square test
    - Find critical values: Z-values / T-values using formula table
    - Draw conclusion: H0/Ha right (accept / reject null hypothesis)
      * **Test > Value means** **Ha is right reject Null hypothesis.**
    - Type-1 error: reject Ho when reality is true
    - Type-2 error: accept Ho when reality is false
  + **Confidence Interval:**
    - Point Estimate +- margin of error
    - Upper bound / lower bound