Statistics Notes for week -1

What is Statistics ?

Statistics covers the study of the science of collection, organization, analysis, interpretation, and presentation of numerical as well as categorical data. (or in other words Qualitative and Quantitative data)

Statistics are used in virtually **all scientific disciplines** such as the physical and social sciences, as well as in business, the humanities, government, and manufacturing.

Goal of statistics Is to study a large collection of people or Objects (Eg: opinion poll for a political party or finding out what proportion of the people support a particular political party)

But, the challenge is it's Infeasible, Expensive and time consuming

Solution: Survey only a few elements and draw inference about all elements from this smaller group

Key Terminologies:

Population: Total collection of Objects that we want to study

Sample: It's a subgroup of the population that we study to draw inference about the population

Parameter: is what behavior we want to study in the sample to make inference about population

Statistic: Proportion, Mean, Median, standard deviation, variance when computed from a sample is called a statistic

Selecting a Sample: A Good Sample is a representative of the population and hence the resulting statistics will be useful.

Different sampling strategies are:

- Simple random sampling
- Stratified sampling
- Clustered sampling

Designing an experiment:

Purpose of the experiment is to study if administering the drug for 30 days reduce sugar level.

How do I design an experiment to collect the data:

- Select a group of volunteers/Subjects
- Measure there sugar levels today
- Administer the drug
- Measure the sugar level after 30 days

What is wrong in the above experiment:

Group might contain people who have the following habits:

- Physical exercise
- Smoking
- Alcohol

This habits will influence the sugar levels directly, hence will contribute to a misleading outcome.

Note: While studying the effect of one variable (Medicine) on another (Sugar Level) we must ensure that we nullify the effect of lurking variables like Smoking, alcohol, exercise etc.

We achieve this using randomized control experiments.

- Explanatory, response and lurking variables
- Treatments, control groups, placebo
- Single blind and double blind experiments

Describing and summarizing data

User ID	Age Group	Genre	# of hours
0001	20-25	Drama	60
0002	25-30	Romantic	30
0003	20-25	Thriller	25

In tabular format its very difficult to answer simple questions

- What is the minimum/maximum number of hours spend by age group 20-25 in watching movies
- Are there more users in the lower range (0-5hrs) or higher range (89-110hrs)
- Is the data clustered at the center (Most users in the range of 45-67 hrs) and very few at the 2 extremes

Other alternative is creating visualization of data using plots and graphs:

Plotting Graphs / Histogram

Compute mean, median, mode, variance & Std Deviation

Drawing plots & computing summary allows us to quickly get a feel for the data

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables kurtosis, and skewness.

Descriptive Statistics;

- Relative frequency charts
- Frequency polygons
- Histograms
- Stem & leaf plots
- Box plots
- Scatter plots
- Measures of centrality & Spread

Why do we need probability theory?

probability theory, a branch of mathematics concerned with the analysis of random phenomena. The outcome of a random event cannot be determined before it occurs, but it may be any one of several possible outcomes. The actual outcome is considered to be determined by chance.

The word *probability* has several meanings in ordinary conversation. Two of these are particularly important for the development and applications of the mathematical theory of probability. One is the interpretation of probabilities as relative frequencies, for which simple games involving coins, cards, dice, and roulette wheels provide examples.

A Sampling strategy is said to be truly random (unbiased) if every element in the population has equal chances of becoming a part of sample

Probability is a Branch of mathematics that deals with chances

How many ways can I create sample of 2 for a population of 10 , 90 possible ways

If you observe some trend in this small sample what is the chance that you will observe the same trend in other samples or the entire population, for this also we need probability theory.

Topics we will cover in this course:

- Introduction to probability theory
- Sample spaces
- Events
- Axioms of probability
- Desecrate and continuous random variables
- Bernoulli Uniform and Normal distribution

Giving guarantees for estimates

Given a Population (10 Students) and creating a different Samples of (2 Students)

We can arrive at a estimated mean height

“The mean itself has a probability distribution (different values of mean have different chances of being observed)”

“If the mean computed from a single sample is x can you give an interval such that you are 95% sure that the mean of the population lies in this interval”

Topics to be covered:

- Point Estimates
- Distribution of sampling statistics
- Interval estimates

Hypothesis & How to test it

Bowling speed(mph)

100 85 87 98 90 81

Hypothesis: this means bowling speed of Bumrah is greater than 90 mph

This is true based on sample

However: we are estimating from sample (what if Bumrah got lucky and this sample was good)

Mean itself has a distribution (different samples, different means)

"I reject the hypothesis that the mean speed of Bumrah is greater than 90 mph because there is a 25%" chance that I might get a sample in which the mean speed is greater than 90 mph even if the true mean is less than 85 mph"

Hypothesis Testing:

Single Population, Two Populations, Multiple populations

Z-Tests, T-Tests, Analysis of Variance (ANOVA)

Modelling relationship between variables

"What is the relationship between number of days of treatment and sugar level?"

Dealing with uncertainty as m and c are estimated from sample and not population

"Are we 99% sure that the value of m estimated from sample lies with in a small neighborhood around the true value of m"

Statistical Modelling:

Assume a simple relationship between the variables

y = mx + c (where y = decrease in sugar level, x = number of days of treatment)

Relationship is linear

Linear Regression

- Estimating parameters
- Estimating confidence bands
- Measuring goodness of fit

Hypothesis: In Cricket, the five ways of getting dismissed are equally likely

How does this Model fit the data ?

Estimate probabilities form data (say last 100 dismissals: Sample)

Are the variations observed in the sample significate or due to random chance ?

Chi – Square Test

- Determine Goodness of Fit
- Determine if 2 variables are independent

Summary: List of Topics

Descriptive Statistics

- ✓ Different types of data
- ✓ Different types of plots
- ✓ Measure of centrality and Spread

Probability Theory

- ✓ Sample Specs, events, axioms
- ✓ Discrete and continuous RVs
- ✓ Bernoulli, Uniform, Normal dist
- ✓ Sampling strategies

Inferential Statistics

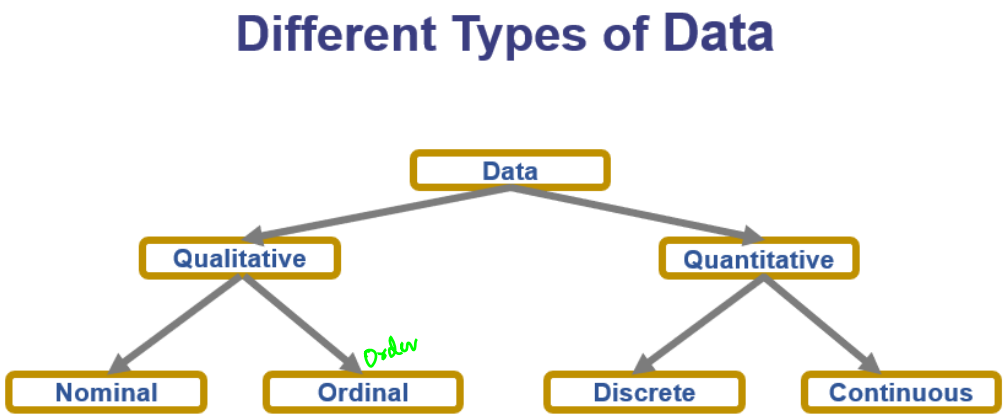
- ✓ Interval Estimators
- ✓ Hypothesis testing (z-test, t-test)
- ✓ ANOVA, Chi-square test
- ✓ Linear Regression

Descriptive Statistics

Different types of data

The questions we are trying to answer:

- What are different data types ?
- How do we describe qualitative data ?
- How do we describe quantitative data ?
- How do we describe relation between attributes ?



Qualitative or categorical attributes are those which describe the object under consideration using a finite set of discrete classes

Nominal: There is no natural ordering of these attributes

Ordinal: There is a natural ordering in these attributes

Eg: S < M < L

Some examples for Qualitative data

	Nominal	Ordinal
Employee	Gender (M/F)	Income Range (Low, Med, High)
Health care	Disease (contagious, Non contagious)	Health Risk (Low, Med, High)
Agriculture	Crop Type (Paddy, Sugar cane)	Farm Type (Low, Med, High)
Government	Nationality	Opinion(agree, neutral, strongly agree)

Quantitative data: attributes that can be measured/ represented only by numeric value and used to count or measure certain properties of a population

Different Types of Data: Quantitative Data – Types of Numbers

Whole Numbers:	0,1,2,3…. (No Fractions, No Negatives)
Integers:	-5, -4, -3, -2, -1, 0, 1, 2, 3 (No Fractions)
Rational Numbers:	Ratio of 2 integers (½, 1/3, ¼, 2/1, 3/1 ….)
Irrational Numbers:	Cannot be expressed as ratio of 2 integers (Π, ^2
Real Numbers:	Rational and Irrational numbers

Discrete attributes are those quantitative attributes that can take only a finite number of numerical value (Integers)

Continuous attributes refer to quantitative attributes which can take on factional values

Some examples for Quantitative data

	Continuous	Discrete
Employee	Income tax, Gross Salary	# Projects, # Family members
Health care	Cholesterol level, sugar level	Days of treatment, weeks of pregnancy
Agriculture	Total yield / acres	# Farmers # of Crops
Government	GDP, GST	# of Districts # of Languages

Different Types of Data: Why bother about data types?

The type of statistical analysis depends on the type of variable

Qualitative attributes:

Qualitative attributes:

X What is the average color of all shirts in my catalog

X What is the average nationality of all students in the class

✓ What is the frequency of color red in my catalog

X Regression Analysis (because its analysis b/n 2 numbers

✓ Analysis of Variance (ANOVA)

✓ Chi-Square test

Quantitative (Discrete) attributes:

- ✓ What is the average value in the dataset?
- ✓ What is the Spread of the data?
- ✓ What is the frequency of a given value?
- ✓ What is the frequency of a given value?
- ✓ Regression Analysis

Quantitative (Continuous) attributes:

- ✓ What is the average value in the dataset?
- ✓ What is the Spread of the data?
- X What is the frequency of a given value?
- ✓ Regression Analysis

How to Describe Qualitative Data

Match #	Runs	Mins	SR	Pos	Dismissal	Oppn	Date
0	0	0	0.00	5	Caught	Pakistan	18-12-89
1	0	2	0.00	5	Caught	New Zealand	1-3-90
2	36	51	92.3	6	Caught	New Zealand	6-3-90
3	10	15	63.33	5	Caught	Sri Lanka	25-4-90
4	20	31	60.00	7	Run Out	Pakistan	27-4-90

5	**19**	**38**	**54.28**	**4**	**Bowled**	**England**	**18-7-90**
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**

The above table contains Sachin Tendulkars ODI data, Here the columns Position played in, Type of Dismissal and Opposition represents qualitative data.

The kind of question we would like to answer in relation to Qualitative data are:

How often did Sachin get bowled ?

Against which team did he score centuries?

That means we are interested in the Frequency / Number of repetition of qualitative data elements.

There are different ways we can describe qualitative data.

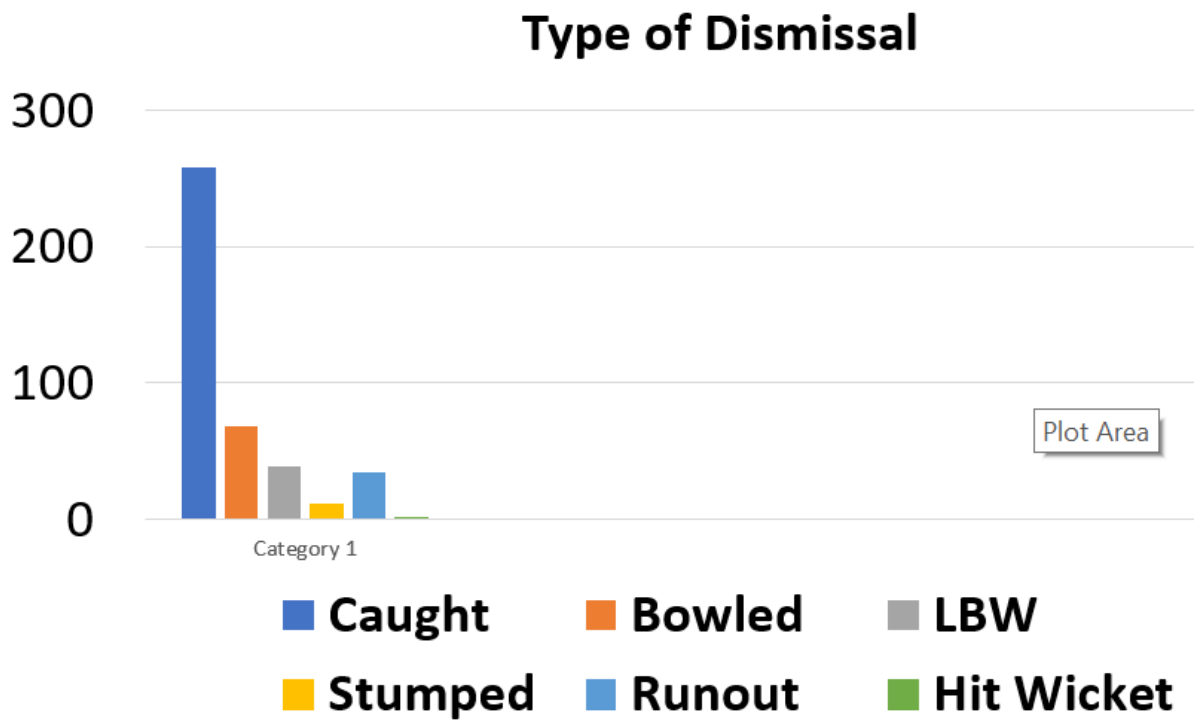
1) **Frequency Table:** Frequency refers to the number of times an event or a value occurs. A frequency table is a table that lists items and shows the number of times the items occur.

Creating a frequency table

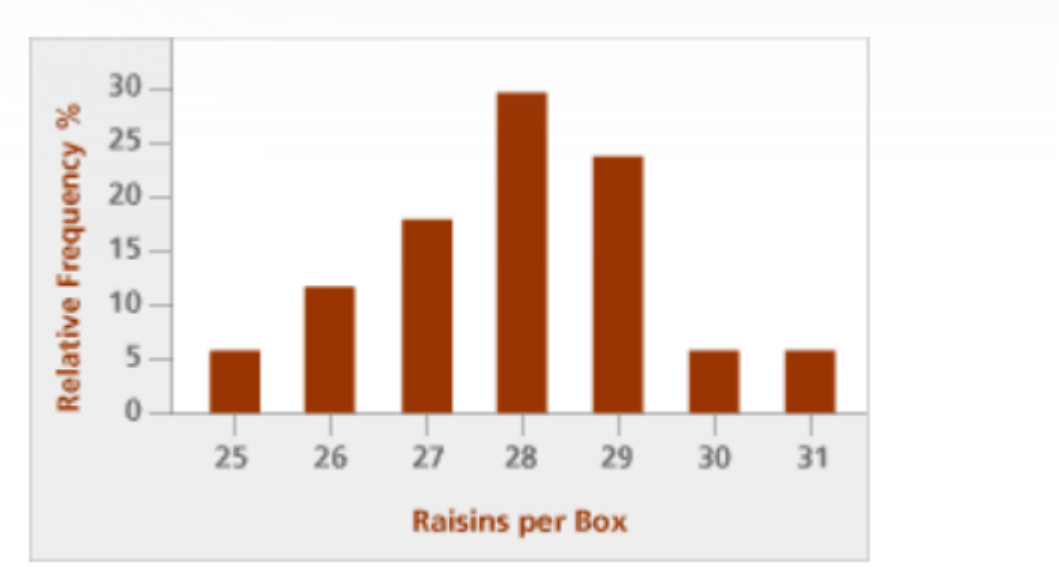
Step 1: Make three columns. The first column carries the data values in ascending order (from lesser to large values).
Step 2: The second column contains the number of times the data value occurs using tally marks. Count for every row in the table. Use tally marks for counting.
Step 3: Count the number of tally marks for each data value and write it in the third column.
For example, Rita maintains the record of the number of customers that visit her shop daily using the frequency table and tally marks.

Day	Number of customers	Frequency
Monday	𝍸 𝍸 𝍸 | | |	18
Tuesday	𝍸 𝍸 | | |	13
Wednesday	𝍸 𝍸 𝍸 𝍸	20
Thursday	𝍸 𝍸 | | | |	14
Friday	𝍸 𝍸 𝍸 𝍸 |	21
Saturday	𝍸 𝍸 𝍸 𝍸 𝍸 | |	27
Sunday	𝍸 𝍸 𝍸 𝍸 𝍸 |	26

2) **Frequency plots:** A frequency plot is a graph that shows the pattern in a set of data by plotting how often particular values of a measure occur.
Horizontal Axis: values of the categorical attribute
Vertical Axis: Counts of these values
Height of the bar is proportional to count
You can Sort the values by their counts for better visualization



3) Relative Frequency Plots: A relative frequency bar graph looks just like a frequency bar graph except that the units on the vertical axis are expressed as percentages. In the raisin example, the height of each bar is the relative frequency of the corresponding raisin count, expressed as a percentage

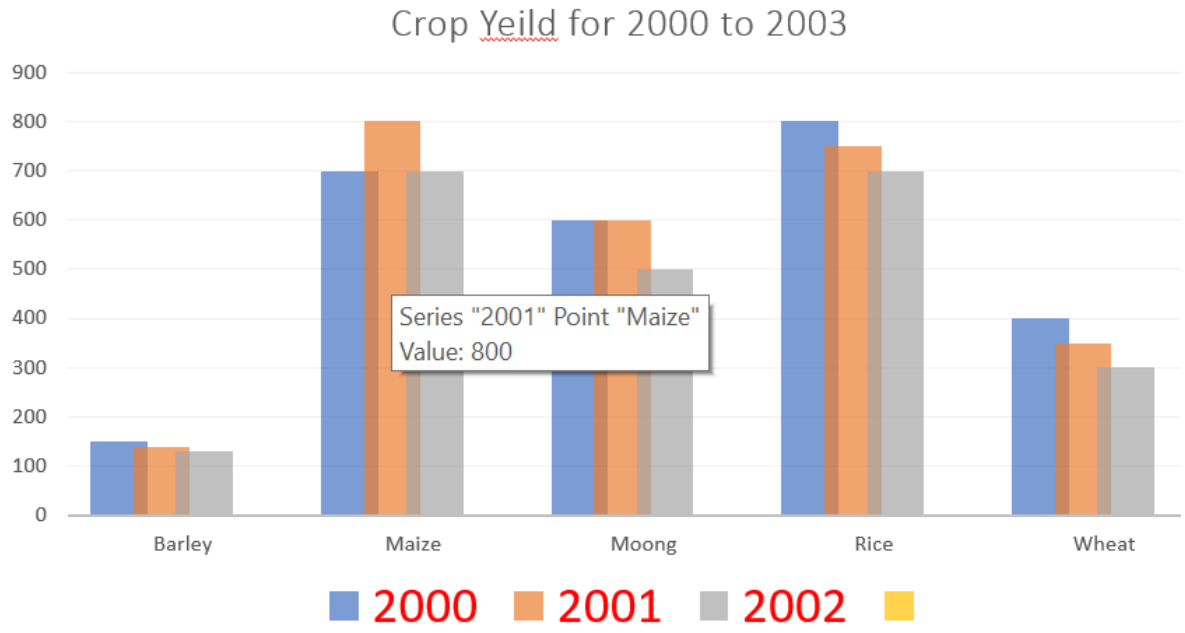


Relative Frequencies are easier to interpret then absolute Frequencies

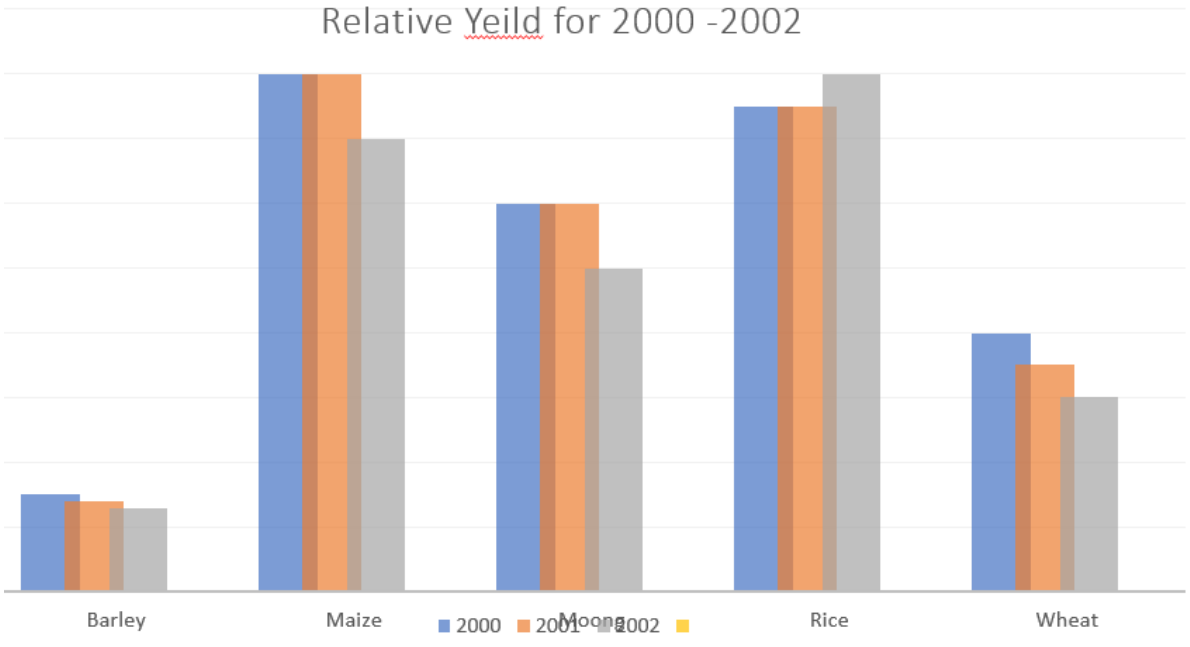
- Horizontal Axis: values of the categorical attribute
- Vertical Axis: Relative percentages
- Height of the bar is proportional to Percentage

Opposition	Centuries	Relative Frequency
Australia	9	(9/49)=0.184
Sri Lanka	8	(8/49)=0.163
South Africa	5	(5/49)=0.102
Pakistan	5	(5/49)=0.102
Zimbabwe	5	(5/49)=0.102
New Zealand	5	(5/49)=0.102
Kenya	4	(4/49)=0.081
West Indies	4	(4/49)=0.081
England	2	(2/49)=0.041
Namibia	1	(1/49)=0.020
Bangladesh	1	(1/49)=0.020
Total	**49**	(49/49)=0

4) Grouped Frequency Charts: are used to compare different sets of data, it is similar to a frequency chart, but here different data sets are grouped together and plotted next to each other so that its easier to compare there frequency for each of the categorical values/ intervals.



5) Grouped Relative Frequency Charts: they are similar to Grouped Frequency charts the only change here is instead of plotting the absolute frequency on the y axis we plot the percentage.



Describing quantitative data

While describing **qualitative data**, the **primary question** we were trying to answer is: What is the frequency of different categories of data?

To answer the above question we learnt about Frequency Table, Frequency Plots, Relative Frequency Plots, Grouped Frequency Charts, Grouped Relative Frequency Charts etc.

Now the question is "Does the same question make sense for **quantitative data**"

To answer this lets dive into an example and we will focus on Sachin Tendulkar's ODI data

Match #	Runs	Mins	SR	BF	4s	6s	Pos	Dismissal	Oppn	Date	Match ID
0	0	0	0.00	2	0	0	5	Caught	Pakistan	18-12-89	ODI # 593
1	0	2	0.00	2	0	0	5	Caught	New Zealand	1-3-90	ODI # 612
2	36	51	92.3	39	5	0	6	Caught	New Zealand	6-3-90	ODI # 616
3	10	15	63.33	12	0	0	5	Caught	Sri Lanka	25-4-90	ODI # 623
4	20	31	60.00	25	1		7	Run Out	Pakistan	27-4-90	ODI # 625
5	19	38	54.28	35	1	0	4	Bowled	England	18-7-90	ODI # 634
6	31	31	119.23	26	3	1	6	Bowled	England	20-7-90	ODI # 635
7	53	83	129.26	41	7	2	5	Bowled	Sri Lanka	1-12-90	ODI # 646
…	…	…	…	…..	…..	…..	…	…	…	…	…..

Now there are several quantitative attributes here eg: runs, mins, Strick rate, Ball faced, number of 4s, number of 6s etc.

In the above data Strick Rate is a continuous data.

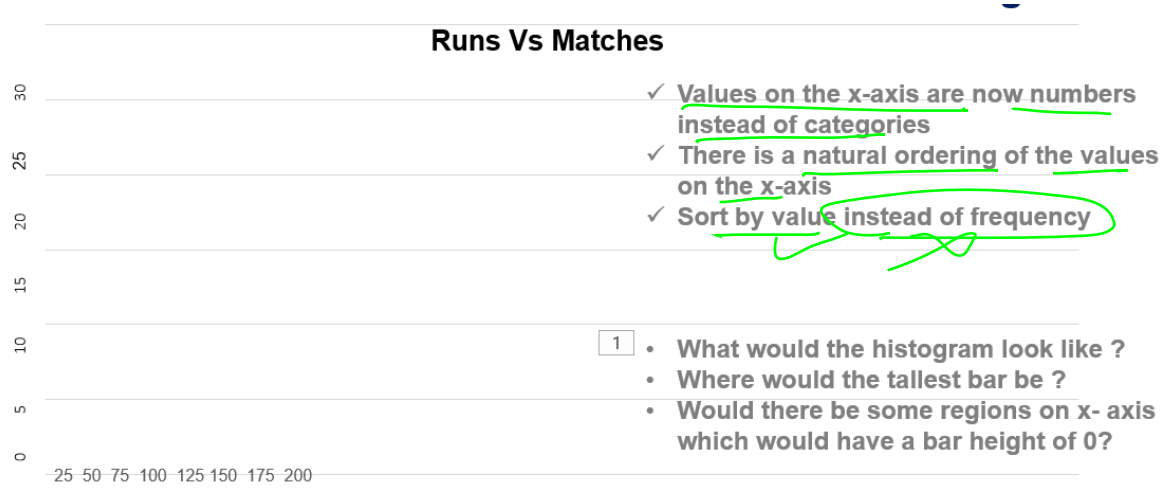
Let's focus on the desecrate data first, If we consider runs scored does it make sense to ask how many times did he get out on 0 or 49 or 99, In other words "What is the frequency of value of 0, 49, 99?"

The answer is it does make sense, we would be interested in knowing how many times he got out at 0 or 49 or 99 right ?

So what we are asking is what is the frequency of values of 0. 49. 99? Which is the same as the question we asked in the qualitative data.

But here we are dealing with Quantitative Discrete data and at least in our example of Sachin Tendulkar s ODI stats we know that the runs scored range from 0 to 200 and we are interested in knowing the frequency of some of these values like 0, 49, 99 etc..

So the main point is that the same question that we asked for Qualitative data still holds good and to answer this question we are going to introduce what is known as histograms



Histogram has Runs / values on the X axis and on the Y axis you have the frequencies / # of matches again the intension here is to draw bars for each of the ticks on the x axis and the height of the bar would represent the number of times that he got out scoring so many runs. Like 0 or 1 or 49 and so on..

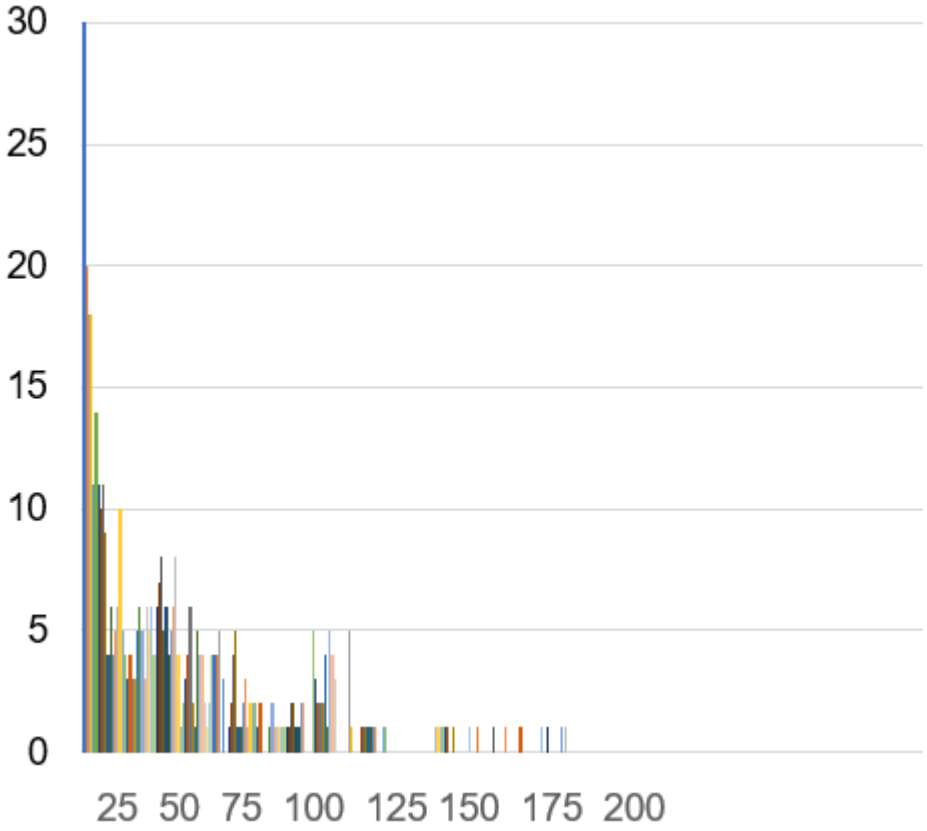
Remember that when we were plotting the qualitative data, we sorted the values based on the Y axis, for example in our agriculture data, based on the number of frequency.

But, here since we are plotting the values in the X axis and this values have natural numbers, which are in an order the plots are already sorted on the X axis.

Now, lets see how the histogram looks like and before looking at the histogram

Being fans of Sachin, lets ask ourselves a few questions:

- What would the histogram look like ?
- Where would the tallest bar be ?
- Would there be some regions on x- axis which would have a bar height of 0?



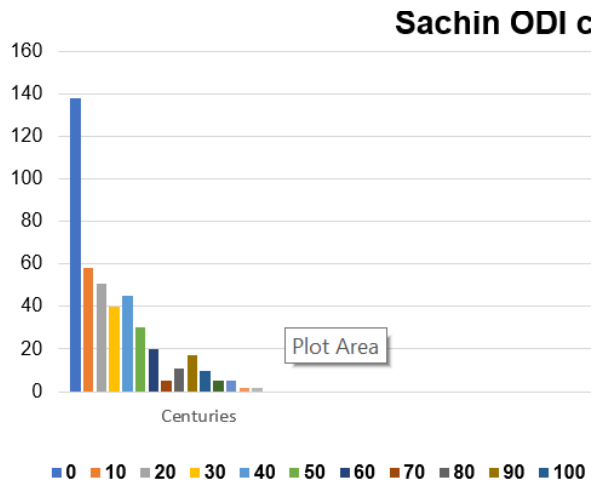
Most of the tall bars are in the beginning of the plot, indicating that he was vulnerable in the beginning of the innings

Looks like between 1 and zero the only lucky numbers for Sachine were 56, 58, 59, 75, 76 and 92

We are talking here of data that has values upto 200 and already the bars look overlapped.

If we try plotting the data for agriculture then this will be unreadable.

- ✓ Here its diffecult to answer some of the questions like: How many times was he dismissed in 90s or at single digit score
- ✓ Solution: Group values into bins: 0-9, 10-19….Each bin will now show the sum of the frequencies of all values in it.



The plot is now much easier to visualize although it hides some details (Such as how many times did he score exactly 0 or 1 or 100?)

Now the next question will be what is the right bin size ?

What about a bin size of 5?

Too many bins + it does not serve the grouping properly

So can we just use a larger bin- size? 20 or 40

As we increase the bin size the granularity is compromised with very few details, so both extremes are bad.

A right bin size will depend on the spread of the data, domine you are working on and also the questions your trying to answer.

Ideal bin size reveals meaningful patterns (neither hides nor reveals too many details)

Left-end-inclusion convention. A class interval contains its left end boundary but not its right end boundary.

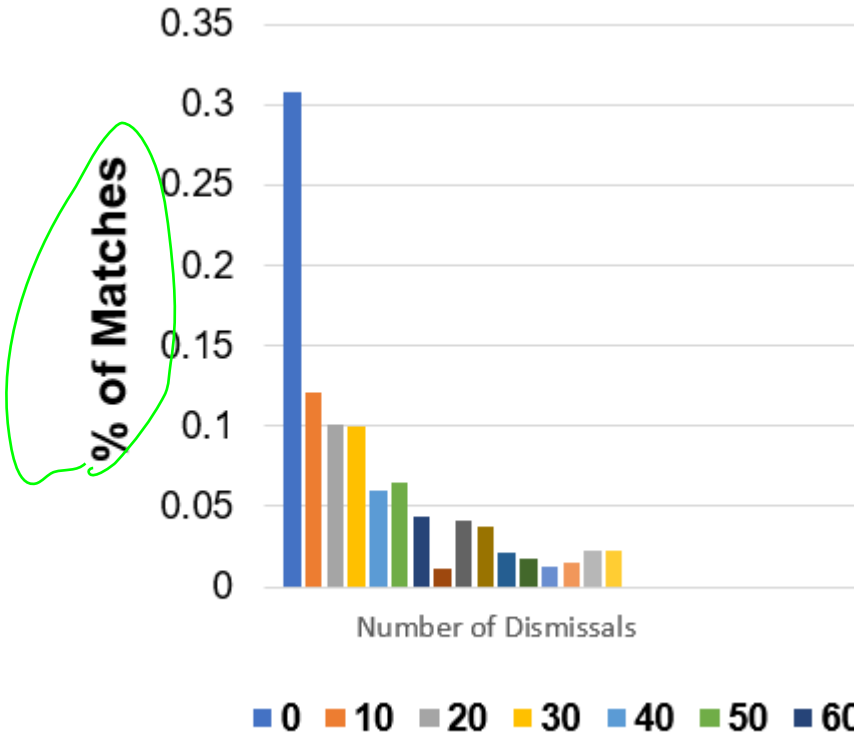
Steps to draw histogram:

- Sort the values in increasing Order (already sorted because X axis represents discrete or continuous data that is already sorted)
- Choose the class intervals such that all values are covered.
- Compute frequency of each interval
- Draw bars for each interval (height of the bar are proportional to the frequency computed)

Relative Frequency Histograms: instead of frequency we have percentage in y axis.

It helps answer questions like: In what percentage of matches did Sachin score less then 10 runs

Class interval	Frequency	Relative Frequency
0-10	143	(143/463)=0.3088
10-20	56	(56/463)=0.1209
20-30	47	(47/463)=0.1015
30-40	46	(46/463)=0.0993
40-50	28	(28/463)=0.0604
50-60	30	(30/463)=0.0647
60-70	20	(20/463)=0.0431
70-80	5	(5/463)=0.0107
80-90	10	(10/463)=0.041
90-100	18	(18/463)=0.0388
100-110	10	(10/463)=0.0215
Total	**463**	(463/463)=0



Steps to draw Relative Frequency histograms:

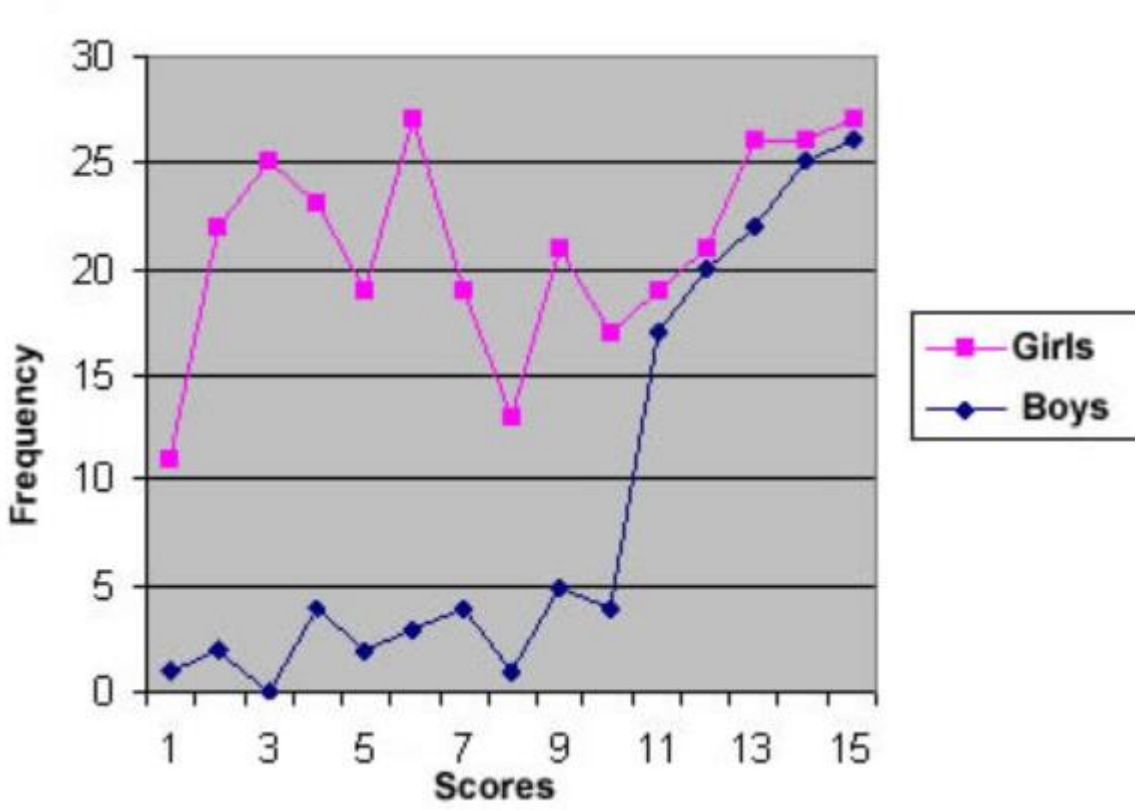
- Sort the values in increasing order
- Choose the class interval such that all values are covered (in particular min and max values should be covered, its ok if some of the intervals do not have values.
- Compute frequency of each interval
- Compute relative frequency of each interval
- Draw bars for each interval (such that the height of the bars are proportional to the relative frequencies computed in the previous step)

Frequency Polygons: is a line graph representation of a set of scores from a frequency table. The horizontal x-axis is represented by the scores on the

scale and the vertical y-axis is represented by the frequencies. It is used to compare the over all trend of different data sets / players / crops

Steps to draw Frequency Polygons:

- Sort the values
- Choose the class intervals
- Compute the frequency of each interval
- Compute the midpoint of each interval
- Plot the frequency above the midpoint



Relative frequency polygons: A **relative frequency** polygon has peaks that represent the percentage of total data points falling within the interval.

To construct a relative frequency polygon:

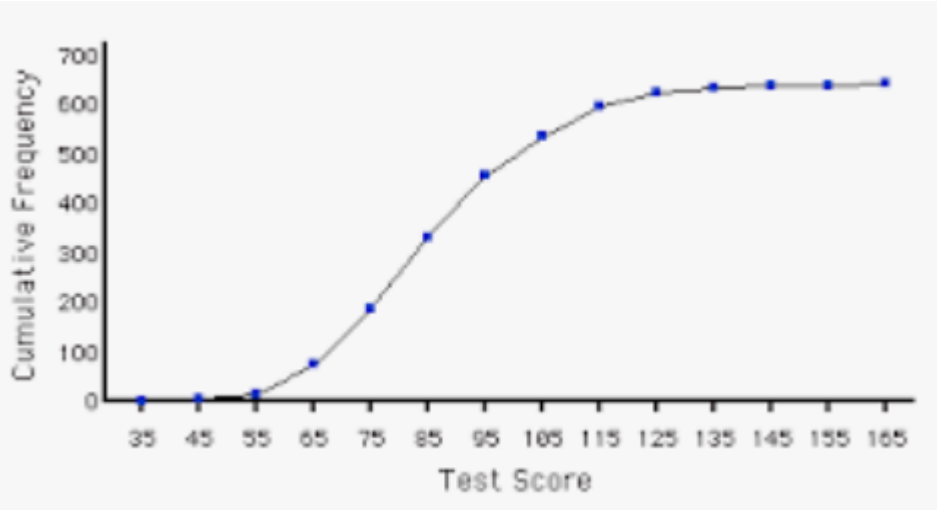
a. Construct a frame just as you would for a histogram.
b. Label the vertical axis from 0 – 100%, and the horizontal axis with the intervals you have chosen.

c. Sum the number of points in each interval, divide the sum of each interval by the total number of data points, and multiply by 100. The result is the percentage of the total number of data points that is represented by each interval. Mark a point representing the percentage along the midline of the interval.
d. Once all points have been accounted for, connect the points and color in the area under the line.
e. If you are graphing a second set of data, repeat the process.

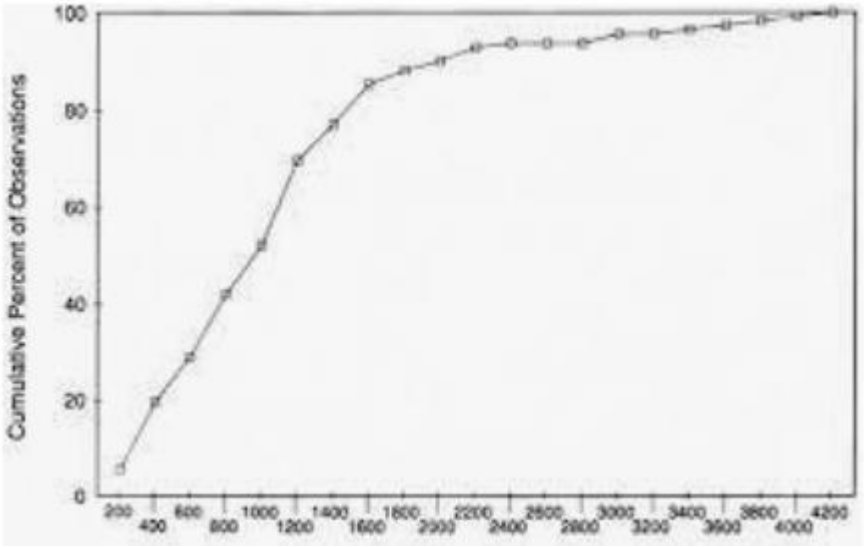
Cumulative Frequency polygon: is a type of frequency polygon that shows cumulative frequencies. In other words, the cumulative frequency are added on the graph from left to right.

For each class interval also add the sum of the frequencies of all class intervals before it

Easy to compare multiple sets of data



Cumulative Relative Frequency Polygon:: is a type of frequency polygon that shows cumulative frequencies. In other words, the cumulative percentages are added on the graph from left to right.



Typical Trends in Histograms

- How far are the values in the data spread out
- Is the data density high in certain intervals
- Are there gaps in the data
- Are there any outliers in the data

Uses of Histograms in Machine Learning

Histograms are also used in Machine Learning for various purposes.

1. **Identifying Discriminatory features**

Let's say we are trying to build a machine learning system that takes a lot of information about a patient such as Age, Height,

Weight, Cholesterol, Sugar Level, and so on and this Machine Learning system decides if the patient has health risk or not.

If we go to a Doctor, the Doctor will not ask so many readings, he might ask those which are the few important readings to know about maybe Cholesterol is important in deciding if a patient has Heart risk or any other health condition or not.

So, a Doctor might know what are the important factors/attributes to look for but a Machine Learning system does not know this in advance meaning what the important attributes are the system doesn't know in advance and the features that help us identify the health risk in this case or in general the features that help us understand the output as a function of the input are termed as **Discriminatory features**.

The data set would look like the below, where each column is one attribute and each row represent the data for one patient, their past records, we know whether had a health condition or not(let's say we got this data from some hospital we collaborated with)

Now if we want to understand which are the features(attributes) we should really look at or in other words, which are the features that really matter in deciding if someone has a health risk or not

One thing we could do is to split this data set into two sets, one is for the patients having a risk(past records) and the other is

about the patients not having a risk; for each of these datasets we could individually draw the histogram or frequency polygon.

From the above plot, we can say that for people who don't have risk, at least the max heart rate for them appears to be on the lower side, and for people who are at risk, the max heart rate seems to be at the higher end.

Just like we have the frequency polygon drawn for the Heart Rate attribute in the above plot, we could have it for any other attribute as well such as Cholesterol or Sugar or anything of the sort for that matter. If that attribute really matters in predicting the output, then the histogram or the frequency polygon would look like the above plot — where for one set say no risk patients, the values cluster at one end, and for the other set, the values cluster at another end. And the corresponding feature would be a good input feature to the machine learning system.

It is not necessary that the people who have health risk or does not have the health risk are taller or shorter. It might be true for some cases may be for 'arthritis' or something of that sort where people having certain height might be more susceptible but for a heart condition it might not really matter and in that scenario if we see a histogram like the above one where for both the sets there is no clear distinction between the trends, they see more or less similar, that means Height may not be a good feature to help the Machine Learning algorithm to identify or segregating or discriminating patients with risk from patients with no risk.

2. Analysing output scores

Histograms also help in analyzing output scores of the Machine Learning system. Nowadays everybody wants to build a Chatbot and unfortunately, despite the hype around the chatbot, they are nowhere being close to satisfactory. So, if we ask a Chatbot “What’s the temperature outside?”, we want it to say something like “It is very hot, the temperature is around 33 degrees celsius” and not give back some random response.

If we chat with some chatbot for 2–3 turns, we start getting some random responses, they are not able to maintain context longer than 3 or 4 turns and they are not really at a level where they can replace humans.

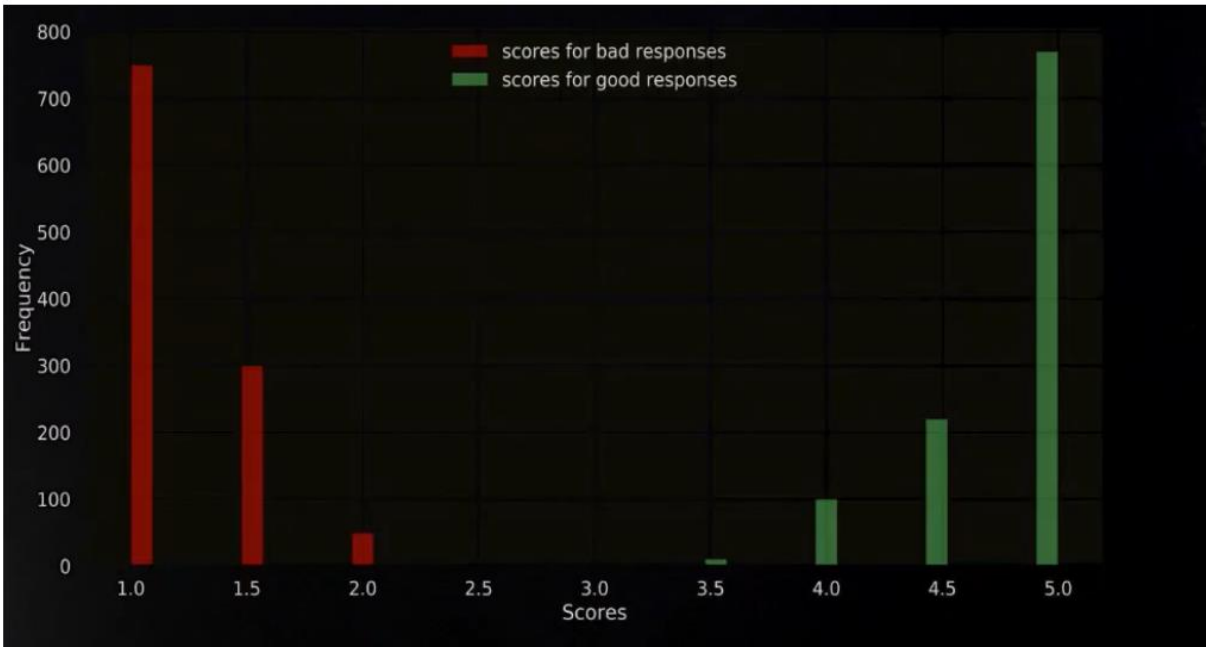
Now if we want to deploy a Chatbot in a call center environment and we know that the first 2–3 turns it handles okay but beyond that, it may or may not handle it properly.

So, what would be useful there is to develop a Machine Learning system that looks at the answer given by the Chatbot and the ML system should be able to decide if the Bot’s response is good or not. If it is good, then let the Bot continue talking to the customer and if the ML system says it’s a bad response then a human should take over and not let the Bot

continue because it has lost context, it is not been able to reply correctly.

Suppose someone develops such a system, now the question is how to check if it is good or not?

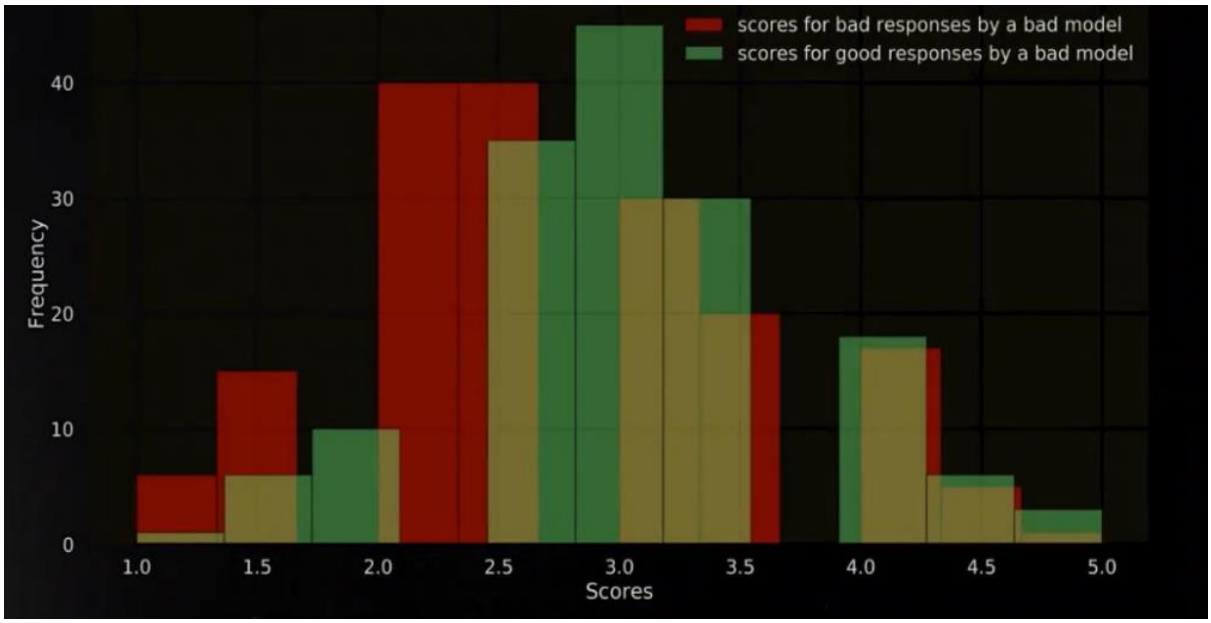
We take some bad responses and some good responses and we see what is the score that the Machine Learning system assigns to it. So, let's assume that the Machine Learning algorithm assigns it a score from 1 to 5, so this is what we expect the histogram to ideally look like:



For all the bad responses, we want the score to be less than or equal to 2 whereas, for all the good responses, we want the score to be greater than 3, maximum of them we would

like them to get a high score. This is the ideal situation that we expect from a Machine Learning system.

Let's say we get the below type of histogram:



The red bars in the above histogram is for the bad responses and the green bars is for the good responses and from the above plot we can see say that the Algorithm for which we got this plot does not have a good discriminatory power because there are some good responses which got a score in the range of 0 to 2, that means the Algorithm is classifying good responses as bad responses and there are some bad responses which got a very high score in the range of 4 to 5, and if we look closely, for most responses, the tallest bar for the good as well as the bad responses are in the middle range of score of 2.5 to 3.5 which is just like saying okay for everything and that's what the system/algorithm is doing.

So, this is a very interesting way of analyzing what the machine learning system is outputting for different kinds of inputs that we have or the score that it is generating.