Mithilesh singh

Statistics

Statistics is the science of learning from data.

It is a set of equations that allows us to solve complex problems.
"Statistics is the science of making decisions under uncertainty."

It is a mathematical science including methods of collecting, organizing and analyzing data in such a way that meaningful conclusions can be drawn from them.

EDA ----- exploratory Data Analysis skill

Types of Statistics:

1. Descriptive statistics deals with the simple processing of data. Like Mean, Median, Mode, Variance, Standard Deviation. (Without attempting to draw any inferences about population from it).

2. Inferential statistics is a scientific discipline that uses mathematical tools to make forecasts and projections by analyzing the given data. (attempting to draw any inferences about the population from it).

Statistics are widely use in almost all fields engineering, economics, biology, social sciences, business, agriculture, Prediction of any events , Elections, communications, Medicine.

Statistical Thinking

EXAMPLE 1:- “Vacation in Goa , you are Renting a basic Normal Bike for one day local sight-seeing , do you want to buy the bike insurance for Rs. 500/per day ?”

Example 2:- My friend wise grandfather smoked 5 packs a day and drank a quart of scotch a day, but was always healthy and died peacefully in his sleep when he was 90, Do you conclude that all the health warnings about cigarettes are wrong?

Example 3:- A celebrity advertised Amul ice-cream in Month May last year, due to that ice-cream sales increase 25% in following three months. Thus the advertisement was effective.

Example 4:- The more Liquor shop in the city , the more crime there is , so the liquor shop is the responsible for crime?

Measures of Central Tendency: Definition & Examples

A **measure of central tendency** is a single value that represents the center point of a dataset. This value can also be referred to as “the central location” of a dataset.

In statistics, there are three common measures of central tendency:

- **The mean**
- **The median**
- **The mode**

Each of these measures finds the central location of a dataset using different methods depending on the type of data .

Why Are Measures of Central Tendency Useful?

Consider the following scenario:

A young couple is trying to decide a rented home in a city and the most they can spend is 180000/- annum. Some locations in the city have expensive houses, some have cheap houses, and others have medium-priced houses. They want to easily narrow down their search to specific location that are within their budget.

If the couple just looked at the individual home rent in each location, they might have a tough time determining which location best fit their budget because they might see something like this:

Location *A* Home Rent: 140k, 190k, 265k, 115k, 270k, 240k, 250k, 180k, 160k, 200k, 240k, 280k, …

Location *B* Home Rent: 140k, 290k, 155k, 165k, 280k, 220k, 155k, 185k, 160k, 200k, 190k, 140k, 145k, …

Location *C* Home Rent: 140k, 130k, 165k, 115k, 170k, 100k, 150k, 180k, 190k, 120k, 110k, 130k, 120k, …

However, if they knew the *average* (e.g. a measure of central tendency) home Rent in each location, then they could narrow down their search much quicker: -

Average location *A* Rent: 220k

Average Location *B* Rent: 190k

Average location *C* Rent: 140k

A measure of central tendency is useful because it provides us with a single value that describes the "center" of a dataset.

This helps us understand a dataset much more quickly compared to simply looking at all of the individual values in the dataset.

Mean

The most commonly used measure of central tendency is **the mean**.

Mean = (sum of all values) / (total # of values)

Student	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Age in year	8	15	22	21	12	9	11	27	14	13

Median

The **median** is the middle value in a dataset.

arranging all the values in ascending order and finding the middle value. If there are an odd number of values, the median is the middle value. If there are an even number of values, the median is the average of the two middle values.

Student	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Age in year	8	15	22	21	12	9	11	27	14	13

The Mode

The **mode** is the value that occurs most often in a dataset. A dataset can have no mode (if no value repeats), one mode, or multiple modes.

For example, the following dataset has no mode:

Player	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Home Runs	8	9	11	12	13	14	15	21	22	27

The following dataset has one mode: **15**. This is the value that occurs most frequently.

Player	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Home Runs	8	9	11	12	13	**15**	**15**	21	22	27

The following dataset has three modes: **8, 15, 19**. These are the values that occur most frequently.

Player	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Home Runs	**8**	**8**	11	12	**15**	**15**	17	**19**	**19**	27

The mode can be a particularly helpful measure of central tendency when working with categorical data.

When to Use the Mean, Median, and Mode

Mean: Finds the average value in a dataset.

Median: Finds the middle value in a dataset.

Mode: Finds the most frequently occurring value in a dataset.

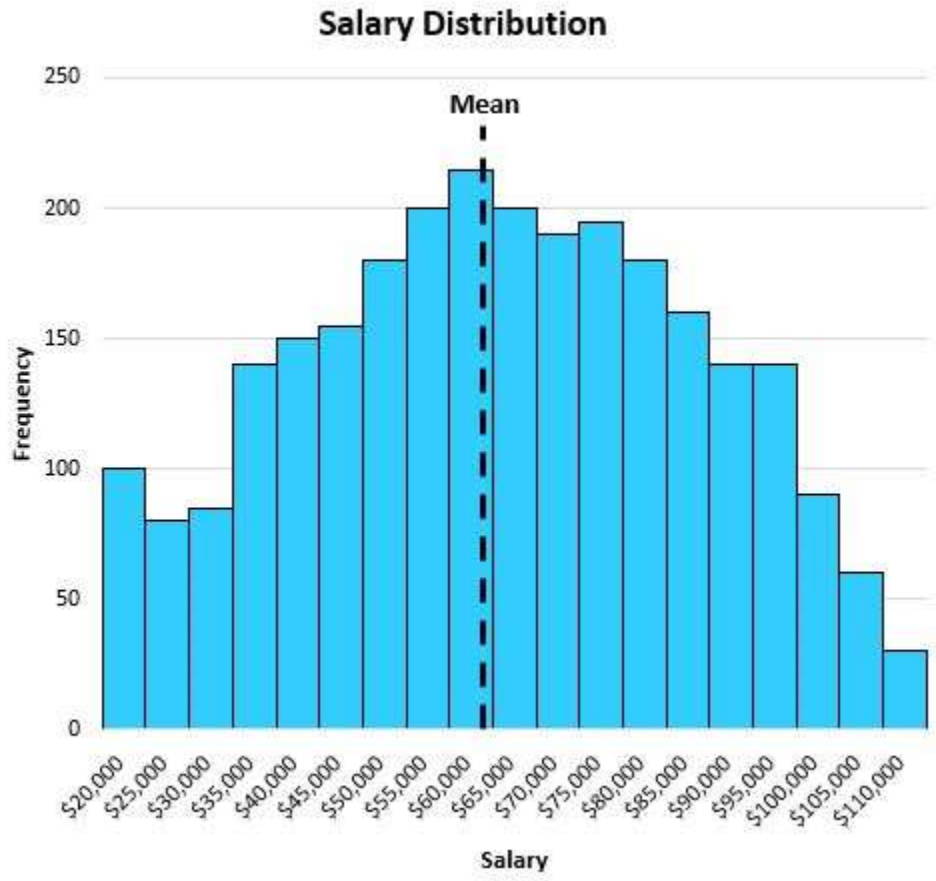
When to use the mean

It is best to use the mean when the distribution of the data is fairly symmetrical and there are no outliers.

For example, suppose we have the following distribution that shows the salaries of individuals in a certain town:

Since this distribution is fairly symmetrical (i.e. if you split it down the middle, each half would look roughly equal) and there are no outliers (i.e. no extremely high salaries

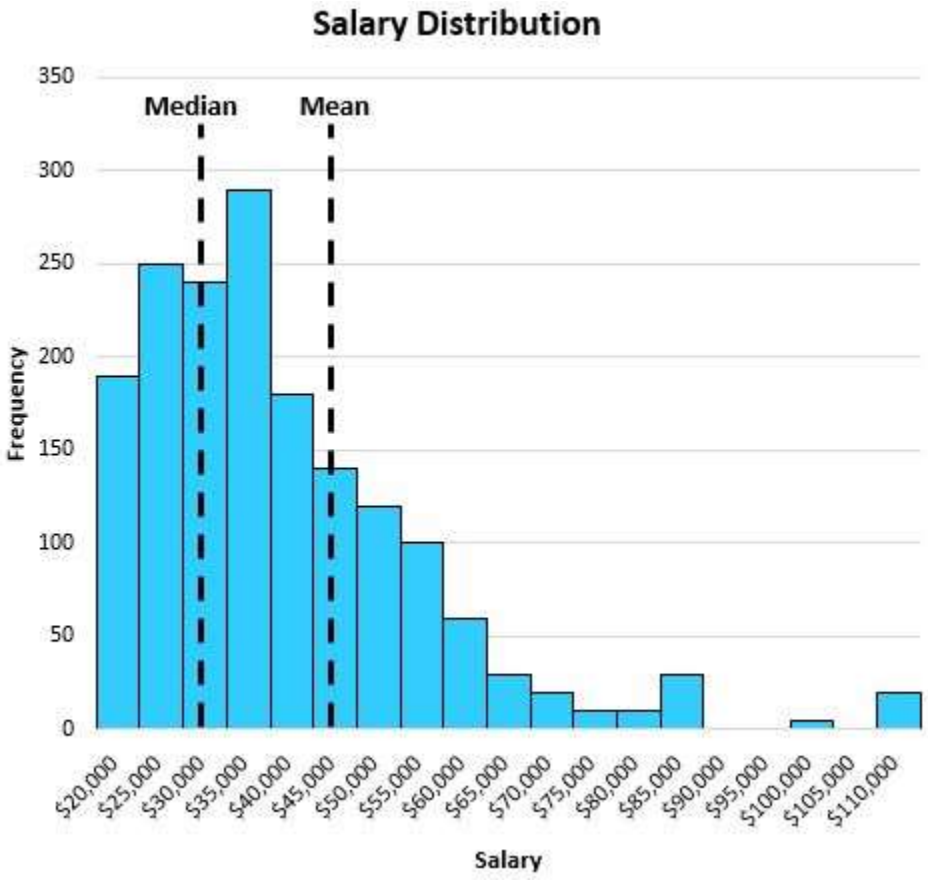
The mean turns out to be $63,000, which is roughly located in the center of the distribution:



When to use the median

It is best to use the median when the distribution of the data is either skewed or there are outliers present.

Skewed data:

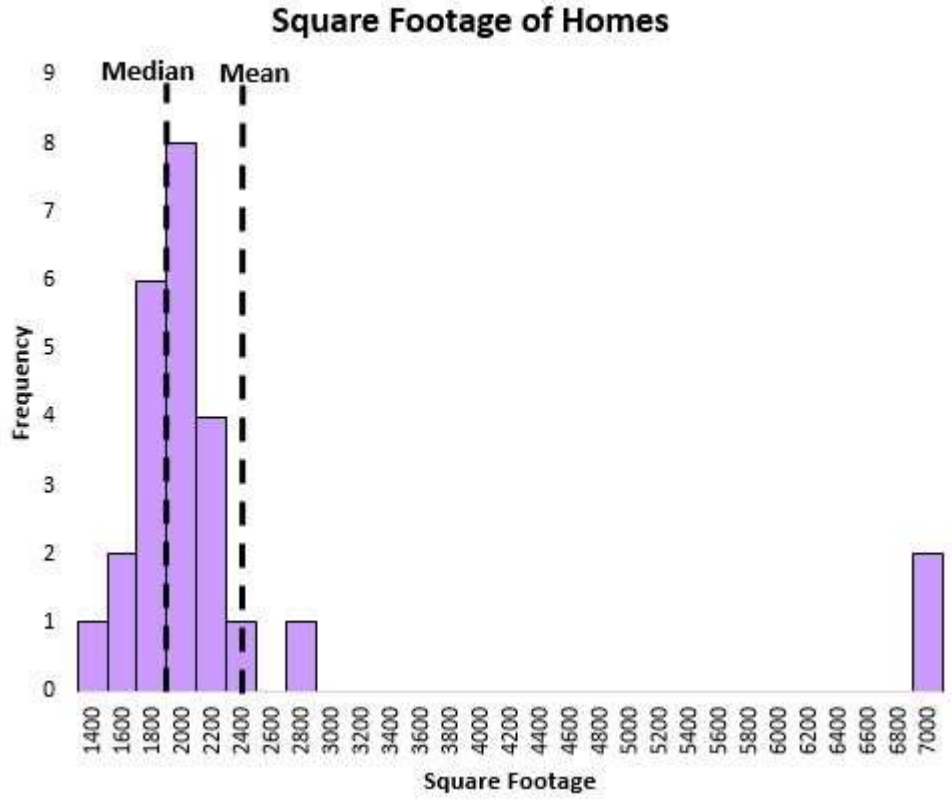


The median does a better job of capturing the “typical” salary of an individual than the mean.

In this particular example, the mean tells us that the typical individual earns about $47,000 per year in this town while the median tells us that the typical individual only earns about $32,000 per year, which is much more representative of the typical individual.

Outliers:

The median also does a better job of capturing the central location of a distribution when there are outliers present in the data. For example, consider the following chart that shows the square footage of houses on a certain street:

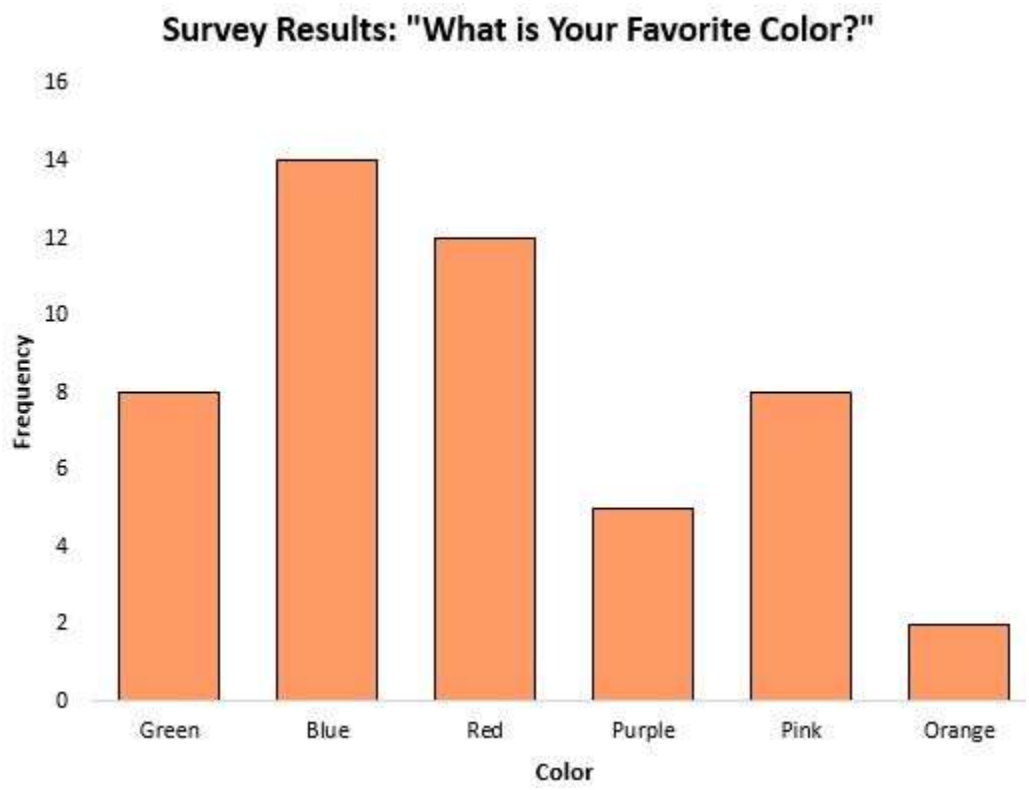


The mean is heavily influenced by a couple extremely large houses, while the median is not. Thus, the median does a better job of capturing the “typical” square footage of a house on this street compared to the mean.

When to use the mode

It is best to use the mode when you are working with categorical data and you want to know which category occurs most frequently.

You conduct a survey about people's preferences among three choices for a website design :-



working with categorical data then it's not even possible to calculate the median or mean, mode is the measure of central tendency.

Note: It's important to note that if a dataset is *perfectly* normally distributed, then the mean, median, and mode are all the same value.

Measure of Spread / Dispersion

1. Standard deviation: - Standard deviation is the measurement of average distance between each quantity and mean. That is, how data is spread out from mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values

Sample SD

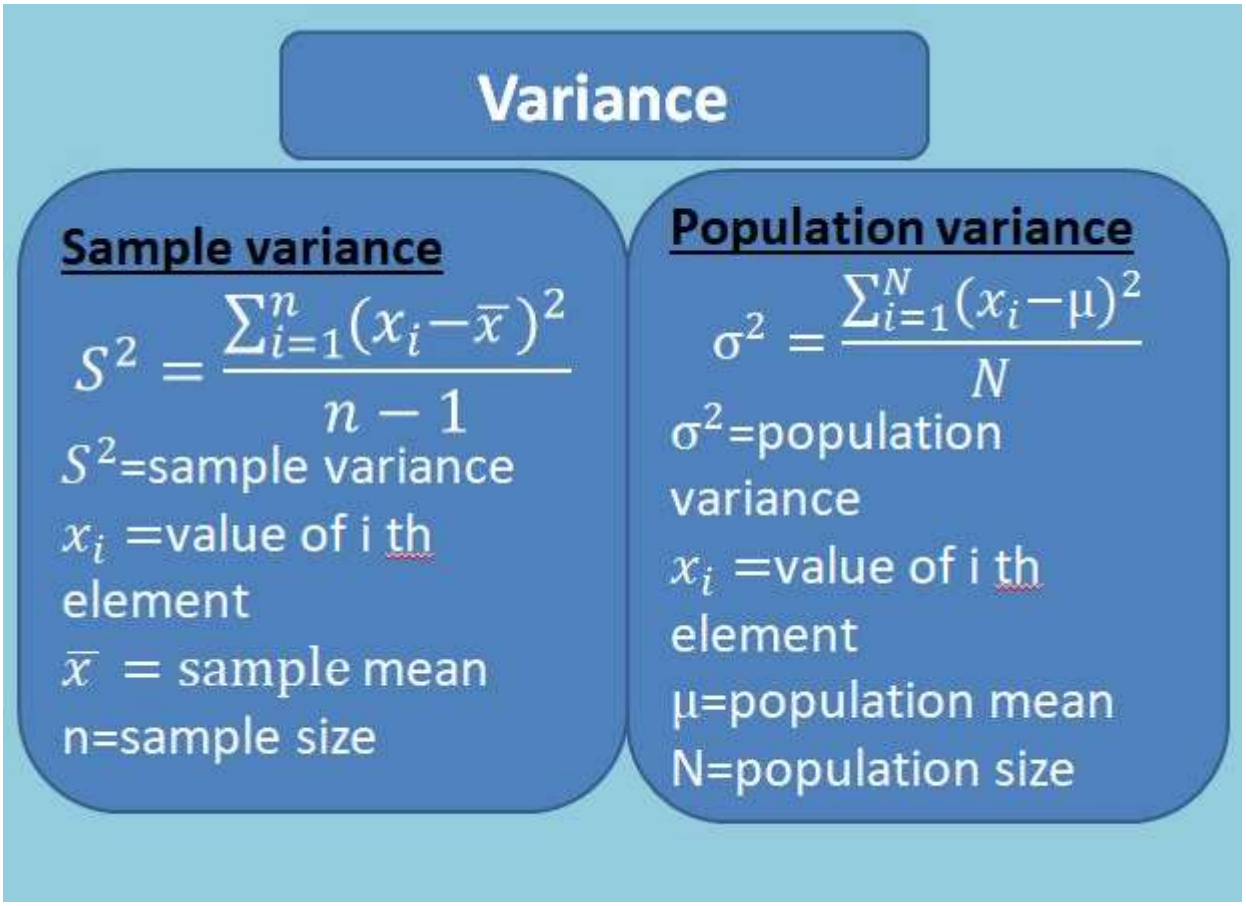
Population SD

Standard Deviation Formula	
Sample	**Population**
$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$
X – The Value in the data distribution *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*	*X – The Value in the data distribution* *μ – The population Mean* *N – Total Number of Observations*

2. Variance: - Variance is a square of average distance between each quantity and mean. That is the square of standard deviation.

Sample Variance

Population Variance



The concept of variance has a central role in statistics because it is used to calculate other statistics like ANOVA, R-Squared, hypothesis testing, statistical inference, and more.

example calculations of standard deviation.

Calculate the standard deviation for the sample age data of the 15 students: -

22	23	25	27	28	35	32	28	30	40	24	26	27	29	31

Calculations: -

STEP – 1: Calculate the Mean

- The mean age of the 15 students is,
- Mean (Age) = (22 + 23 + 25 + 27 + 28 + 35 + 32 + 28 + 30 + 40 + 24 + 26 + 27 + 29 + 31) / 15 => (427 / 15)
- **Mean (Age) = 28.47**

STEP – 2: Calculate the Standard Deviation

- Let X be the Age of the 15 Students Sample,
- Then the Standard Deviation of sample X is,

$$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$$

- Let us calculate the standard deviation.

X	$X - \bar{x}$	$(X - \bar{x})^2$
22	-6.47	41.86
23	-5.47	29.92
25	-3.47	12.04
27	-1.47	2.16
28	-0.47	0.22
35	6.53	42.64
32	3.53	12.46
28	-0.47	0.22
30	1.53	2.34
40	11.53	132.94
24	-4.47	19.98
26	-2.47	6.1
27	-1.47	2.16
29	0.53	0.28
31	2.53	6.4
	$\sum(X - \bar{x})^2$	**311.72**

The total number of observations, n = 15.

$$\sum(X - \bar{x})^2 = 311.72, \text{ n} = 15$$

$$s = \sqrt{\frac{311.72}{15-1}} = \sqrt{\frac{311.72}{14}} = \sqrt{22.47}$$

$$s = 4.72$$

Standard Deviation Importance
Hospital Bed Example

A hospital has 100 bed facility. The day-wise demand for beds in the month of Jan 2020 is given below. If there is more demand on any particular day, the hospital will have to refer the patient to some other hospital.

Jan 2020						
Sun	Mon	Tue	Wed	Thu	Fri	Sat
			1	2	3	4
			150	90	120	110
5	6	7	8	9	10	11
65	70	80	130	90	75	70
12	13	14	15	16	17	18
100	110	140	90	90	80	75
19	20	21	22	23	24	25
75	120	85	90	80	115	115
26	27	28	29	30	31	
90	120	140	150	95	90	

Mean statistic

The mean for the above data is 100.
Given, the hospital has 100 bed facility.

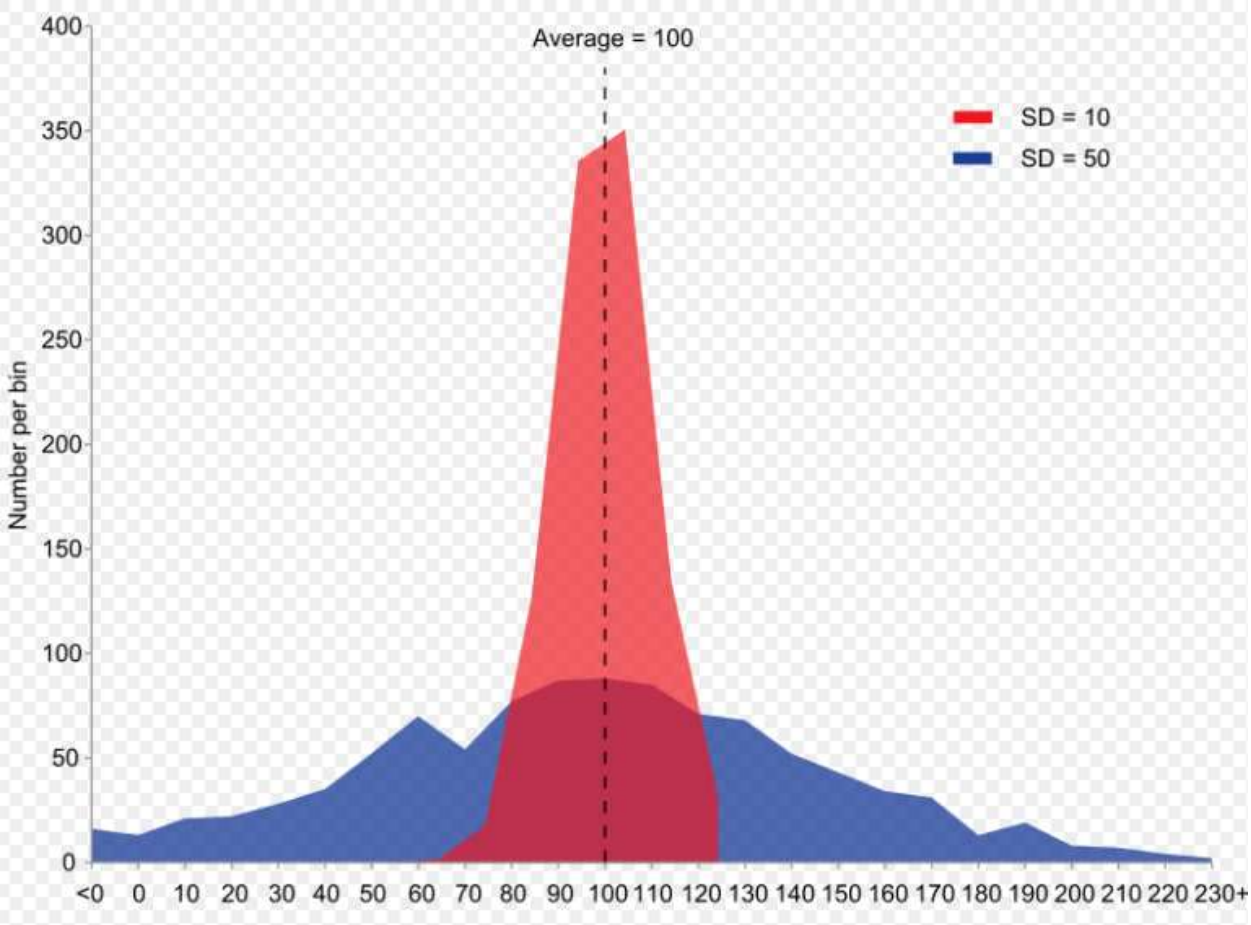
Standard Deviation – The Standard Deviation is 24.5 for the above data.

Interpretation

An exceedingly high standard deviation in demand suggests that:

- the hospital staff and its facilities would be too stretched on some days of the month
- and it would be too relaxed on some other days of the month.

Example:- Cricket Match player



Measure types(variable)

The basic distinction is between

QUANTITATIVE DATA :-(for which one asks “how much?”)

Data that measure in Numbers. Like Height, weight, Score, Salary.

CATEGORICAL DATA (for which one asks “what type?”).

Type of city, color , Department, Education field.

(1) Quantitative variables

(a)Discrete number- counted as whole no.

exp:- No. of kids 2, 3, Training session 1,2,3 its always integer not a float.(2.5)

(b)Continuous: - Number can be infinite precision.

Weight :- 105, 89, 73.5 kg,

Example-you can sleep 7 hr 30 min & 20 sec, distance can be 70.97 meters

(2) Categorical variables

Can not be measure against each other like color (blue, red, yellow)

(4) Categorical:- The only measure of central tendency can be use is *the mode*.

Types:-

Categorical- Ordinal

-Result rank A++, A+, A, B+.. here it can be compare among them.

-PhD, Master, Bachelor ,12 th, 10th Pass.

Some examples of variables that can be measured on an ordinal scale include:

- **Satisfaction:** unsatisfied, neutral, satisfied, very satisfied
- **Socioeconomic status:** Low income, medium income, high income
- **Workplace status:** Entry Analyst, Analyst I, Analyst II, Lead Analyst
- **Degree of pain:** Small amount of pain, medium amount of pain, high amount of pain

Categorical-nominal

Some examples of variables that can be measured on a nominal scale include:

- **Gender:** Male, female
- **Eye color:** Blue, green, brown
- **Blood type:** O-, O+, A-, A+, B-, B+, AB-, AB+
- **City you live:** Mumbai, Bangalore, Delhi, Chennai, Kolkata

Examples of types of data	
Quantitative	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of attacks of asthma per week
Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse Disagree, neutral, agree	Sex (male/female) Alive or dead Blood group O, A, B, AB