Frequency Distribution.

A list of values × corresponding frequencies.

Class Width
Difference between two "lower class limits"

Lower class limit
Smallest value belonging to a class.

Upper class limit
Largest value in a class.

steps -

1. Determine # of classes / groups / bins (N): 8
2. Class width (Bin width): $\frac{\text{max value} - \text{min value}}{N}$ eg:- $\frac{44-18}{8} = 3.25$ = 4 * Round up.
3. Start w/ smallest value: 18
4. Create classes with class width

(18-21) (22-25) (26-29) (46-49)
18 22 26 30 44
inclusive, not inclusive

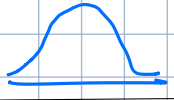
Class midpoint
$\frac{\text{Upper class limit} + \text{lower class limit}}{2}$ = 19.5

Class Boundaries
used to separate classes without gaps (in histograms)

18-21, 22-25 → 21.5 is the class boundary.
25.5, 29.5 ...

Frequency distribution tables. → % of points in each group.
Cumulative frequency distribution → Addition till current.

"NORMAL" →



Data characteristics. CENTRE, VARIATION, DISTRIBUTION, OUTLIERS & CHANGES OVER TIME.

① Center - what value is most of the data surrounding.

3 ways. (mean vs average) ↳ interesting.

MEAN - Arithmetic average. : $\frac{\Sigma x_i}{n}$. $\bar{x}$

$\bar{x}$ = sample mean
μ = population mean

$\mu = \frac{\Sigma x_i}{N}$.

Median - It's the middle value of your dataset.
- Must be in order.

If odd number of values, the median is the middle value.
If even number of values, the median is average of two middle values.

Median is less prone to outliers. If there are outliers in the data, the mean is affected heavily.

Hence, its median household income and home value etc. People like bill gates will shoot up the mean.

mode - The most commonly occuring data value.

skewness -  ← skewed

(2) Variation

Eg:- Bank lines ← Time you have to wait in a bank

#1	6, 6, 6	$\bar{x} = 6$
#2	4, 7, 7	$\bar{x} = 6$
#3	1, 3, 14	$\bar{x} = 6$

The mean is the same for all but all three are spread differently.

ways to measure variation.

1. Range - max value - min value
   Easy to find. doesn't take into account all values.

2. Standard deviation - Average distance your data values are from the mean
   Standard deviation is never going to be negative.
   - can be zero. - meaning there is no deviation. all entries are the same.
   - greatly affected by outliers.
   since $(x - \bar{x})$ can result in +ve and -ve numbers hence we square it.
   ⇒ this overestimates the variation. since we are dividing with smaller number.

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

or

$$s = \sqrt{\frac{n\,\Sigma(x^2) - (\Sigma x)^2}{n(n-1)}}$$ ← we need to calculate mean here -

Std. dev of 1, 3, 14 → 7

Std. dev of a "sample" has $(n-1)$ as denominator because the purpose of a sample is to over-estimate the properties of a population.

Standard deviation of a population $= \sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{N}}$

3. Variance. $= s^2 / \sigma^2$ ↑ std. dev. = data is more spread out.

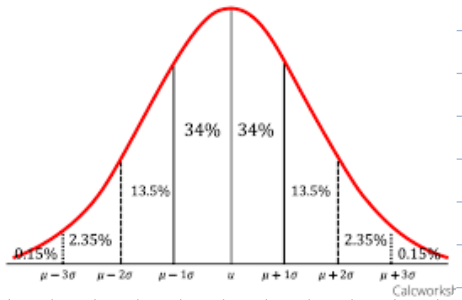
std. dev = $\sqrt{\text{variance}}$.

closely grouped data will have a small std. dev.
spread-out data will have a large std. dev.

If the data is normally distributed, we can use the empirical rule.

68% of the data will fall with 1 std. from the mean.
95% of the data will fall within 2 std. dev.
99.7% of the data within 3 std. dev.
It is a never ending bell shaped curve. The rareness of finding the data beyound 3 std. dev increases.

Data within 2 std. dev from mean = Usual data.



Standard deviations cannot be compared. A numerically higher std. dev does not mean the one data point is more spread than the other. Every data value has its own "unit"

Coefficient of variation. → compares a samples std. dev to its mean.

$C.V = \frac{S}{\bar{x}} \times 100$. (in %)

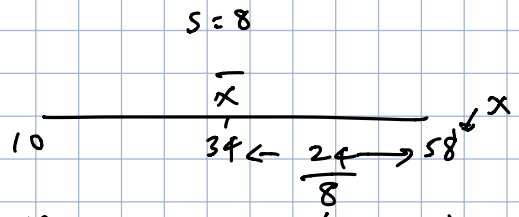
One way to do it. Not very effective.

Measures of relative standing (comparing measures between and within data sets)

Z-score - The number of std. dev away from the mean a data value is.

Sample: $z = \frac{x - \bar{x}}{s}$

Population: $z = \frac{x - \mu}{\sigma}$

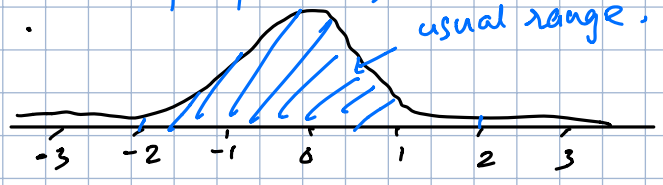


Allows comparison of the variation in two different samples/population.

Tall example. which one is relatively taller. Z-score - 2.14 & 1.82.
↑ taller.

Z-score can be negative (value to the left of mean)
Z-score at the mean is zero.

Usual range - -2 to 2
Unusual range - less than -2 & greater than 2



Tells you how rare a data point is? Helps in hypothesis testing.
↳ can be used to detect outliers.
Often called as z-score normalization in machine learning.

Quartiles - divides the data into four parts.
Q_1 → Bottom 25% of sorted data.
Q_2 → same as median → Bottom 50% of sorted data.
Q_3 → Bottom 75% of the sorted data.

You just need to know how to calculate median.

Ex:- 1, 3, 6, 10, 15, 21, 28, 35.

Q_1 $M = 12.5$ Q_3.

median of 1st list — median of 2nd list.

Percentile - divides the data into 100 parts. → 1%.
which means you cut the data 99 times.
There are 99 percentiles. There cannot be 100 percentile.

just like there is no Q4.

An exam score of 95%ile means you scored better than 95% of the people who took the test. It compares your performance with everyone else.

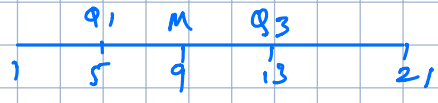
Percentile of X = $\frac{\#\text{ of values less than } X}{\#\text{ of values}} \times 100$.

$P_{25} = Q_1$, $P_{50} = Q_2$, $P_{75} = Q_3$.

Inter Quartile Range (IQR) = $Q_3 - Q_1$. Represents the middle 50% of the data.

↳ middle 50% of the data.

5 # summary :-
min.
Q_1
median
Q_3
maximum



Outlier :- Find I.Q.R. (=8) multiply it by 1.5.
↳ 1.5(IQR). = 12

lower bound = Q_1 minus 1.5(IQR) = -7
upper bound = Q_3 plus 1.5(IQR) = 25

Anything that does not fall in this range is mathematically called an outlier.