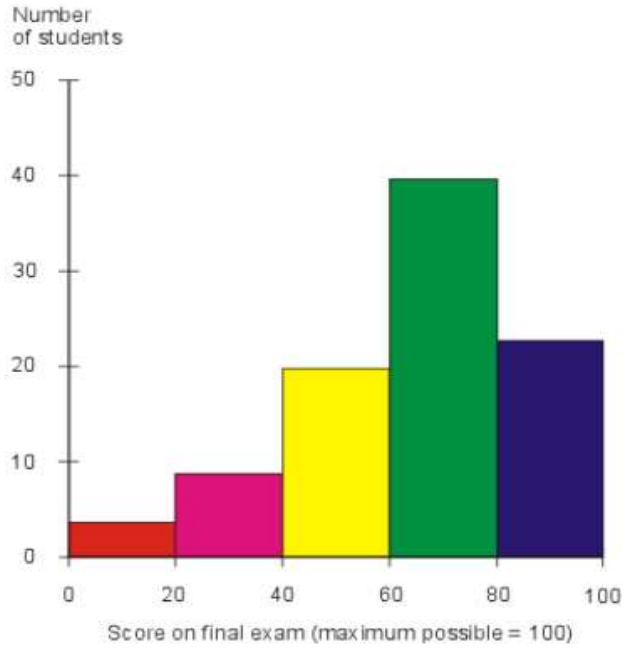
Basics of Statistics

Histograms

Frequency of occurrence of specific phenomena which lie within a specific range of values arranged in consecutive and fixed intervals.

E.g.- Is a histogram showing the results of a final exam given to a hypothetical class of students. Each score range is denoted by a bar of a certain color.



Frequency Distribution

A frequency distribution tells
how frequencies are distributed over values.
Frequency distributions are mostly used for
summarizing categorical variables.

	id	fname	sex	major
1	7042	Piper	female	Sociology
2	7104	Nicole	female	Anthropology
3	8016	Samuel	male	Other
4	8088	Logan	male	Psychology
5	8100	Alexa	female	Anthropology
6	9002	Scarlett	female	Sociology
7	9035	Wyatt	male	Economy

E.g.- We had 183 students fill out a questionnaire. One of the questions was which study major they're following.

Frequency Distribution Table

What's currently your (primary) major?		N	Percent
Psychology		62	33.9%
Economy		35	19.1%
Sociology	FREQUENCIES → ARE DISTRIBUTED OVER	33	18.0%
Anthropology	← VALUES	37	20.2%
Other		16	8.7%
Total		183	100.0%

Descriptive Statistics

Helps us in summarizing a single metric.

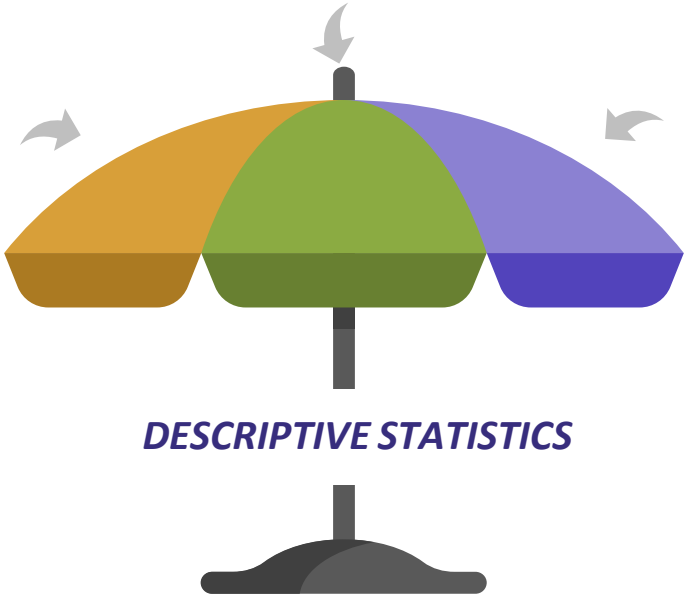
Following are the aspects used together to summarize your metric

Central value of metric

Mean & Median

Spread of values in the metric

Max, Min, Range,
Inter Quartile Range



Variations in the metric

Standard Deviation,
Coefficient of variation

Mean vs Median

1
2
3
4
5
6
7
8
9
100

Mean **is not** a robust measure, **it is** affected by the presence of extreme values

Median **is a** robust measure & **not** affected by the presence of extreme values

Mean = 14.5 (misleading)
Median = 5.5

Robust Mean used along with median gives the complete story

Types of Robust Mean

Trimmed Mean

Drop 10% of observations from each end and calculate the mean.

In the previous example use –
(2+3+4+5+6+7+8+9)/8 = 5.5

Winsorized Mean

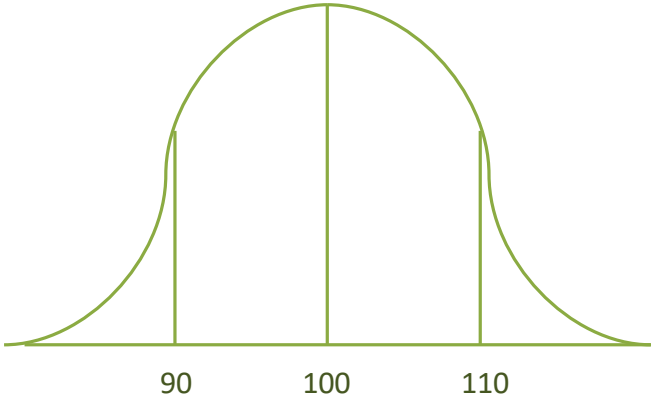
Change any value less than 10% to be equal to the 10%. Change any value more than 90% to be equal to the 90%

In the previous example use –
(2+2+3+4+5+6+7+8+9+9)/10 = 5.5

Standard Deviation Vs Coefficient of variation

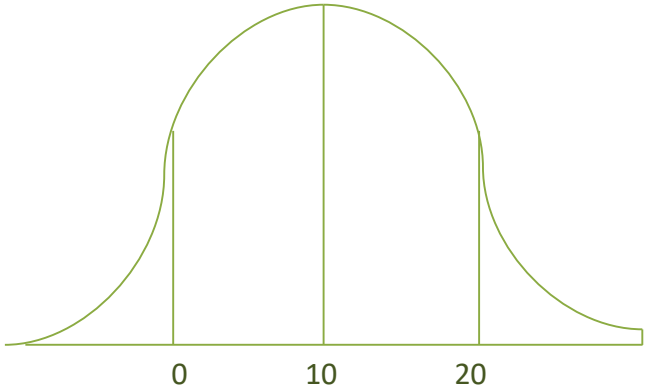
Which stock will you invest in?

Stock 1



M= $100
σ= $5

Stock 2

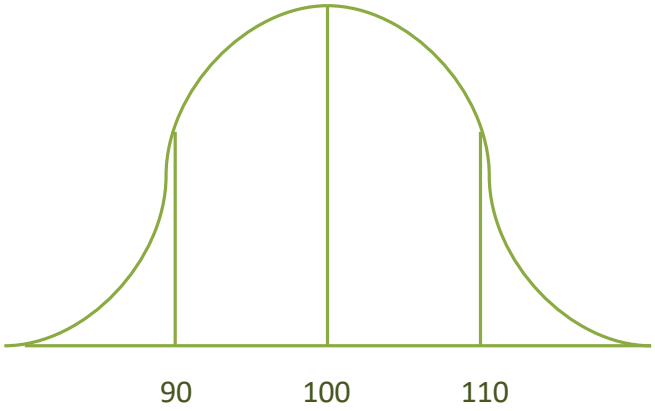


M= $10
σ= $5

Standard Deviation Vs Coefficient of variation

Which stock will you invest in?

Stock 1



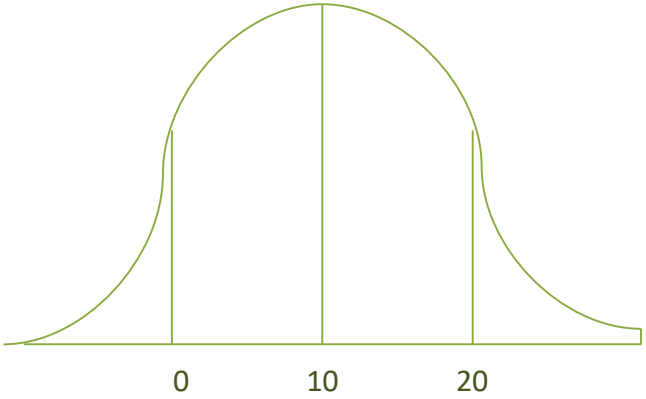
M= $100
σ= $5

CV = 5/100 = 5%

Standard Deviation
Measures Average change from the mean

Coefficient of variation
Measures Standard Deviation per unit Mean

Stock 2



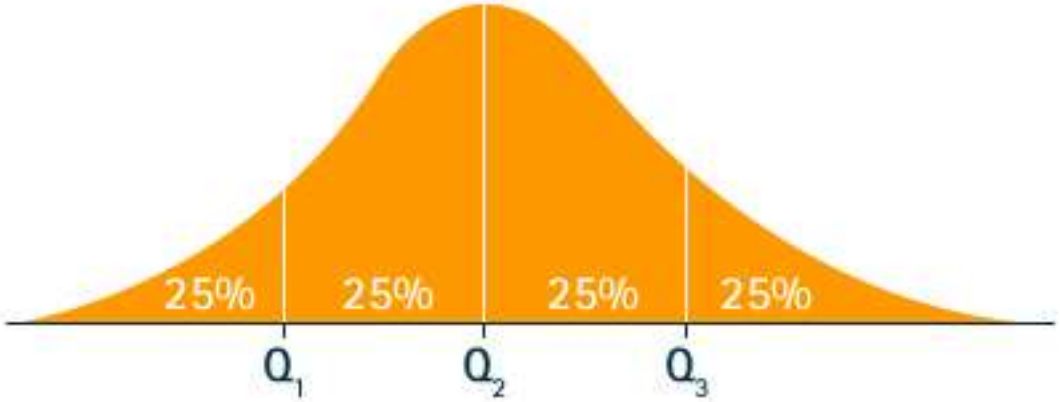
M= $10
σ= $5

CV = 5/10 = 50%

QUARTILES

The median that you just learned about would give you the middle of the data. But, what if you wanted to look closer at the top 10% of the data? Suppose you want to create a segment of most valued customers where "most valued" customer is defined as among the top 10% of buyers by average value.

Quartile statistics can help you with this. A quartile is when you take the data in the histogram and partition it into four equally sized groups. Then you can analyze the data in a particular quartile.



MEASURES OF DISPERSION

Store number	Sales (thousands)	Squared deviation from mean
1	$x_1 = 10$	$(10 - 12)^2 = 4$
2	$x_2 = 8$	$(8 - 12)^2 = 16$
3	$x_3 = 14$	$(14 - 12)^2 = 4$
4	$x_4 = 20$	$(20 - 12)^2 = 64$
5	$x_5 = 11$	$(11 - 12)^2 = 1$
6	$x_6 = 9$	$(9 - 12)^2 = 9$
Totals	$\sum x_i = 72$	$\sum (x_i - \mu)^2 = 98$

The mean is $\mu = 72/6 = 12$

The variance is $\sigma^2 = 98/6 \approx 16.33$

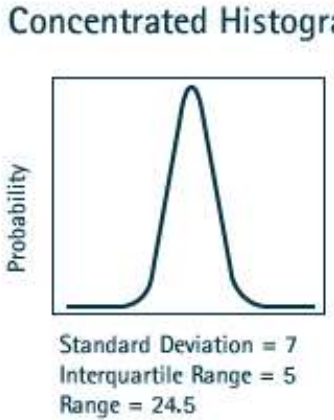
The standard deviation is $\sigma = \sqrt{98/6} \approx 4.04$

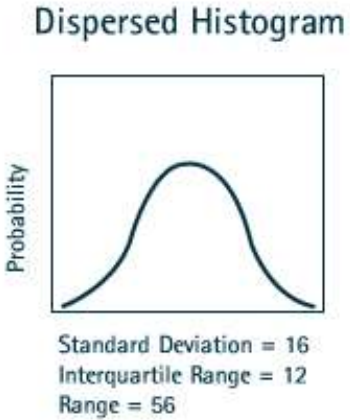
The *sample* variance is $s^2 = 98/(6 - 1) = 19.6$

$$Q3 - Q1 = IQR$$

Maximum – Minimum = Range

WHY SHOULD YOU CARE ABOUT DISPERSION?





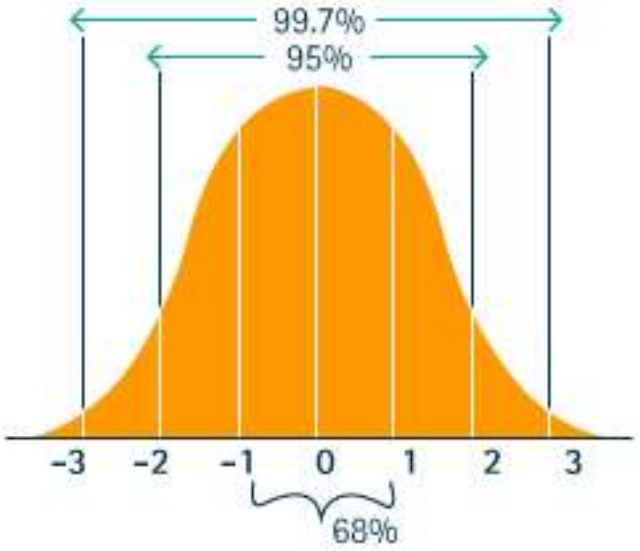
Measures of dispersion enable us to answer the question: If we use a different sample would we come up with different conclusions? Standard deviation allows us to answer that question.

One reason you should care about dispersion in a histogram is for marketing. Marketing addresses situations where people have different wants and needs. A business can segment the market depending on the wants and needs for groups of customers. A business needs to characterize markets as heterogeneous or disperse.

EMPIRICAL RULE : EXAMPLE

To interpret the standard deviation use the Empirical Rule which states that with data from a normal distribution, approximately:

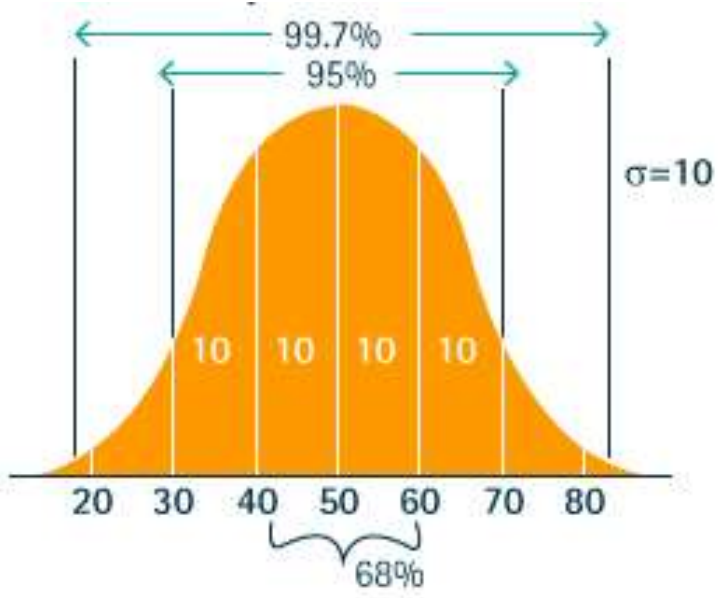
- **68% of the observations will fall within 1 standard deviation of the mean.**
- **95% of the observations will fall within 2 standard deviations of the mean.**
- **99.7% of the observations will fall within 3 standard deviations of the mean.**



EMPIRICAL RULE : EXAMPLE

Let’s take a look at an example. At Company XYZ, employees had to take a test. The mean score is 50 and the standard deviation is 10.

- 1 standard deviation from the mean are the scores of 40 and 60. 68% of the test takers would score between 40 and 60.
- 2 standard deviations from the mean are the scores of 30 and 70. 95% of the test takers would score between 30 and 70.
- 3 standard deviations from the mean are the scores of 20 and 80. 99.7% of the test takers would have a score between 20 and 80.



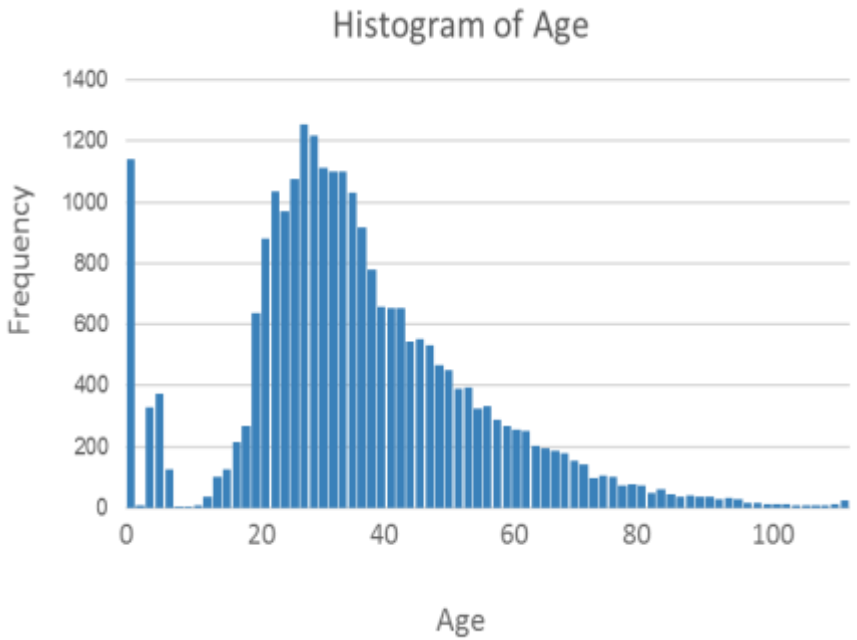
Histograms

A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data.

WHAT IRREGULARITY DO YOU SEE HERE?

FREQUENCY DISTRIBUTION OF AGE

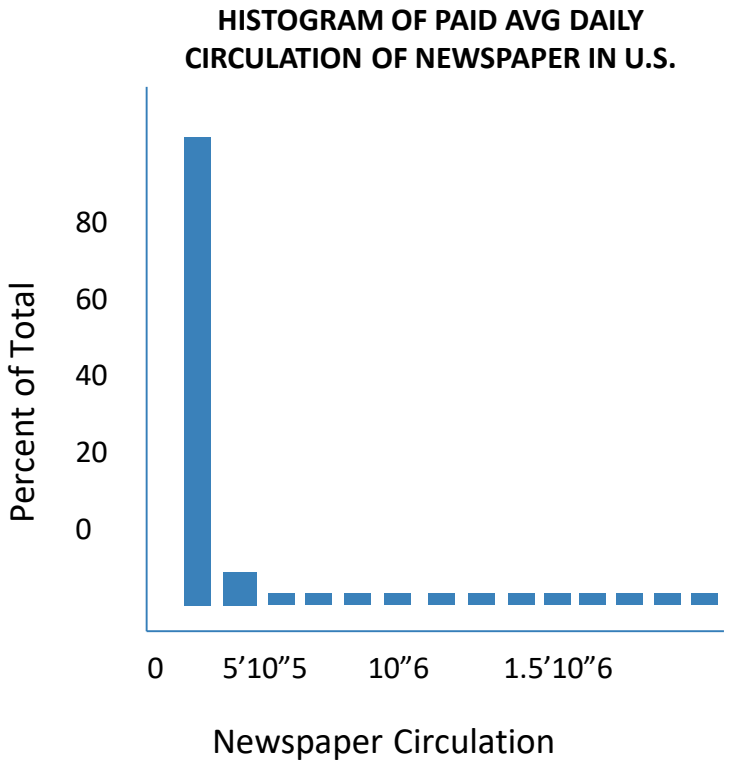
Age	Freq	Age	Freq	Age	Freq	Age	Freq	Age	Freq	Age	Freq	Age	Freq	Age	Freq	Age	Freq
0	1142	15	101	24	1254	33	653	42	325	51	179	60	47	69	29	78	6
1	6	16	125	25	1219	34	655	43	333	52	153	61	60	70	16	79	13
2	330	17	216	26	1112	35	543	44	287	53	144	62	45	71	15	99	26
3	374	18	267	27	1099	36	552	45	267	54	98	63	37	72	13		
4	127	19	636	28	1102	37	533	46	256	55	106	64	42	73	14		
9	1	20	881	29	1030	38	467	47	250	56	102	65	35	74	12		
10	4	21	1035	30	916	39	449	48	204	57	73	66	35	75	6		
13	6	22	971	31	781	40	391	49	194	58	77	67	27	76	9		
14	37	23	1076	32	659	41	393	50	185	59	72	68	34	77	6		



Interpretation of Histograms

What is your interpretation of this histogram

How can you get a better understanding of the underlying story

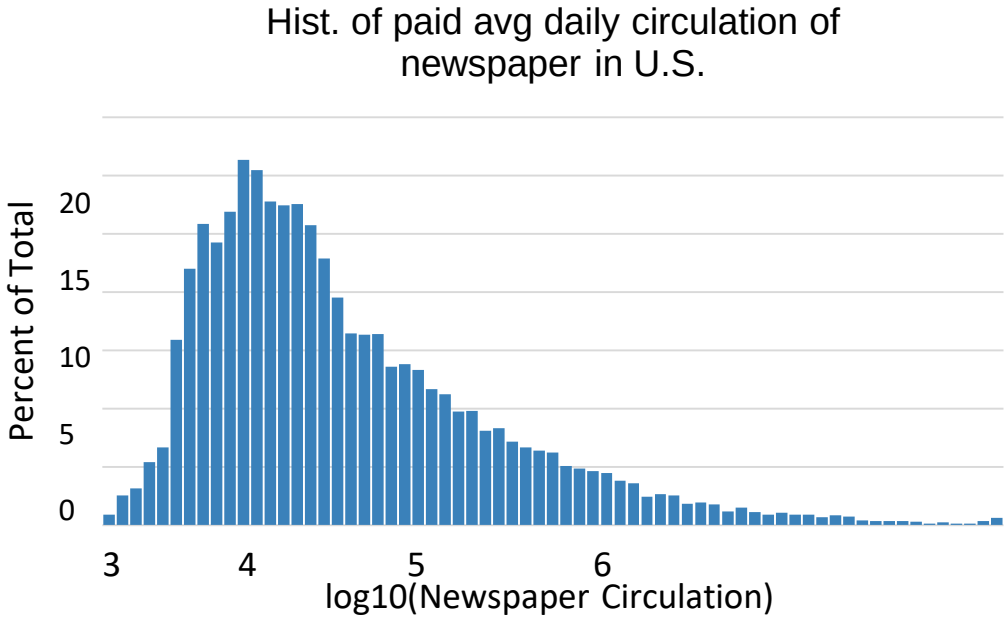


Interpretation of Histograms

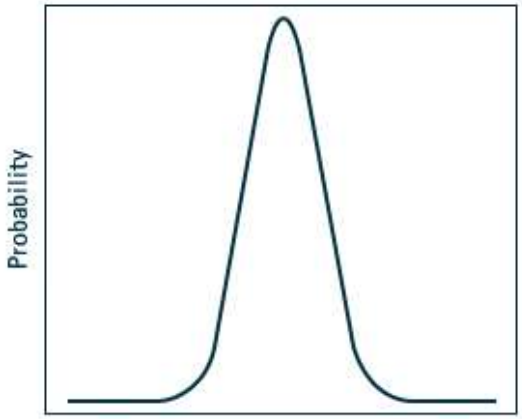
How to fix a right skewed distribution?

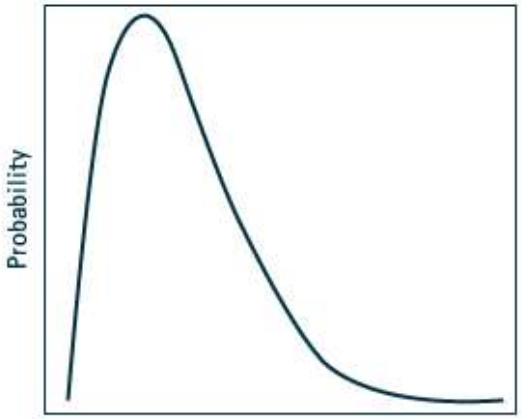
You have to transform the variable on the x-axis by taking the logarithm of the variable.

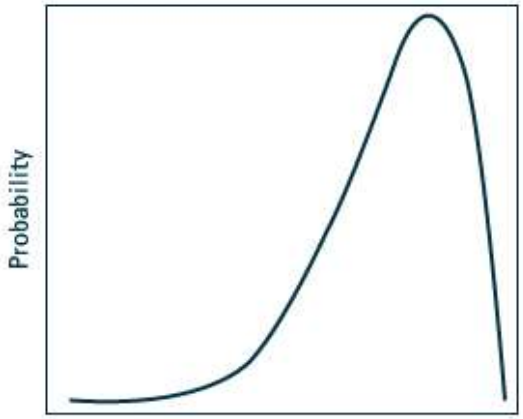
=> Will give a ***normal distribution histogram***



Interpretation of Histograms

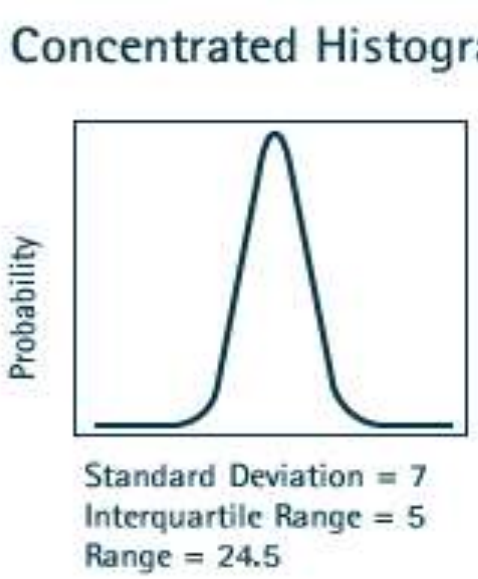


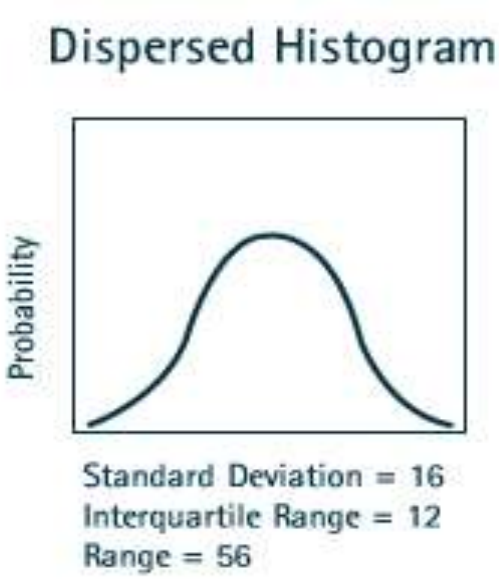




Which histogram represents salary distribution?

Interpretation of Histograms





Why should you care about dispersion?

Percentiles

A **percentile** (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls.

- Percentiles are commonly used to report scores in tests, like the SAT, GRE and LSAT.
- If you know that your score is in the 90th percentile, that means you scored better than 90% of people who took the test.

- The 25th percentile is also called the **first quartile**.
- The 50th percentile is the **median**
- The 75th percentile is also called the **third quartile**.

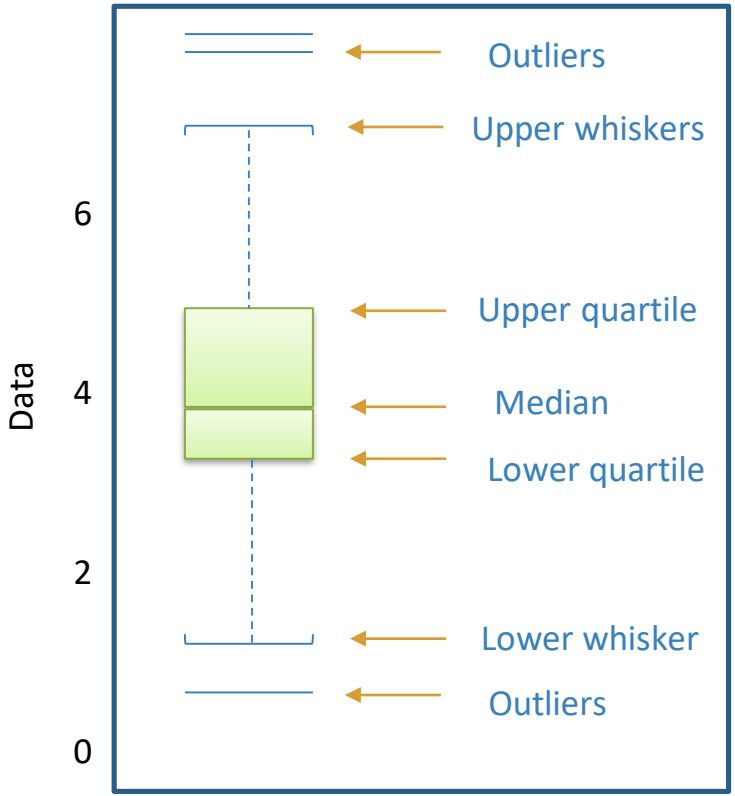
Box & Whisker Plot

A. Is the top view of a histogram

B. Provides details like you get in a histogram but not all the detail

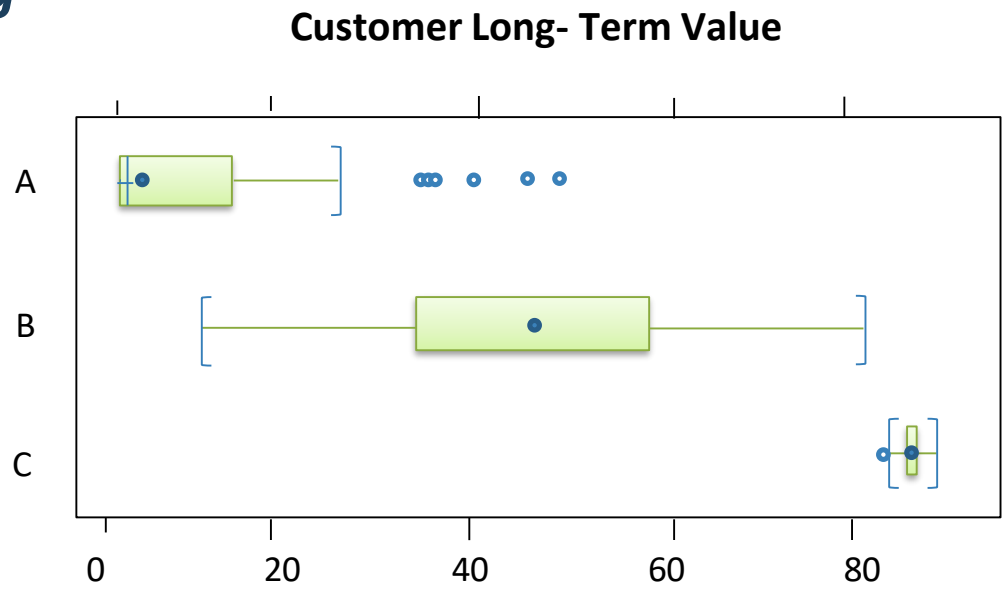
C. At the same time, it provides richer summary that mean/ median

D. Makes it easier to compare performance of a metric across multiple segments



Box & Whisker Plot Interpretation

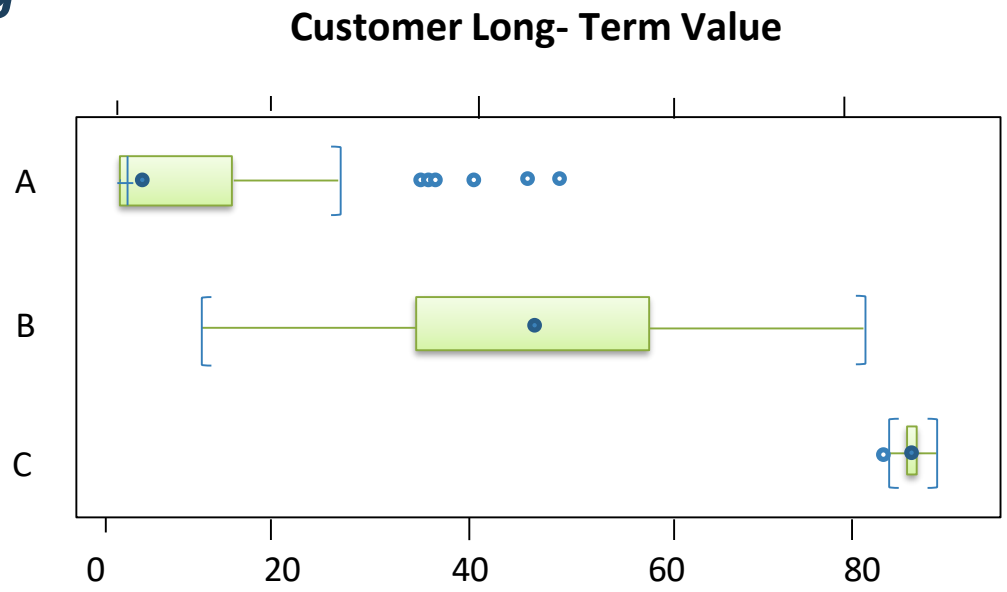
Interpreting



Which segment is highly homogenous

Box & Whisker Plot Interpretation

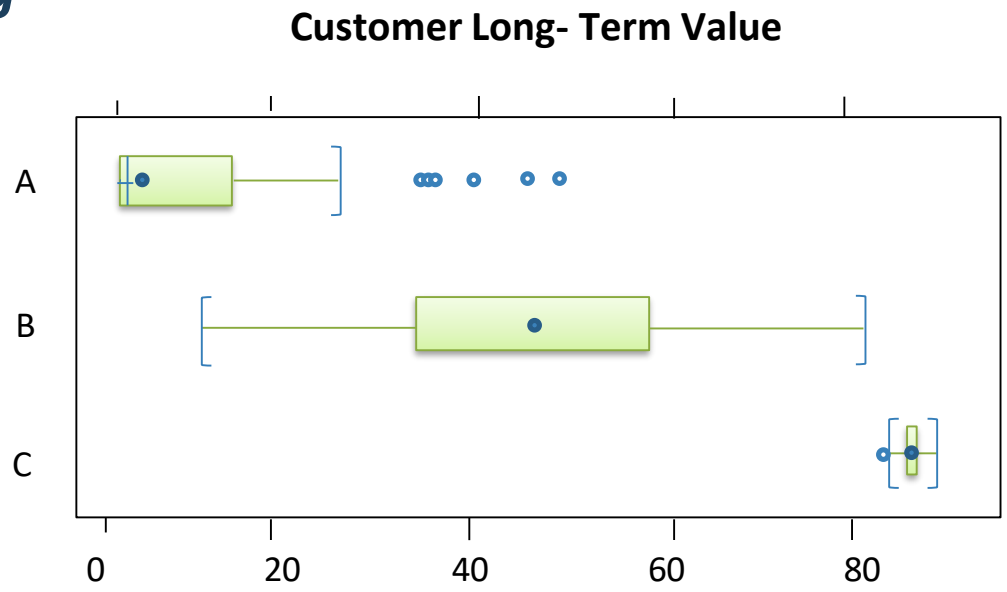
Interpreting



Which one is skewed

Box & Whisker Plot Interpretation

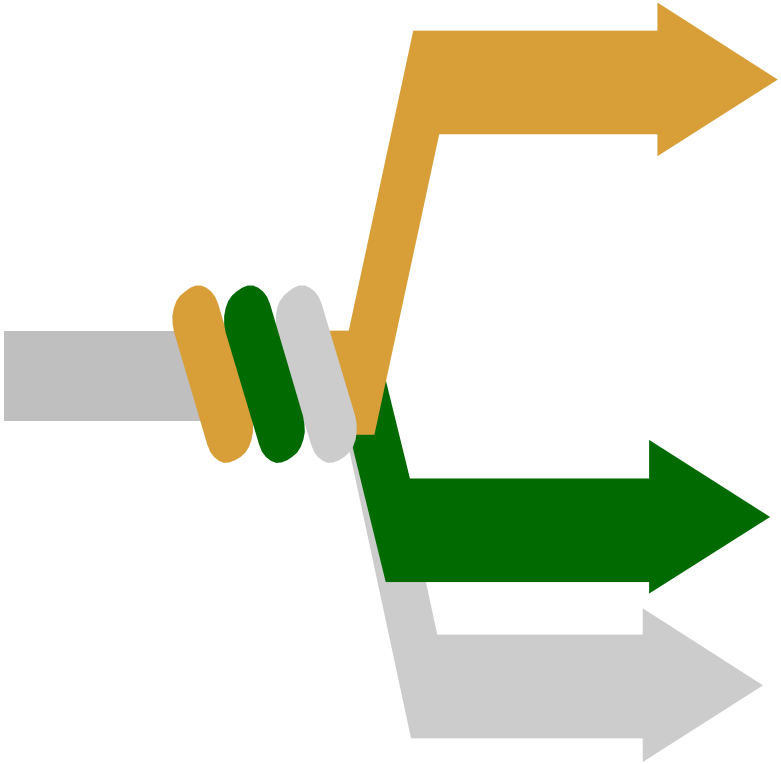
Interpreting



Which is symmetrical

Using Descriptive Analytics Toolkit on your metric

For every metric that you have defined for your issues



Realize the **underlying story in the metric by performing:**
One variable analysis using **descriptive toolkit**–

- Descriptive Statistics
- Histogram
- Box Plot

Look for **need for metric transformation, identify problems like** –

- Outliers
- Missing value

Realize if the distribution is **dispersed** or **concentrated**

Probability

A **probability distribution** is a list of all of the possible outcomes of a random variable along with their corresponding probability values.

It indicates the likelihood of an event or outcome.

Following is the notation to describe probabilities:

p(x) = the likelihood that random variable takes a specific value of x.

The sum of all probabilities for all possible values must equal 1.
The probability for a particular value or range of values must be between 0 and 1.

E.g.-

Outcome of die roll	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6

Sampling

Population

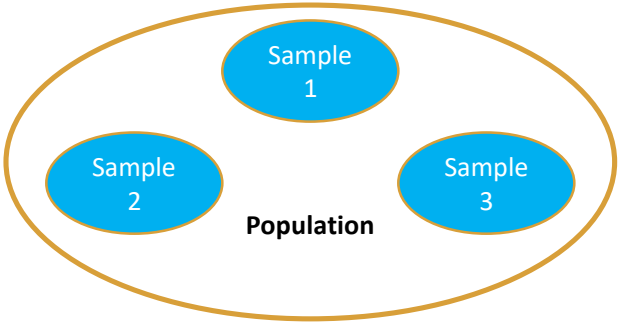
- Not to be confused with literal meaning of "population" which means number of people living in a defined geographical region.
- The "population" in statistics includes all members of a defined group that we are studying or collecting information on for data driven decisions.

Example:
- Current inflation rates of EU countries.
- All the votes casted in an electoral poll.

Sample

- It is a part of the "population".
- Can be biased or un-biased (also know as random sample).

Example:
- Current inflation rates of EU countries having per capita income of less than 20000 Euros per annum.
- A portion of votes collected to predict the election outcome through "Exit Poll".



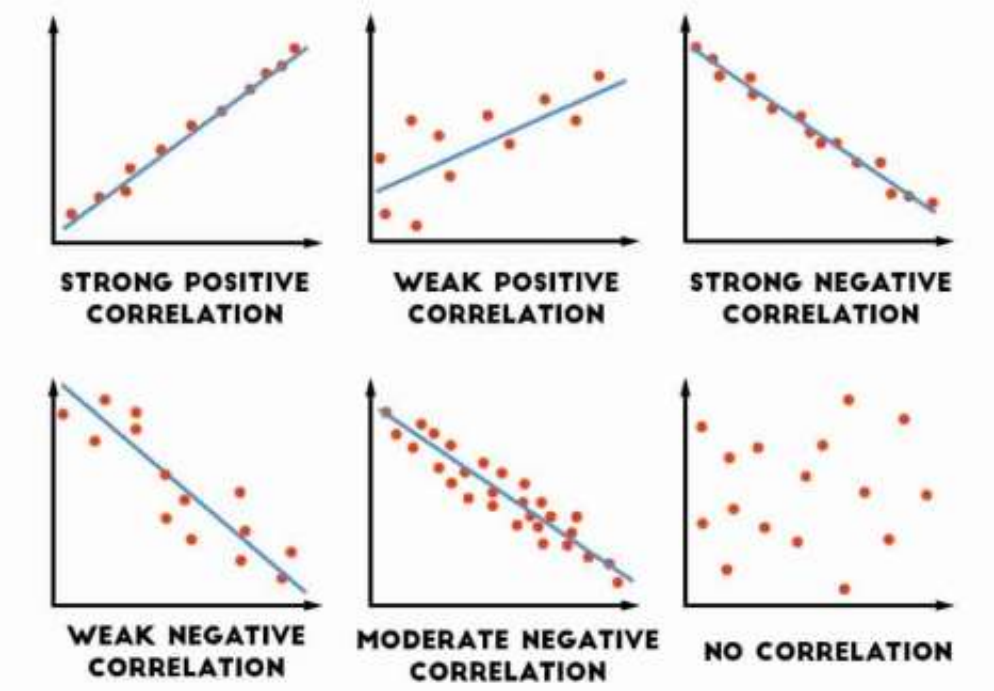
Sampling is a process in which a predetermined number of observations are taken from a larger population.

There are two major types of sampling:
- Simple random sampling
- Stratified sampling

What is Correlation

Correlation is a statistical technique that can show whether and how strongly pair of variables are related.

It’s a standardized metric. The value is between -1 to 1.



https://www.spss-tutorials.com/pearson-correlation-coefficient/

Thanks!

Any questions ?

Next steps

- Attempt quiz –
- Share feedback –