

FMS

Fundamentals of Mathematical Statistics

Venkata Reddy Konasani

Contents

- Central Tendency
- Variance
- Percentiles
- Quartiles
- Outlier Detection
- Box-plot
- Probability
- Conditional Probability
- Binomial Distribution
- Normal Distribution
- Confidence Intervals

Descriptive statistics

Descriptive statistics

- The basic descriptive statistics to give us an idea on the variables and their distributions
- Permit the analyst to describe many pieces of data with a few indices
- Central tendencies
 - Mean
 - Median
- Dispersion
 - Range
 - Variance
 - Standard deviation

Central tendencies: Mean and Median

Central tendencies

- Mean
 - The arithmetic mean
 - Sum of values / Count of values
 - Gives a quick idea on average of a variable

Mean Calculation

Selling_Price		
3.35		
4.75		
7.25	Sum	=SUM(C2:C302)
2.85	Count	=COUNT(C2:C302)
4.6	Average	=F4/F5
9.25		
6.75	With Formula	=AVERAGE(C2:C302)
6.5		
8.75		
7.45		

Sum	1403.05
Count	301
Average	4.661296
With Formula	4.661296

What is the Overall Average cost per item?

Cost_per_item	Items
60	4
70	6
80	5
90	3
100	2

Overall cost / overall items

Calculations:

- $60 \times 4 = 240$
- $70 \times 6 = 420$
- $80 \times 5 = 400$
- $90 \times 3 = 270$
- $100 \times 2 = 200$

Sum of $X \times f$: $240 + 420 + 400 + 270 + 200 = 1530$

Sum of frequencies $\sum f$: $4 + 6 + 5 + 3 + 2 = 20$

$$\text{Mean} = \frac{1530}{20} = 76.5$$

Grouped Frequency Table (Continuous Data)

- Consider a grouped frequency table for the ages of employees in a company
- What is the average age of the employees?

Age Group	Frequency (f)
20 - 29	10
30 - 39	18
40 - 49	12
50 - 59	8
60 - 69	2

Grouped Frequency Table (Continuous Data)

Age Group	(Midpoint X)	Frequency (f)
20 - 29	24.5	10
30 - 39	34.5	18
40 - 49	44.5	12
50 - 59	54.5	8
60 - 69	64.5	2

To calculate the mean:

1. Take the midpoint for each class.
2. Multiply the midpoint by the frequency.
3. Sum these products.
4. Divide by the total frequency.

$$\text{Mean} = \frac{\sum(X \times f)}{\sum f}$$

Calculations:

- $24.5 \times 10 = 245$
- $34.5 \times 18 = 621$
- $44.5 \times 12 = 534$
- $54.5 \times 8 = 436$
- $64.5 \times 2 = 129$

Sum of $X \times f$: $245 + 621 + 534 + 436 + 129 = 1965$

Sum of frequencies $\sum f$: $10 + 18 + 12 + 8 + 2 = 50$

$$\text{Mean} = \frac{1965}{50} = 39.3$$

Guess the mean

- Guess the mean of Indian income

Guess the mean

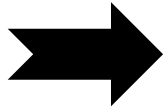
- 1.5, 1.7, 1.9, 0.8, 0.8, 1.2, 1.9, 1.4, 9, 0.7, 1.1

Median

- Mean is not a good measure in presence of outliers
- For example Consider below data vector
 - 1.5, 1.7, 1.9, 0.8, 0.8, 1.2, 1.9, 1.4, 9, 0.7, 1.1
- 90% of the above values are less than 2, but the mean of above vector is 2
- There is an unusual value in the above data vector i.e 9
- It is also known as outlier.
- Mean is not the true middle value in presence of outliers. Mean is very much effected by the outliers.
- We use median, the true middle value in such cases
- Sort the data either in ascending or descending order

Median

1.5		0.7
1.7		0.8
1.9		0.8
0.8		1.1
0.8		1.2
1.2		1.4
1.9		1.5
1.4		1.7
9		1.9
0.7		1.9
1.1		9



- Mean of the data is 2
- Median of the data is 1.4
- Even if we have the outlier as 90, we will have the same median
- Median is a positional measure, it doesn't really depend on outliers
- When there are no outliers then mean and median will be nearly equal
- When mean is not equal to median it gives us an idea on presence of outliers in the data

LAB- Mean and Median Calculations

Kms_Driven
27000
43000
6900
5200
42450
2071
18796
33429
20273
42367
2135
51000

Mean	40461
Median	32000

Dispersion Measures : Variance and Standard Deviation

Dispersion

- Just knowing the central tendency is not enough.
- Two variables might have same mean, but they might be very different.
- Look at these two variables. Profit details of two companies A & B for last 14 Quarters in MMs

															Mean
Company A	43	44	0	25	20	35	-8	13	-10	-8	32	11	-8	21	15
Company B	17	15	12	17	15	18	12	15	12	13	18	18	14	14	15

- Though the average profit is 15 in both the cases
- Company B has performed consistently than company A.
- There was even losses for company A
- Measures of dispersion become very vital in such cases

Variance and Standard deviation

- Dispersion is the quantification of deviation of each point from the mean value.
- Variance is average of squared distances of each point from the mean
- Variance is a fairly good measure of dispersion.
- Variance in profit for company A is 352 and Company B is 4.9

Value	Value-Mean	(Value-Mean)^2
43	28	784
44	29	841
0	-15	225
25	10	100
20	5	25
35	20	400
-8	-23	529
13	-2	4
-10	-25	625
-8	-23	529
32	17	289
11	-4	16
-8	-23	529
21	6	36
15.0		352

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Value	Value-Mean	(Value-Mean)^2
17	2	4
15	0	0
12	-3	9
17	2	4
15	0	0
18	3	9
12	-3	9
15	0	0
12	-3	9
13	-2	4
18	3	9
18	3	9
14	-1	1
14	-1	1
15.0		4.9

Standard Deviation

- Standard deviation is just the square root of variance
- Variance gives a good idea on dispersion, but it is of the order of squares.
- Its very clear from the formula, variance unites are squared than that of original data.
- Standard deviation is the variance measure that is in the same units as the original data

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

LAB: Variance and Standard deviation

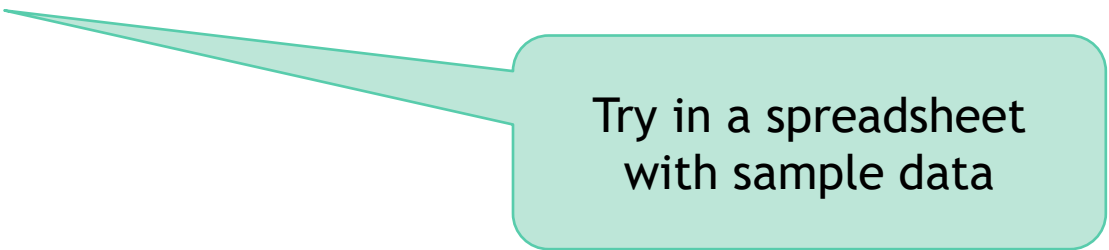
- Calculate the Variance and SD of selling price in Fuel_type “Petrol” vs “Other”
- Where is higher variance in selling price ?

Sample Question in the Exam

Q83

A spreadsheet contains the monthly salaries of 100 faculty members at an IIM. If you were to **increase** the salaries of all the faculty members by a generous **Rs.1,44,000**, then what would happen to the **standard deviation** of the salaries?

- ☐ Remains unchanged
- ☐ Will increase by Rs.12,000
- ☐ Will decrease by Rs.12,000
- ☐ Will increase by Rs.1,44,000



Try in a spreadsheet
with sample data

Percentiles & Quartiles

Percentiles

- A student attended an exam along with 1000 others.
 - He got 68% marks? How good or bad he performed in the exam?
 - What will be his rank overall?
 - What will be his rank if there were 100 students overall?
- For example, with 68 marks, he stood at 90th position. There are 910 students who got less than 68, only 89 students got more marks than him
- He is standing at 91 percentile.
- Instead of stating 68 marks, 91% gives a good idea on his performance
- Percentiles make the data easy to read

Percentiles

- p^{th} percentile: p percent of observations below it, $(100 - p)\%$ above it.
- Marks are 40 but percentile is 80%, what does this mean?
- 80% of CAT exam percentile means
 - 20% are above & 80% are below
- Percentiles help us in getting an idea on outliers.
- For example the highest income value is 400,000 but 95th percentile is 20,000 only. That means 95% of the values are less than 20,000. So the values near 400,000 are clearly outliers

LAB : Percentiles

Kms_Driven	
27000	
43000	
6900	
5200	
42450	
2071	
18796	
33429	
20273	
42367	
2135	
51000	
15000	

Percentile	Value
0	=PERCENTILE.INC(\$C\$2:\$C\$302,E4)
0.1	=PERCENTILE.INC(\$C\$2:\$C\$302,E5)
0.2	=PERCENTILE.INC(\$C\$2:\$C\$302,E6)
0.3	=PERCENTILE.INC(\$C\$2:\$C\$302,E7)
0.4	=PERCENTILE.INC(\$C\$2:\$C\$302,E8)
0.5	=PERCENTILE.INC(\$C\$2:\$C\$302,E9)
0.6	=PERCENTILE.INC(\$C\$2:\$C\$302,E10)
0.7	=PERCENTILE.INC(\$C\$2:\$C\$302,E11)
0.8	=PERCENTILE.INC(\$C\$2:\$C\$302,E12)
0.9	=PERCENTILE.INC(\$C\$2:\$C\$302,E13)
1	=PERCENTILE.INC(\$C\$2:\$C\$302,E14)

LAB : Percentiles

Percentile	Value
0	500
0.1	6000
0.2	12900
0.3	18000
0.4	24524
0.5	32000
0.6	38600
0.7	45000
0.8	53460
0.9	69341
1	500000

Percentile	Value
0.9	69341
0.91	71000
0.92	72000
0.93	77427
0.94	78000
0.95	87934
0.96	89000
0.97	127000
0.98	142000
0.99	304707
1	500000

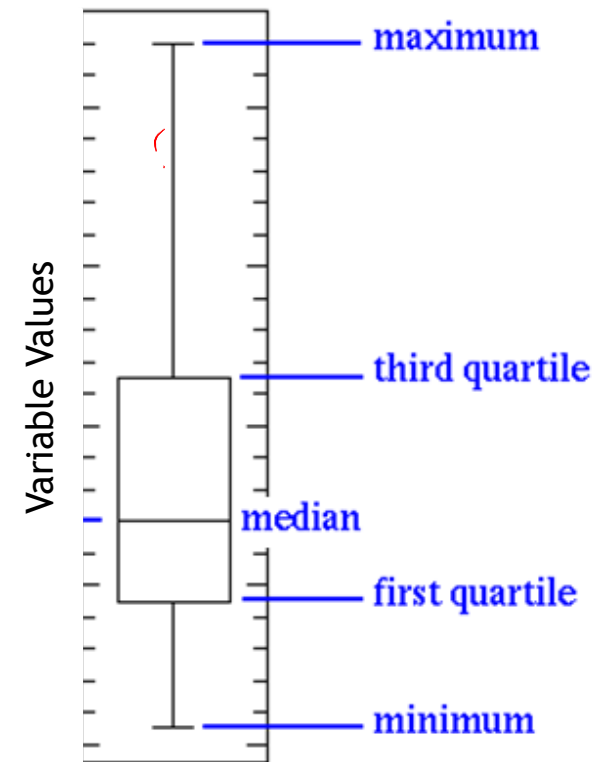
Quartiles

- Percentiles divide the whole population into 100 groups whereas quartiles divide the population into 4 groups
- $p = 25$: First Quartile or Lower quartile (LQ)
- $p = 50$: second quartile or Median
- $p = 75$: Third Quartile or Upper quartile (UQ)

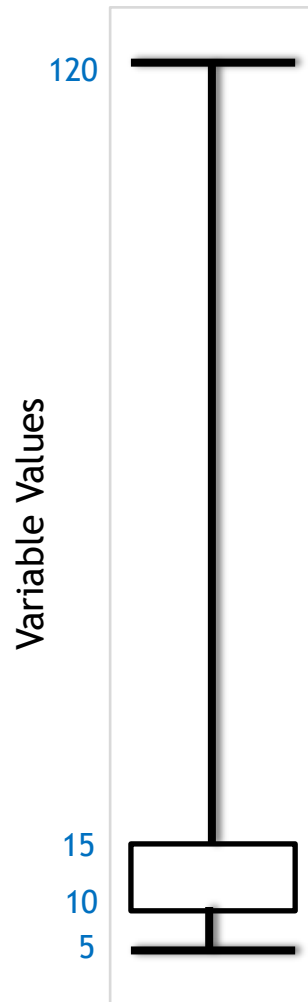
Box plots and outlier detection

Box plots and outlier detection

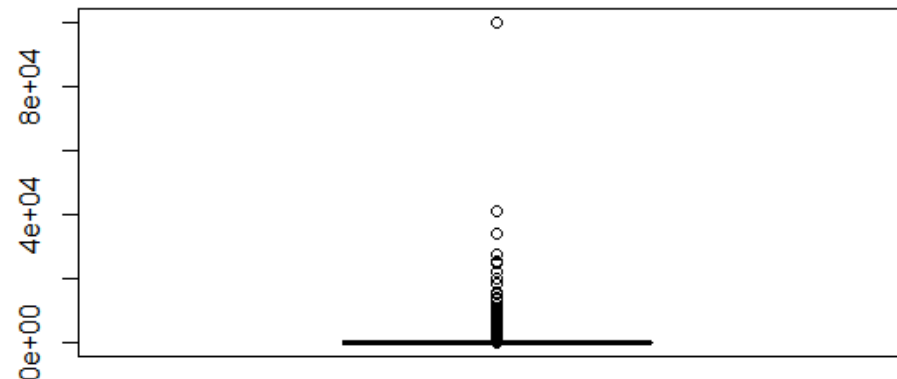
- Box plots have box from LQ to UQ, with median marked.
- They portray a five-number graphical summary of the data Minimum, LQ, Median, UQ, Maximum
- Helps us to get an idea on the data distribution
- Helps us to identify the outliers easily
- 25% of the population is below first quartile,
- 75% of the population is below third quartile
- If the box is pushed to one side and some values are far away from the box then it's a clear indication of outliers



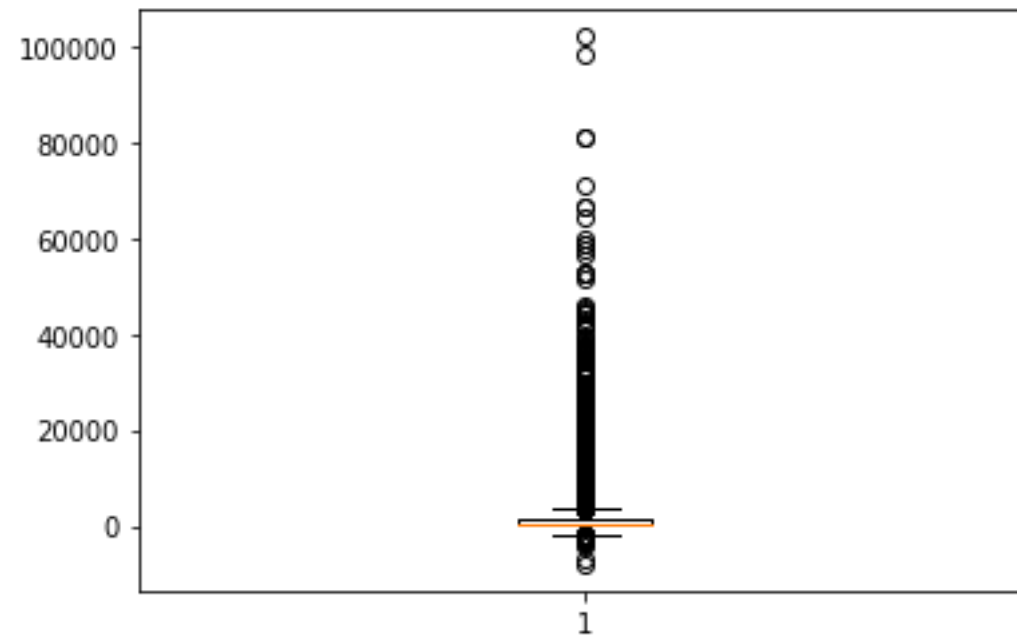
Box plots and outlier detection



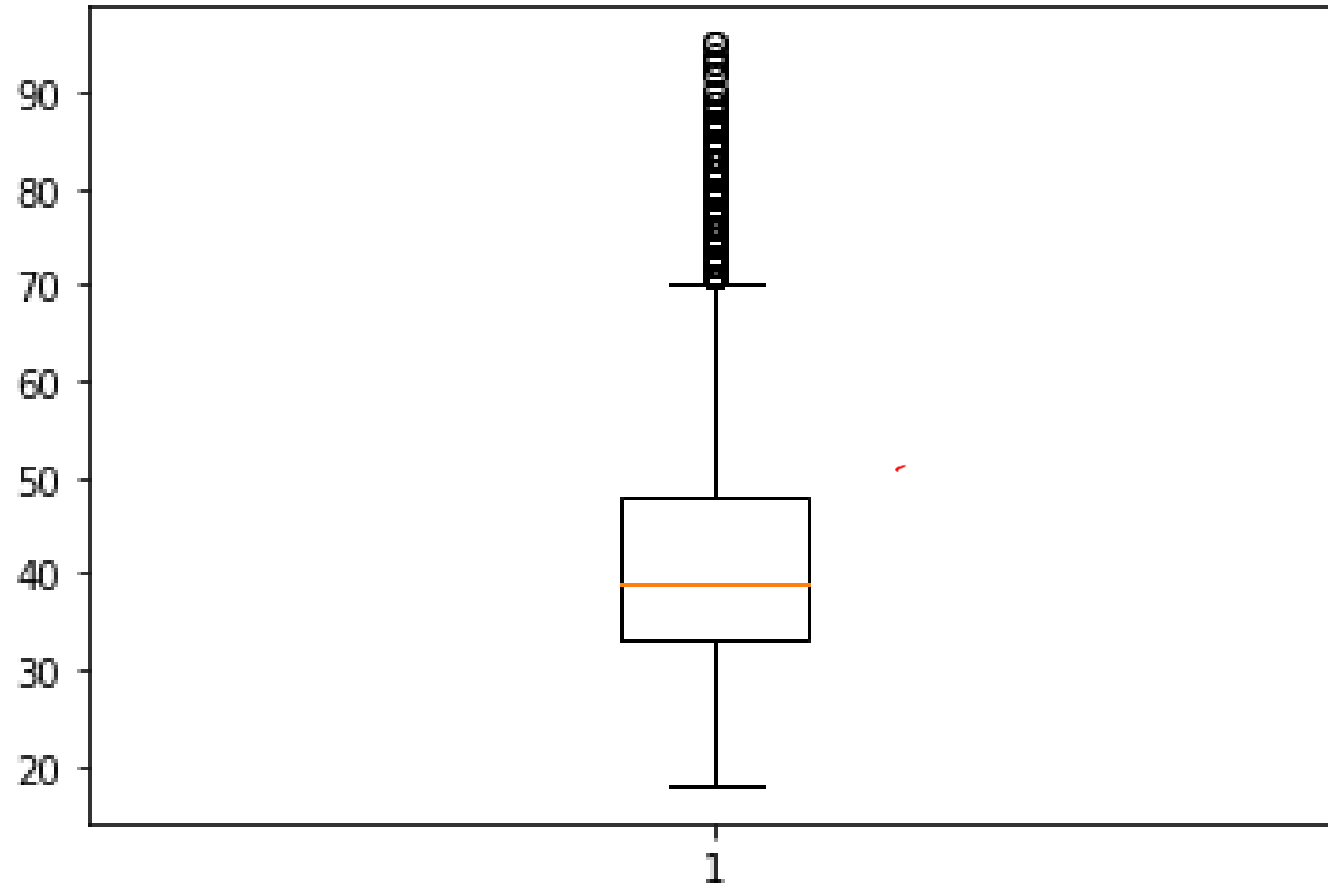
- Some set of values far away from box, is gives us a clear indication of outliers.
- In this example the minimum is 5, maximum is 120, and 75% of the values are less than 15
- Still there are some records reaching 120. Hence a clear indication of outliers
- Sometimes the outliers are so evident that, the box appear to be a horizontal line in box plot.



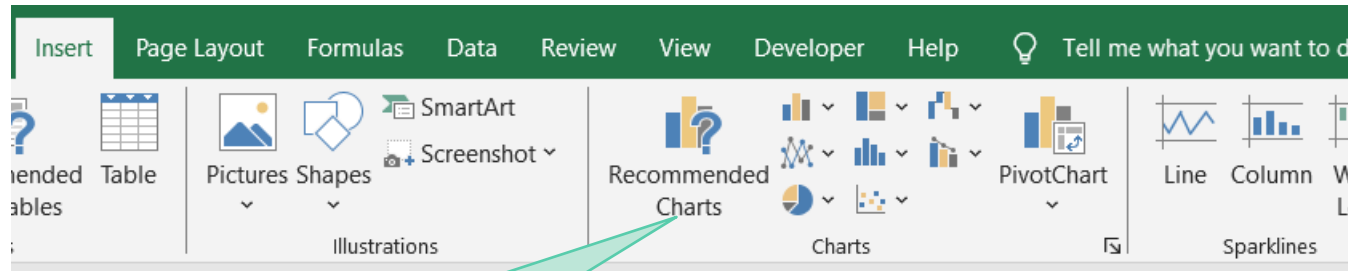
Output - Balance



Output - Age

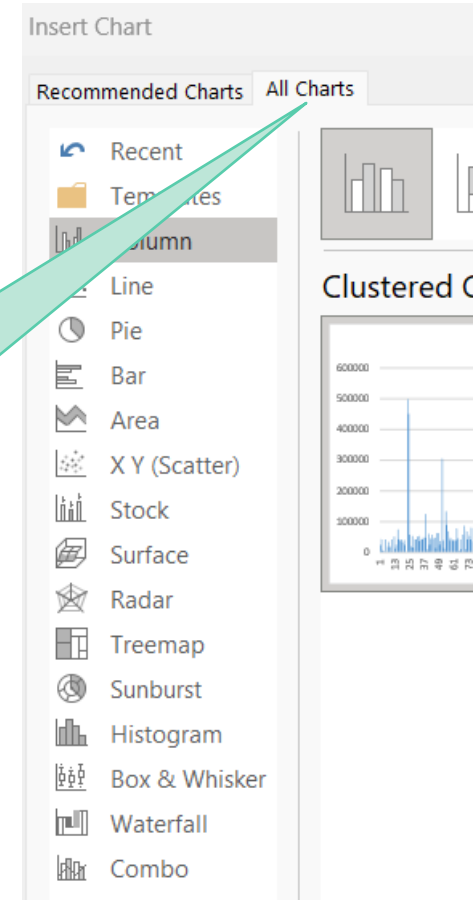


LAB: Box Plot

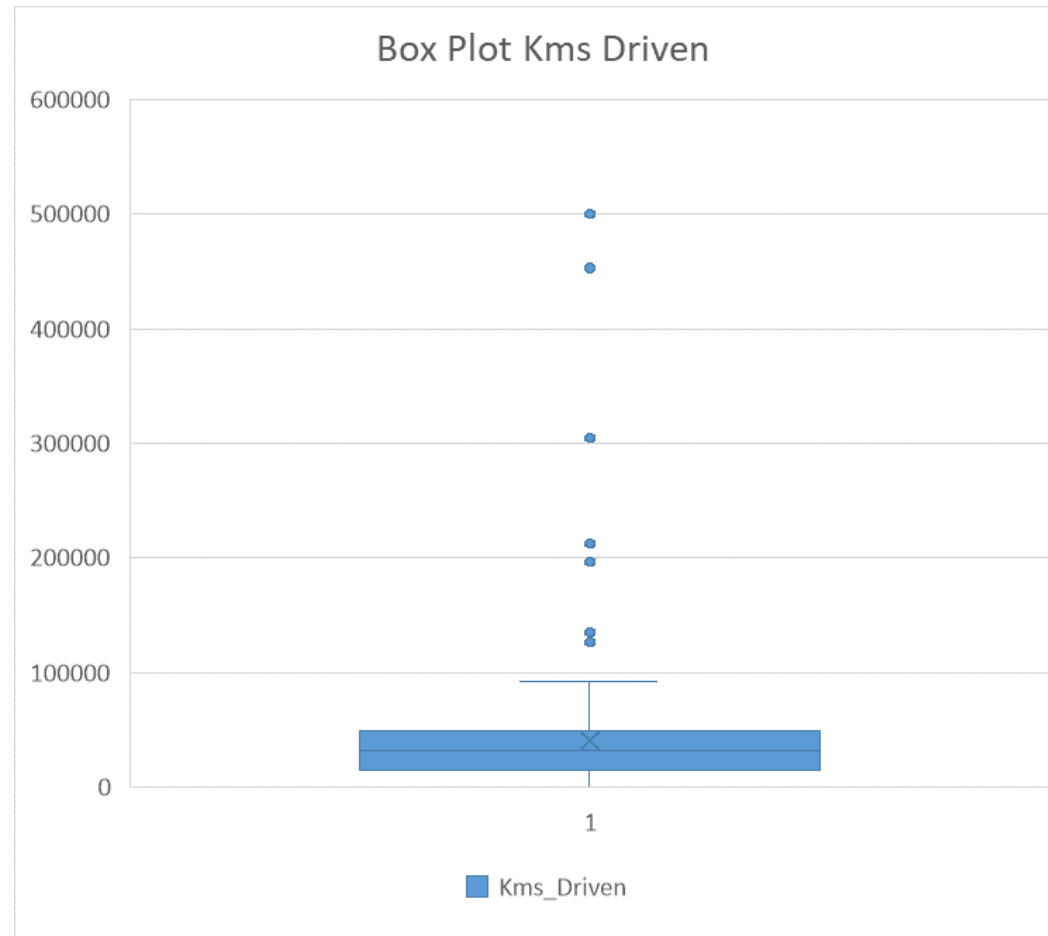


1) Select the data and click on recommended charts

2) Click on All charts then Box & Whisker

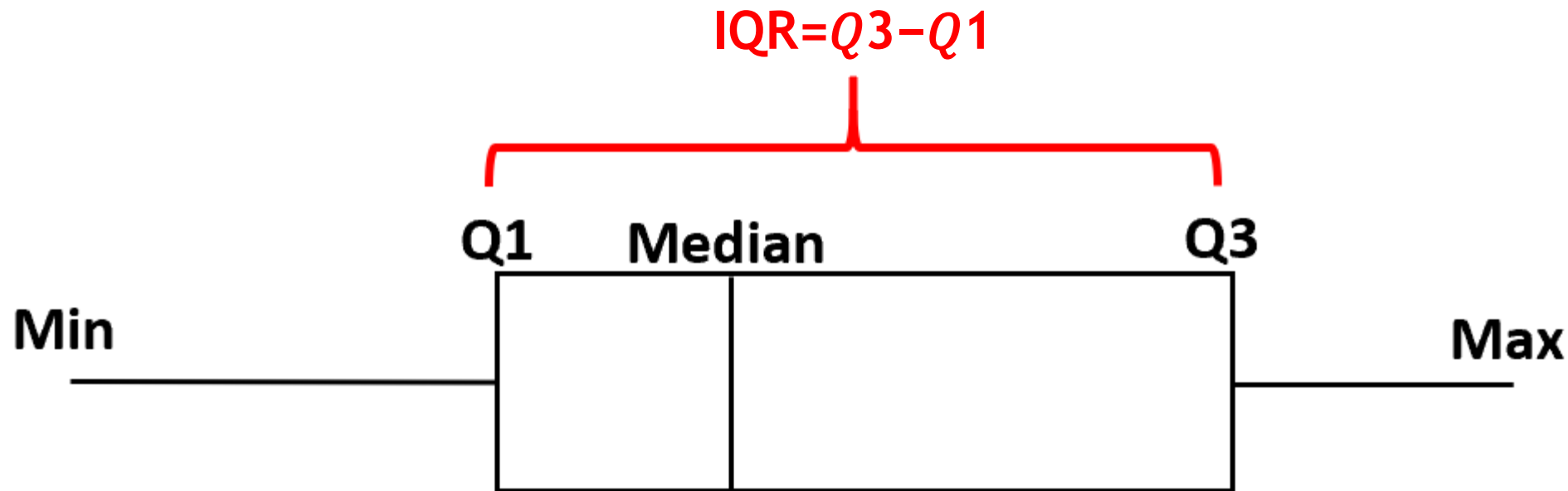


LAB: Box Plot



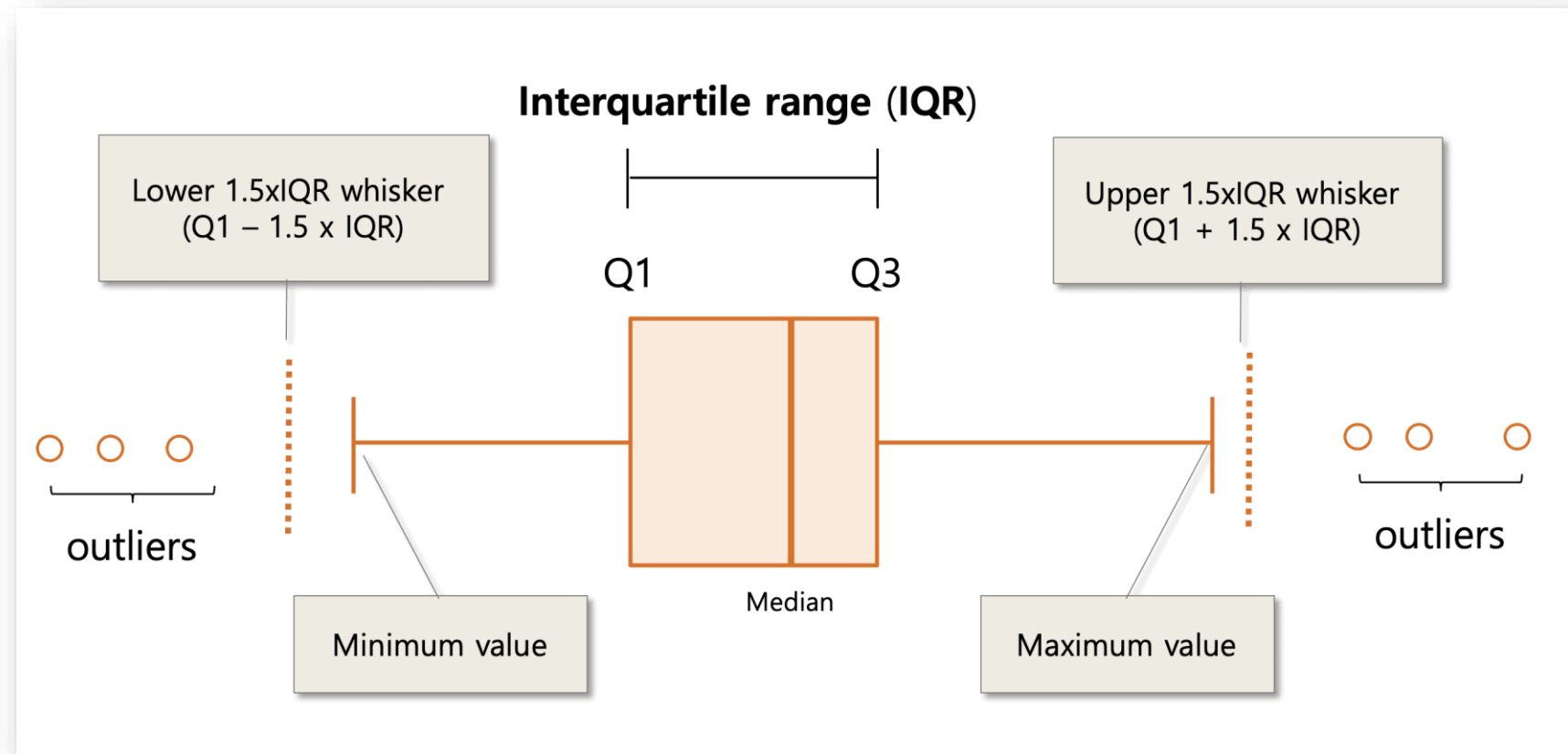
Interquartile Range (IQR)

- The Interquartile Range (IQR) in a box plot represents the range within which the central 50% of the data lies.
- It is the difference between the third quartile (Q3) and the first quartile (Q1):



Outliers Definition

- Outliers are typically defined as points that
 - Fall below $Q1 - 1.5 \times IQR$
 - Fall above $Q3 + 1.5 \times IQR$



LAB: IQR and Outliers

Percentile	Kms Driven Value
0	=PERCENTILE.INC(\$C\$2:\$C\$302,E4)
0.25	=PERCENTILE.INC(\$C\$2:\$C\$302,E5)
0.5	=PERCENTILE.INC(\$C\$2:\$C\$302,E6)
0.75	=PERCENTILE.INC(\$C\$2:\$C\$302,E7)
1	=PERCENTILE.INC(\$C\$2:\$C\$302,E8)

IQR	=F7-F5
Upper Fence	=F7+1.5*F11
Lower Fence	=F5-1.5*F11

LAB: IQR and Outliers

Percentile	Kms Driven Value
0	=PERCENTILE.INC(\$C\$2:\$C\$302,E4)
0.25	=PERCENTILE.INC(\$C\$2:\$C\$302,E5)
0.5	=PERCENTILE.INC(\$C\$2:\$C\$302,E6)
0.75	=PERCENTILE.INC(\$C\$2:\$C\$302,E7)
1	=PERCENTILE.INC(\$C\$2:\$C\$302,E8)

IQR	=F7-F5
Upper Fence	=F7+1.5*F11
Lower Fence	=F5-1.5*F11

Percentile	Kms Driven Value
0.00	500
0.25	15000
0.50	32000
0.75	49000
1.00	500000

IQR	34000
Upper Fence	100000
Lower Fence	-36000

Sample Question

Q47

A spreadsheet contains the monthly salaries of 100 faculty members at an IIM. You are also given the following data:

Quartile 1	131000
Quartile 2	148000
Quartile 3	168000

Then, the outliers on the "upper end" begin at what value?



Probability

Conditional Probability (Given prior information)

- What is the probability that any given day is a Saturday ?
- Given that it's a weekend, what is the probability of Saturday?
- Given yesterday was Friday, what's the probability of Saturday?

- Given Prior information, the probability changes.
- The denominator changes.

Exam Questions

The pivot table below depicts the **plan subscriptions** in 297 customers of a telecom company who have churned (meaning, have switched to another Telecom company).

Churn?	Yes		
Count of Customer	Voice Mail Plan -->		
International Plan	no	yes	Grand Total
no	191	26	217
yes	61	19	80
Grand Total	252	45	297

Given that a churned customer did **NOT** have an International plan, what is the probability that they had a Voice Mail plan?

- ☐ 26 / 45
- ☐ 19 / 80
- ☐ 26 / 217
- ☐ 61 / 80

Exam Questions

Churn?	Yes		
Count of Customer	Voice Mail Plan -->		
International Plan	no	yes	Grand Total
no	191	26	217
yes	61	19	80
Grand Total	252	45	297

Given that a churned customer did **NOT** have an Voice Mail plan, what is the probability that they had an International Plan?

- ☐ 191 / 252
- ☐ 191 / 297
- ☐ 191 / 252
- ☐ 61 / 252

The pivot table below depicts the **plan subscriptions** in 297 customers of a telecom company who have churned (meaning, have switched to another Telecom company).

Churn?	Yes		
Count of Customer	Voice Mail Plan -->		
International Plan	no	yes	Grand Total
no	191	26	217
yes	61	19	80
Grand Total	252	45	297

There are **2094** customers in the original dataset from which this pivot table has been constructed, by restricting the attention to only those customers who have churned.

In a random sample of **100** customers from the original dataset, approximately how many are **statistically likely** to churn?

- ☐ 14
- ☐ 33
- ☐ 7
- ☐ 19

Marginal Probability

	Pass	Fail	Total
Male	30	20	50
Female	40	10	50
Total	70	30	100

- Probability of being Male $P(\text{Male}) =$
- Probability of being Female $P(\text{Female}) =$
- Probability of pass $P(\text{Pass}) =$
- Probability of fail $P(\text{Fail}) =$

Marginal Probability

	Pass	Fail	Total
Male	30	20	50
Female	40	10	50
Total	70	30	100

- Probability of being Male $P(\text{Male}) = 50/100 = 0.5$
- Probability of being Female $P(\text{Female}) = 50/100 = 0.5$
- Probability of pass $P(\text{Pass}) = 70/100 = 0.7$
- Probability of fail $P(\text{Fail}) = 30/100 = 0.3$

These are known as marginal probabilities

Joint Probability

	Pass	Fail	Total
Male	30	20	50
Female	40	10	50
Total	70	30	100

- Probability $P(\text{Male and Pass}) =$
- Probability $P(\text{Female and Pass}) =$
- Probability $P(\text{Male and Fail}) =$
- Probability $P(\text{Female and Fail}) =$

Joint Probability

	Pass	Fail	Total
Male	30	20	50
Female	40	10	50
Total	70	30	100

- Probability $P(\text{Male and Pass}) = 0.3$
- Probability $P(\text{Female and Pass}) = 0.4$
- Probability $P(\text{Male and Fail}) = 0.2$
- Probability $P(\text{Female and Fail}) = 0.1$

Is Exam Result independent of Gender?

Section-A	Pass	Fail	
Male	10	40	50
Female	50	0	50
	60	40	100

Section-B	Pass	Fail	
Male	30	20	50
Female	30	20	50
	60	40	100

Independence Theory in Probability

- In probability theory, independence refers to the relationship between two or more events where the occurrence of one event does not influence the occurrence of the other event(s).
- Two events are said to be independent if knowing the outcome of one event does not change the probability of the other event.

Independence Theory in Probability

Two events A and B are **independent** if the probability that both events occur is the product of their individual probabilities. Mathematically, events A and B are independent if:

$$P(A \cap B) = P(A) \times P(B)$$

where:

- $P(A \cap B)$ is the probability that both A and B happen (joint probability).
- $P(A)$ is the probability of event A .
- $P(B)$ is the probability of event B .

If this equality holds, the events are independent. If not, the events are dependent.

Is Exam Result independent of Gender?

Section-A	Pass	Fail	
Male	10	40	50
Female	50	0	50
	60	40	100

Section-B	Pass	Fail	
Male	30	20	50
Female	30	20	50
	60	40	100

Is Exam Result independent of Gender?

Section-A	Pass	Fail	
Male	10	40	50
Female	50	0	50
	60	40	100

- Probability $P(\text{Male}) = 0.5$
- Probability $P(\text{Pass}) = 0.6$
- Probability $P(\text{Male and Pass}) = 0.1$
- $P(\text{Male and Pass}) \neq P(\text{Male}) * P(\text{Pass})$
- Exam Result independent of Gender? - **No**

Is Exam Result independent of Gender?

- Probability $P(\text{Male}) = 0.5$
- Probability $P(\text{Pass}) = 0.6$
- Probability $P(\text{Male and Pass}) = 0.3$
- $P(\text{Male and Pass}) = P(\text{Male}) * P(\text{Pass})$
- Exam Result independent of Gender? - Yes

Section-B	Pass	Fail	
Male	30	20	50
Female	30	20	50
	60	40	100

LAB: Independence test

	Employed	Unemployed	Total
Graduate	80	20	100
Non-Graduate	50	50	100
Total	130	70	200

Is employment status independent of education level?

LAB: Independence test

	Employed	Unemployed	Total
Graduate	80	20	100
Non-Graduate	50	50	100
Total	130	70	200

Is employment status independent of education level? - No

Random Variable

Random Variable

- A random variable is a variable that represents the outcomes of a random process or experiment. It assigns numerical values to the results of that experiment
- Examples of Random Variables:
 - **Number of Heads in 10 Coin Flips (Discrete):** The number of heads obtained when flipping a coin 10 times could be 0, 1, 2, ..., up to 10.
 - **Sum of Two Dice Rolls (Discrete):** The possible values are the sums of the faces of two dice, ranging from 2 to 12.
 - **Number of Customers Arriving at a Store in an Hour (Discrete):** This could be any non-negative integer such as 0, 1, 2, etc.
 - **Number of Defective Items in a Batch (Discrete):** Out of 100 items, the number of defective ones could be 0, 1, 2, ..., 100.
 - **Lifetime of a Light Bulb (Continuous):** The time a light bulb lasts before it burns out, measured in hours, could be any positive real number

Random Variable

- **Height of a Person (Continuous):** Height in centimeters is a continuous random variable that can take any value within a range, like between 100 and 200 cm.
- **Waiting Time at a Bus Stop (Continuous):** The amount of time (in minutes or hours) a person waits for the bus, which could be any non-negative real number.
- **Score on a Standardized Test (Discrete):** The total score a student gets on a standardized exam, which can only take certain numerical values (e.g., 200, 201, ..., 800).
- **Daily Rainfall Amount (Continuous):** The amount of rain (in millimeters) that falls in a day is a continuous random variable.
- **Number of Emails Received in a Day (Discrete):** This could be any whole number depending on the day, such as 0, 5, 20, etc.
- **Score in a Basketball Game:** The total points your favorite team scores in a game is a random variable. It could be 80, 95, or any number depending on how the game goes.
- **Temperature Tomorrow:** The temperature outside tomorrow is a random variable. You know it could be, say, 20° C or 25° C, but you can't be sure until tomorrow arrives.

What are **not** random variables?

- Number of planets
- Distance between Bangalore and Delhi
- The Number of Days in a Week
- Year of Birth of a person
- The Number of Hours in a Day
- The Number of Pages in a Book
- The Temperature of Boiling Water at Sea Level
 - The boiling point of water at sea level is 100°C (212°F) and is a fixed, non-random value under those conditions.

Probability Distribution

Probability Distribution

- A **probability distribution** describes how the values of a random variable are distributed.
- It tells you what the possible outcomes of an experiment are and how likely each outcome is.
- In other words, it gives the probabilities of all possible values that a random variable can take.

Probability Distribution

- Imagine rolling a fair six-sided die. The possible outcomes (values) of the random variable X are 1, 2, 3, 4, 5, and 6. Each of these outcomes has an equal probability of $1/6$ (since the die is fair).

Outcome (X)	Probability ($P(X)$)
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

Probability Distribution of heads-Tossing

1 Coin

Outcome	Probability
0 Heads	0.5
1 Head	0.5

Probability Distribution of heads-Tossing

2 Coins

Outcome (Coins)	Number of Heads X	Probability $P(X)$
HH	2	$\frac{1}{4} = 0.25$
HT	1	$\frac{1}{4} = 0.25$
TH	1	$\frac{1}{4} = 0.25$
TT	0	$\frac{1}{4} = 0.25$

Probability Distribution of heads-Tossing

2 Coins

Outcome	Probability
0 Heads	0.25
1 Head	0.5
2 Heads	0.25

Probability Distribution of heads-Tossing

3 Coins

Outcome (Coins)	Number of Heads X	Probability $P(X)$
HHH	3	$\frac{1}{8} = 0.125$
HHT	2	$\frac{1}{8} = 0.125$
HTH	2	$\frac{1}{8} = 0.125$
HTT	1	$\frac{1}{8} = 0.125$
THH	2	$\frac{1}{8} = 0.125$
THT	1	$\frac{1}{8} = 0.125$
TTH	1	$\frac{1}{8} = 0.125$
TTT	0	$\frac{1}{8} = 0.125$

Probability Distribution of heads-Tossing

3 Coins

Number of Heads X	Frequency	Probability $P(X)$
0	1	$\frac{1}{8} = 0.125$
1	3	$\frac{3}{8} = 0.375$
2	3	$\frac{3}{8} = 0.375$
3	1	$\frac{1}{8} = 0.125$

Probability Distribution of heads-Tossing

4 Coins

Outcome (Coins)	Number of Heads X	Probability $P(X)$
HHHH	4	$\frac{1}{16} = 0.0625$
HHHT	3	$\frac{1}{16} = 0.0625$
HHTH	3	$\frac{1}{16} = 0.0625$
HTHH	3	$\frac{1}{16} = 0.0625$
THHH	3	$\frac{1}{16} = 0.0625$
HHTT	2	$\frac{1}{16} = 0.0625$
HTHT	2	$\frac{1}{16} = 0.0625$
HTTH	2	$\frac{1}{16} = 0.0625$
THHT	2	$\frac{1}{16} = 0.0625$
THTH	2	$\frac{1}{16} = 0.0625$
TTHH	2	$\frac{1}{16} = 0.0625$
HTTT	1	$\frac{1}{16} = 0.0625$
THTT	1	$\frac{1}{16} = 0.0625$
TTHT	1	$\frac{1}{16} = 0.0625$
TTTH	1	$\frac{1}{16} = 0.0625$
TTTT	0	$\frac{1}{16} = 0.0625$



Probability Distribution of heads-Tossing

4 Coins

Number of Heads X	Frequency	Probability $P(X)$
0	1	$\frac{1}{16} = 0.0625$
1	4	$\frac{4}{16} = 0.25$
2	6	$\frac{6}{16} = 0.375$
3	4	$\frac{4}{16} = 0.25$
4	1	$\frac{1}{16} = 0.0625$

Binomial Distribution

Binomial Distribution

- The **binomial distribution** is a probability distribution that models the number of successes in a fixed number of independent binary (yes/no or success/failure) trials, where each trial has the same probability of success.

Binomial Probability Formula:

The probability of getting exactly k successes in n trials is given by the binomial formula:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

Probability Distribution of heads-Tossing

3 Coins

Outcome (Coins)	Number of Heads X	Probability $P(X)$
HHH	3	$\frac{1}{8} = 0.125$
HHT	2	$\frac{1}{8} = 0.125$
HTH	2	$\frac{1}{8} = 0.125$
HTT	1	$\frac{1}{8} = 0.125$
THH	2	$\frac{1}{8} = 0.125$
THT	1	$\frac{1}{8} = 0.125$
TTH	1	$\frac{1}{8} = 0.125$
TTT	0	$\frac{1}{8} = 0.125$

Probability Distribution of heads-Tossing

3 Coins

Number of Heads X	Frequency	Probability $P(X)$
0	1	$\frac{1}{8} = 0.125$
1	3	$\frac{3}{8} = 0.375$
2	3	$\frac{3}{8} = 0.375$
3	1	$\frac{1}{8} = 0.125$

$$P(X = 0) = \binom{3}{0} \cdot (0.5)^0 \cdot (0.5)^3 = 1 \cdot 1 \cdot 0.125 = 0.125$$

$$P(X = 1) = \binom{3}{1} \cdot (0.5)^1 \cdot (0.5)^2 = 3 \cdot 0.5 \cdot 0.25 = 0.375$$

$$P(X = 2) = \binom{3}{2} \cdot (0.5)^2 \cdot (0.5)^1 = 3 \cdot 0.25 \cdot 0.5 = 0.375$$

$$P(X = 3) = \binom{3}{3} \cdot (0.5)^3 \cdot (0.5)^0 = 1 \cdot 0.125 \cdot 1 = 0.125$$

Lab: 3 Coins Experiment

n(3 Coins)	3
p	0.5

Number of Heads	Probabailty
0	=BINOM.DIST(C8,\$D\$4,\$D\$5,FALSE)
1	=BINOM.DIST(C9,\$D\$4,\$D\$5,FALSE)
2	=BINOM.DIST(C10,\$D\$4,\$D\$5,FALSE)
3	=BINOM.DIST(C11,\$D\$4,\$D\$5,FALSE)

Lab: 4 Coins Experiment

n(4 Coins)	4
p	0.5

Number of Heads	Probabailty
0	=BINOM.DIST(C8,\$D\$4,\$D\$5,FALSE)
1	=BINOM.DIST(C9,\$D\$4,\$D\$5,FALSE)
2	=BINOM.DIST(C10,\$D\$4,\$D\$5,FALSE)
3	=BINOM.DIST(C11,\$D\$4,\$D\$5,FALSE)
4	=BINOM.DIST(C12,\$D\$4,\$D\$5,FALSE)

Example-2: Quality Control in a Factory

- Suppose a factory produces light bulbs, and historically, 2% of the bulbs are defective.
 - You randomly select 10 bulbs for inspection, and you want to find the probability of finding exactly 1 defective bulb.
 - You randomly select 10 bulbs for inspection, and you want to find the probability of finding exactly 2 defective bulbs.

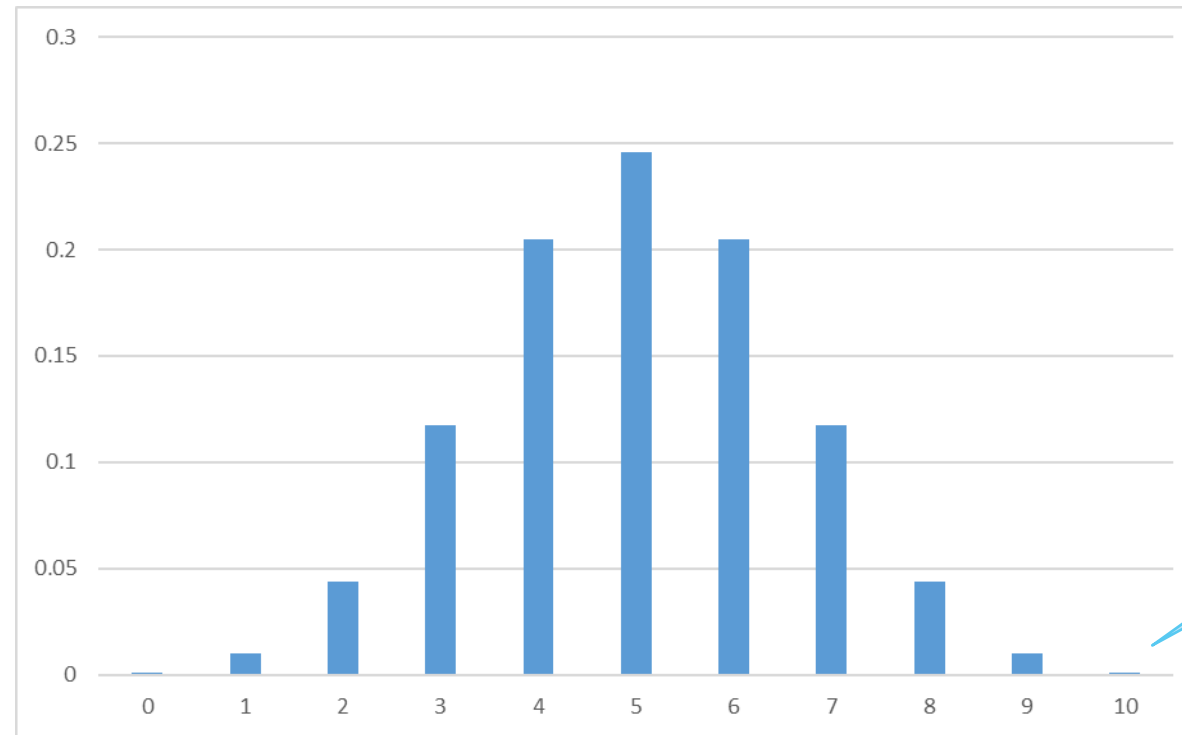
Example-2 : Quality Control in a Factory

- Number of trials (n): 10 (the number of bulbs selected)
- Probability of success (p): 0.02 (the probability that a bulb is defective)
- Number of successes (k): 1 (the number of defective bulbs)

The probability of finding exactly 1 defective bulb is:

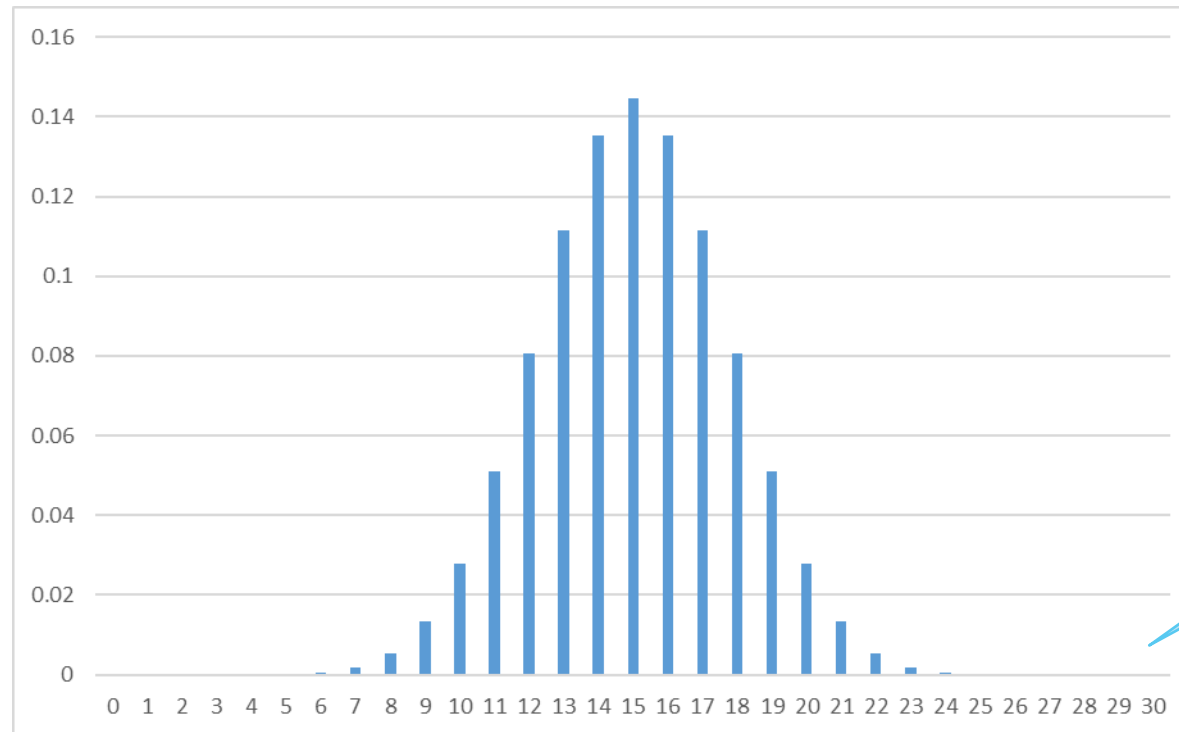
$$P(X = 1) = \binom{10}{1} \cdot (0.02)^1 \cdot (0.98)^9 = 10 \cdot 0.02 \cdot 0.8347 \approx 0.166$$

Binomial distribution when 'n' is large



$n=10$

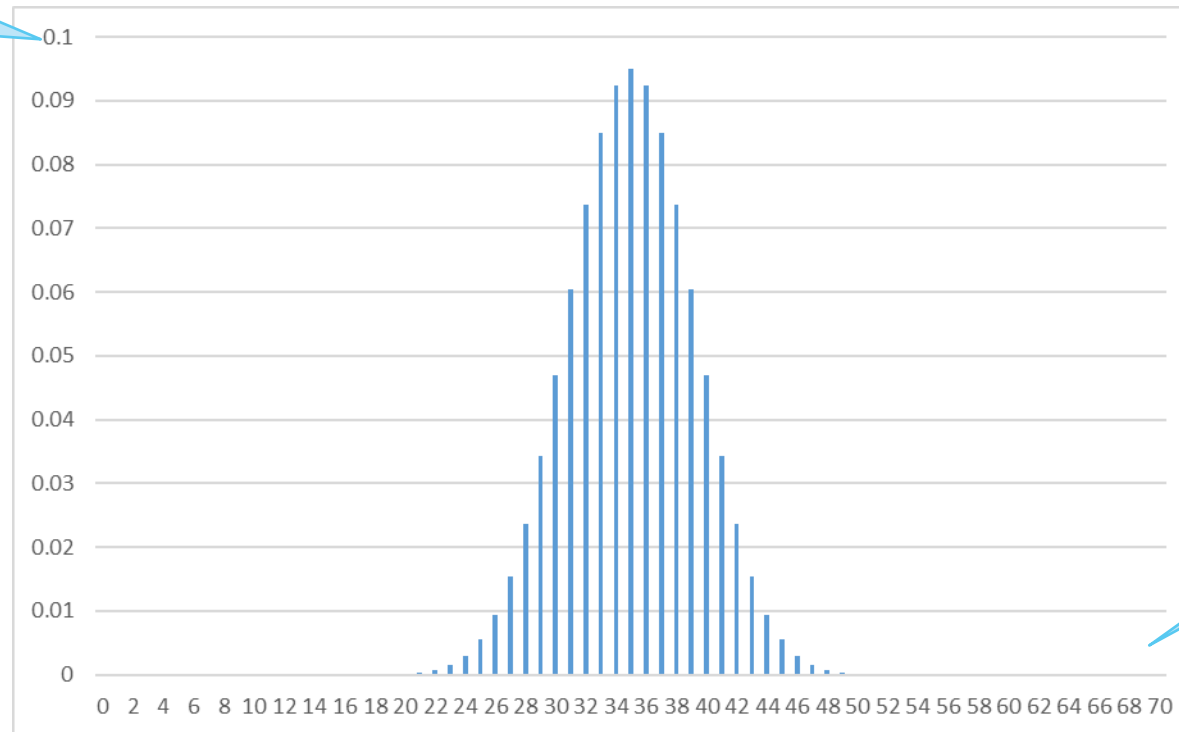
Binomial distribution when 'n' is large



$n=30$

Binomial distribution when 'n' is large

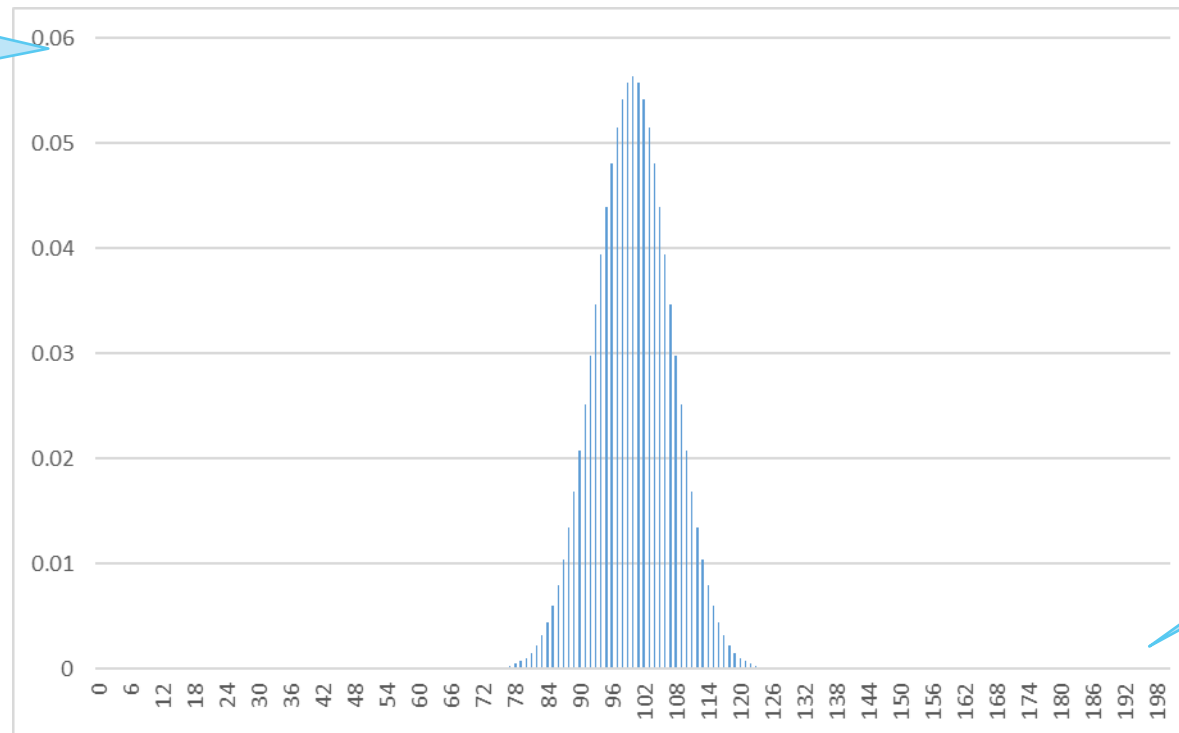
Probability of a
success event
reduces



$n=70$

Binomial distribution when 'n' is large

Probability of a success event reduces, further



$n=200$

Normal Distribution

Normal Distribution

- The normal distribution, also known as the Gaussian distribution or bell curve, is a continuous probability distribution that is symmetrical around its mean.
- It is one of the most important and commonly used distributions in statistics because many natural phenomena tend to follow this pattern.
- Symmetrical: The distribution is perfectly symmetrical about the mean.
- Bell-shaped: The curve has a peak at the mean and tapers off equally on both sides.
- Mean, Median, Mode: In a normal distribution, the mean, median, and mode are all equal and located at the center of the distribution.

Examples of normal distributions

- **Heights of Adults:** The distribution of adult heights typically follows a normal distribution, clustering around the mean.
- **IQ Scores:** IQ scores in the general population are normally distributed with a mean of 100 and a standard deviation of 15.
- **Blood Pressure:** The distribution of systolic blood pressure measurements in a population often follows a normal distribution.
- **Body Temperatures:** The body temperatures of healthy individuals are normally distributed around a mean of 98.6°F.
- **Reaction Times:** Human reaction times to stimuli are often normally distributed, with most people having average response times.
- **Speed of Cars on a Highway:** The speeds of cars on a highway often follow a normal distribution centered around the speed limit.

Probability Density Function (PDF) of Normal Distribution:

The formula for the probability density function of a normal distribution with mean μ and standard deviation σ is:

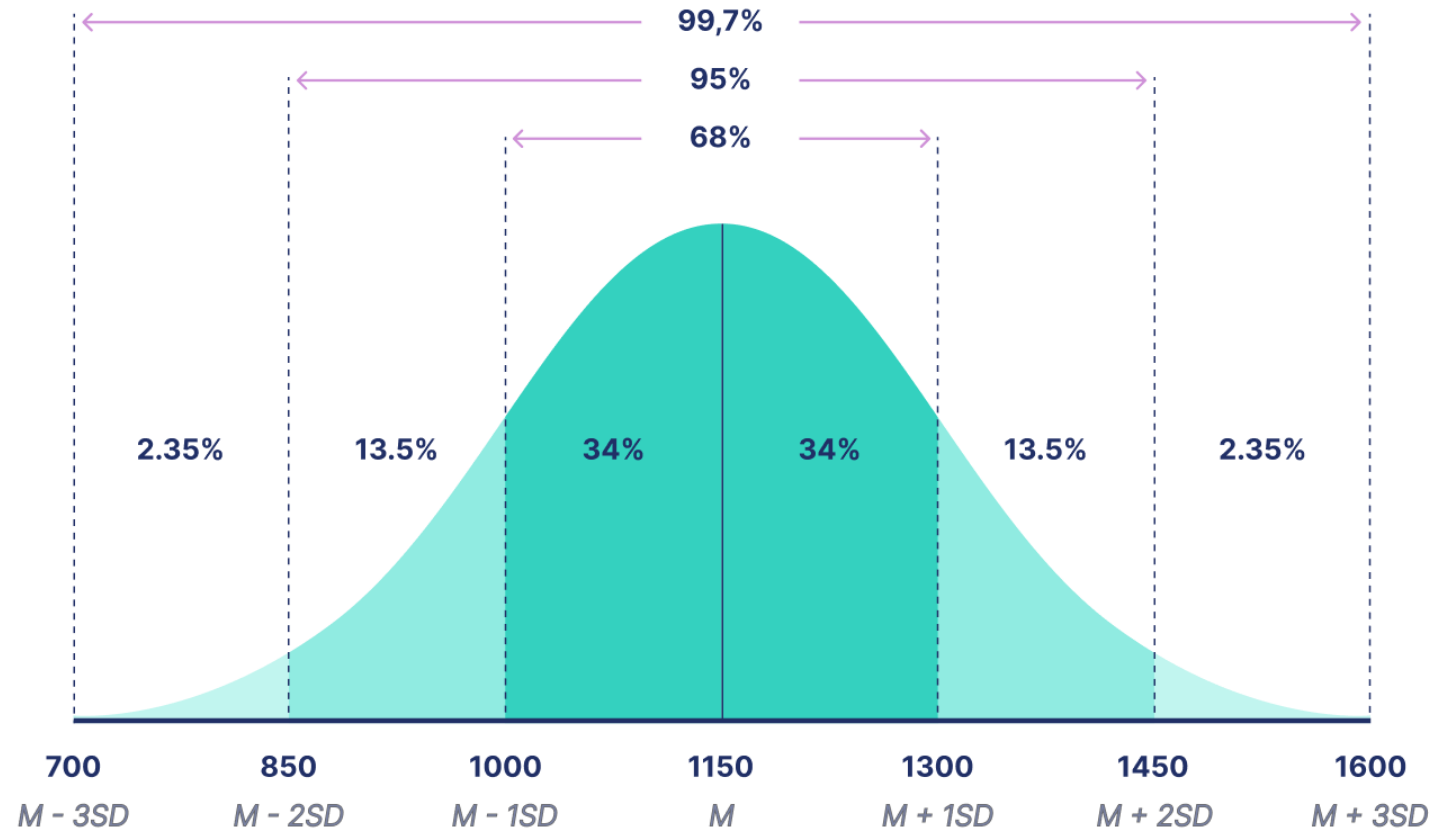
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where:

- μ is the mean,
- σ is the standard deviation,
- e is the base of the natural logarithm.

68 – 95 – 99.7 Rule

- Approximately 68% of the data falls within one standard deviation from the mean
- 95% within two standard deviations and
- 99.7% within three standard deviations.
- Also known as the **Empirical Rule**



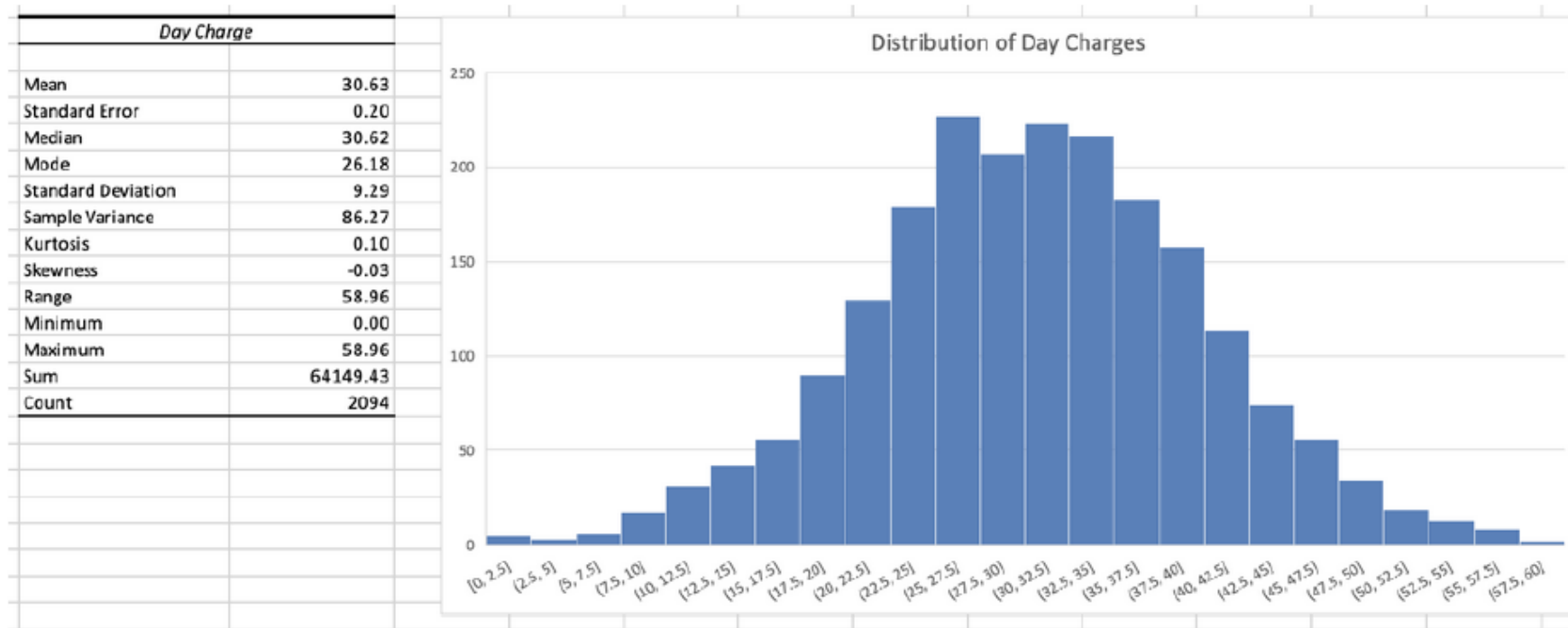
68 – 95 – 99.7 Rule

- Let's say that IQ scores are normally distributed with:
- **Mean (μ) = 100**
- **Standard Deviation (σ) = 15**

- **Within 1 Standard Deviation (68% of Data)**
 - 68% of people have IQ scores between 85 and 115.
- **Within 2 Standard Deviations (95% of Data)**
 - 95% of people have IQ scores between 70 and 130.
- **Within 3 Standard Deviations (99.7% of Data)**
 - 99.7% of people have IQ scores between 55 and 145.

LAB : Exam Questions

The distribution of Day Charges for customers of a telecom company is described below:



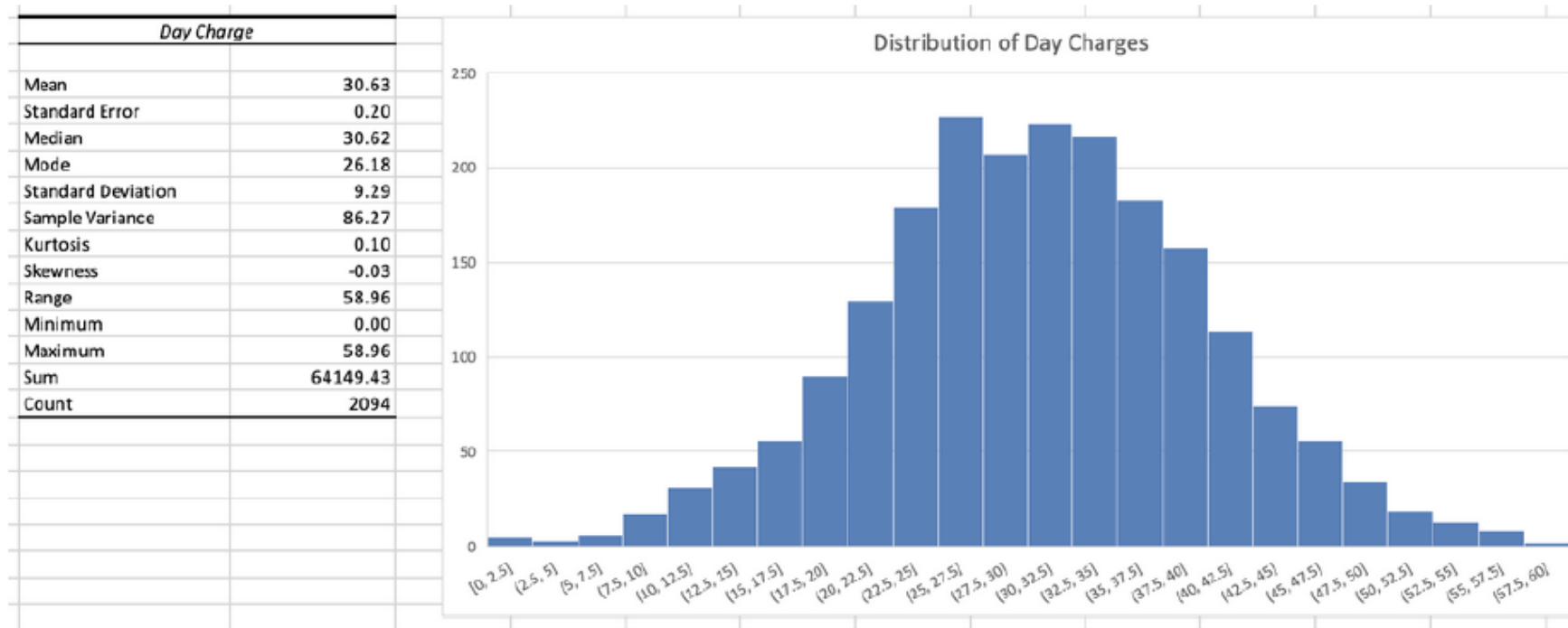
Apply the
Empirical Rule

Assume that the charges are **normally** distributed with the parameters as shown in the table

What is the approximate interval where the **middle 95% of values** are distributed?

LAB : Exam Questions

The distribution of Day Charges for customers of a telecom company is described below:



Mean 30.63

SD 9.29

Mean-(2*SD) 12.05

Mean+(2*SD) 49.21

Assume that the charges are **normally** distributed with the parameters as shown in the table.

What is the approximate interval where the **middle 95% of values** are distributed?

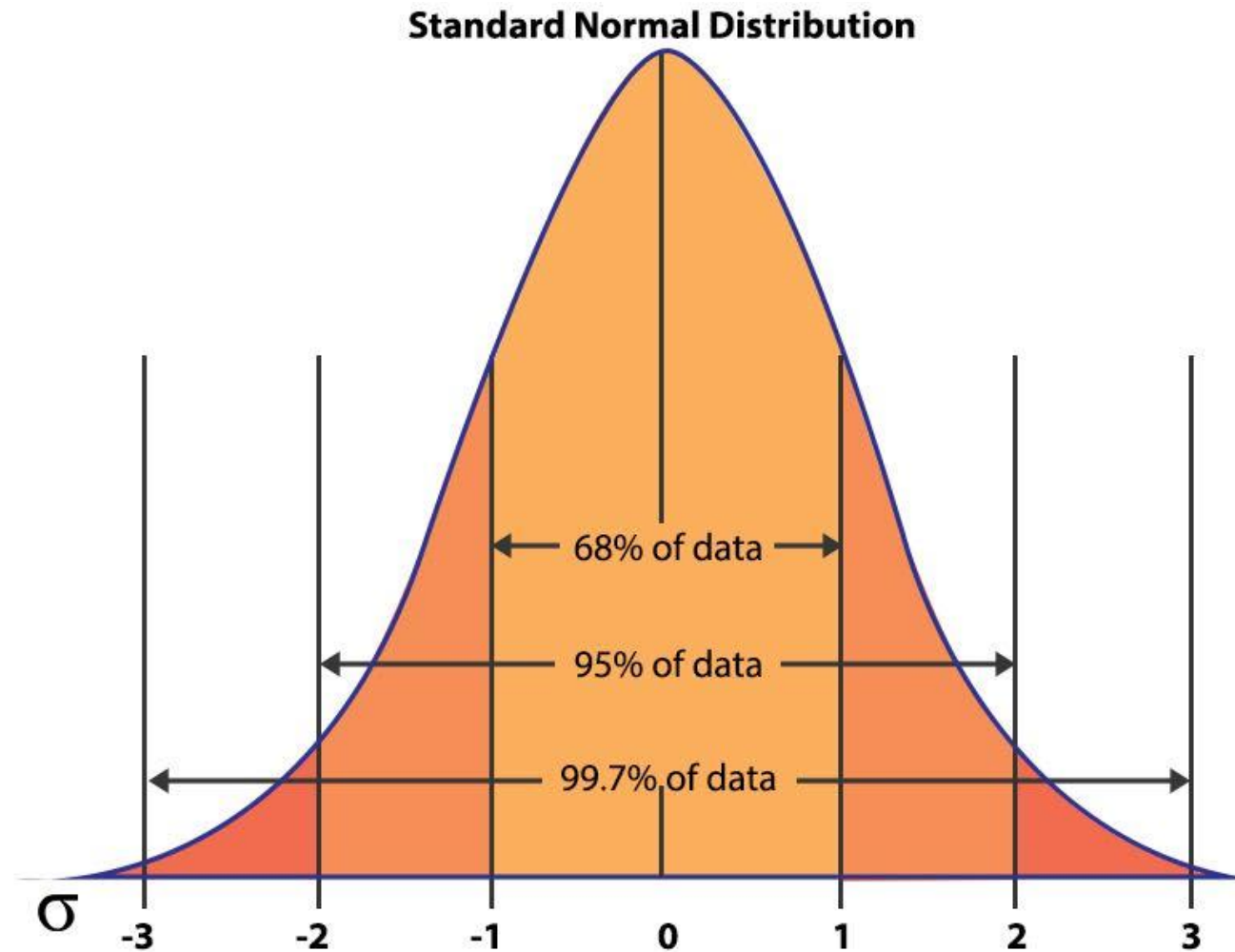
Standard Normal Distribution

- The standard normal distribution is a special case of the normal distribution.
- It is a normal distribution that has been standardized, meaning it has a mean of 0 and a standard deviation of 1.

To convert a normal distribution value X with mean μ and standard deviation σ to a standard normal distribution (z-score), use the formula:

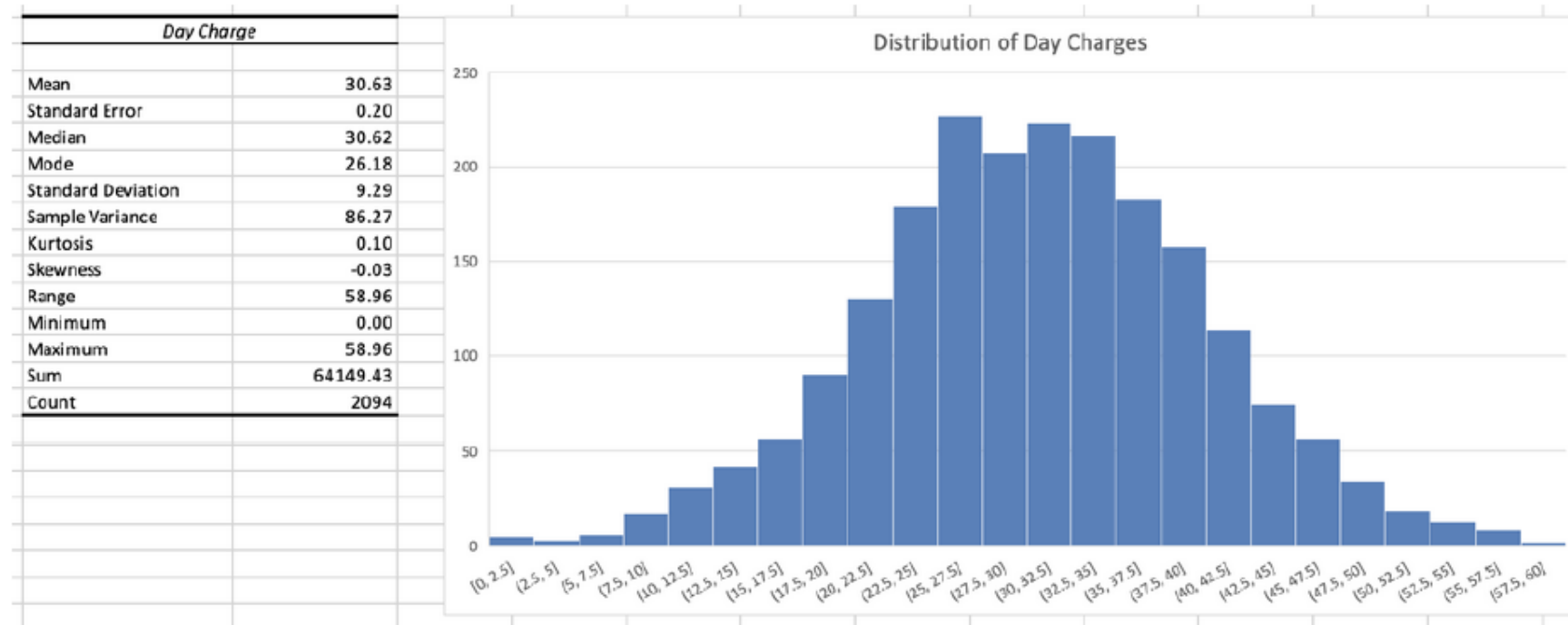
$$Z = \frac{X - \mu}{\sigma}$$

Standard Normal Distribution



LAB : Exam Questions

The distribution of Day Charges for customers of a telecom company is described below:

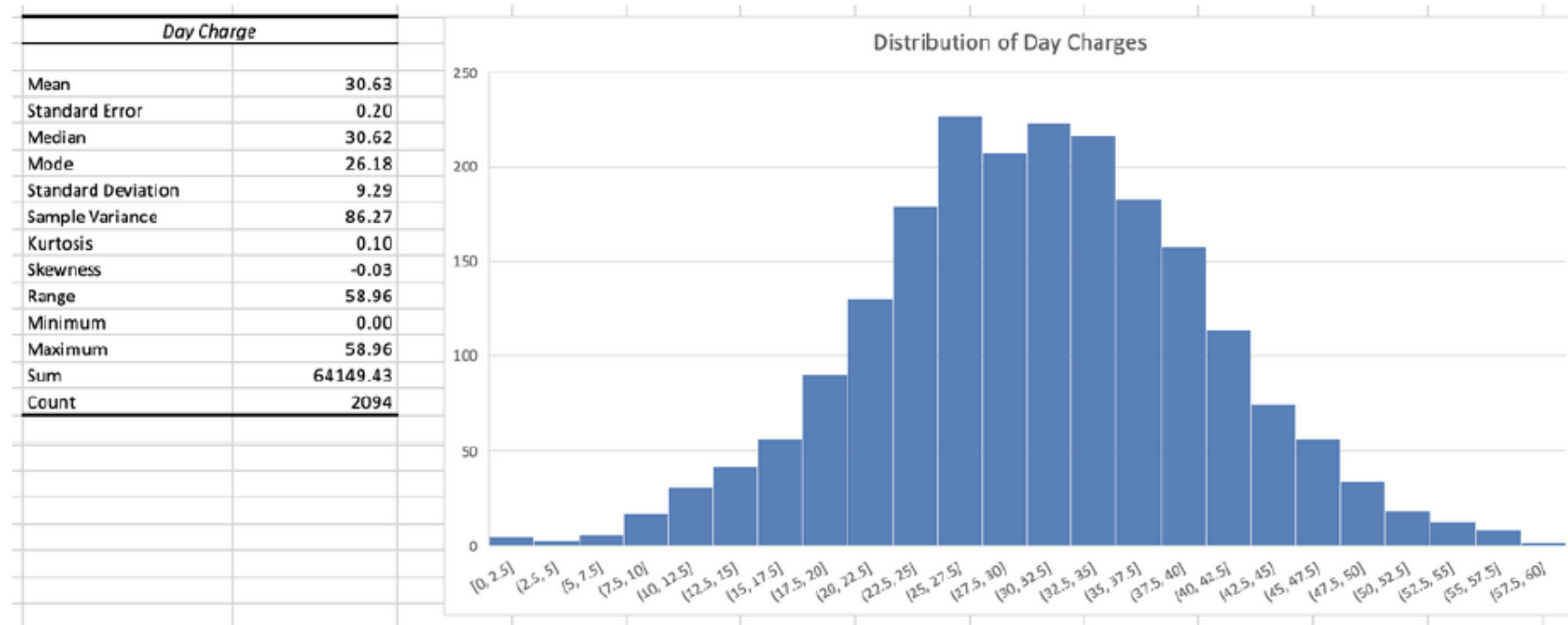


Assume that the charges are **normally** distributed with the parameters as shown in the table.

What approximately is the **97.5th** percentile of the charges?

LAB : Exam Questions

The distribution of Day Charges for customers of a telecom company is described below:



Mean 30.63

SD 9.29

Mean-(2*SD) 12.05

Mean+(2*SD) 49.21

Assume that the charges are **normally** distributed with the parameters as shown in the table.

What approximately is the **97.5th** percentile of the charges?

Mean in Probability Distributions

Expected Value (mean)

- The expected value (EV) of a random variable is a measure of the central tendency of its probability distribution.
- It represents the long-run average outcome of a random process if it were repeated many times.
- In simple terms, it's the "**average**" value you would expect over many trials of the random process.

Formula for Expected Value

For a **discrete random variable**, the expected value $E(X)$ is calculated as:

$$E(X) = \sum_i [x_i \cdot P(x_i)]$$

Where:

- x_i are the possible values that the random variable X can take,
- $P(x_i)$ is the probability associated with each value x_i .

Formula for Expected Value

Consider a fair six-sided die. The possible outcomes when rolling the die are 1, 2, 3, 4, 5, and 6, each with a probability of $\frac{1}{6}$.

The expected value $E(X)$ of rolling the die is:

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$

$$E(X) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = \frac{21}{6} = 3.5$$

Expected Value Calculation

Suppose a lottery ticket costs \$5, and the prize for winning is \$100. The probability of winning is 0.01, and the probability of losing is 0.99.

Let X be the net profit from the lottery ticket. The possible values of X are:

- If you win: $X = 100 - 5 = 95$
- If you lose: $X = -5$

The expected value of the net profit is:

$$E(X) = 95 \cdot 0.01 + (-5) \cdot 0.99$$

$$E(X) = 0.95 + (-4.95) = -4.00$$

Thus, the expected value of the net profit is **-4.00**. This means that, on average, you can expect to lose \$4 per lottery ticket in the long run.

Expected Value Calculation

Hours Worked (Range)	Midpoint (X)	Frequency	Probability $P(X)$
0 - 10	5	4	0.10
11 - 20	15	6	0.15
21 - 30	25	10	0.25
31 - 40	35	12	0.30
41 - 50	45	8	0.20

Similar to
calculation of
mean

To calculate the expected value $E(X)$, we multiply each midpoint by its corresponding probability and sum the results:

$$E(X) = 5 \cdot 0.10 + 15 \cdot 0.15 + 25 \cdot 0.25 + 35 \cdot 0.30 + 45 \cdot 0.20$$

$$E(X) = 0.5 + 2.25 + 6.25 + 10.5 + 9 = 28.5$$

LAB:Exam Question

The following table represents the frequency distribution of **Day Charges** (in USD) for **2094** customers of a telecom outfit.

Day Charges	Count of Day Charge
0-10	30
10-20	220
20-30	743
30-40	781
40-50	278
50-60	42
Grand Total	2094

Declare a random variable **M** as the midpoint of the charges. To illustrate, for the first slab of 0-10 USD, **M** = 5.

What is the **Expected Value** of this variable **M**?

- ☐ 30.65
- ☐ 13.05
- ☐ 32.88
- ☐ 38.23

LAB: Exam Question

Day Charges	Mid Point(x)	Count(n)	Probability P(X)	$x \cdot p(x)$
0-10	5	30	1%	0.1
10-20	15	220	11%	1.6
20-30	25	743	35%	8.9
30-40	35	781	37%	13.1
40-50	45	278	13%	6.0
50-60	55	42	2%	1.1
		2094		30.65

Standard Deviation in Probability Distributions

SD Formula

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

Previous
Formula

$$\sigma = \sqrt{\sum P(X) \cdot (X - \mu)^2}$$

New Formula

SD Calculation

Commute Time (Minutes)	Midpoint (X)	Frequency	Probability P(X)	$X \cdot P(X)$	$(X - \mu)^2$	$P(X) \cdot (X - \mu)^2$
0 - 10	5	20	0.1	0.5	462.25	46.225
11-20	15	40	0.2	3	132.25	26.45
21 - 30	25	60	0.3	7.5	2.25	0.675
31 - 40	35	50	0.25	8.75	72.25	18.0625
41 - 50	45	30	0.15	6.75	342.25	51.3375
		200		26.5	Var	142.75
					SD	11.95

Exam Question

Day Charges	Count of Day Charge
0-10	30
10-20	220
20-30	743
30-40	781
40-50	278
50-60	42
Grand Total	2094

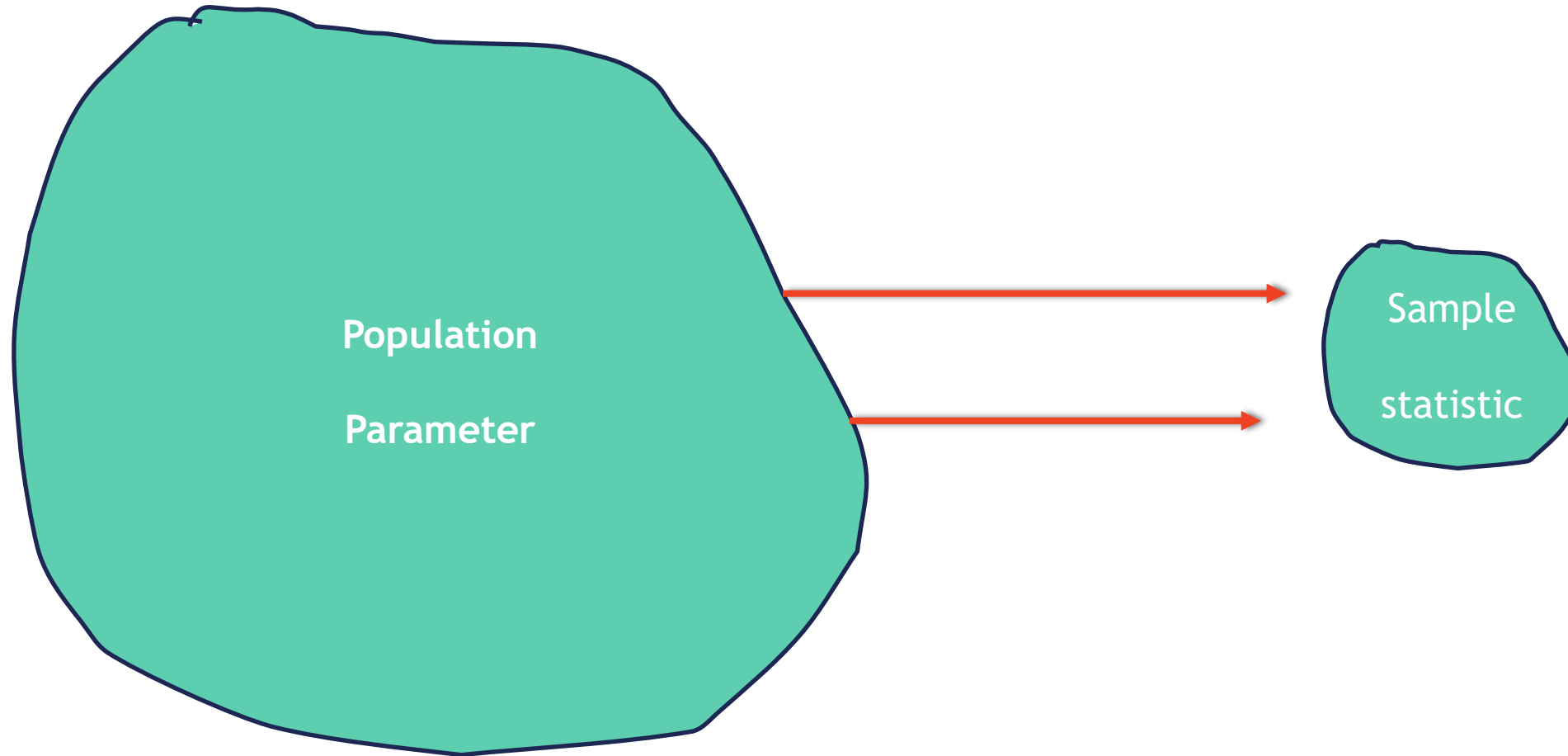
Declare a random variable **M** as the midpoint of the charges. To illustrate, for the first slab of 0-10 USD, **M** = 5.

What is the **Standard Deviation** of this variable **M**?

- ☐ 75.86
- ☐ 92.8
- ☐ 9.63
- ☐ 8.71

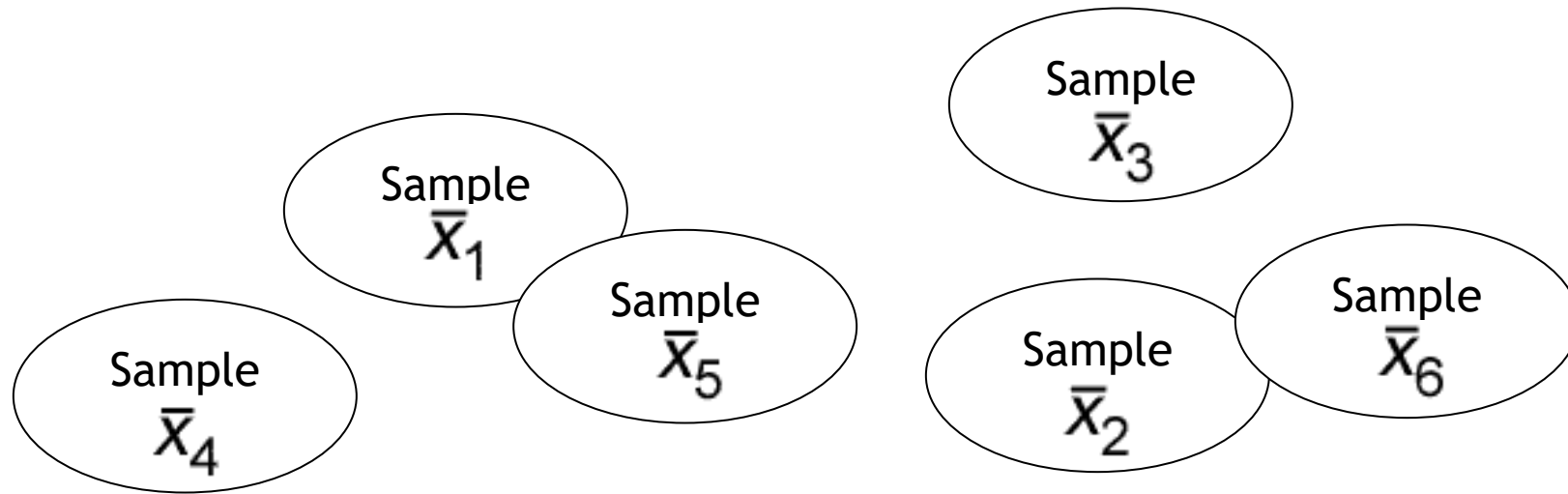
Sampling Distribution

Terminology



Sampling distribution

- A sampling distribution is the probability distribution of a sample statistic that is formed when samples of size n are repeatedly taken from a population.
- If the sample statistic is the sample mean, then the distribution is the sampling distribution of sample means.

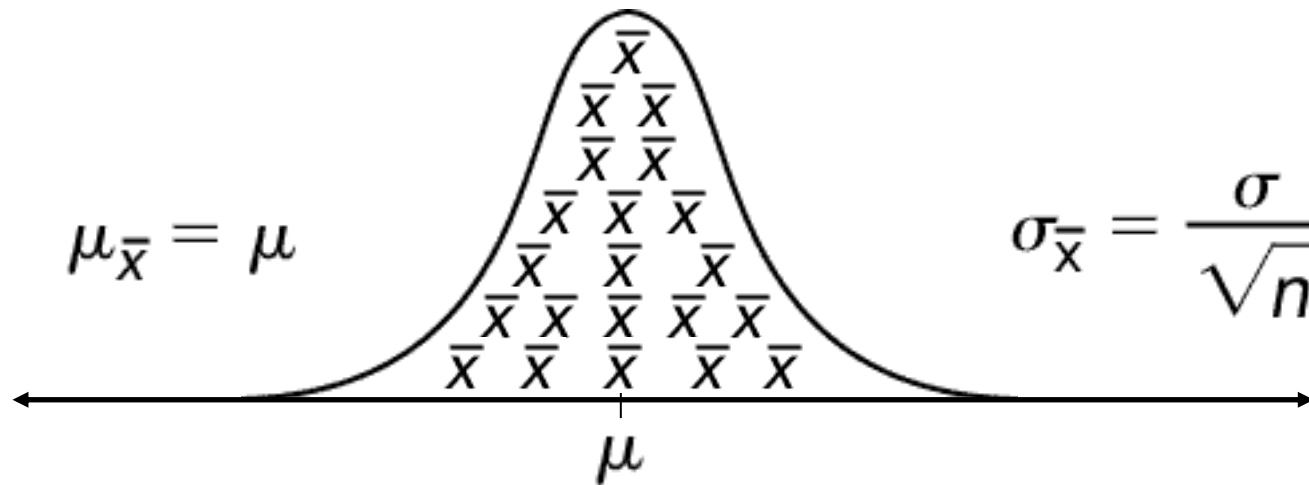


The sampling distribution consists of the values of the sample means, $\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4, \bar{X}_5, \bar{X}_6, \dots$

Central Limit theorem

If a sample n (30) is taken from a population with ***any type distribution*** that has a mean = μ and standard deviation = σ

the ***sample means*** will have a normal distribution $\mu_{\bar{x}} = \mu$ and s.d $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$



Central Limit theorem

- It states that, regardless of the shape of the population distribution, the distribution of the sample mean (or sum) of a sufficiently large number of independent and identically distributed random variables will tend to be normal (bell-shaped) as the sample size increases.
- In other words: As the sample size becomes large, the sampling distribution of the sample mean approaches a normal distribution, regardless of the population's distribution.
- This CLT will be later used in Confidence intervals

Confidence Intervals

Confidence Intervals

- A confidence interval is a range of values, derived from sample data, that is likely to contain the true population parameter (such as the population mean or proportion) with a specified level of confidence.
- It gives us an estimate of where the true value of the population parameter lies based on the sample statistic.
- We can talk about population parameter based on the Sample statistic

Examples: Sample → Population

- Based on a sample of 100 students, a 95% confidence interval for the population mean height is [165.2 cm, 170.8 cm].
- Based on a sample of 50 cars, a 95% confidence interval for the population mean fuel efficiency is [30.1 MPG, 32.5 MPG].
- Using a sample of 50 batteries, a 99% confidence interval for the population mean lifespan is [8.7 hours, 9.3 hours].
- From a survey of 150 residents, a 95% confidence interval for the population mean monthly income is [\$3,500, \$4,200].

Formula for Confidence Interval

For a **population mean** with a known or large sample size, the confidence interval is calculated as:

$$\text{Confidence Interval} = \text{Sample Mean} \pm Z \times \left(\frac{\sigma}{\sqrt{n}} \right)$$

Where:

- **Sample Mean (\bar{X})**: The mean of the sample data.
- **Z**: The Z-score corresponding to the desired confidence level (e.g., 1.96 for 95% confidence).
- **σ** : The population standard deviation (or the sample standard deviation when population standard deviation is unknown and the sample size is large).
- **n**: The sample size.

Confidence Intervals

- Confidence intervals rely on the assumption that the sampling distribution of the sample statistic (such as the mean) is approximately normal.
- Because of the CLT, we can use the normal distribution to construct confidence intervals for the population mean, even if the underlying population distribution is not normal.

Example

Suppose a factory produces light bulbs, and the mean lifespan of a random sample of 100 light bulbs is found to be **800 hours** with a standard deviation of **30 hours**. You want to calculate a **95% confidence interval** for the population mean lifespan.

1. **Sample Mean (\bar{X}):** 800 hours
2. **Standard Deviation (σ):** 30 hours
3. **Sample Size (n):** 100
4. **Confidence Level:** 95% ($Z = 1.96$)

Example

1. **Sample Mean** (\bar{X}): 800 hours
2. **Standard Deviation** (σ): 30 hours
3. **Sample Size** (n): 100
4. **Confidence Level**: 95% ($Z = 1.96$)

$$\text{Confidence Interval} = \text{Sample Mean} \pm Z \times \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$\text{Confidence Interval} = 800 \pm 1.96 \times \left(\frac{30}{\sqrt{100}} \right)$$

$$\text{Confidence Interval} = 800 \pm 1.96 \times 3 = 800 \pm 5.88$$

The **95% confidence interval** is [794.12, 805.88]. This means we are 95% confident that the true population mean lifespan is between 794.12 and 805.88 hours.

Lab: Exam Question

Q54

The producer of the hit movie *Random Tales* wants to know how much money they would pay per plate for a limited-seat dinner with Lakshay Khanna and Twishaa Chopra, the celebrated protagonist couple of the movie. The proceeds of the dinner shall go to fund an NGO.

The NGO decides to carry out a survey. Polling 81 moviegoers at random, the average pledge amount turns out to be Rs.5,400 per plate. The standard deviation is Rs.180.

Given that these dinners shall be held at several cities, what is the approximate 95% confidence interval for the per-plate proceeds that the NGO can expect?

- ☐ INR 5,040 to INR 5,760
- ☐ INR 5,380 to INR 5,420
- ☐ INR 5,360 to INR 5,440
- ☐ INR 5,220 to INR 5,580

Lab: Exam Question

1. Mean $\bar{X} = 5,400$
2. Standard Deviation $\sigma = 180$
3. Sample Size $n = 81$
4. Z-value for 95% Confidence Level = 1.96

$$\text{Confidence Interval} = \text{Sample Mean} \pm Z \times \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$\text{Lower Bound} = 5,400 - 1.96 \times \left(\frac{180}{\sqrt{81}} \right)$$

$$\text{Upper Bound} = 5,400 + 1.96 \times \left(\frac{180}{\sqrt{81}} \right)$$



Thank You
