

Statistics

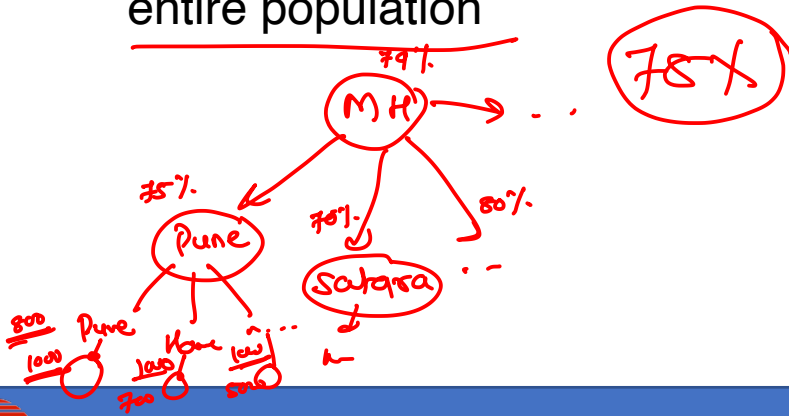
Proportion

- **Population proportion**

- Fraction of the total that possesses a certain attribute
- Fraction of population that has certain characteristic
- E.g.
 - let's say you had 1,000 people in the population and 200 of those people are literate
 - Fraction of people who are literate is $p = 200/1000 = 0.2$

- **Sample proportion**

- In the real world, you usually don't know facts about the entire population and so you use sample data to estimate p
- This sample proportion is written as \hat{p}
- It is exactly same as population proportion except that the data will be used from sample and not from the entire population

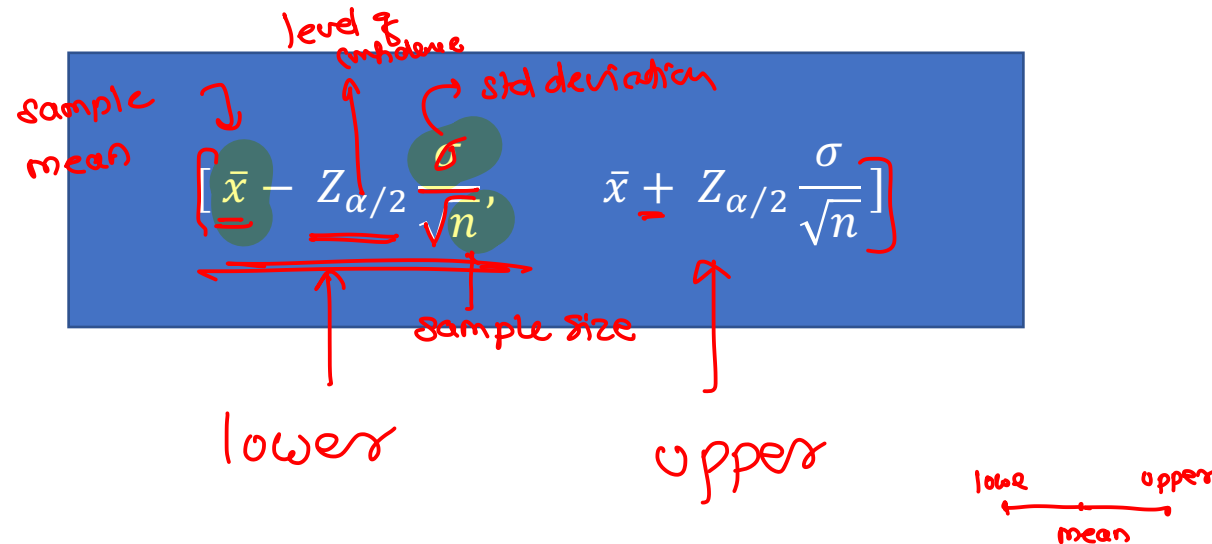


Confidence interval

56.51

- An interval that will contain a population parameter a specified proportion of the time
- A Confidence Interval is a **range of values** we are **fairly sure** our **true value** lies in
- A confidence interval is the probability that a value will fall between an upper and lower bound of a probability distribution
- The confidence interval can take any number of probabilities, with the most common being 90% or 95% or 99%

90% →
95% →
99% →



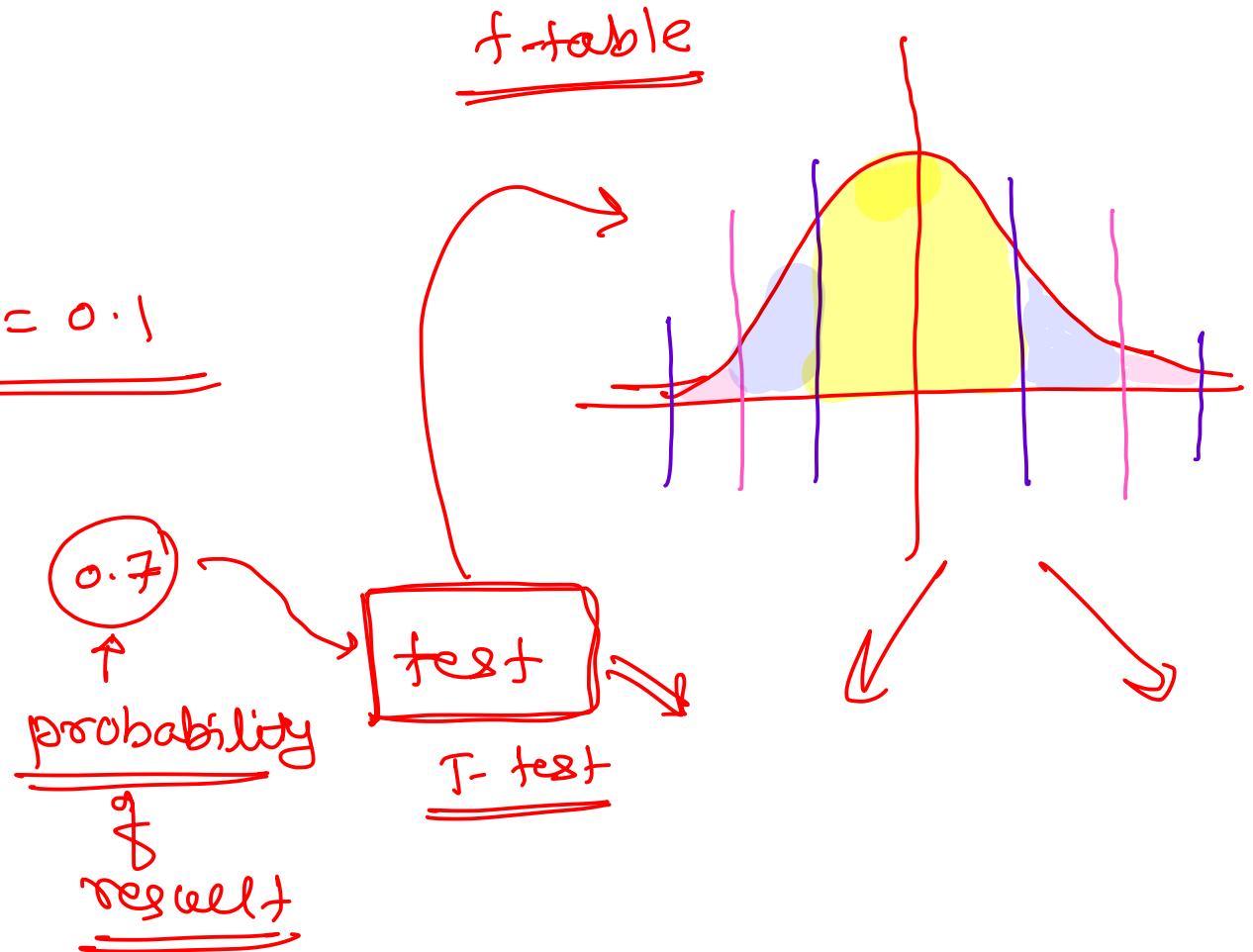
Level of Confidence

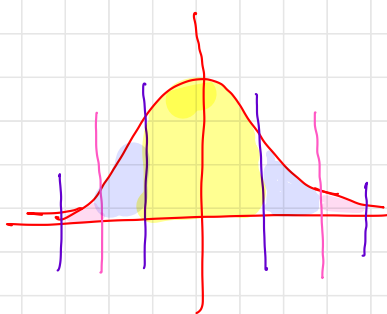
- Denoted by $1 - \alpha$
- α is the value between 0 and 1
- For confidence level 95%, α will be 5%

90% : $100 - 90 = \alpha$, $1 - 0.9 = 0.1$

95% :

99% :

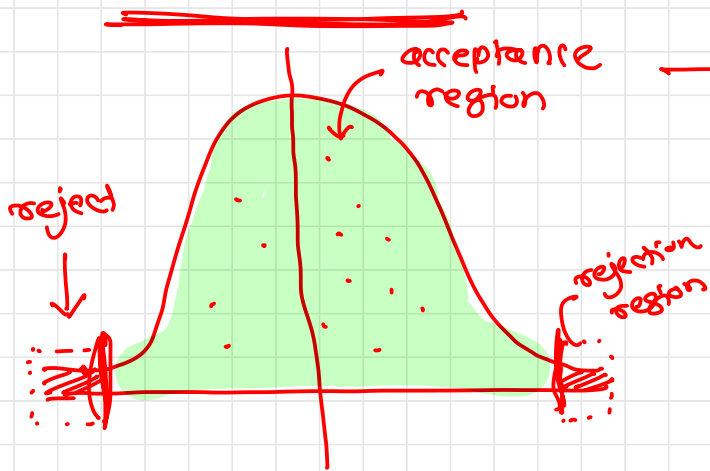




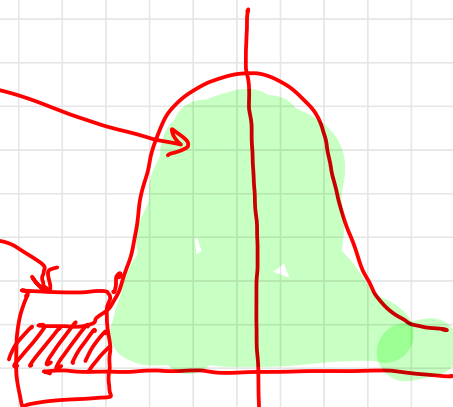
standard table

$$90\% = \alpha = \frac{0.1}{2}$$

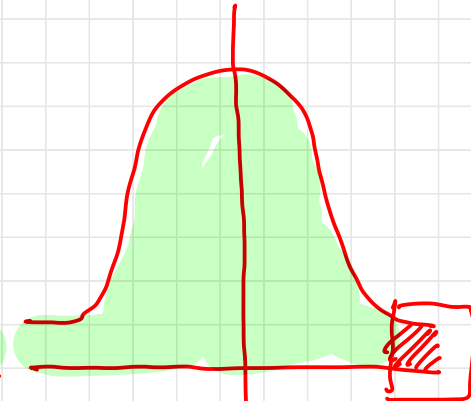
2tailed test



1tailed test



left tailed



right tailed test

Example

- We measure the heights of 40 randomly chosen men, and get a mean height of 175cm. We also know the standard deviation of men's heights is 20cm.

create a confidence interval for μ

$$\bar{x} = 175$$

$$\sigma = 20$$

$$n = 40$$

$$\left[\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

90% \Rightarrow

95% \Rightarrow

99% \Rightarrow



Margin of Error

- The margin of error is the range of values below and above the sample statistic in a confidence interval
- A margin of error tells you how many percentage points your results will differ from the real population value



Hypothesis



Hypothesis → Inferential

- A hypothesis is a proposed explanation for a phenomenon → statement = it is going to rain
- A statistical hypothesis, sometimes called confirmatory data analysis, is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables
- A statement about the parameters describing a population (not a sample)

ml-ops

Hypothesis → statement

→ t-test
→ z-test
→ f-test
→ ANOVA



Hypothesis testing

- Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter
- The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis
- Is used to infer the result of a hypothesis performed on sample data from a larger population
- Is the application of statistical methods to real-world questions



Terminologies

- Null Hypothesis (H_0)

- It is an assumption or a statement about the parameter (an analyst will try to nullify)
- It is presumed to be true until statistical evidence nullifies it for an alternative hypothesis
- We test the likelihood of this statement being true in order to accept or reject our alternate hypothesis
- Can include =, <= or >= sign
- E.g.
 - X is equal to Y

covishield is a vaccine for covid

- Alternative hypothesis (H_1 or H_a)

- A statement that can directly contradicts the null hypothesis
- We determine whether or not to accept this statement based on likelihood of the null hypothesis
- Can include !=, < or > sign
 - X is not equal to Y

covishield is NOT a vaccine for covid



How does it work ?

- In hypothesis testing an analyst tests a statistical sample, with the goal of accepting or rejecting null hypothesis
- The test tells the analyst whether or not his ^{null} primary hypothesis is true
- If it isn't true, the analyst formulates a new hypothesis to be tested, repeating the process until data reveals a true hypothesis
- Procedure
 - The first step is for the analyst to state the two hypotheses so that only one can be right
 - The next step is to formulate an analysis plan, which outlines how the data will be evaluated
 - The third step is to carry out the plan and physically analyze the sample data
 - The fourth and final step is to analyze the results and either accept or reject the null hypothesis



Real Example

- A person wants to test that a penny has exactly a 50% chance of landing on heads
- The null hypothesis: it lands on heads $\hat{=} H_0$
- The alternative hypothesis: It does not land on heads $\hat{=} H_1$
- Mathematically, the null hypothesis would be represented as $H_0 : P = 0.5$
- Mathematically, the alternative hypothesis would be represented as $H_a : P \neq 0.5$
- A random sample of 100 coin flips is taken from a random population of coin flippers, and the null hypothesis is then tested
- If it is found that the 100 coin flips were distributed as 40 heads and 60 tails, the analyst would assume that a penny does not have a 50% chance of landing on heads and would reject the null hypothesis and accept the alternative hypothesis
- Afterward, a new hypothesis would be tested, this time that a penny has a 40% chance of landing on heads.

→ num = 100
'int num = 100
→ num: 'int = 100 ✓



Hypothesis Testing

- If the null hypothesis is rejected, then we say the data supports another mutually exclusive alternate hypothesis
- We never **PROVE** a hypothesis
- There are different types of tests used for testing hypothesis
 - Parametric tests
 - Relies on theoretical distributions of the test statistic under the null hypothesis and assumptions about the distribution of the sample data
 - Non Parametric tests
 - Referred to as “Distribution Free” as they do not assume that data are drawn from any particular distribution



One sided and two sided tests

- Hypothesis tests can be one or two sided (tailed)
- One tailed tests are directional
 - $H_0: \mu_1 - \mu_2 \leq 0$
 - $H_A: \mu_1 - \mu_2 > 0$
- Two tailed tests are not directional
 - $H_0: \mu_1 - \mu_2 = 0$
 - $H_A: \mu_1 - \mu_2 \neq 0$



P-values

- Calculate a test statistic in the sample data that is relevant to the hypothesis being tested
- After calculating a test statistic we convert this to a Pvalue by comparing its value to distribution of test statistic's under the null hypothesis
- Measure of how likely the test statistic value is under the null hypothesis
- $P\text{-value} \leq \alpha \Rightarrow \text{Reject } H_0 \text{ at level } \alpha$
- $P\text{-value} > \alpha \Rightarrow \text{Do not reject } H_0 \text{ at level } \alpha$



When to reject H_0

- Level of significance, α : Specified before an experiment to define rejection region
- Rejection region: set of all test statistic values for which H_0 will be rejected



Errors in Hypothesis Testing

	H0 True	H0 False
Do not reject H0	Correct Decision $1 - \alpha$	Incorrect Decision β
Reject H0	Incorrect Decision α	Correct Decision $1 - \beta$



T-Test

- It is a parametric test which tells you how significant the differences between groups are
- In other words, it lets you know if those differences (measured in means/averages) could have happened by chance
- T-tests are called so, because the test results are all based on t-values
- A t-test looks at the following values to determine the probability of difference between two sets of data
 - t-statistic
 - the t-distribution values
 - the degrees of freedom



T-Test: T test statistics

- T-values are an example of test statistics
- A test statistic is a standardized value that is calculated from sample data during a hypothesis test
- The procedure that calculates the test statistic compares your data to what is expected under the null hypothesis
- To perform t-test calculation we need following values
 - The Difference between the mean values from each data set (called the mean difference)
 - The standard deviation of each group
 - The number of data values of each group



T-Test: Degree of freedom

- Refers to the values in a study that has the freedom to vary and are essential for assessing the importance and the validity of the null hypothesis
- Computation of these values usually depends upon the number of data records available in the sample



T-Test: Assumptions

- The first assumption is concerned with the scale of measurement. Here assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale.
- The second assumption is regarding simple random sample. The Assumption is that the data is collected from a representative, randomly selected portion of the total population.
- The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.
- The fourth assumption is a that reasonably large sample size is used for the test. Larger sample size means the distribution of results should approach a normal bell-shaped curve.
- The final assumption is the homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.



T-Test: Types

- One sample t-test
 - Used to compare a sample mean with a known population mean or some other meaningful, fixed value
- Two sample t-test
 - Used to compare two means from independent groups
- Paired t-test
 - Used to compare two means that are repeated measures for the same participants — scores might be repeated across different measures or across time

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$$



T-Test: Example

- A coffee shop relocates to Italy and wants to make sure that all lattes are consistent. They believe that each latte has an average of 4 Oz of espresso. If this is not the case, they must increase or decrease the amount. A random sample of 25 lattes shows a mean of 4.6 Oz of espresso and a standard deviation of .22 Oz. Use $\alpha = .05$ and run a one sample t-test to compare with the known population mean (use two tailed method)

Sample Mean M = 4.6 oz

Population Mean μ = 4 oz.

Sample standard deviation = 0.22 oz.

Sample size n = 25



Mann Whitney U Test

- Also known as Wilcoxon Rank Sum Test
- This test can be used to investigate whether two *independent* samples were selected from populations having the same distribution
- Uses ranking to determine the result



Mann Whitney U Test: Steps

- Assign numeric ranks to all the observations (put the observations from both groups to one set), beginning with 1 for the smallest value
- Now, add up the ranks for the observations which came from sample 1. The sum of ranks in sample 2 is now determinate, since the sum of all the ranks equals $N(N + 1)/2$ where N is the total number of observations
- Calculate u values

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}$$

$$U_2 = R_2 - \frac{n_2(n_2 + 1)}{2}$$

- Use the smaller value from u_1 and u_2
- Lookup the u value in the u-table



Mann Whitney U Test: Example

- Treatment A: 3, 4, 2, 6, 2, 5
- Treatment B: 9, 7, 5, 10, 6, 8



ANOVA

- Analysis of Variance
- Used to test the means from three or more samples
- Used to answer the question:
 - What is the probability that two samples come from populations that have the same variance?
 - What is the probability that three or more samples come from the same population?



ANOVA: Rational

- Basic idea is to partition total variation of the data into two sources
 - Variation within levels (groups)
 - Variation between levels (groups)
- If H_0 is true the standardized variances are equal to one another

$$F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$



$$F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

- Where
 - SSG = Sum of Squares Groups
 - SSE = Sum of Squares Error
 - df_{groups} = degrees of freedom (groups)
 - df_{error} = degrees of freedom (error)



ANOVA Example

sample

2
3
7
2
6

sample

10
8
7
5
10

sample

10
13
14
13
15



sample

2	- 4	=	-2^2	4
3	- 4	=	-1^2	1
7	- 4	=	3^2	9
2	- 4	=	-2^2	4
6	- 4	=	2^2	4
				<hr/>
				22


sample

10	- 8	=	2^2	4
8	- 8	=	0^2	0
7	- 8	=	-1^2	1
5	- 8	=	-3^2	9
10	- 8	=	2^2	4
				<hr/>
				18

sample

10	- 13	=	-3^2	9
13	- 13	=	0^2	0
14	- 13	=	1^2	1
13	- 13	=	0^2	0
15	- 13	=	2^2	4
				<hr/>
				14

Sum of Squares Within Groups = $22 + 18 + 14 = 54$

observation		mean	observation - mean	(observation - mean) ²	
2	-	8.3	= -6.3	40.1	
3	-	8.3	= -5.3	28.4	
7	-	8.3	= -1.3	1.8	
2	-	8.3	= -6.3	40.1	
6	-	8.3	= -2.3	5.4	
10	-	8.3	= 1.7	2.7	
8	-	8.3	= -0.3	0.1	
7	-	8.3	= -1.3	1.8	
5	-	8.3	= -3.3	11.1	
10	-	8.3	= 1.7	2.8	
10	-	8.3	= 1.7	2.8	
13	-	8.3	= 4.7	21.8	
14	-	8.3	= 5.7	32.1	
13	-	8.3	= 4.7	21.8	
15	-	8.3	= 6.7	44.4	
				257.3	Total Sum of Squares

Sum of Squares Between Groups

2
3
7
2
6
10
8
7
5
10
10
13
14
13
15

mean

2
3
7
2
6

mean

10
8
7
5
10

mean

10
13
14
13
15

mean

1. $\text{mean} - \text{mean}$

$\text{mean} - \text{mean}$

$\text{mean} - \text{mean}$

2. $(\text{mean} - \text{mean})^2$

$(\text{mean} - \text{mean})^2$

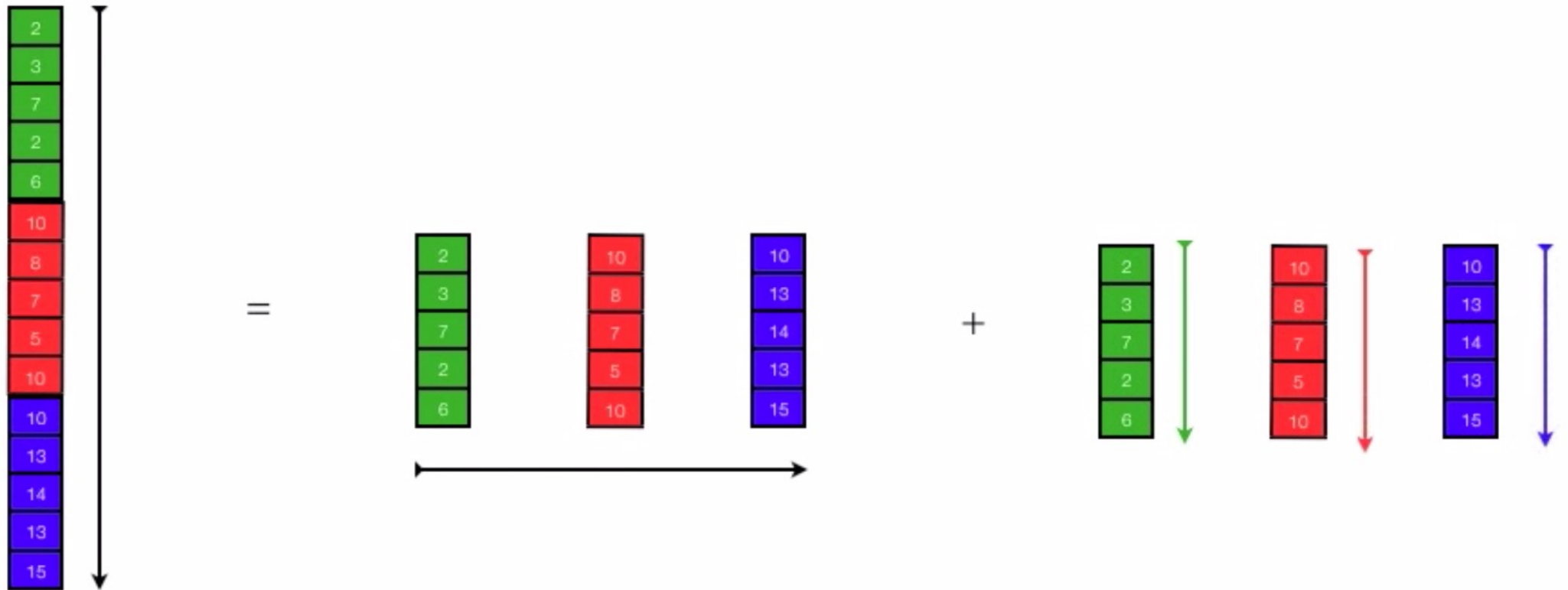
$(\text{mean} - \text{mean})^2$

3. $(\text{mean} - \text{mean})^2 + (\text{mean} - \text{mean})^2 + (\text{mean} - \text{mean})^2$

4. $(\text{mean} - \text{mean})^2 + (\text{mean} - \text{mean})^2 + (\text{mean} - \text{mean})^2 \times 5$

$= (18.1 + 0.1 + 21.8) * 5$
 $= 40.7 * 5$
 $= 203.3$

Property of ANOVA



$$\begin{array}{rclclcl} \text{Total Sum of Squares} & = & \text{Sum of Squares Between Groups} & + & \text{Sum of Squares Within Groups} \\ 257.3 & = & 203.3 & + & 54 \end{array}$$



F Distribution

$$\frac{\text{Sum of Squares Between Groups}}{\text{degrees of freedom}} = \frac{203.3}{2} = 101.667$$

$$F = \frac{101.667}{4.5} = 22.59$$

$$\frac{\text{Sum of Squares Within Groups}}{\text{degrees of freedom}} = \frac{54}{12} = 4.5$$

