Student T Test

Introduction

- A t-test compares the average values of two data sets and determines if they came from the same population
- Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement

$H_0 = \bar{x}_1 = \bar{x}_2$, $H_1 = \bar{x}_1 \neq \bar{x}_2$

- It assumes a null hypothesis that the two means are equal
- Using the formulas, values are calculated and compared against the standard values
- The assumed null hypothesis is accepted or rejected accordingly
- If the null hypothesis qualifies to be rejected, it indicates that data readings are strong and are probably not due to chance

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 Formula

- Calculating a t-test requires three fundamental data values
  - Difference between the mean values from each data set, or the mean difference $\bar{x}$
  - Standard deviation of each group σ
  - Number of data values of each group n
- This comparison helps to determine the effect of chance on the difference, and whether the difference is outside that chance range
- The t-test questions whether the difference between the groups represents a true difference in the study or merely a random difference
- The t-test produces two values as its output:
  - T-value or T-Score → p-value
  - Degrees of freedom

$H_0 = \bar{x} = \mu$

T-Value or T-Score

- The t-value, or t-score, is a ratio of the difference between the mean of the two sample sets and the variation that exists within the sample sets
- The numerator value is the difference between the mean of the two sample sets
- The denominator is the variation that exists within the sample sets and is a measurement of the dispersion or variability
- This calculated t-value is then compared against a value obtained from a critical value table called the T-distribution table
- Higher values of the t-score indicate that a large difference exists between the two sample sets
- The smaller the t-value, the more similarity exists between the two sample sets

Degrees of Freedom

- Degrees of freedom refer 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 set

Paired Sample T-Test

- The correlated t-test, or paired t-test, is a dependent type of test and is performed when the samples consist of matched pairs of similar units, or when there are cases of repeated measures
- This method also applies to cases where the samples are related or have matching characteristics, like a comparative analysis involving children, parents, or siblings

$$T = \frac{mean1 - mean2}{\frac{s(diff)}{\sqrt{n}}}$$

- Where
  - mean1 and mean2 = The average values of each of the sample sets
  - s(diff) = The standard deviation of the differences of the paired data values
  - n = The sample size (the number of paired differences)
  - Degrees of freedom = n -1

Equal Variance or Pooled T-Test

- The equal variance t-test is an independent t-test and is used when the number of samples in each group is the same, or the variance of the two data sets is similar

$$T = \frac{mean1 - mean2}{\frac{(n1-1)*var1^2 + (n2-1)var2^2}{n1+n2-2} * \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

- Where
  - mean1 and mean2 = Average values of each of the sample sets
  - var1 and var2 = Variance of each of the sample sets
  - n1 and n2 = Number of records in each sample set
  - Degrees of Freedom: n1 + n2 - 2

Unequal Variance T-Test

- The unequal variance t-test is an independent t-test and is used when the number of samples in each group is different, and the variance of the two data sets is also different
- This test is also called Welch's t-test

$$T = \frac{mean1 - mean2}{\sqrt{\frac{var1}{n1} + \frac{var2}{n2}}}$$

- Where
  - mean1 and mean2 = Average values of each of the sample sets
  - var1 and var2 = Variance of each of the sample sets
  - n1 and n2 = Number of records in each sample set
- Degrees of Freedom

$$DoF = \frac{\left(\frac{var1^2}{n1} + \frac{var2^2}{n2}\right)^2}{\frac{\left(\frac{var1^2}{n1}\right)^2}{n1 - 1} + \frac{\left(\frac{var2^2}{n2}\right)^2}{n2 - 1}}$$

Which T-Test to use ?

- If two sample sets are same or related => Paired T-Test
- If two sample sets are of same size => Equal Variance T-Test
- If two sample sets have same variance => Equal Variance T-Test
- If two sample sets do not have same variance => Unequal Variance T-Test

Example

- S1 = 19.7, 20.4, 19.6, 17.8, 18.5, 18.9, 18.3, 18.9, 19.5, 21.95
- S2 = 28.3, 26.7, 20.1, 23.3, 25.2, 22.1, 17.7, 27.6, 20.6, 13.7, 23.2, 17.5, 20.6, 18, 23.9, 21.6, 24.3, 20.4, 23.9, 13.3

$\bar{S1} = 19.35$

$\bar{S2} = 21.6$

variance1 = 1.27

variance2 = 19.71

$n1 = 10$

$n2 = 20$

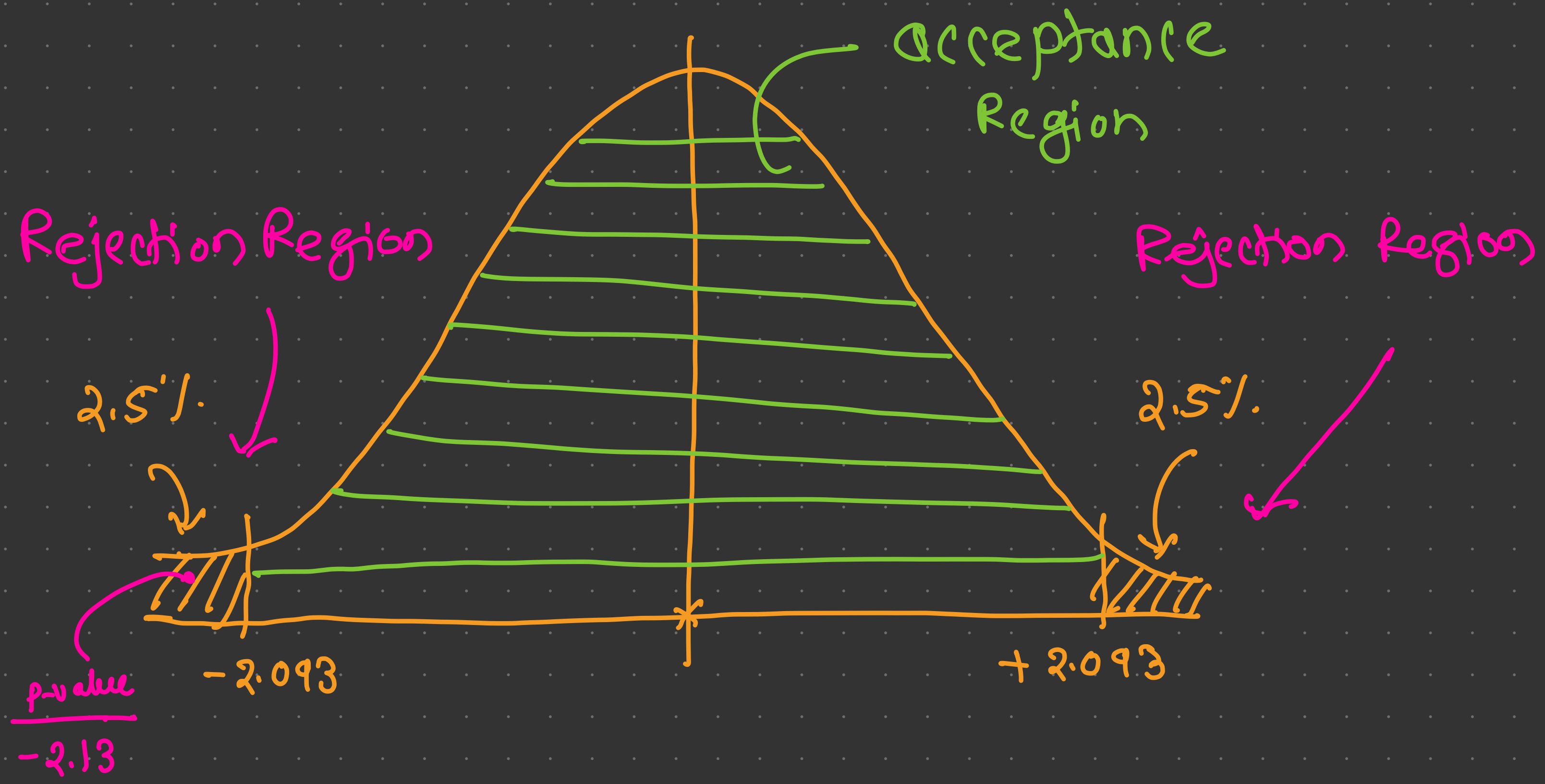
① $H_0 = \bar{S1} = \bar{S2}$, $H_a = \bar{S1} \neq \bar{S2}$

② if α is NOT given, by default $\alpha = 0.05$

③ Since $V1 \neq V2$, we will use unequal variance T-test

④ do the computation, T = −2.13, DoF = 19.31

α = 0.05 , two tailed test



since p-value (-2.13) is falling in Rejection Region, the Null hypothesis is rejected

U-Test

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

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

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

- Where
  - n1 = size of first sample
  - n2 = size of second sample
  - R1 = sum of all observations of first sample
  - R2 = sum of all observations of second sample
- Use the smaller value from u1 and u2
- Lookup the u value in the u-table

Mann Whitney U Test: Example

- S1 = 3, 4, 2, 6, 2, 5
- S2 = 9, 7, 5, 10, 6, 8

Chi-Square Test

Introduction

- The Chi-Square test is a statistical procedure for determining the difference between observed and expected data
- This test can also be used to determine whether it correlates to the categorical variables in our data
- It helps to find out whether a difference between two categorical variables is due to chance or a relationship between them

Test Definition

- A chi-square test is a statistical test that is used to compare observed and expected results
- The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration
- As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables
- A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable
- Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal
- They cannot have a normal distribution since they can only have a few particular values

Use of Chi-Square

- Chi-square is a statistical test that examines the differences between categorical variables from a random sample in order to determine whether the expected and observed results are well-fitting
- Uses of the Chi-Squared test:
  - The Chi-squared test can be used to see if your data follows a well-known theoretical probability distribution like the Normal or Poisson distribution
  - The Chi-squared test allows you to assess your trained regression model's goodness of fit on the training, validation, and test data sets

Limitations

- The chi-square test, for starters, is extremely sensitive to sample size
- Even insignificant relationships can appear statistically significant when a large enough sample is used
- The chi-square can only determine whether two variables are related. It does not necessarily follow that one variable has a causal relationship with the other. It would require a more detailed analysis to establish causality.

Formula

$$x^2 = \frac{\sum(O - E)^2}{E}$$

- Where
    - O = Observed Value
    - E = Expected Value

ANOVA

ANOVA

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1

ANOVA: Rational

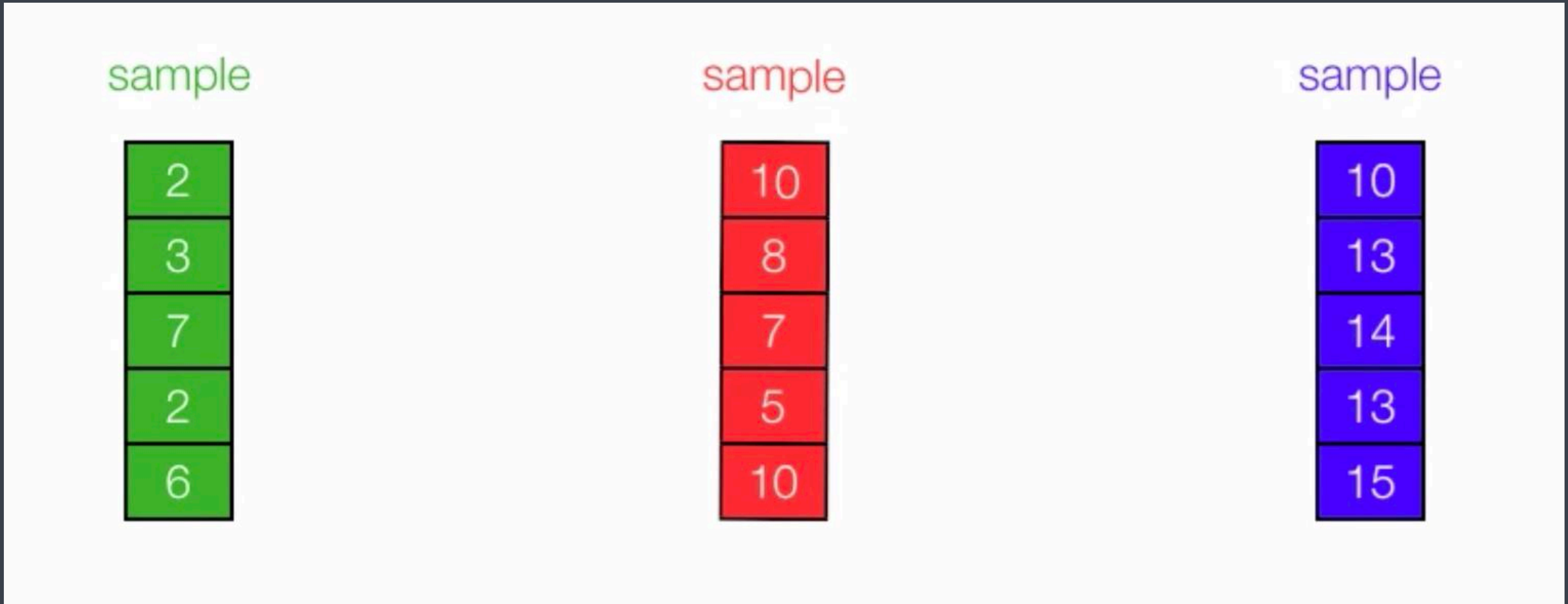
- Basic idea is to partition total variation of the data into two sources
  - Variation within levels (groups)
  - Variation between levels (groups)
- If H0 is true the standardized variances are equal to one another

ANOVA

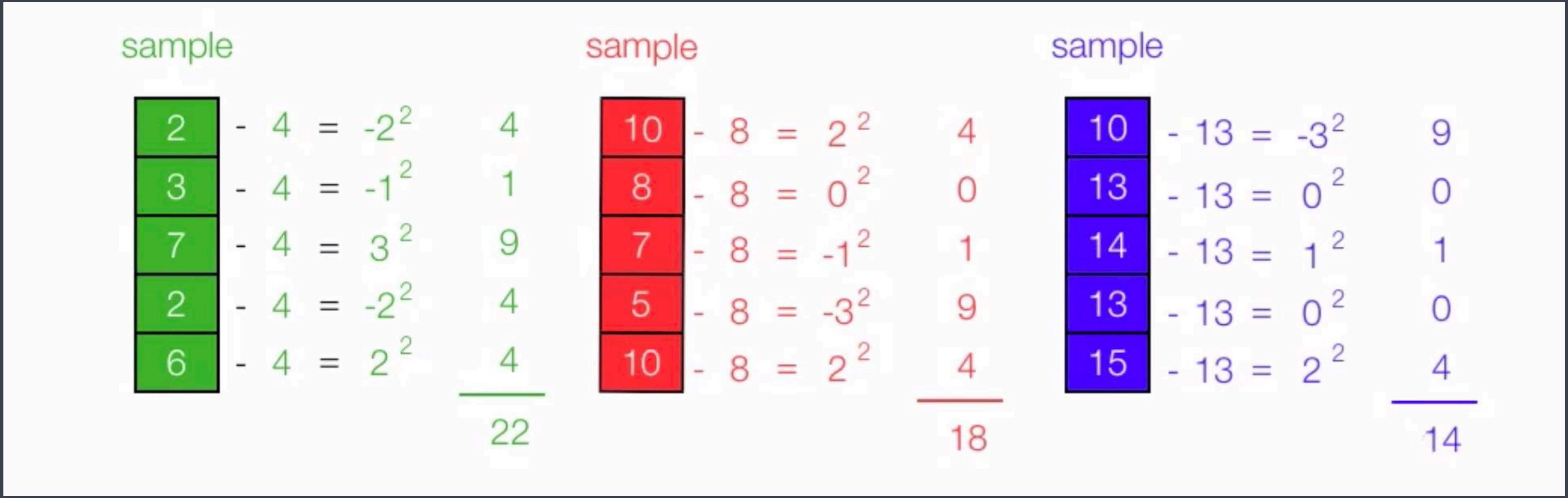
$$F = \frac{Variance\ Between\ Groups}{Variance\ Within\ Groups} = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}}$$

- 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



SSG



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

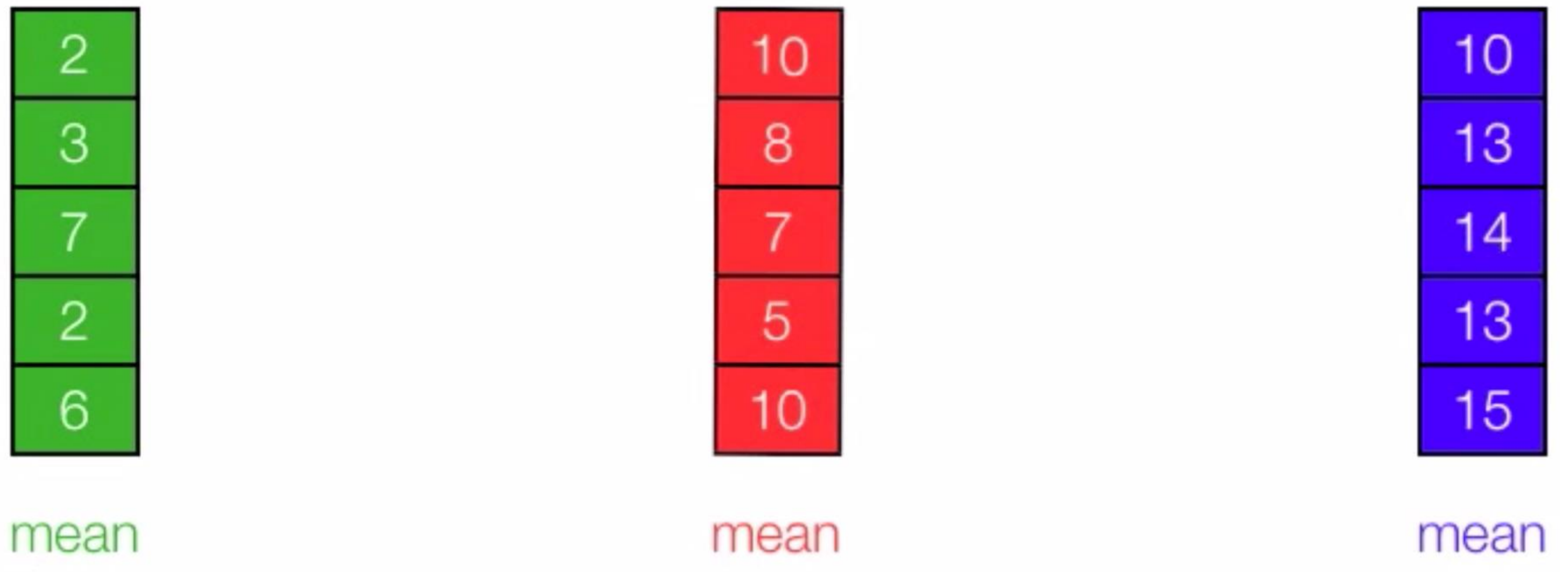
SST

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

mean



1. mean - mean   mean - mean   mean - mean

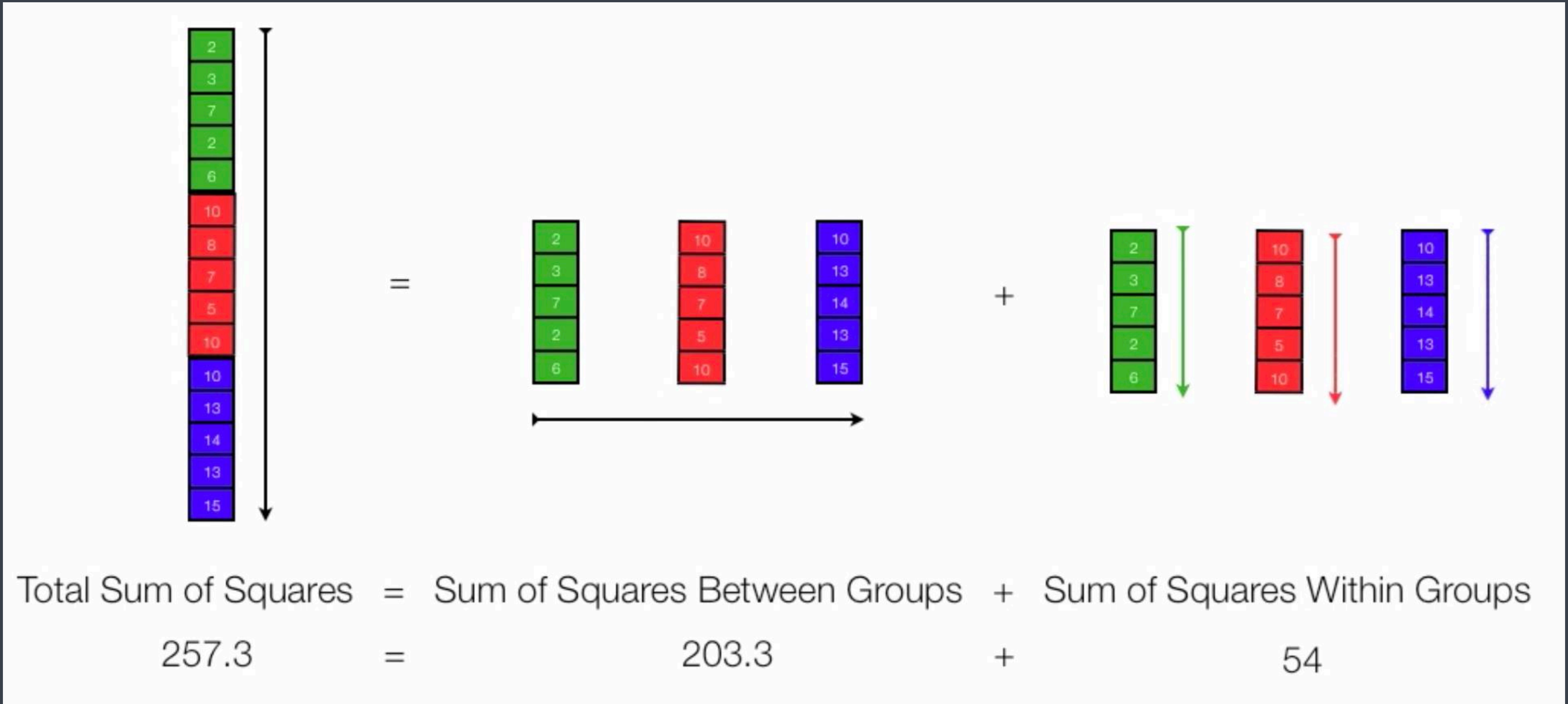
2. $(\text{mean} - \text{mean})^2$ $(\text{mean} - \text{mean})^2$ $(\text{mean} - \text{mean})^2$

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



F Distribution

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

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

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