

Fall 2021

Homework No. 12

ANOVA R Sessions

Name: Ravleen Kaur

Class ID: 43

School ID: 1113138

Course: Statistics for Data Science

Course ID: DTSC-620-W01

Date: 12/01/2021

“Probability Value” (P‐Value )

Definition: The p‐value (or probability value) is probability associated with the calculated test statistic defined as the lowest significance level at which the null hypothesis H0 can be rejected.

1. We start with the P‐value (AUC/prob) greater or equal to α, and look for statistic (needed sample rvv’s values).
2. Large p‐values closer to one provide only weak evidence against the null hypothesis H0.
3. Small p‐values, closer to zero constitute strong evidence against the null hypothesis H0.

Question: Proposition Logic

What sort of a sign is the following sign “⇒” ?

Answer: Sign “⇒” is a sign of implication logical operation. – Implication is “if‐then” logical operation, used frequently embedded in text. • Comment: True value of p cannot imply false q value.

| **p** | **q** | **p** ⇒ **q** |
| --- | --- | --- |
| T | T | T |
| T | F | F |
| F | T | T |
| F | F | T |

Finding P‐values using Z or t Tests

Depending on whether we are doing a z‐test or a t‐test we have:

– Left‐tailed test:

• p‐value = P(z ≤ Zstat) or p‐value=P(t ≤ Tstat) – Right‐tailed test:

• p‐value = P(z ≥ Zstat) or p‐value = P( t ≥ Tstat)

– Two‐tailed test: • p‐value = 2P(z ≥ Zstat ) or p‐value= 2P(t ≥ Tstat)

Hypothesis Test V

**Question:** It has been reported that the average credit card debt for college seniors is μ=$3262. The student senate at a large university feels that their seniors have a debt much less than this (μ<$3262), so it conducts a study of n=50 randomly selected seniors and finds that the average debt is x=$2995, and the population standard deviation is σ=$1100. Assume the test is based on a Type-I error of α=0.05.

**Answer:**

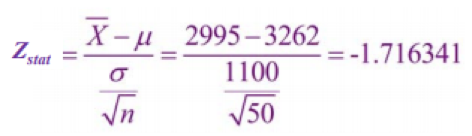
A statistical test uses the data obtained from a sample to compute test‐statistic Zstat.

The numerical value obtained from a statistical test is called the test value.

Based on test‐statistic and test significance level α take a decision about whether the null hypothesis H0: μ=$3262 should be rejected

For left‐tailed test and α=0.05, critical test‐threshold value Zc will be negative (Covering probability 0.05 as AUC).

Using R or Z‐table we find Zc=-1.645 or



As this is a left‐tailed test, our rejection region consists of values of z that are smaller than our critical value of Zc=‐1. 645.Since our test statistic value (Zstat=-1.716341) is less than critical value (Zc=-1.645), we reject the null hypothesis.

**Question.** Let X represent Weschler Adult Intelligence scores (WAIS) where we have X distribution with typical parameters x ~ N(100, 15) Take Statistical Random Sample (SRS) of n = 9 rv’s from

Lake Wobegon population X= {116, 128, 125, 119, 89, 99, 105, 116, 118}

• Calculate: x = 112.8

• Does sample mean provide strong evidence that population mean is μ > 100?

**Answer:** Hypothese

– H0: µ = µ0 = 100 versus

– Ha: µ > 100 (one‐sided)

– Ha: µ ≠ 100 (two‐sided)

Summary: P-Value Test

• There is a similar to z‐test but different, an approach that utilizes a P‐value.

• The P‐Value (or probability value) is the probability of getting a sample statistic (such as the mean) or a more extreme sample statistic in the direction of the alternative hypothesis when the null hypothesis is true.

• The P‐value is the actual AUC under the standard N(0,1) distribution curve of the test values or a more extreme values (further in the tail).

Test Power

Two types of decision errors:

• Type‐I error = erroneous rejection of “true” H0

• Type‐II error = erroneous retention of “false” H0

What is the meaning of the sign “≡”?

• The sign “≡” is used to label logical operation of equivalence.

**Question**:

Why is α known as significance coefficient (or in α\*100% as significance level)?

**Answer:**

• If the Type‐I error of rejecting “a good‐hypothesis” is very important/costly, α must be very small.

• If the error is not very significant not very costly, (no big losses or damages caused by the Type‐I error), α can be made larger.

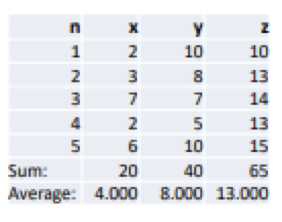
**Question:** Hypothesis is that three sets of samples originate from the same or similar populations.

• Sample means are similar?

**Answer:**

H0: μx = μy = μz

Ha: At least one mean is different from the other two

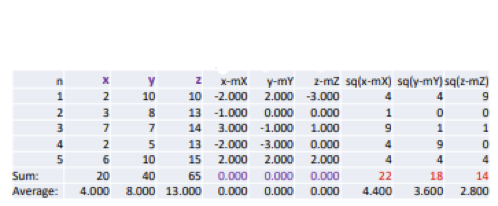


Center all sample values around 0‐average.

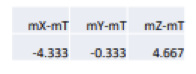
– Centered means that the mean becomes 0.

Find Sum of Squares of centered values and Add all as Sum of Squares Within each Group (SSWg)

Assume that all sample values, x, y and z, originate from the same population.



And then, Batch x, y and z as a; All or total) • Center a around mean mA=mT=8.33. • Find Sum of Squares Total, (SST=257.33) • Find total SST mean mSST=17.16?



Find squares of these deviations, sum them up, and multiply by n=5:

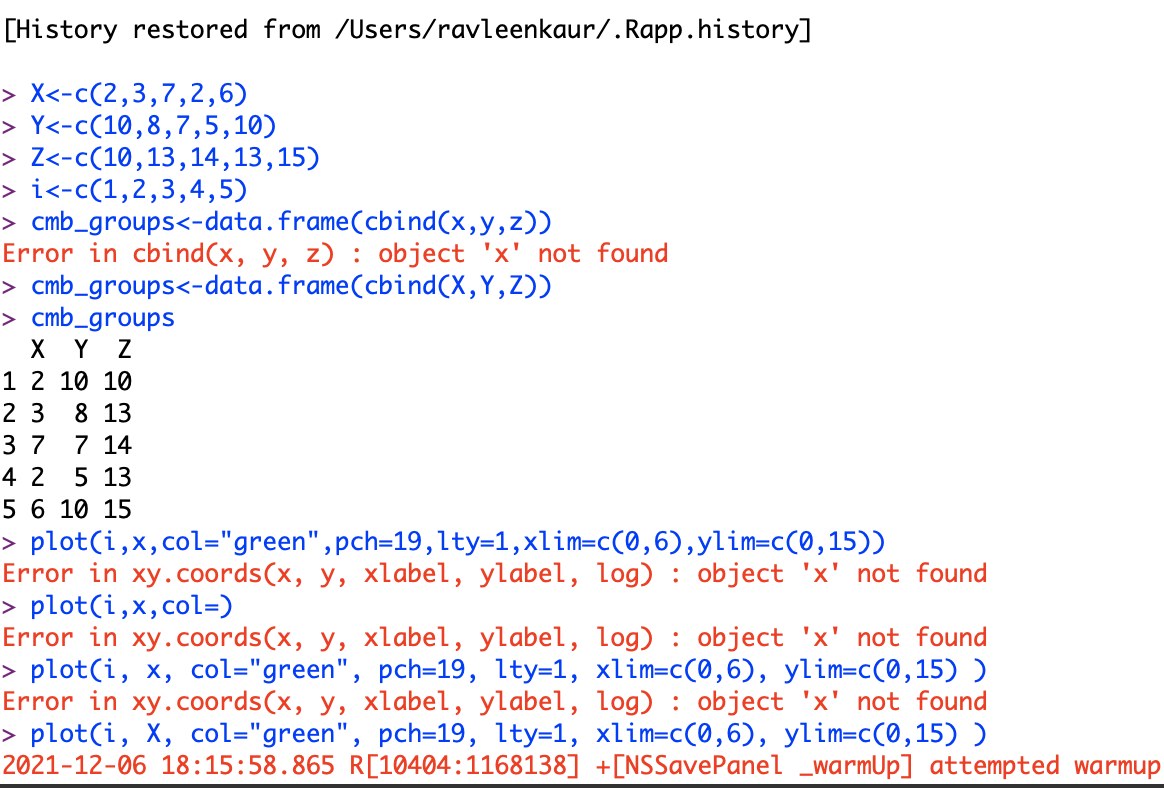
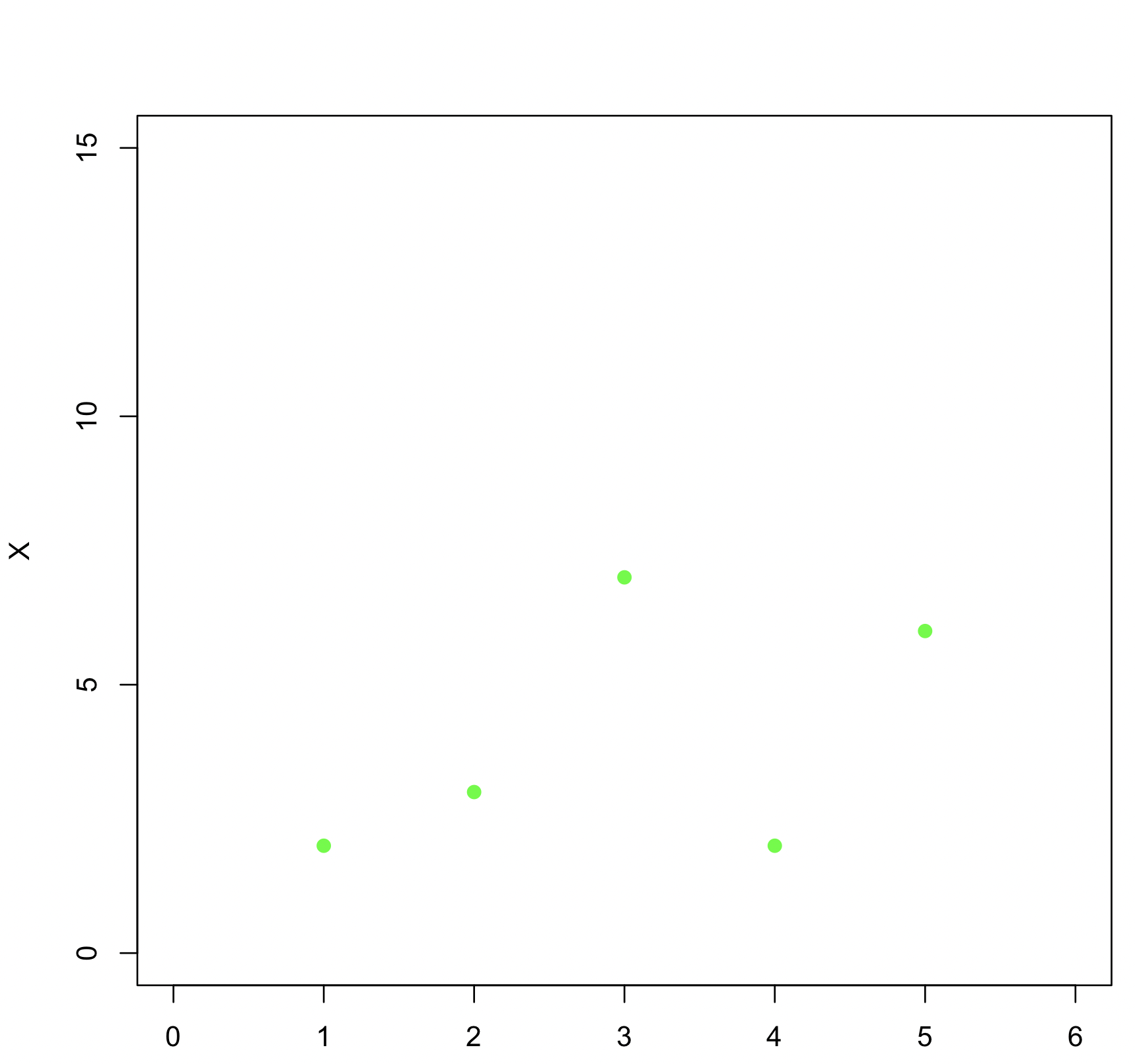


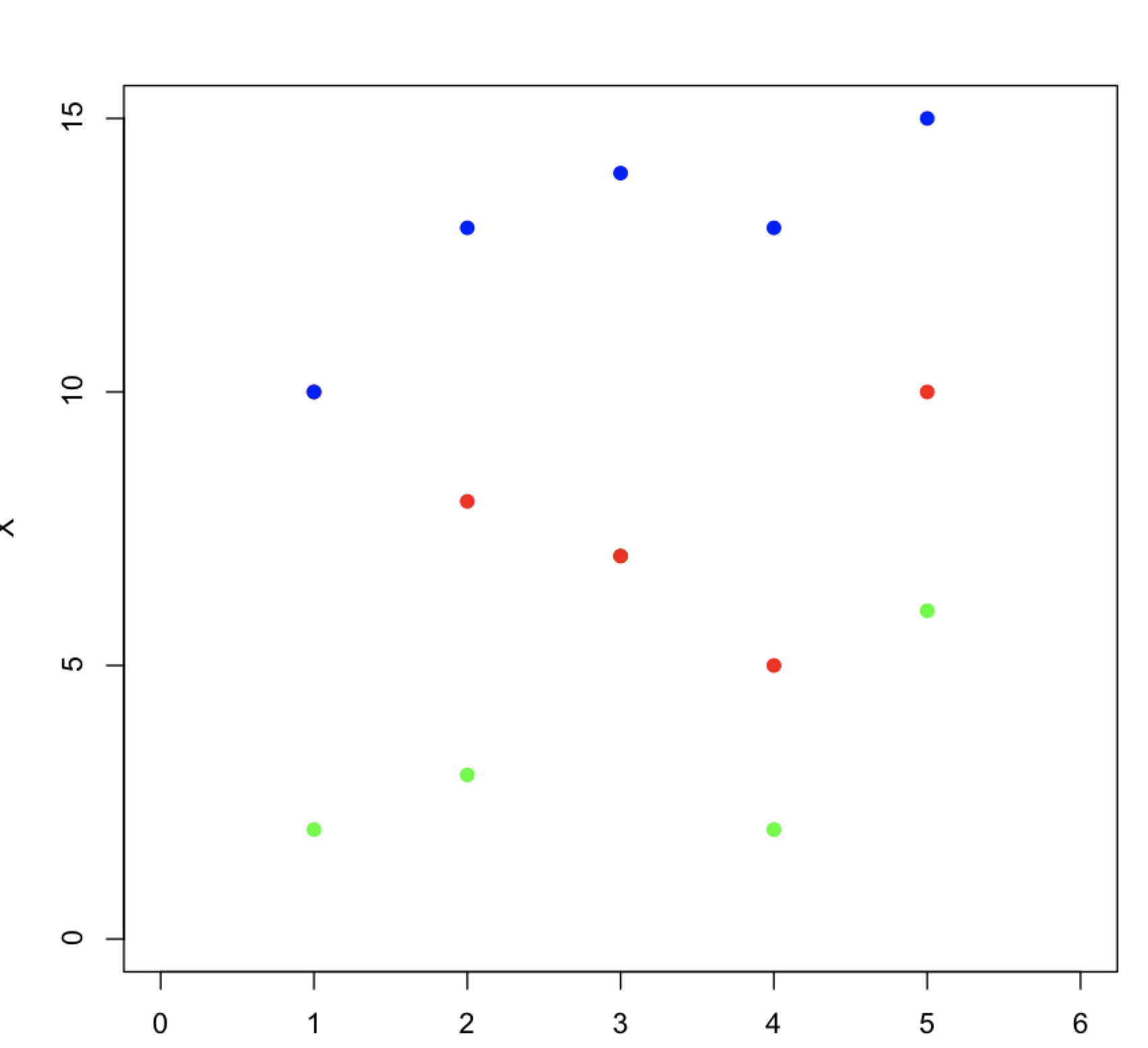
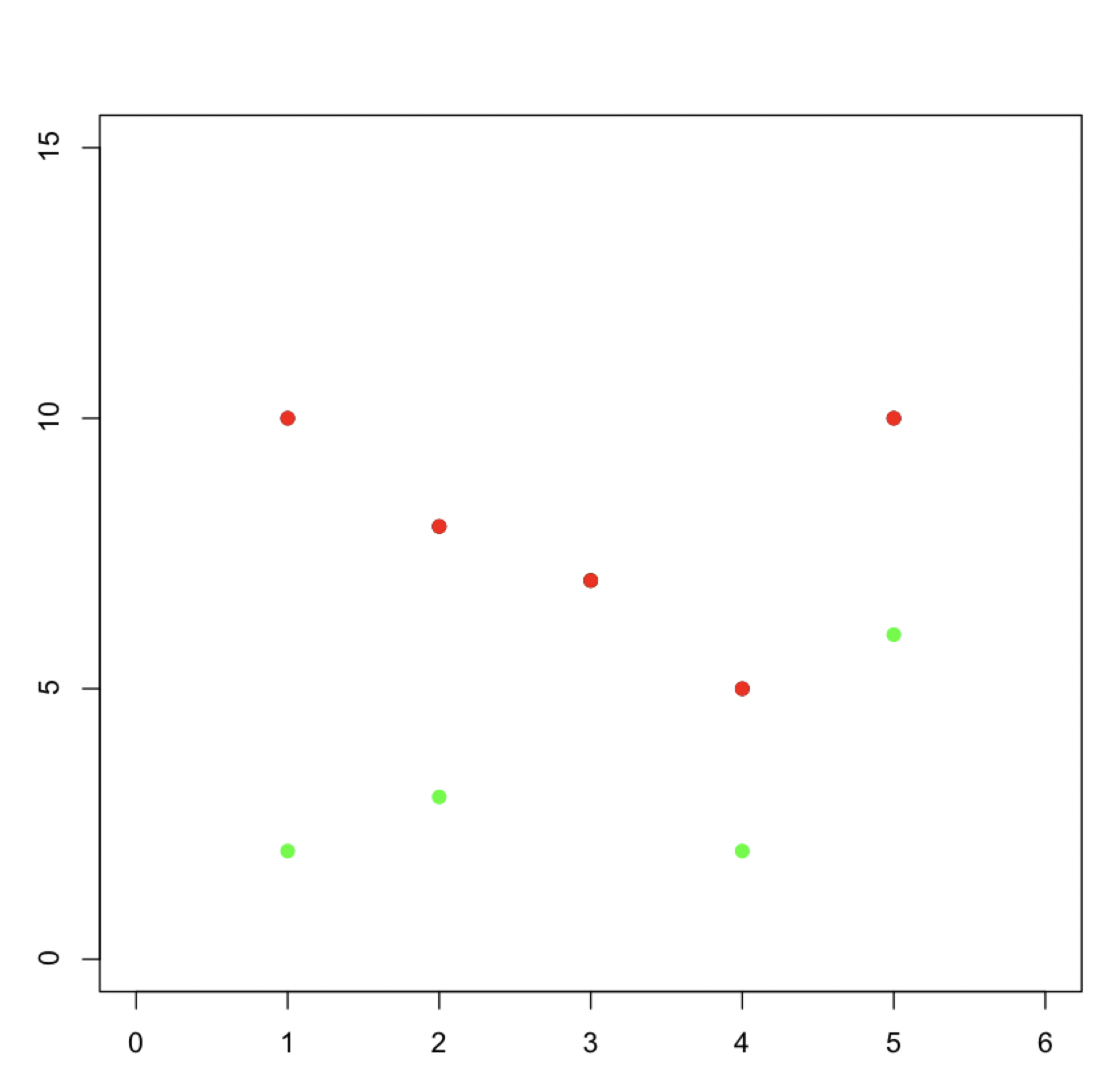
**SS balance**:

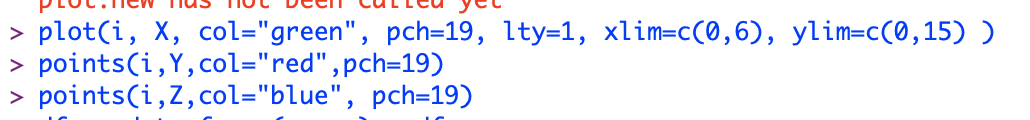
Total Sum of Squares = Sum of Squares Between Groups + Sum of Squares Within Groups SST = SS(Bg) + SS(Wg) 257.33 = 203.333 + 54

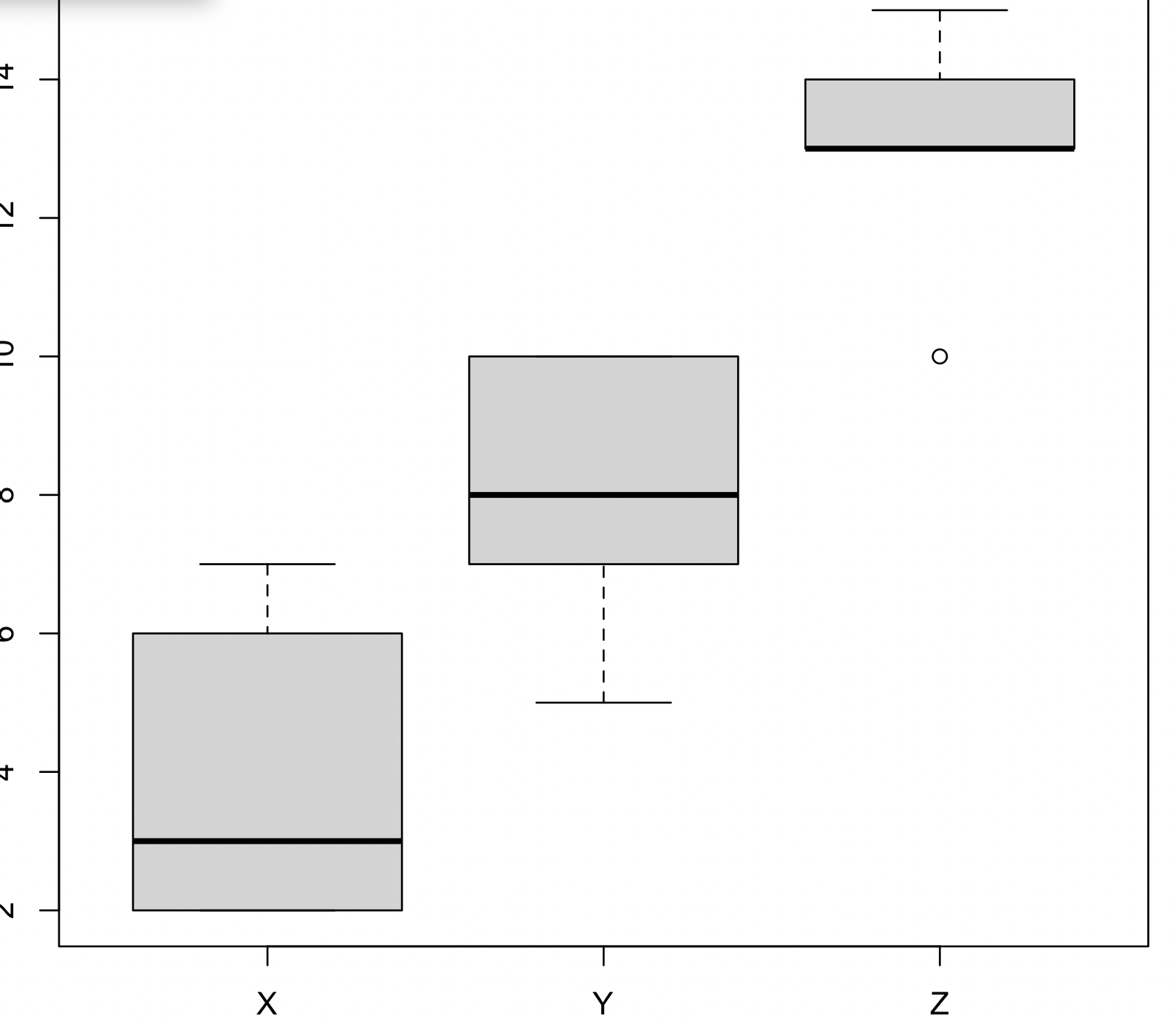
Scatter Plot Data Visualization

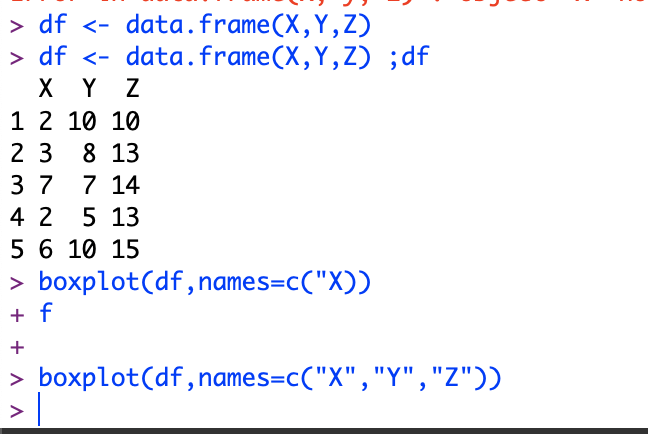
Involving number crunching with ANOVA procedure can be reduced to a few R commands.

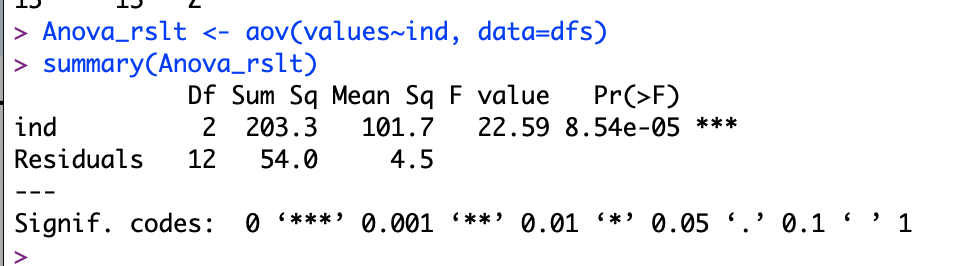
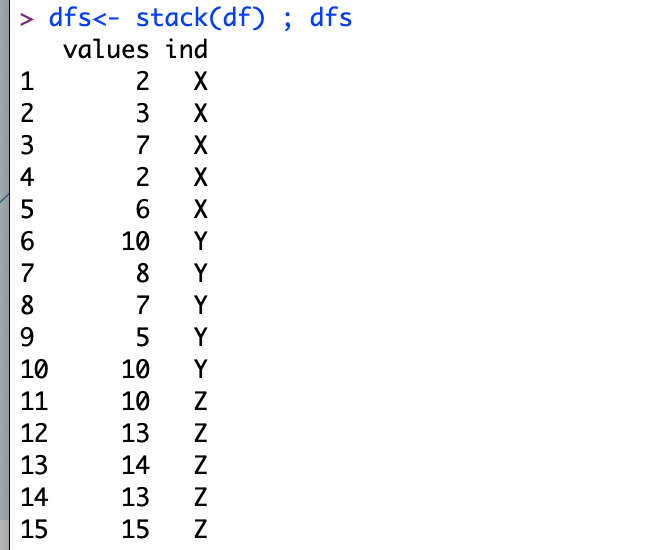


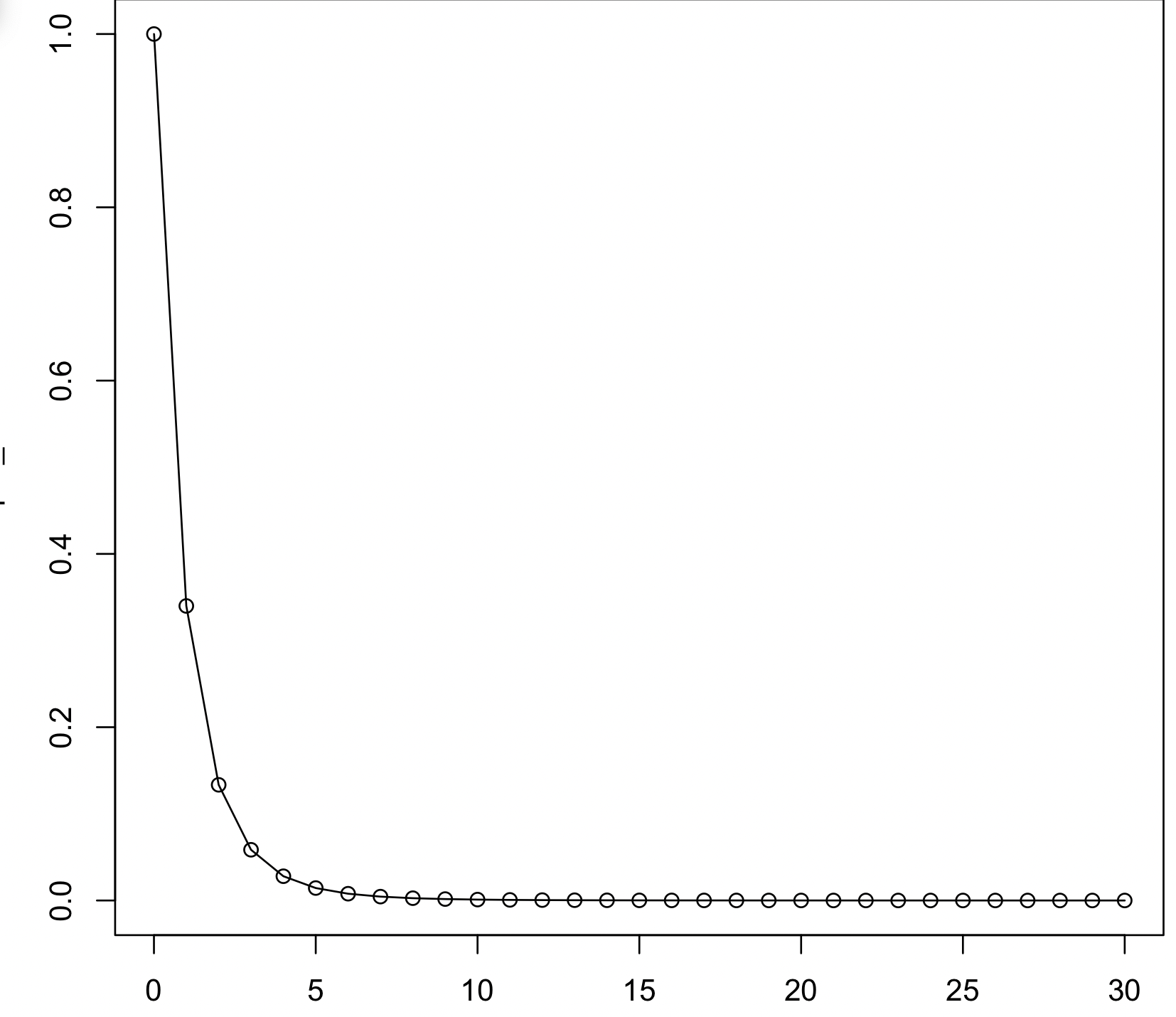


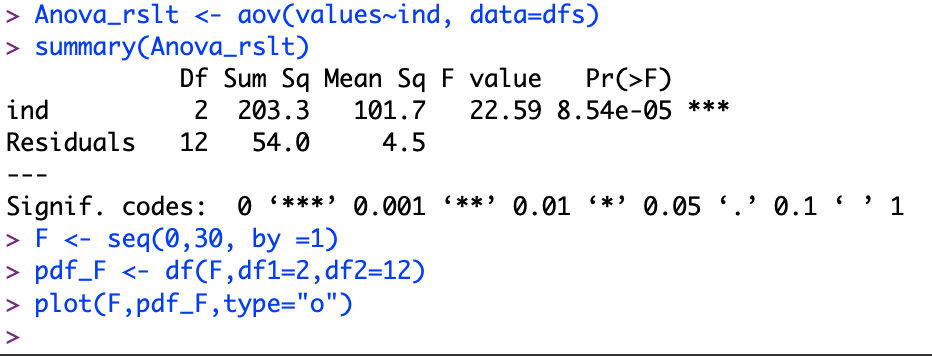












One‐Way ANOVA ‐ Table

ANOVA test results are typically summarized in the so called ANOVA table. • Completing the table with the p‐value • MS – Mean of the Squares

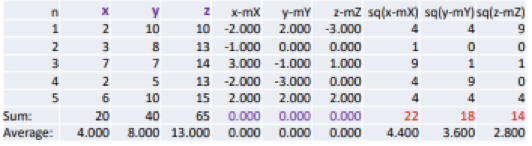
**Question**: Degrees of Freedom • Why is the term “degrees of freedom” or df used in discussions of t‐distributions or F‐distributions?

**Answer: Why is the term “degrees of freedom” or df used in discussions of t‐distributions or F‐distributions?**

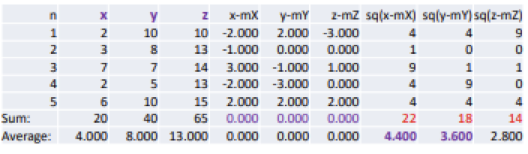
• When a sum of k values produce result r, than with one of the k‐values fixed/pinned‐down, the remaining k‐1 values can (are free to) vary while maintaining the result r. •

With more components added (the larger k) we have more degrees of freedom (the larger is df=k‐1).

**Question.** Which of the 3 groups of samples (x, y or z) have similar sample variance?



**Answer:**

****

Samples X and Y have similar variance. – Observe that 0‐sums are useless.

**Summary: One‐Way ANOVA**

• This was the so called one‐way analysis of variance used to test the claim that three or more population means are equal or not equal.

• This is an extension of the two independent samples t‐test