F- Statistic (ANOVA)

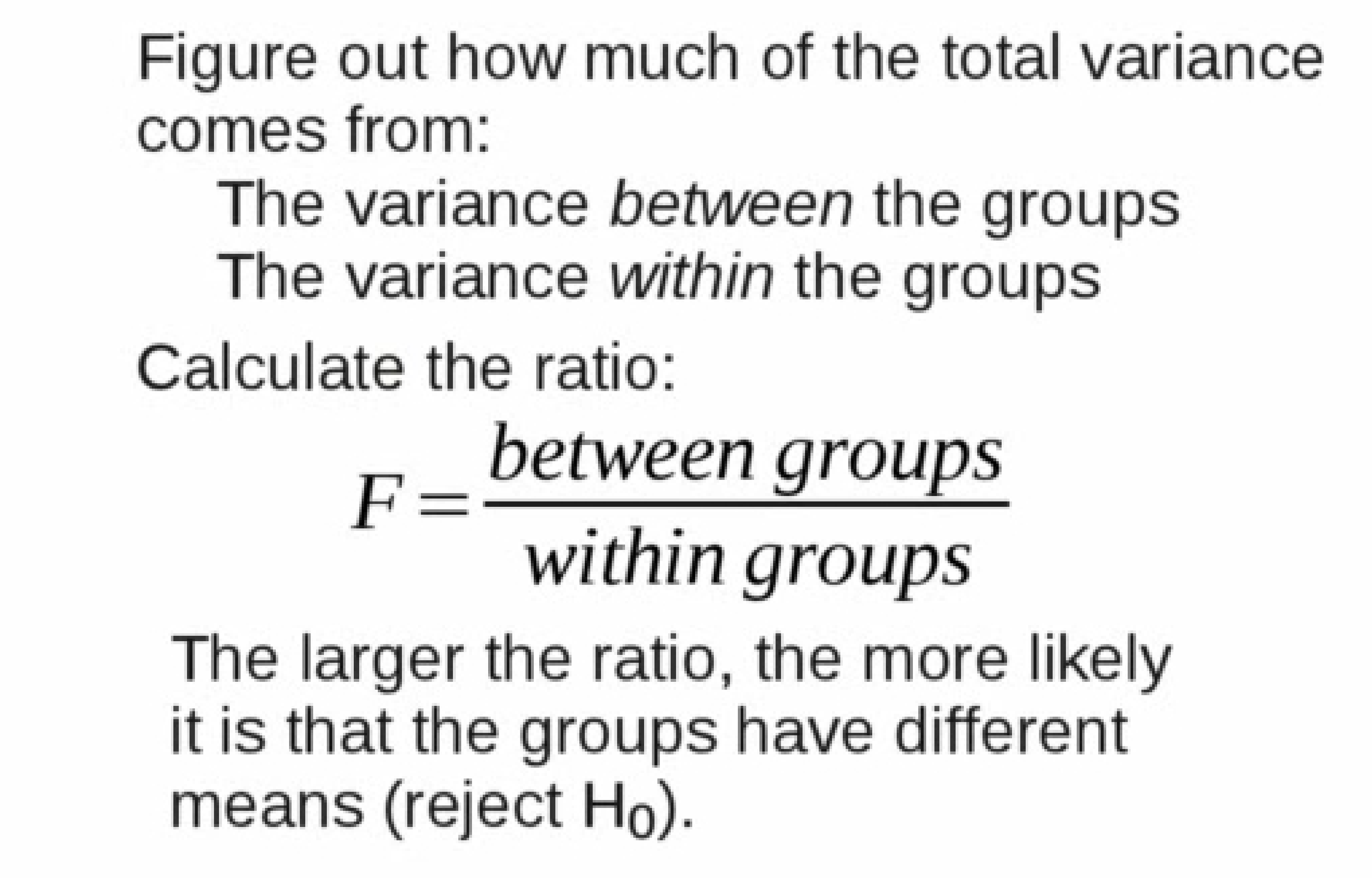


Figure out how much of the total variance comes from:

The variance *between* the groups

The variance *within* the groups

Calculate the ratio:

$$F = \frac{between\ groups}{within\ groups}$$

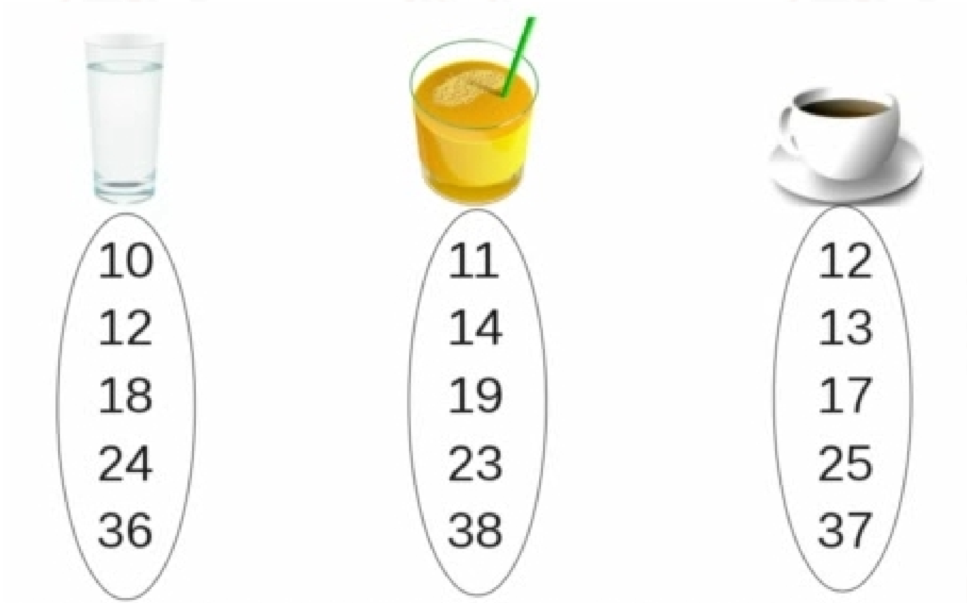
The larger the ratio, the more likely it is that the groups have different means (reject H_0).

Suppose that three groups have given water, fruit juice, coffee. Now we want to test the response between three groups. We have to use F-test as we have more than two groups (otherwise we would have used t-test)

Null Hypothesis: Drink didn't make much difference

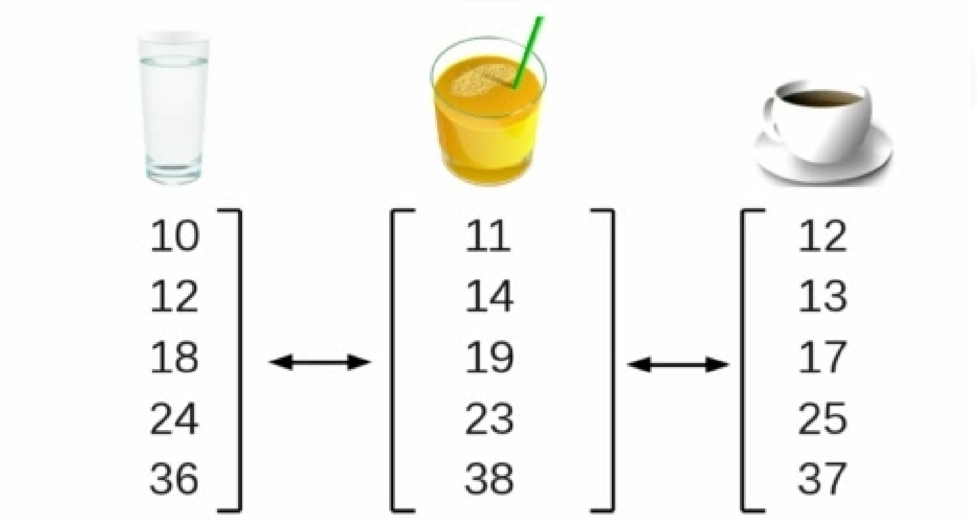
Example 1: Lot of variation with in group, but a little between groups.

This case we will **accept** Null Hypothesis, variation between groups is not much.



10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

There's a lot of variation in each group...

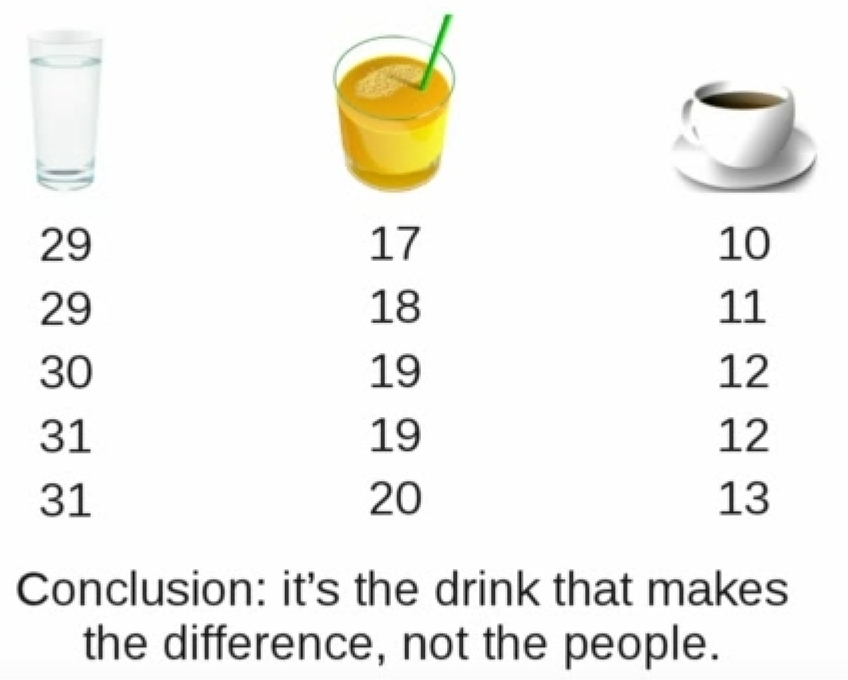


10	11	12
12	14	13
18	19	17
24	23	25
36	38	37

...but each group looks pretty much the same.

Example 2: A little variation with in group, but a lot of variation between groups.

This case we will **reject** the Null Hypothesis

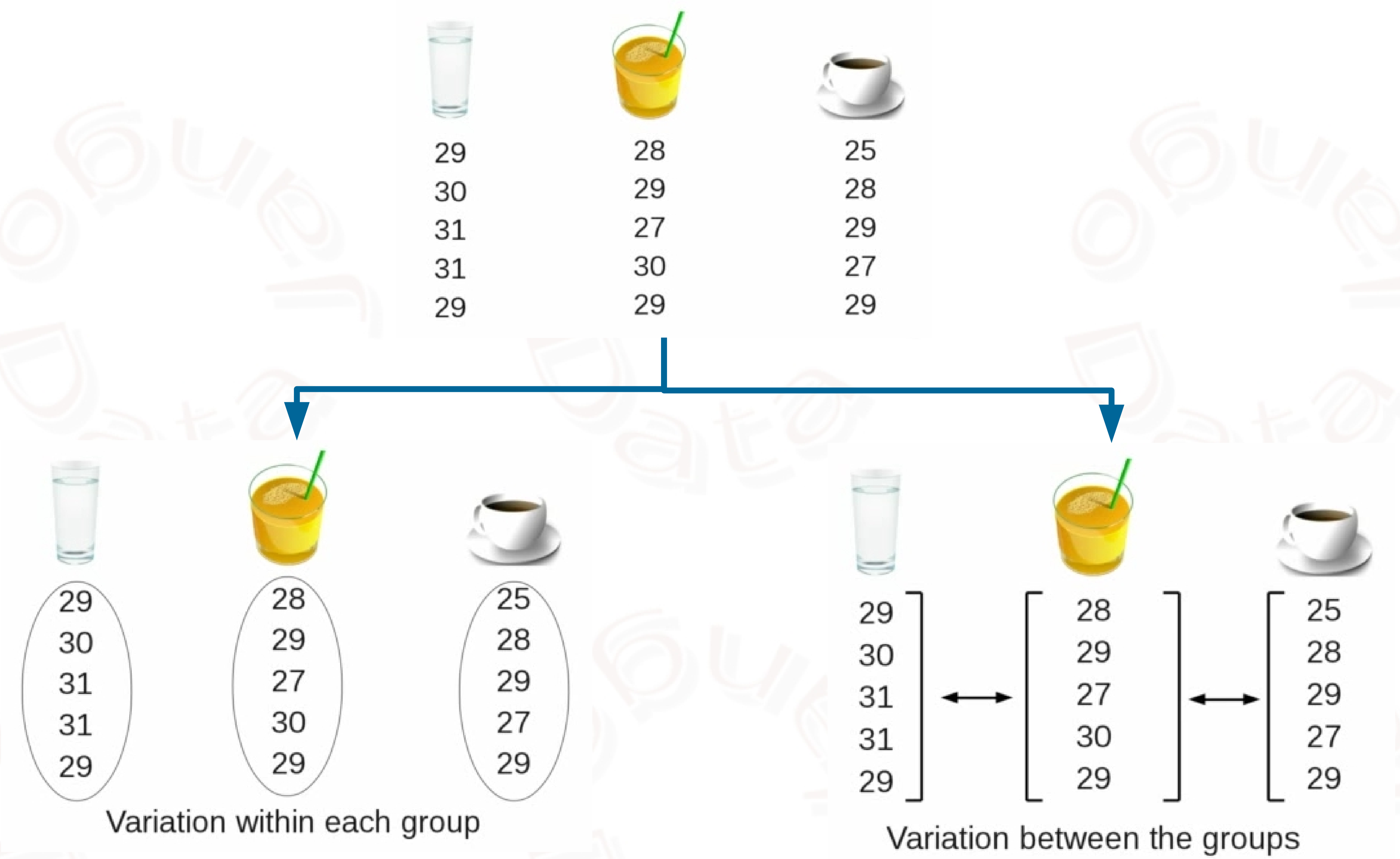


29	17	10
29	18	11
30	19	12
31	19	12
31	20	13

Conclusion: it's the drink that makes the difference, not the people.

Example 3: There is not much variation with in OR between groups.

The result of ANOVA calculation is F(2, 12) = 4.7, the p-value = 0.04 (from F-distribution table) . We can **reject** the Null Hypothesis.



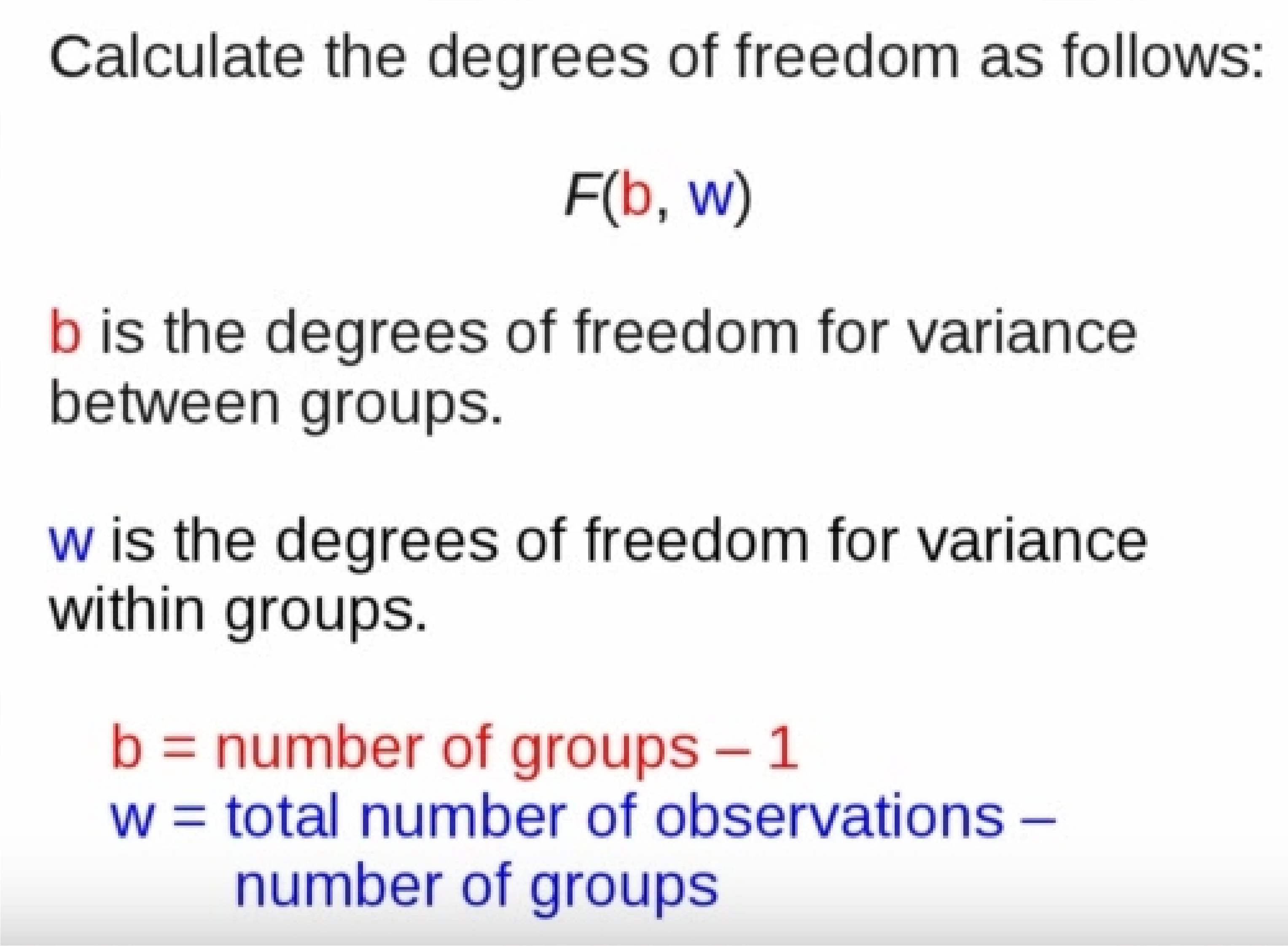
29	28	25
30	29	28
31	27	29
31	30	27
29	29	29

29	28	25
30	29	28
31	27	29
31	30	27
29	29	29

Variation within each group

29	28	25
30	29	28
31	27	29
31	30	27
29	29	29

Variation between the groups



Calculate the degrees of freedom as follows:

$$F(b, w)$$

b is the degrees of freedom for variance between groups.

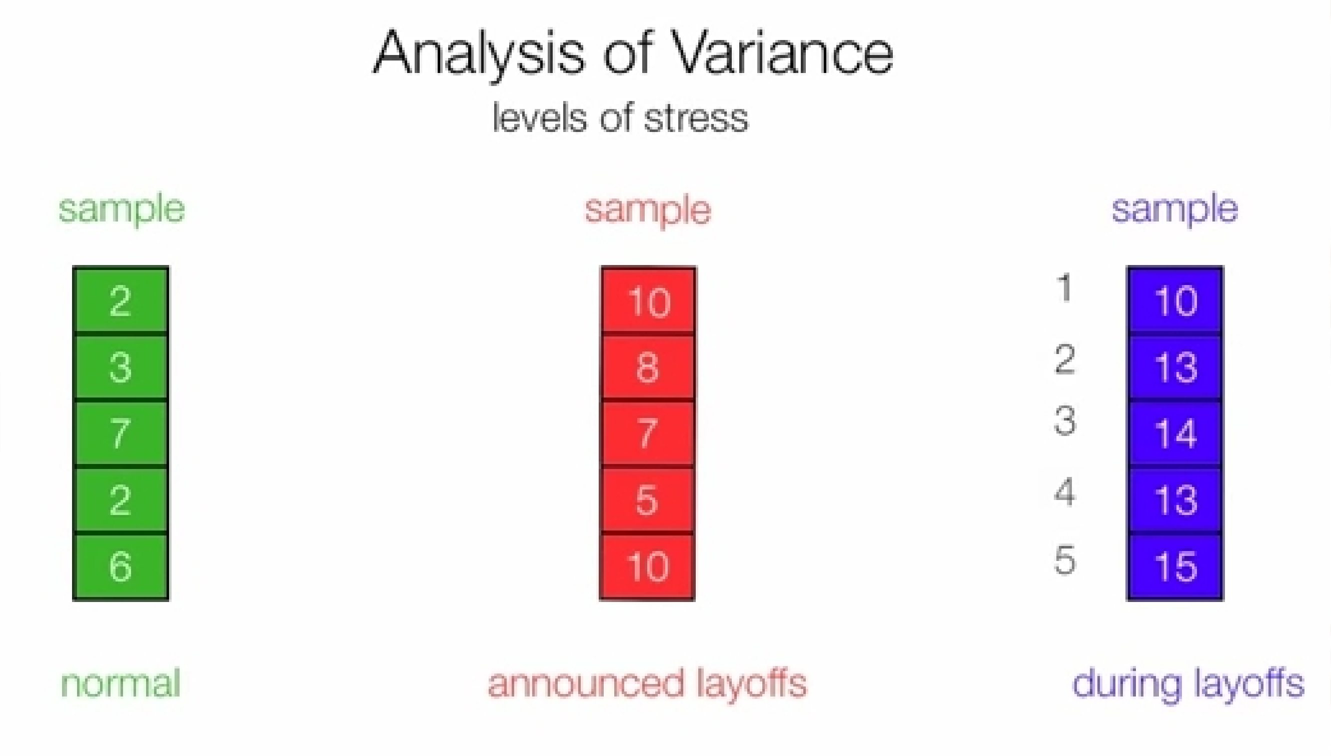
w is the degrees of freedom for variance within groups.

b = number of groups – 1
w = total number of observations – number of groups

ANOVA Calculation example

Sample 1 : Stress under normal condition
Sample 2 : Stress after announced layoffs
Sample 3 : Stress during layoffs
Need to measure the impact of announced layoffs.

Null Hypothesis : No impact of announced layoffs on employee stress



Analysis of Variance

levels of stress

	sample	sample	sample
1	2	10	10
2	3	8	13
3	7	7	14
4	2	5	13
5	6	10	15
	normal	announced layoffs	during layoffs

Analysis of Variance

Sum of Squares Within Groups

sample			sample			sample		
2	- 4 = -2^2	4	10	- 8 = 2^2	4	10	- 13 = -3^2	9
3	- 4 = -1^2	1	8	- 8 = 0^2	0	13	- 13 = 0^2	0
7	- 4 = 3^2	9	7	- 8 = -1^2	1	14	- 13 = 1^2	1
2	- 4 = -2^2	4	5	- 8 = -3^2	9	13	- 13 = 0^2	0
6	- 4 = 2^2	4	10	- 8 = 2^2	4	15	- 13 = 2^2	4
		22			18			14

Sum of **S**quares **W**ithin Groups = 22 + 18 + 14 = 54

Total Sum of Squares = Sum of Squares Between Groups + Sum of Squares Within Groups

54

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

Total **S**um of **S**quares

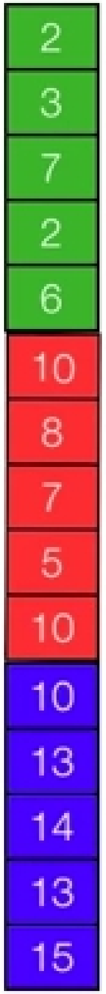
SST = 257.3

Total Sum of Squares = <u>Sum of Squares Between Groups</u> + Sum of Squares Within Groups

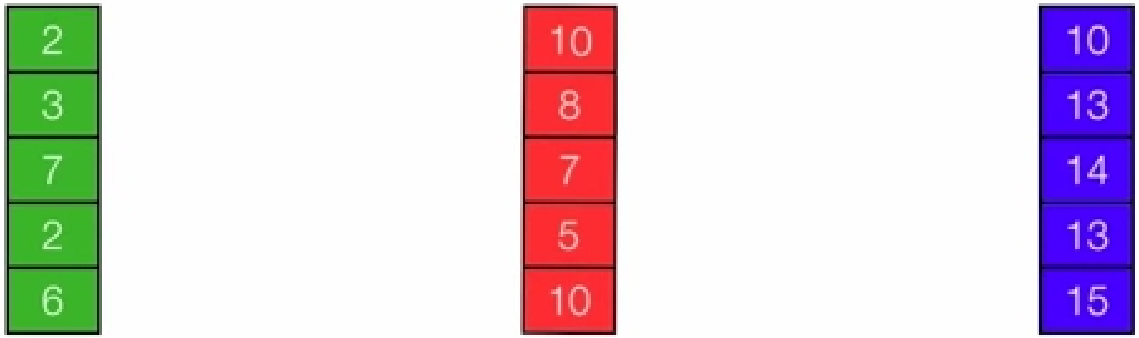
257.3 54

Analysis of Variance

Sum of Squares Between Groups



mean



mean mean mean

1. mean - mean mean - mean mean - mean

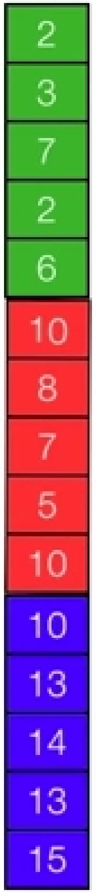
2. $(\text{mean} - \text{mean})^2$ $(\text{mean} - \text{mean})^2$ $(\text{mean} - \text{mean})^2$

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

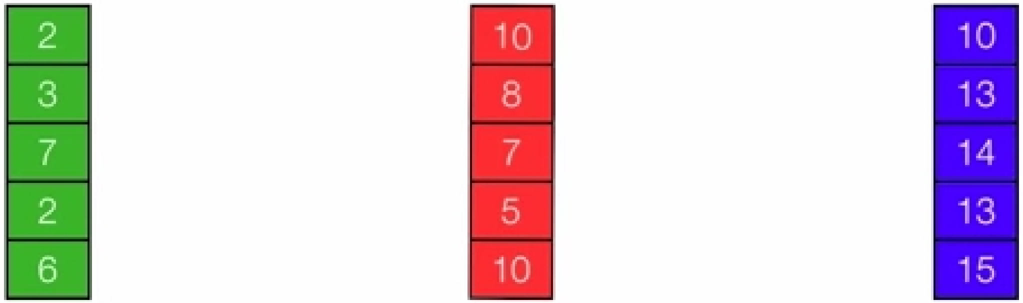
4. $(\text{mean} - \text{mean})^2 + (\text{mean} - \text{mean})^2 + (\text{mean} - \text{mean})^2 \times 5$

Analysis of Variance

Sum of Squares Between Groups



mean 8.3



mean $4 - 8.3 = (-4.3)^2$ $8 - 8.3 = (-.3)^2$ $13 - 8.3 = (4.7)^2$

$18.8 + .1 + 21.8 = 40.7$

$40.7 \times 5 = 203.3$

Total Sum of Squares = Sum of Squares Between Groups + Sum of Squares Within Groups

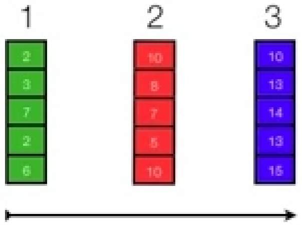
257.3 = 203.3 + 54

Final Calculations

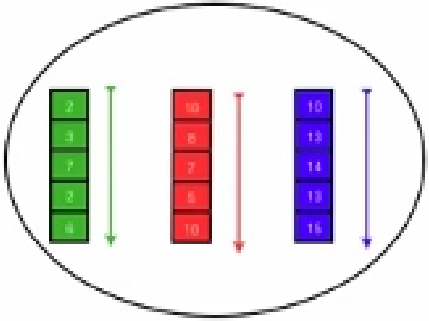
$$\frac{\text{Sum of Squares Between Groups}}{\text{degrees of freedom}} = \frac{203.3}{}$$

Groups - 1

3 - 1 = 2



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



15 - 3 = 12

Observations - Groups

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

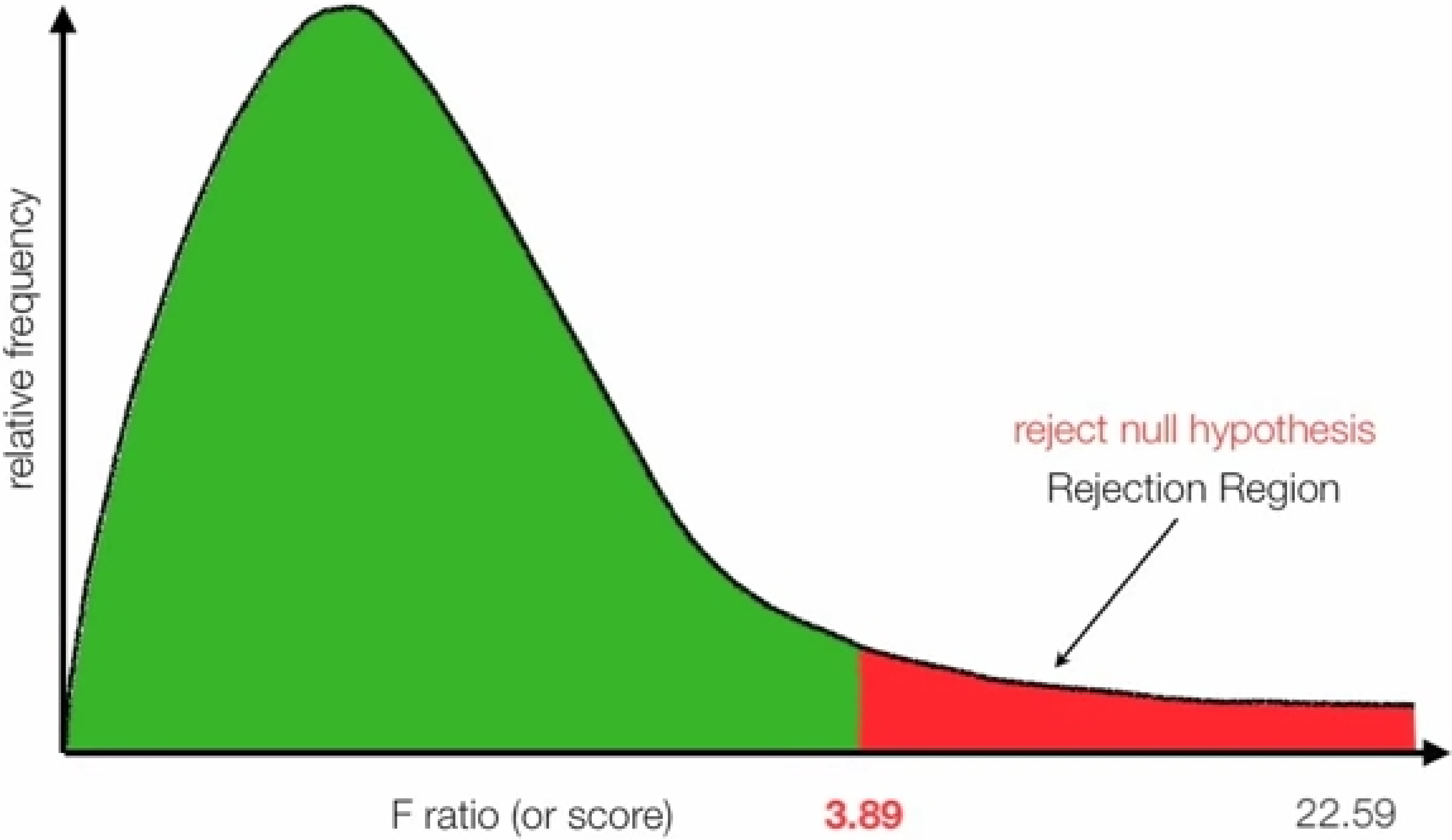
$$\boxed{F = \frac{101.667}{4.5} = 22.59}$$

F Distribution   F (2,12) = 22.59, *p*<.05

degrees of freedom numerator

degrees of freedom denominator

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	246.0	248.0	249.1	250.1
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25

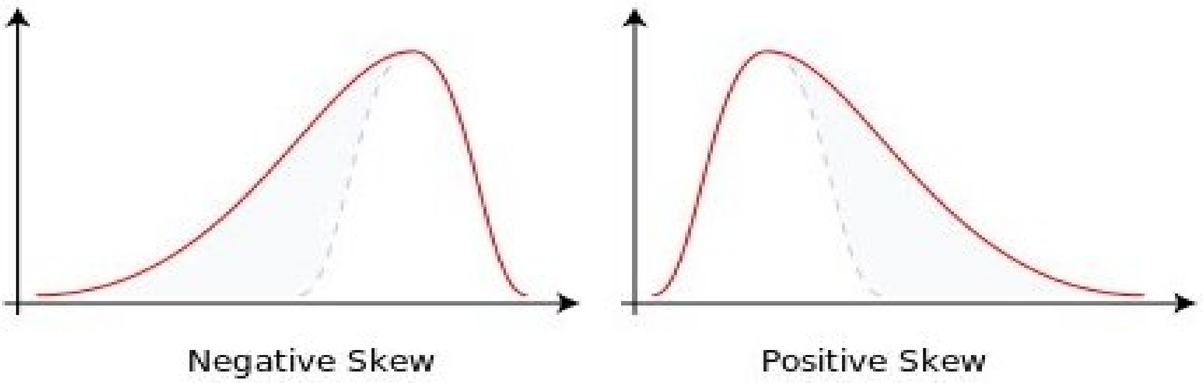

relative frequency
reject null hypothesis
Rejection Region
F ratio (or score)
3.89
22.59

Skewness

Consider the two distributions in the figure just below. Within each graph, the values on the right side of the distribution taper differently from the values on the left side. These tapering sides are called tails, and they provide a visual means to determine which of the two kinds of skewness a distribution has:

negative skew: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be left-skewed, left-tailed, or skewed to the left, despite the fact that the curve itself appears to be skewed or leaning to the right; left instead refers to the left tail being drawn out and, often, the mean being skewed to the left of a typical center of the data. A left-skewed distribution usually appears as a right-leaning curve.

positive skew: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be right-skewed, right-tailed, or skewed to the right, despite the fact that the curve itself appears to be skewed or leaning to the left; right instead refers to the right tail being drawn out and, often, the mean being skewed to the right of a typical center of the data. A right-skewed distribution usually appears as a left-leaning curve.[1]



Skewness in a data series may sometimes be observed not only graphically but by simple inspection of the values. For instance, consider the numeric sequence (49, 50, 51), whose values are evenly distributed around a central value of 50. We can transform this sequence into a negatively skewed distribution by adding a value far below the mean, e.g. (40, 49, 50, 51). Similarly, we can make the sequence positively skewed by adding a value far above the mean, e.g. (49, 50, 51, 60).

The Pearson mode skewness, or first skewness coefficient, is defined as

$$\frac{\text{mean} - \text{mode}}{\text{standard deviation}}.$$

The Pearson median skewness, or second skewness coefficient, is defined as

$$\frac{3\,(\text{mean} - \text{median})}{\text{standard deviation}}.$$

Person MODE vs MEDIAN skewness: Pearson’s first coefficient of skewness uses the mode. Therefore, if the mode is made up of too few pieces of data it won’t be a stable measure of central tendency. For example, the mode in both these sets of data is 9:
1 2 3 4 5 6 7 8 9 9.
1 2 3 4 5 6 7 8 9 9 9 9 9 9 9 9 9 9 9 9 10 12 12 13.
In the first set of data, the mode only appears twice. This isn’t a good measure of central tendency so you would be cautioned not to use Pearson’s coefficient of skewness. The second set of data has a more stable set (the mode appears 12 times). Therefore, Pearson’s coefficient of skewness will likely give you a reasonable result.

Interpretation (Skewness):

- The direction of skewness is given by the sign.
- The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the distribution differs from a normal distribution.
- A value of zero means no skewness at all.
- A large negative value means the distribution is negatively skewed.
- A large positive value means the distribution is positively skewed.

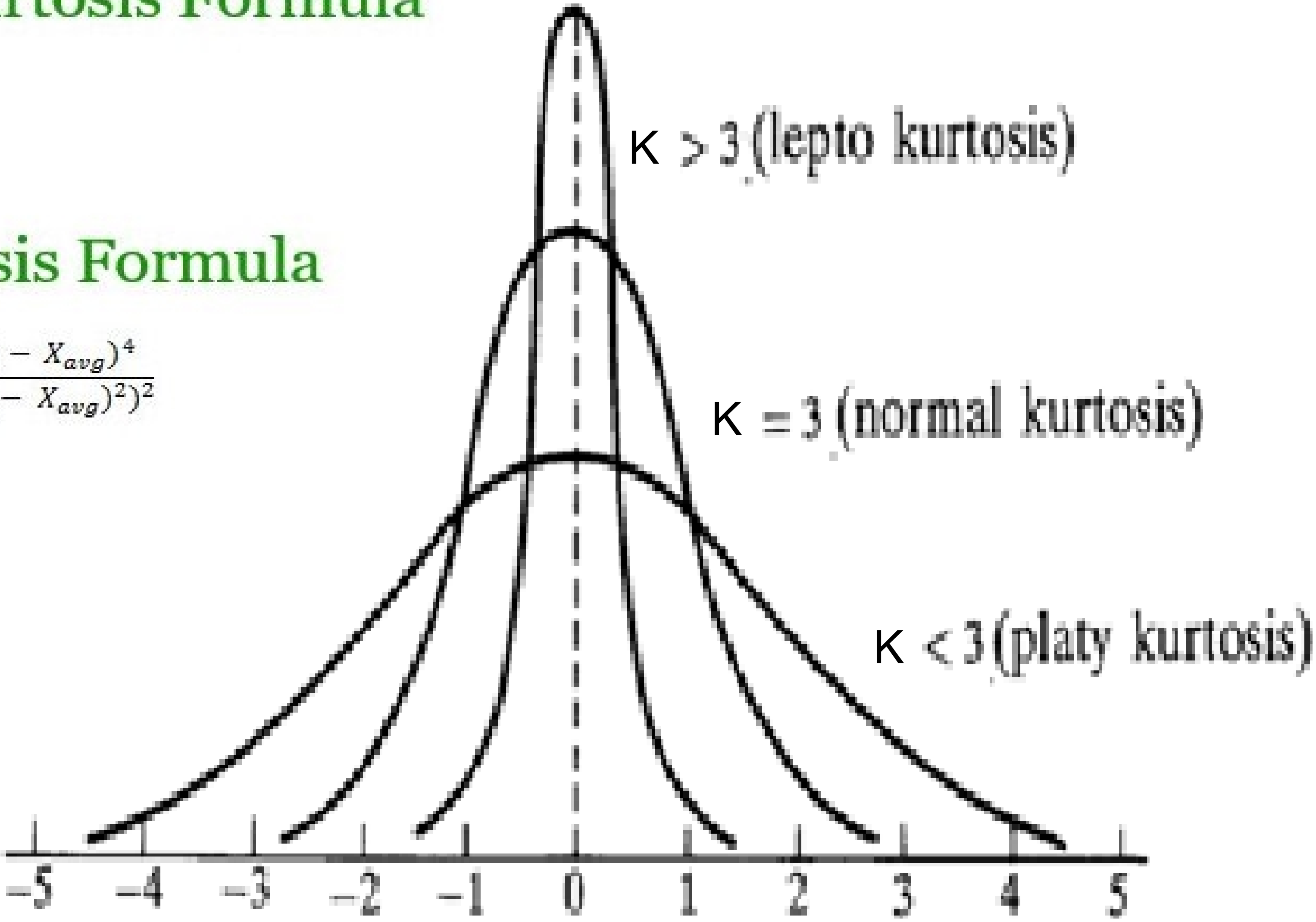
Kurtosis

Population Kurtosis Formula

$$K = n \frac{\sum_{i=1}^{n}(X_i - X_{avg})^4}{(\sum_{i=1}^{n}(X_i - X_{avg})^2)^2}$$

Sample Kurtosis Formula

$$K = \frac{n(n+1)(n-1)}{(n-2)(n-3)} \frac{\sum_{i=1}^{n}(X_i - X_{avg})^4}{(\sum_{i=1}^{n}(X_i - X_{avg})^2)^2}$$


K > 3 (lepto kurtosis)
K = 3 (normal kurtosis)
K < 3 (platy kurtosis)
−5
−4
−3
−2
−1
0
1
2
3
4
5