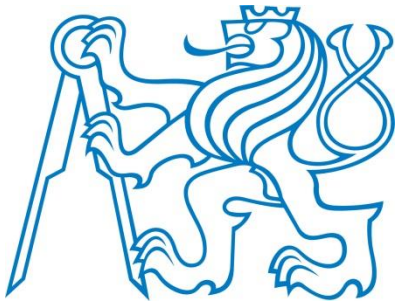
Experimental Data Analysis

in ©MATLAB

Lecture 5:

Analysis of variance, post-hoc testing

Jan Rusz
Czech Technical University in Prague



Why ANOVA?

- Prior to ANOVA, we were able to compare only two populations
  - Independent samples t-test (random)
  - Matched-sample t-test (paired)
- Limiting ourselves to the comparison of two populations is … limiting
- What if we wish to compare the means of several populations?
- ANOVA: ANalysis Of VAriance
- ANOVA represents most common tool used by researcher worldwide to compare the means of different populations

Motivation

Probability that 2 people share the same birthday is quite low 0. 27%.

Probability that in a group of 10 people, 2 people share the same birthday is 11.7%.

Probability that in a group of 20 people, 2 people share the same birthday is 41.1%.

Probability that in a group of 366 people, 2 people share the same birthday is 100%.

We want to find an effect in many groups of people. If we compare each group to others (i.e. many hypothesis) our chances changes. We should correct the level of significance (see lecture 6) or perform an Omnibus test (e.g. ANOVA), a “master test” which give the answer by testing only one hypothesis.

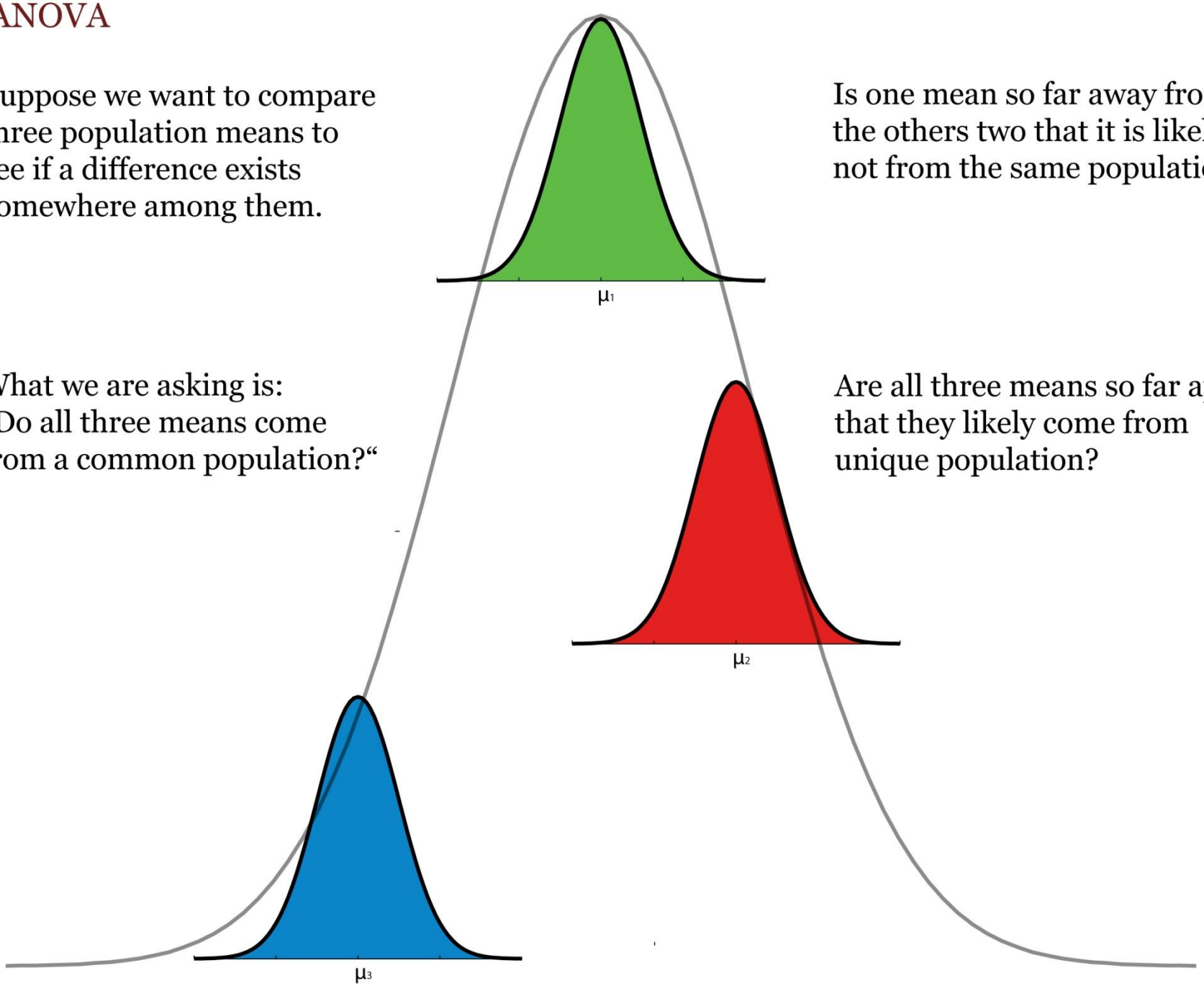
ANOVA

Suppose we want to compare three population means to see if a difference exists somewhere among them.

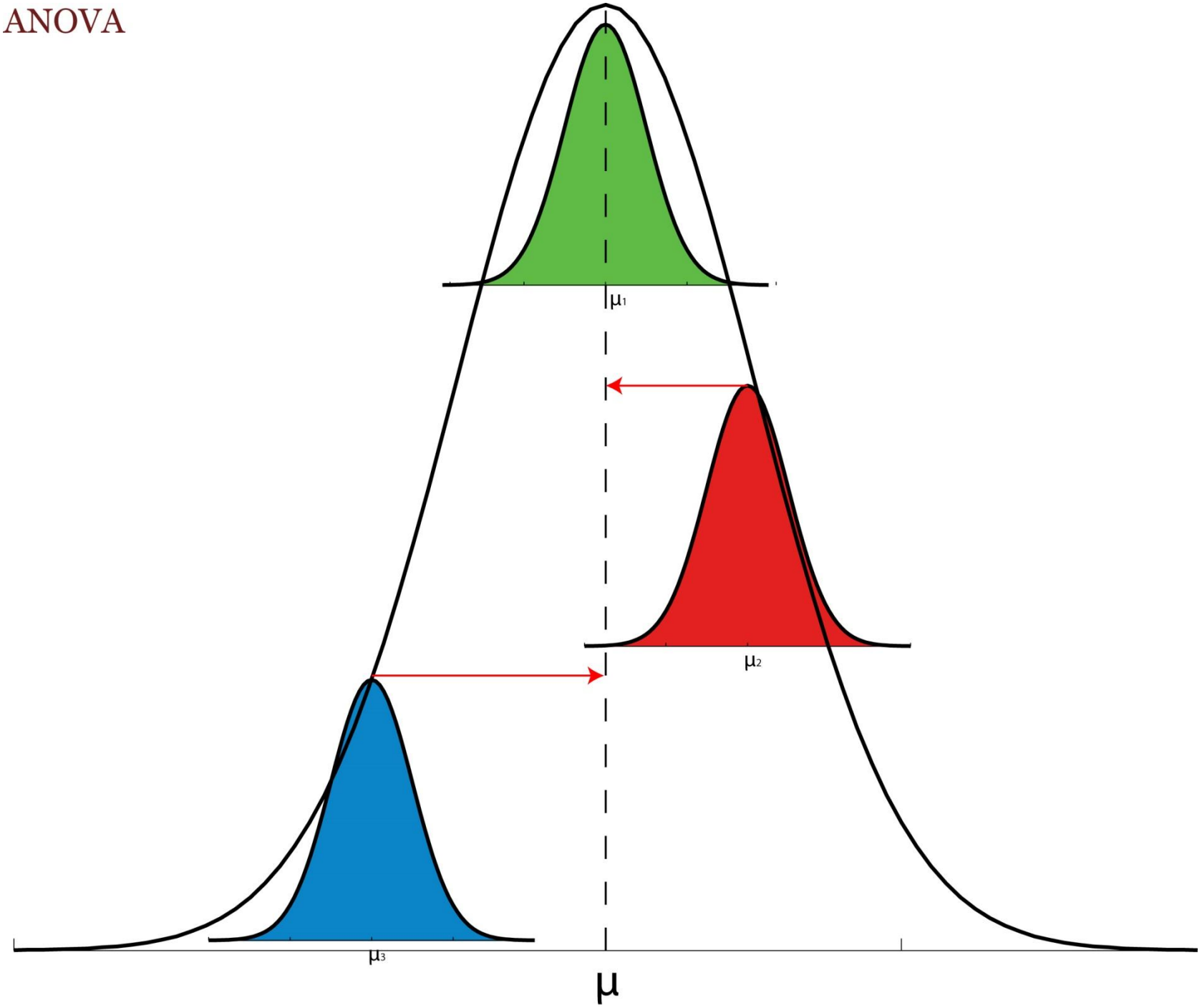
Is one mean so far away from the others two that it is likely not from the same population?

What we are asking is:
„Do all three means come from a common population?“

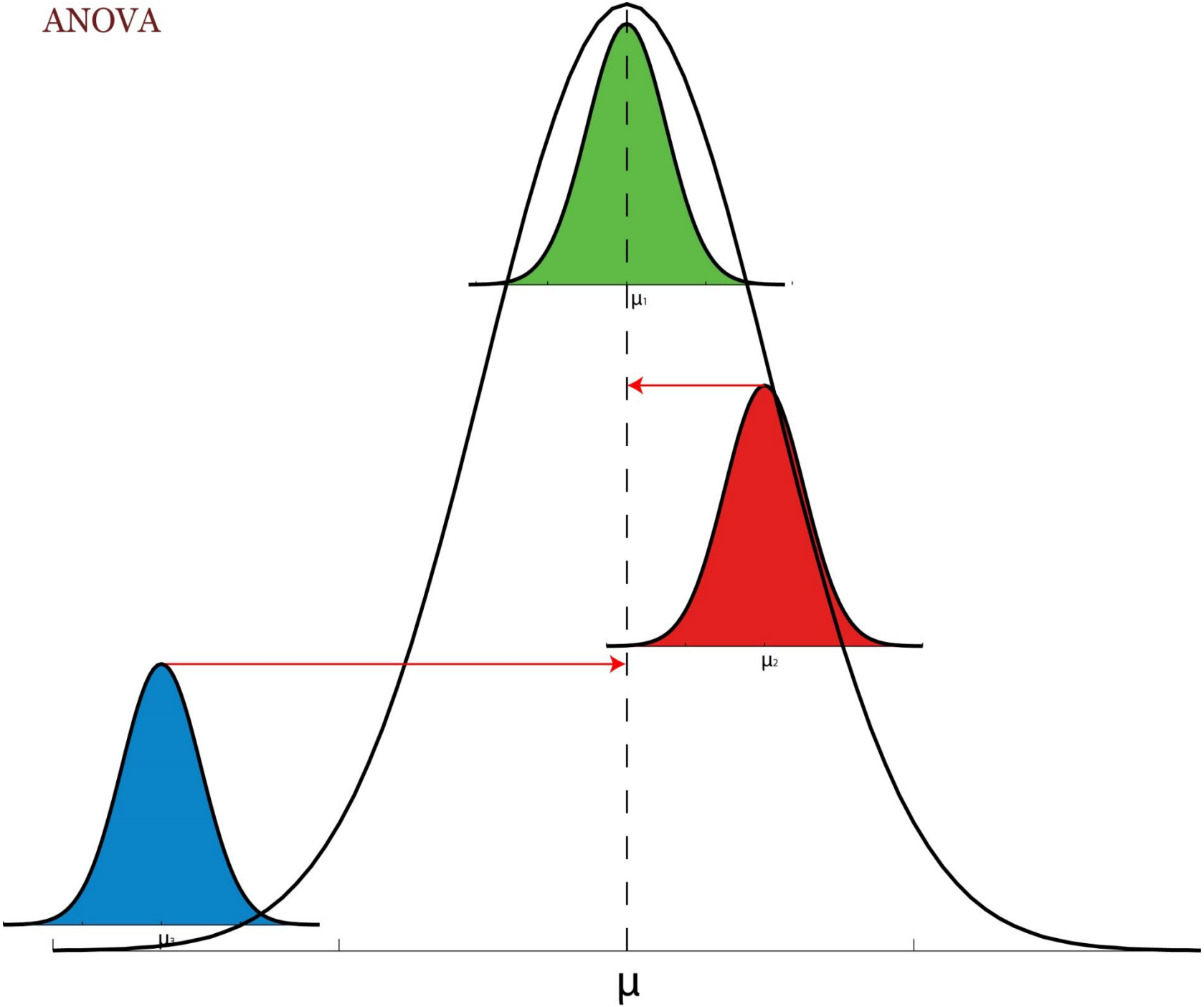
Are all three means so far apart that they likely come from unique population?



ANOVA

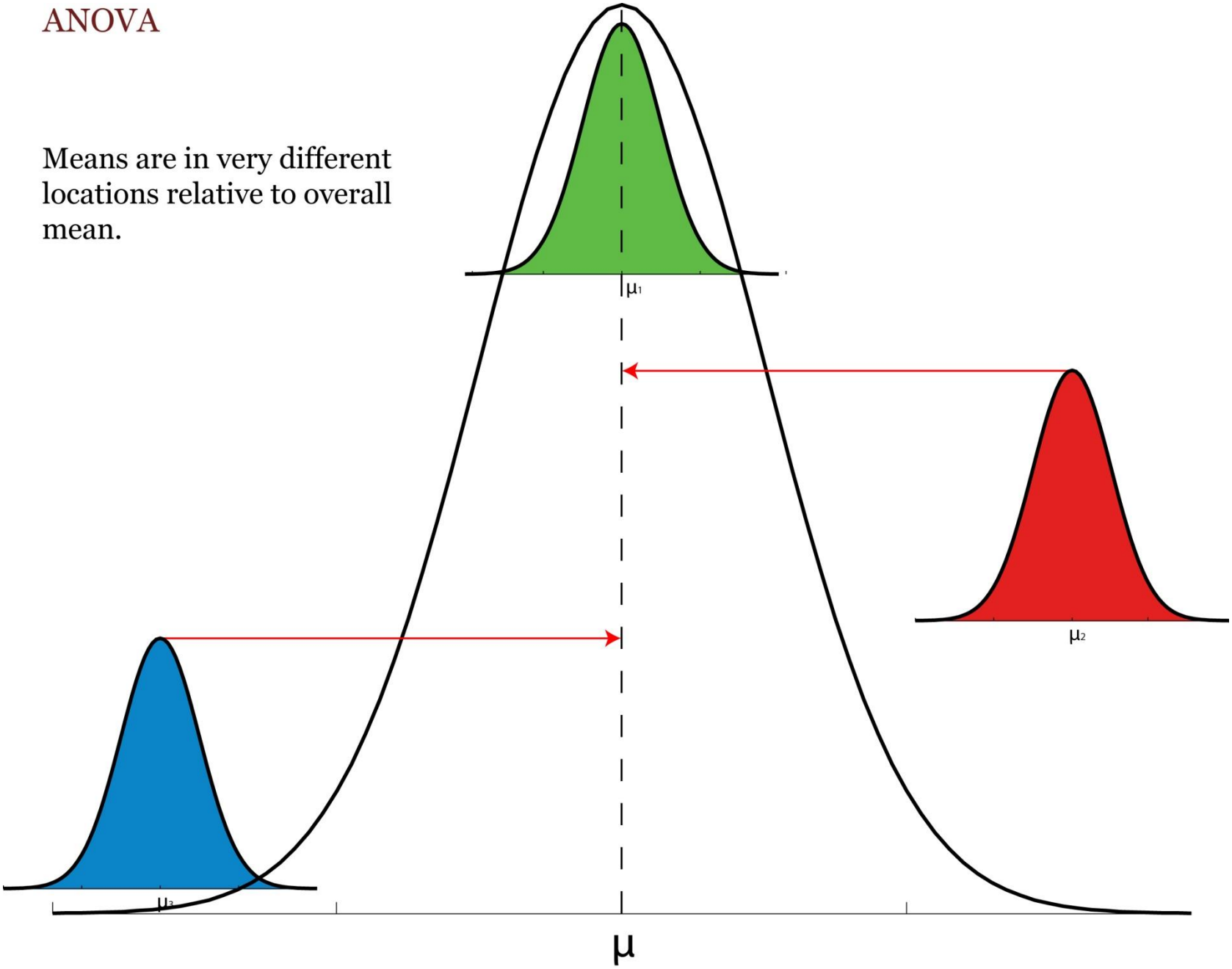


ANOVA



ANOVA

Means are in very different locations relative to overall mean.



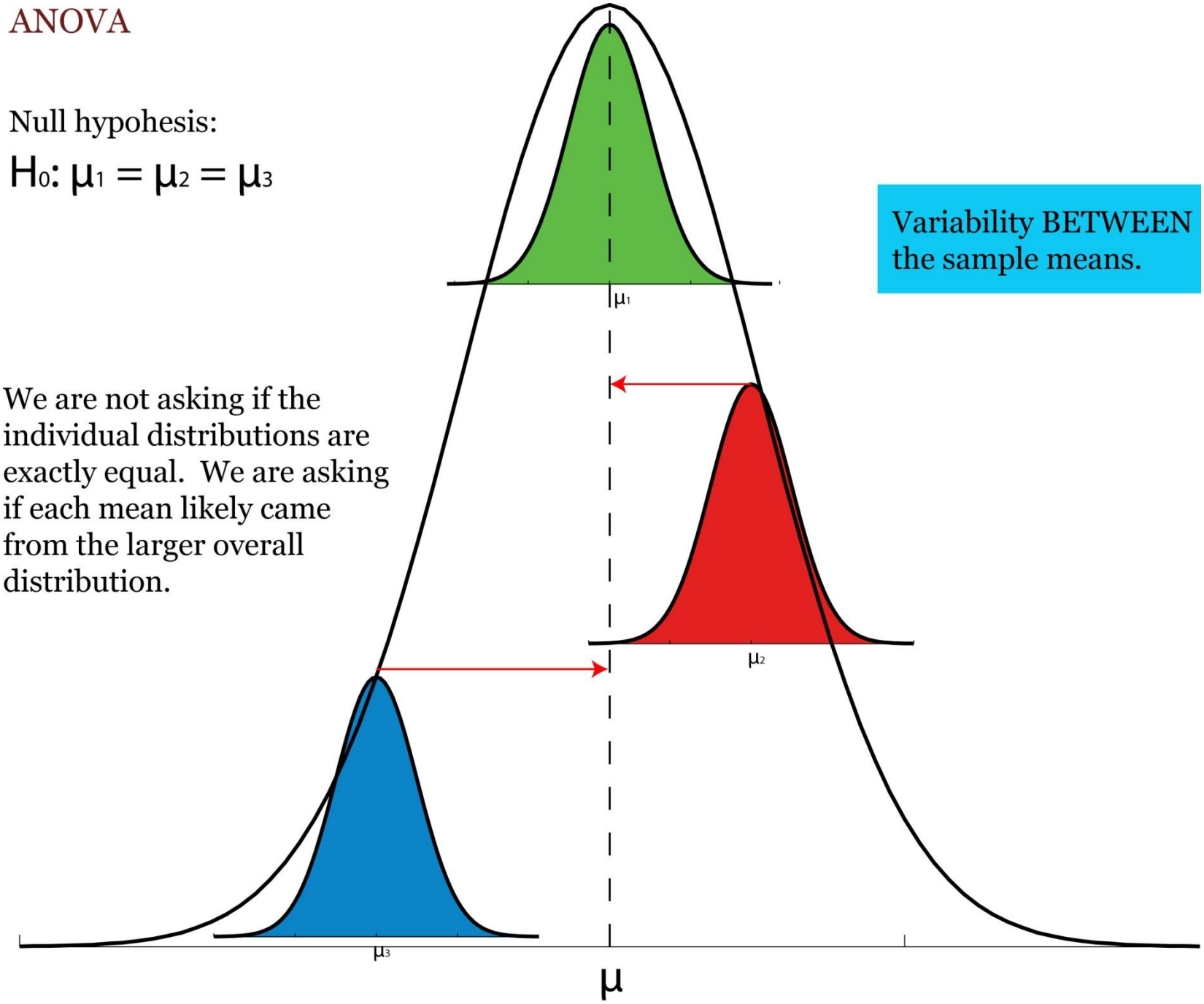
ANOVA

Null hypohesis:

$H_0: \mu_1 = \mu_2 = \mu_3$

Variability BETWEEN the sample means.

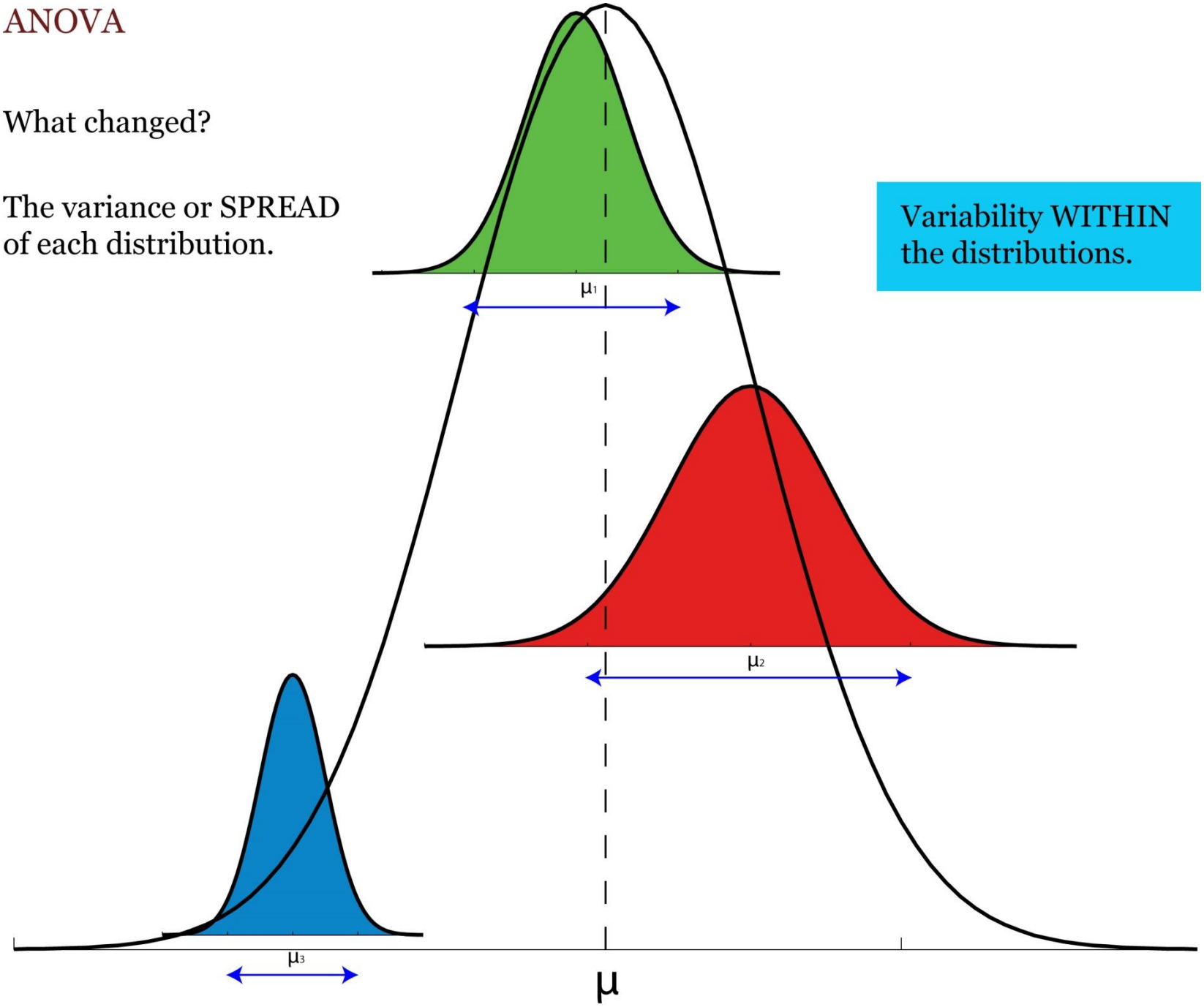
We are not asking if the individual distributions are exactly equal. We are asking if each mean likely came from the larger overall distribution.



ANOVA

What changed?

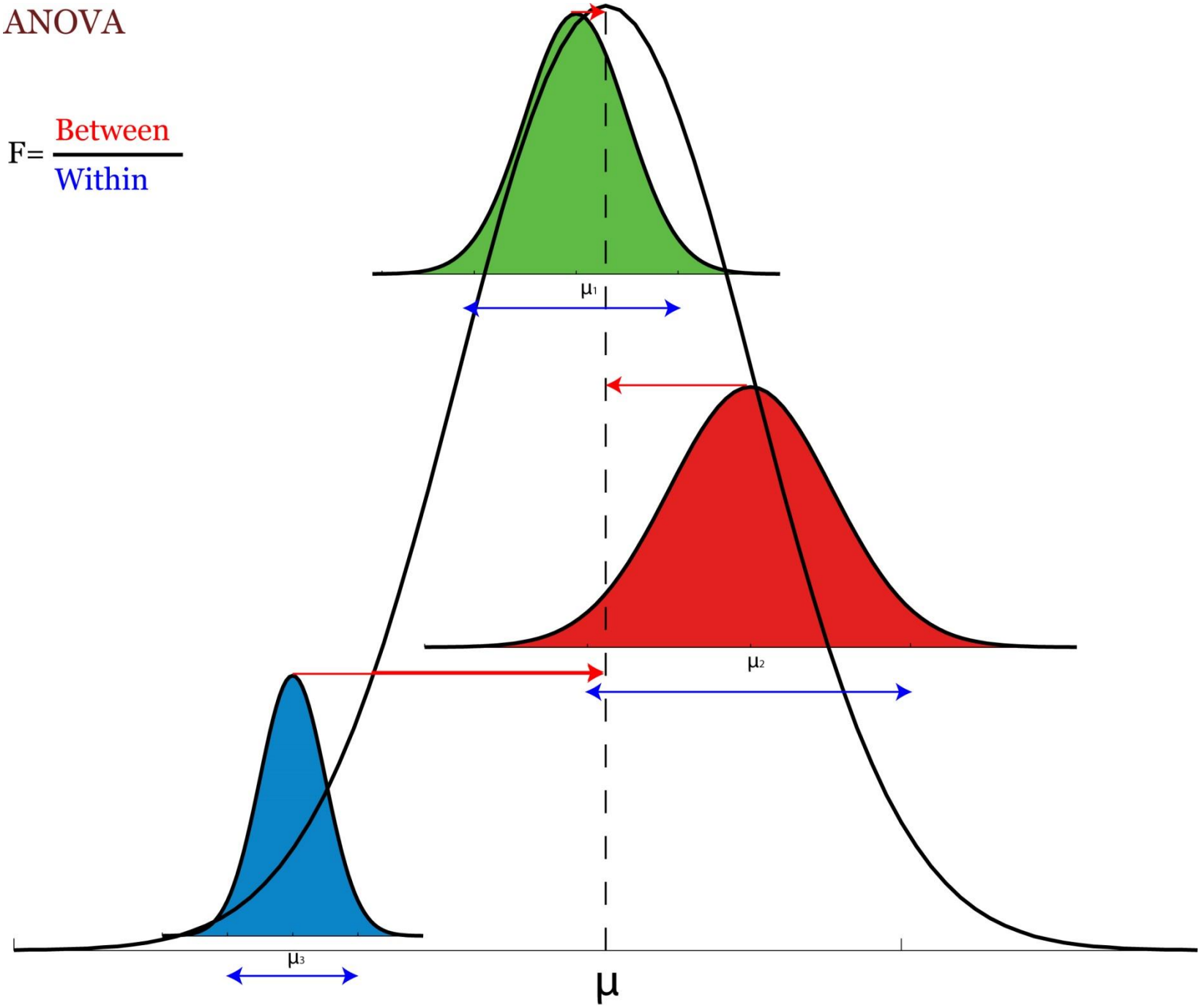
The variance or SPREAD of each distribution.



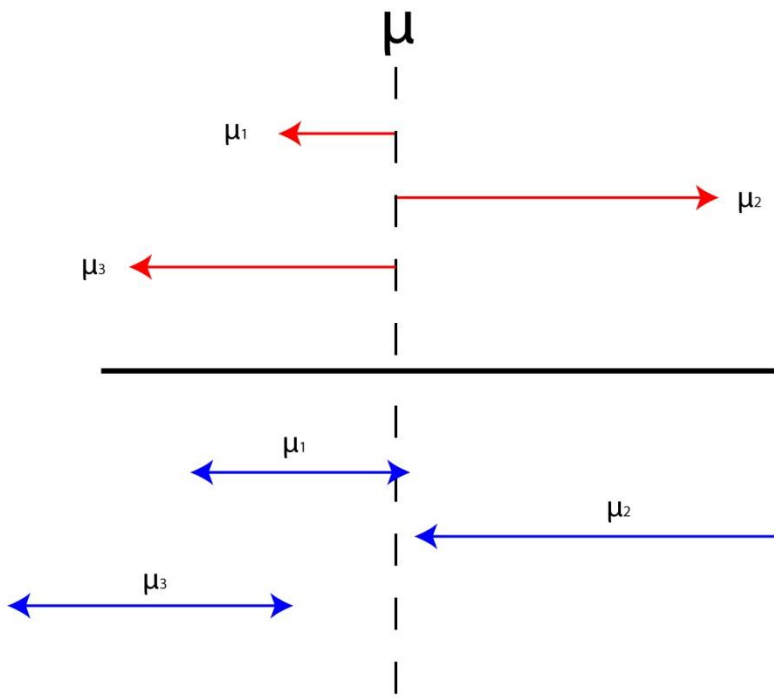
Variability WITHIN the distributions.

ANOVA

$$F = \frac{\text{Between}}{\text{Within}}$$



ANOVA: Analysis of variance is a variability ratio



Variability BETWEEN the means. / Variability WITHIN the distributions.

Distance from overall mean. / Internal spread.

$$= \frac{\text{Variance between}}{\text{Variance within}}$$

$$\frac{\text{LARGE}}{\text{small}} = \text{Reject } H_0$$

At least one mean is an outlier and each distribution is narrow, distinct from each other

$$\frac{\text{Variance between}}{\text{Variance within}}$$

$$\frac{\text{similar}}{\text{similar}} = \text{Fail to reject } H_0$$

Means are fairly close to overall mean and/or distributions overlap, hard to distinguish.

$$\frac{\text{small}}{\text{LARGE}} = \text{Fail to reject } H_0$$

Means are very close to overall mean and/or distribution melt together.

ANOVA: Practical example

The measure of maximum phonation time through sustained vowel phonation well reflects subjects' aerodynamic efficiency of the vocal tract. Normal speaker should be able to perform sustained vowel phonation for more than 15 second. The aim of the research was to verify if there is some impairment of respiratory control across patients with Huntington's disease and Parkinson's disease as compared to healthy control subjects.

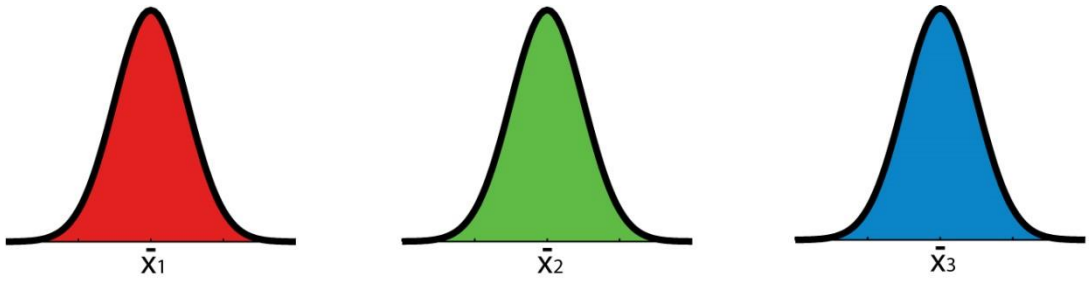
We will conduct this analysis using a One-Way ANOVA technique.

Maximum phonation time

Groups ← Single factor „group“

Huntington's disease	Parkinson's disease	Healthy controls
5	8	29
7	17	22
8	11	18
5	10	18
5	15	13

Random sample within each group.



Huntington's disease	Parkinson's disease	Healthy controls
5	8	29
7	17	22
8	11	18
5	10	18
5	15	13
$\bar{x}_1 = 6$	$\bar{x}_2 = 12$	$\bar{x}_3 = 20$

Overall/Grand Mean:

The mean of all 15 scores taken together.

$\bar{\bar{x}} = 12.67$

ANOVA

- Variance is the average squared deviation (difference) of a data point from the distribution mean
- Take the distance of each data point from the mean, square each distance, add them together, and then find the average
- Instead that we left the part „find the average“ and we receive SUM OF SQUARES (SS)
- SS is variance without finding the average of the sum of the squared deviations

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

← Squared differences

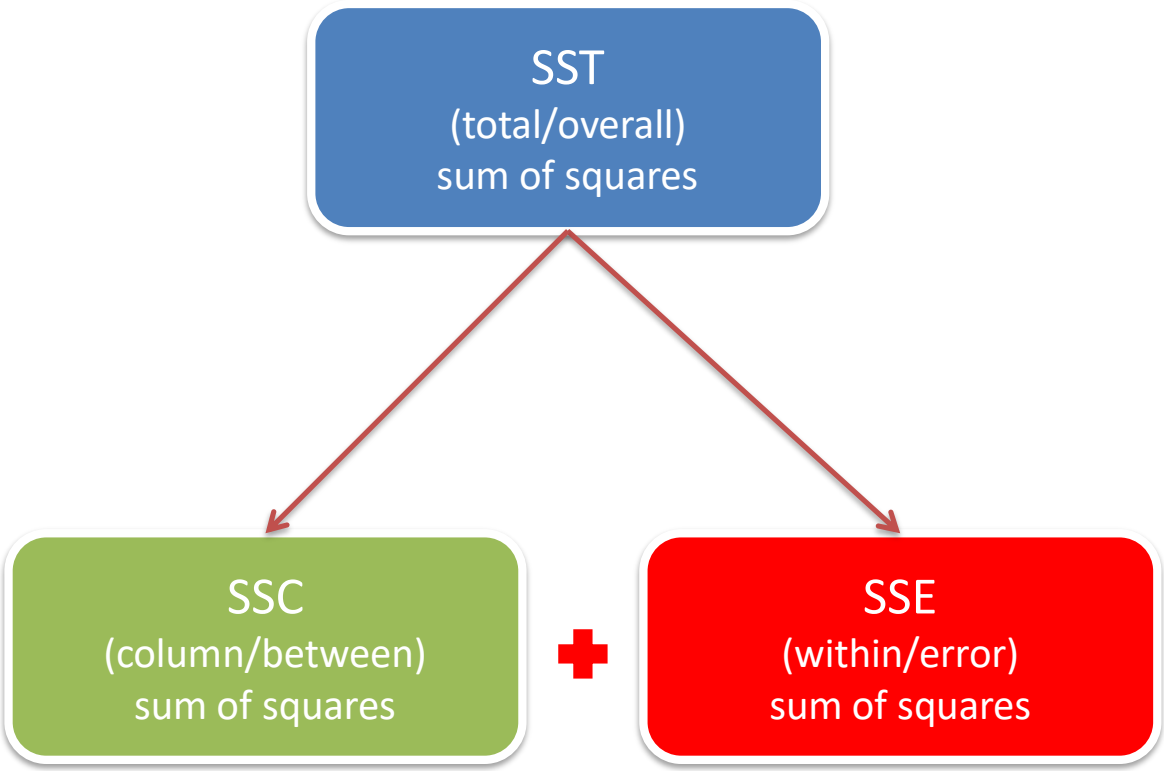
← Averaged

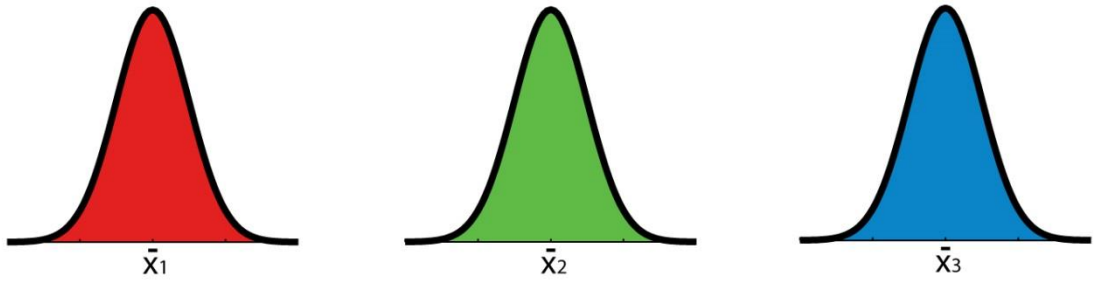
$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sum of Squares

$$SS = \sum(x - \bar{x})^2$$

Sum of Squares





Huntington's disease	Parkinson's disease	Healthy controls
5	8	29
7	17	22
8	11	18
5	10	18
5	15	13
$\bar{x}_1 = 6$	$\bar{x}_2 = 12$	$\bar{x}_3 = 20$

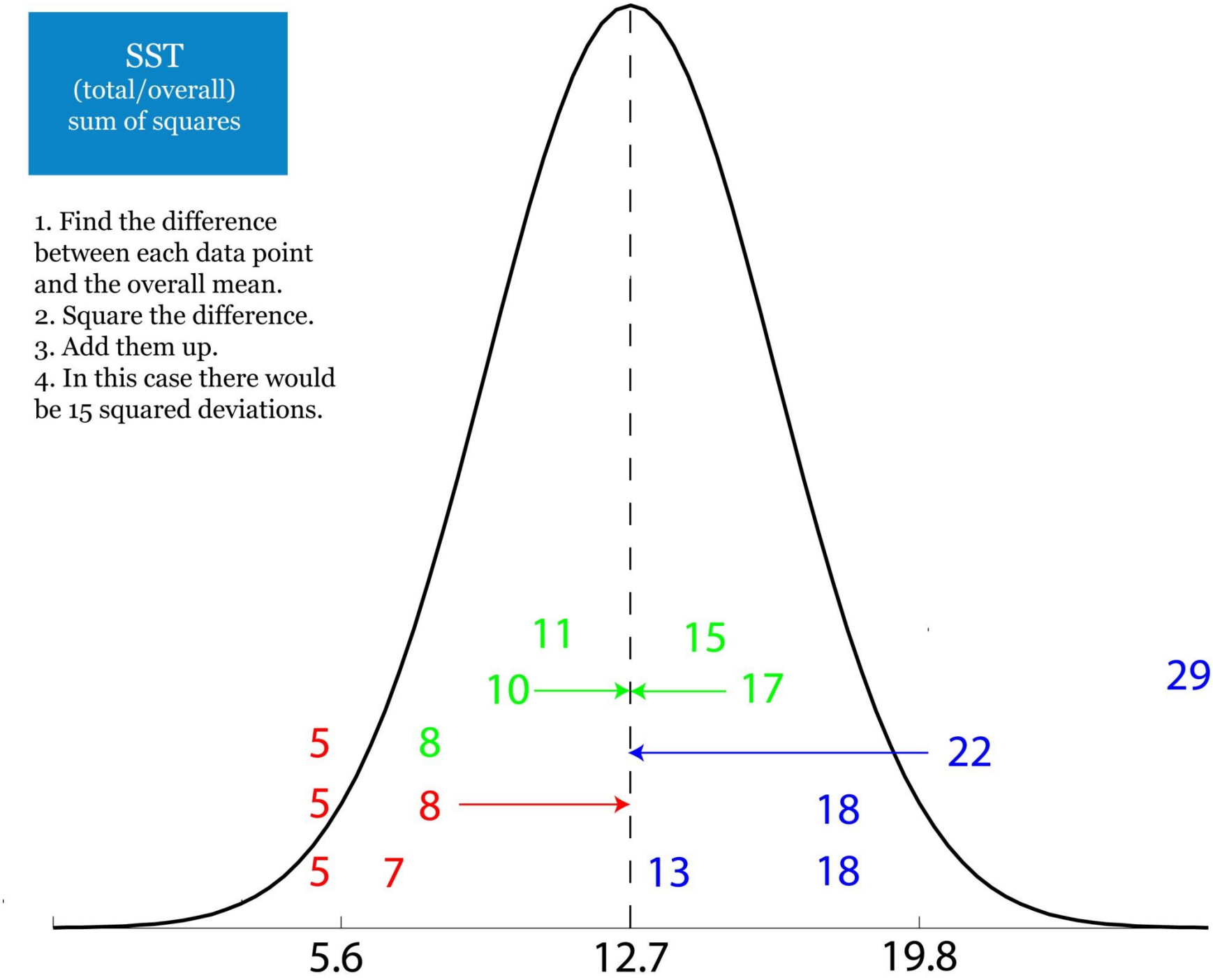
SST
(total/overall)
sum of squares

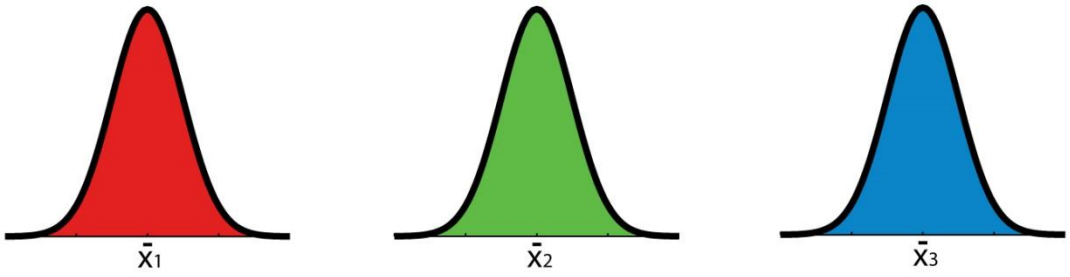
1. Find the difference between each data point and the overall mean.
2. Square the difference.
3. Add them up.

$\bar{\bar{x}} = 12.67$

SST
(total/overall)
sum of squares

1. Find the difference between each data point and the overall mean.
2. Square the difference.
3. Add them up.
4. In this case there would be 15 squared deviations.





Huntington's disease	Parkinson's disease	Healthy controls
5	8	29
7	17	22
8	11	18
5	10	18
5	15	13
$\bar{x}_1 = 6$	$\bar{x}_2 = 12$	$\bar{x}_3 = 20$

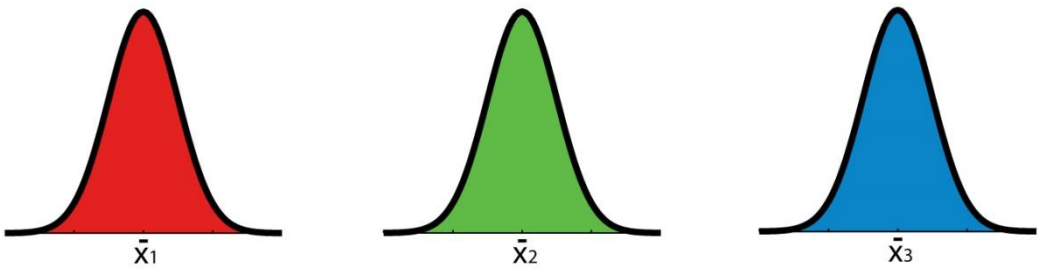
SST
(total/overall)
sum of squares

$$SST = \sum (x - \bar{\bar{x}})^2$$

$$SST = (5 - 12.67)^2 + (5 - 126.7)^2 + \ldots + (13 - 12.67)^2$$

$$SST = 696.9$$

$$\bar{\bar{x}} = 12.67$$



Huntington's disease	Parkinson's disease	Healthy controls
5	8	29
7	17	22
8	11	18
5	10	18
5	15	13
$\bar{x}_1 = 6$	$\bar{x}_2 = 12$	$\bar{x}_3 = 20$

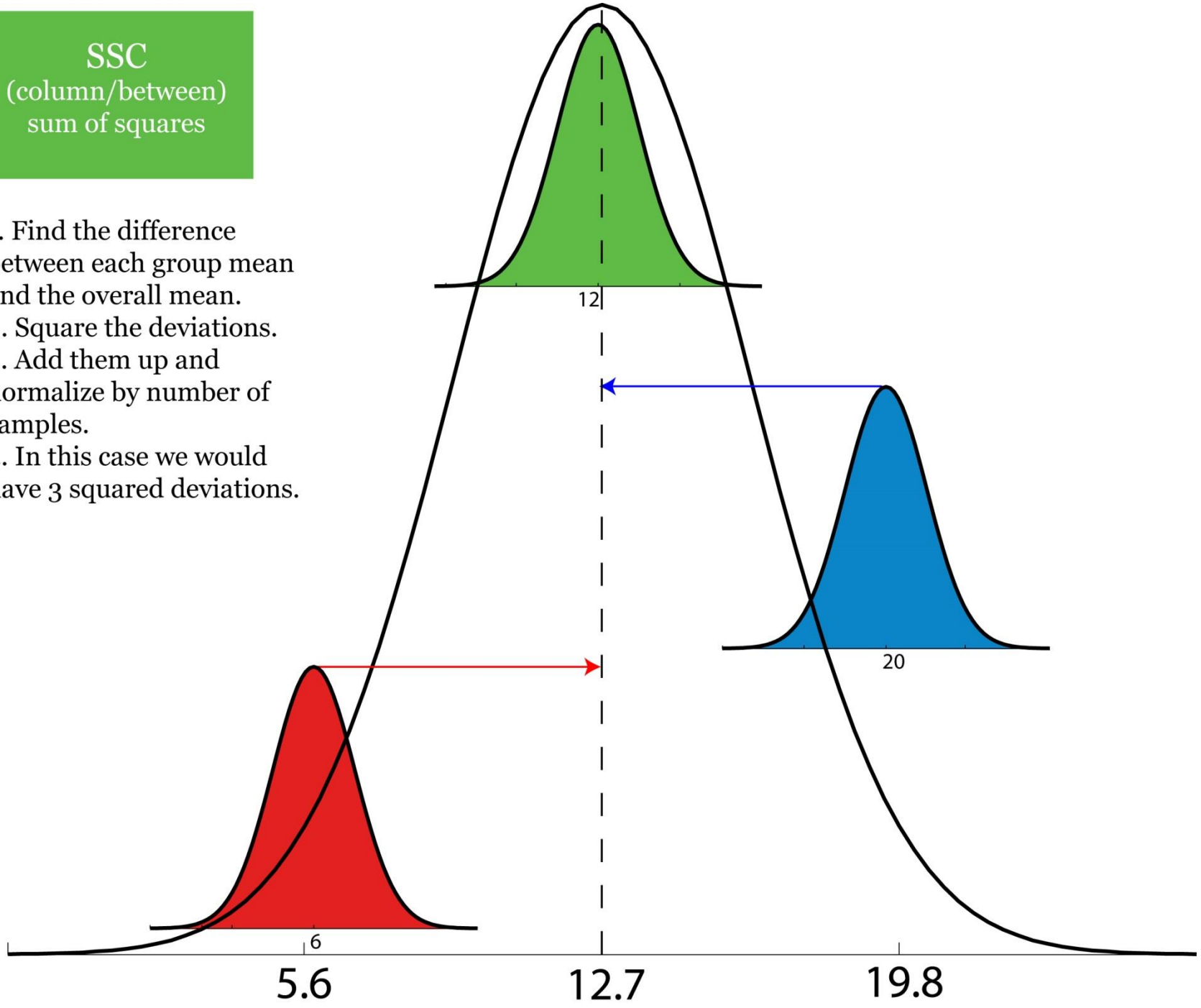
SSC
(column/between)
sum of squares

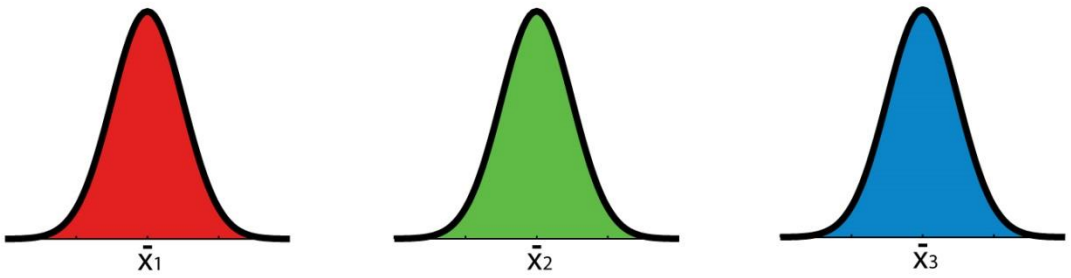
1. Find the difference between each group mean and the overall mean.
2. Square the deviations.
3. Add them up and normalize by number of samples.

$\bar{\bar{x}} = 12.67$

SSC
(column/between)
sum of squares

1. Find the difference between each group mean and the overall mean.
2. Square the deviations.
3. Add them up and normalize by number of samples.
4. In this case we would have 3 squared deviations.





Huntington's disease	Parkinson's disease	Healthy controls
5	8	29
7	17	22
8	11	18
5	10	18
5	15	13
$\bar{x}_1 = 6$	$\bar{x}_2 = 12$	$\bar{x}_3 = 20$

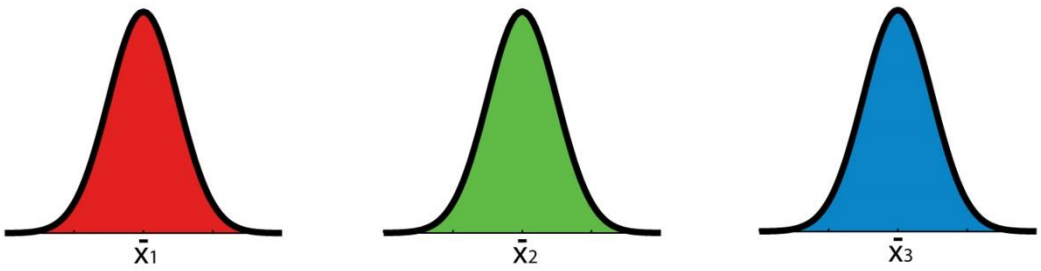
SSC
(column/between)
sum of squares

$$SSC = \sum r(\bar{x} - \bar{\bar{x}})^2$$

$$SSC = [5(6 - 12.67)^2] + [5(12 - 12.67)^2] + [5(20 - 12.67)^2]$$

$$SSC = 492.1$$

$$\bar{\bar{x}} = 12.67$$



Huntington's disease	Parkinson's disease	Healthy controls
5	8	29
7	17	22
8	11	18
5	10	18
5	15	13
$\bar{x}_1 = 6$	$\bar{x}_2 = 12$	$\bar{x}_3 = 20$

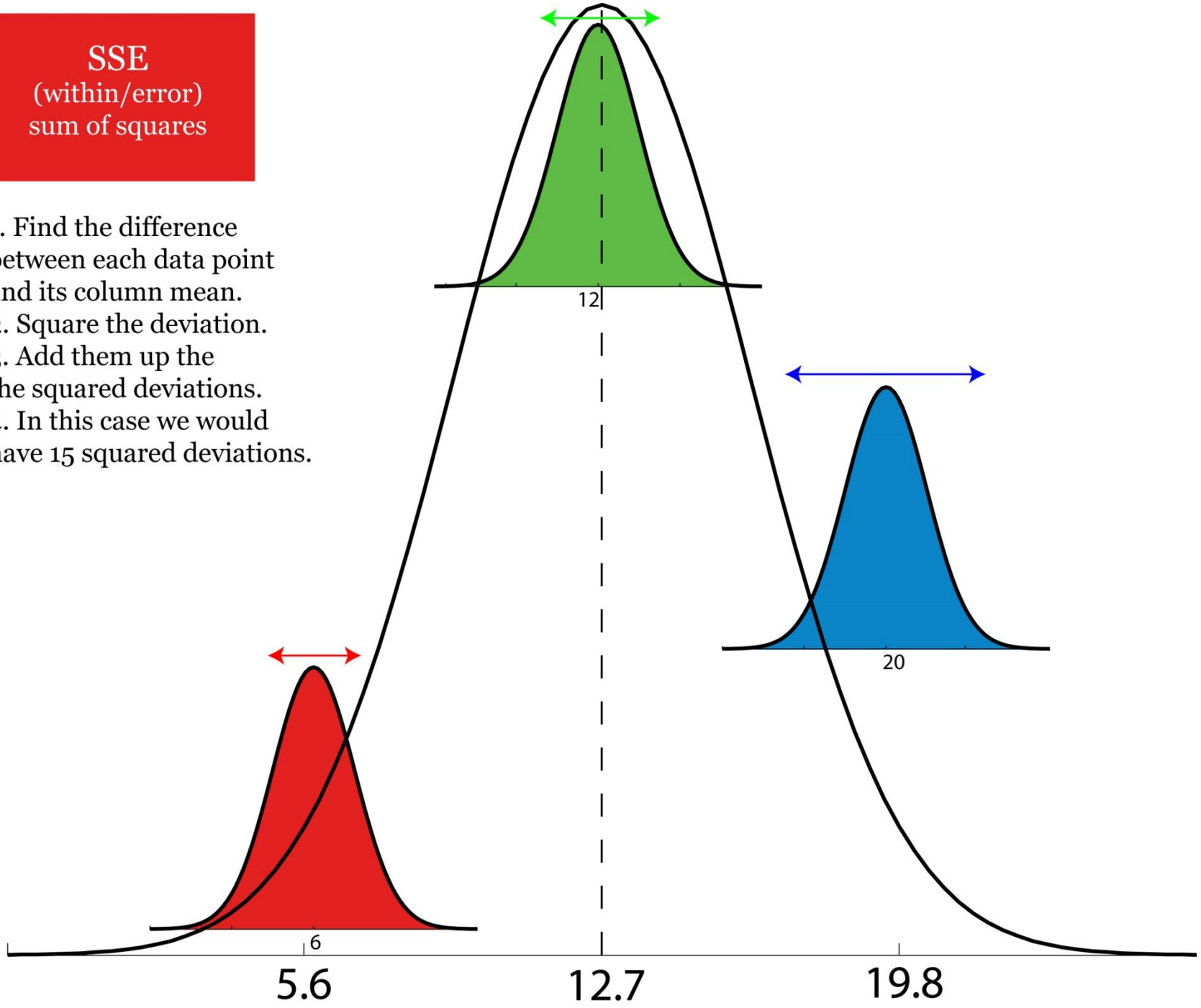
SSE
(within/error)
sum of squares

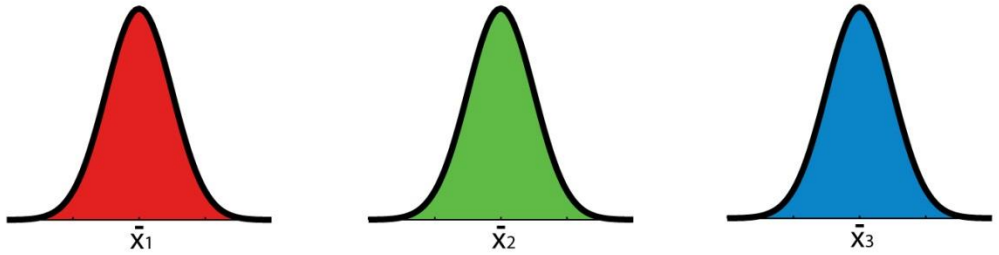
1. Find the difference between each data point and its column mean.
2. Square the deviation.
3. Add them up the the squared deviations.

$\bar{\bar{x}} = 1[illegible]7$

SSE
(within/error)
sum of squares

1. Find the difference between each data point and its column mean.
2. Square the deviation.
3. Add them up the the squared deviations.
4. In this case we would have 15 squared deviations.





Huntington's disease	Parkinson's disease	Healthy controls
5	8	29
7	17	22
8	11	18
5	10	18
5	15	13
$\bar{x}_1 = 6$	$\bar{x}_2 = 12$	$\bar{x}_3 = 20$

SSE
(within/error)
sum of squares

$$SSE = \sum (x - \bar{x})^2$$

$$SSE(HD) = (5-6)^2 + (7-5)^2 + \cdots + (5-6)^2 = 8$$

$$SSE(PD) = (8-12)^2 + (17-12)^2 + \cdots + (15-12)^2 = 54.8$$

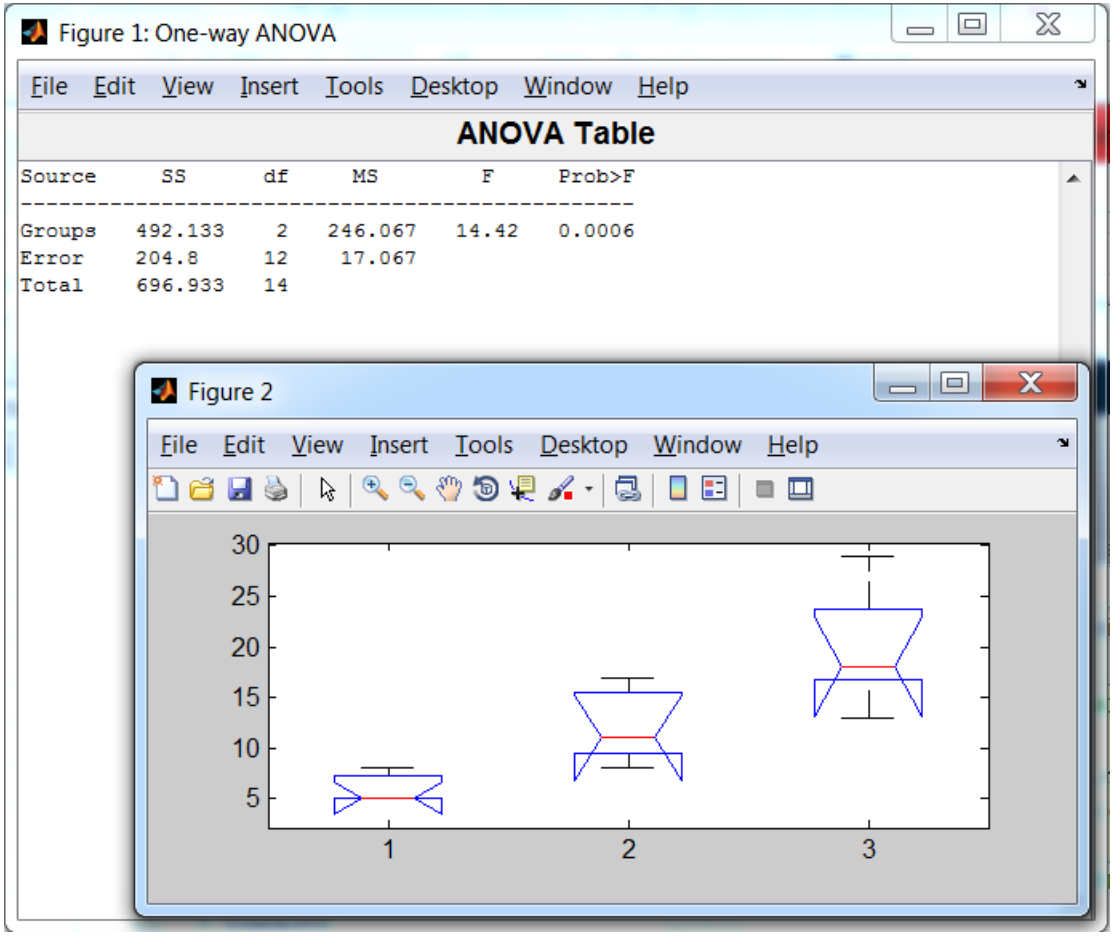
$$SSE(HC) = (29-20)^2 + (22-20)^2 + \cdots + (13-20)^2 = 142$$

$$SSE = SSE(\text{HD}) + \text{SSE(PD)} + \text{SSE(HC)} = 8.0 + 54.8 + 142.0 = 204.8$$

Matlab output to replicate

[p,a,s] = anova1(Sample,Group)

Group	Sample
1	5
1	7
1	8
1	5
1	5
2	8
2	17
2	11
2	10
2	15
3	29
3	22
3	18
3	18
3	13



One-way ANOVA

N = total observations
SSC = Sum of squares (columns/between)
SSE = Sum of squares (within/error)
SST = Sum of squares (total)

r = number of data in each column
C = # columns/treatments
MSC = mean square columns
MSE = mean square error

$$SSC = \sum r(\bar{x} - \bar{\bar{x}})^2$$

$$df_{columns} = C - 1$$

$$MSC = \frac{SSC}{df_{columns}}$$

$$SSE = \sum (x - \bar{x})^2$$

$$df_{error} = N - C$$

$$MSE = \frac{SSE}{df_{error}}$$

$$SST = \sum (x - \bar{\bar{x}})^2$$

$$df_{total} = N - 1$$

$$F = \frac{MSC}{MSE}$$

One-way ANOVA

N = total observations
SSC = Sum of squares (columns/between)
SSE = Sum of squares (within/error)
SST = Sum of squares (total)

r = number of data in each column
C = # columns/treatments
MSC = mean square columns
MSE = mean square error

$SSC = 492.1$

$df_{columns} = 3 - 1 = 2$

$$MSC = \frac{492.1}{2} = 246.1$$

$SSE = 204.8$

$df_{error} = 15 - 3 = 12$

$$MSE = \frac{204.8}{12} = 17.1$$

$SST = SSC + SSE = 696.9$

$df_{total} = 15 - 1 = 14$

$$F = \frac{246.1}{17.1} = 14.4$$

How to express p-value?

p = 1 – fcdf(F, df_columns, df_error);

How to report results of ANOVA?

[$F(df_{columns}, df_{error})$ = F-test, p-value, effect size]
Our case: [$F(2,12) = 14.4$, $p < 0.001$, $\eta^2 = 0.71$]

Approximately: $F \geq 4$ for $p \leq 0.05$

We can rejected H_0!

How to calculate effect size for ANOVA?

$\eta^2 = SSC/SST$
$\eta^2 = 492.1/696.9$
$\eta^2 = 0.71$

Small effect size: $\eta^2 = 0.01$
Medium effect size : $\eta^2 = 0.059$
Large effect size: $\eta^2 = 0.138$

What other information do we need?

Does Huntington's disease significantly differs from Parkinson's disease?

Does Huntington's disease significantly differs from healthy controls?

Does Parkinson's disease significantly differs from healthy controls?

Post-hoc tests

- Used to determine which mean of group of means is/are significantly different from the others
- Depending on research design & research question

Most common methods (also accessible in Matlab)

- Least significant difference (LSD)
- Tukey's honestly significant difference
- Bonferroni
- Scheffe

Least significant difference

- Least conservative method as no adjustment is made for multiple comparisons
- Likely lead to Type I error, especially when performed a lots of comparisons
- Preferred for strongly significant F-test ($p < 0.01$)
- Too liberal, but also have great power

Tukey procedure

- More conservative (and generally more appropriate than LSD)
- Controls the alpha level (less power than LSD)
- Suitable for unequal sample size
- Most commonly applied

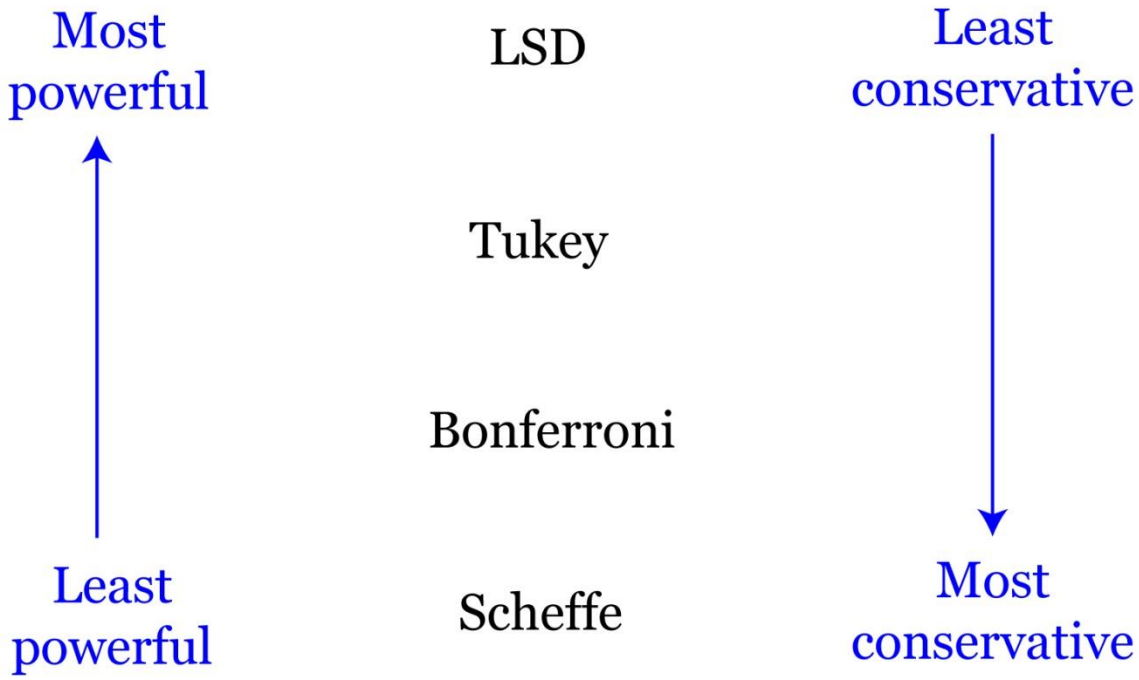
Bonferroni procedure

- Divide alpha by the number of tests
- Sacrifice slightly more power than Tukey but can be applied to any set of comparisons (useful in more situations then Tukey)
- Usually better than Tukey if we want to perform a small number of planned comparisons

Scheffe procedure

- Most conservative (less powerful) of all tests
- Protects against data snooping, i.e. that the researcher decides to perform comparisons *after* looking at the data (as contrasted with *pre-planned* inference, which the researcher plans *before* looking at the data)
- Controls the alpha levels for testing all possible comparisons
- Should be used if you have not planned comparisons in advance

Post-hoc procedure: Usage summary



Use BONFERRONI when interested only in small number of planned comparisons
Use TUKEY when only interested in all (or most) comparisons
Use SCHEFFE when for any unplanned comparisons

Tukey's HSD: Example to our dataset

$$HSD = Q_{\alpha,k,dferror}\sqrt{\frac{MSE}{r}}$$

Q is selected from the table of Studendized range values and k represents number of means/treatments

α = 0.05

k df	2	3	4	5	6	7	8	9	10
1	18.0	27.0	32.8	37.1	40.4	43.1	45.4	47.4	49.1
2	6.08	8.33	9.80	10.88	11.73	12.43	13.03	13.54	13.99
3	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46
4	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47

$$HSD = 3.77\sqrt{\frac{17.1}{5}} = 7$$


vs.

= 12-6 = 6 ✖


vs.

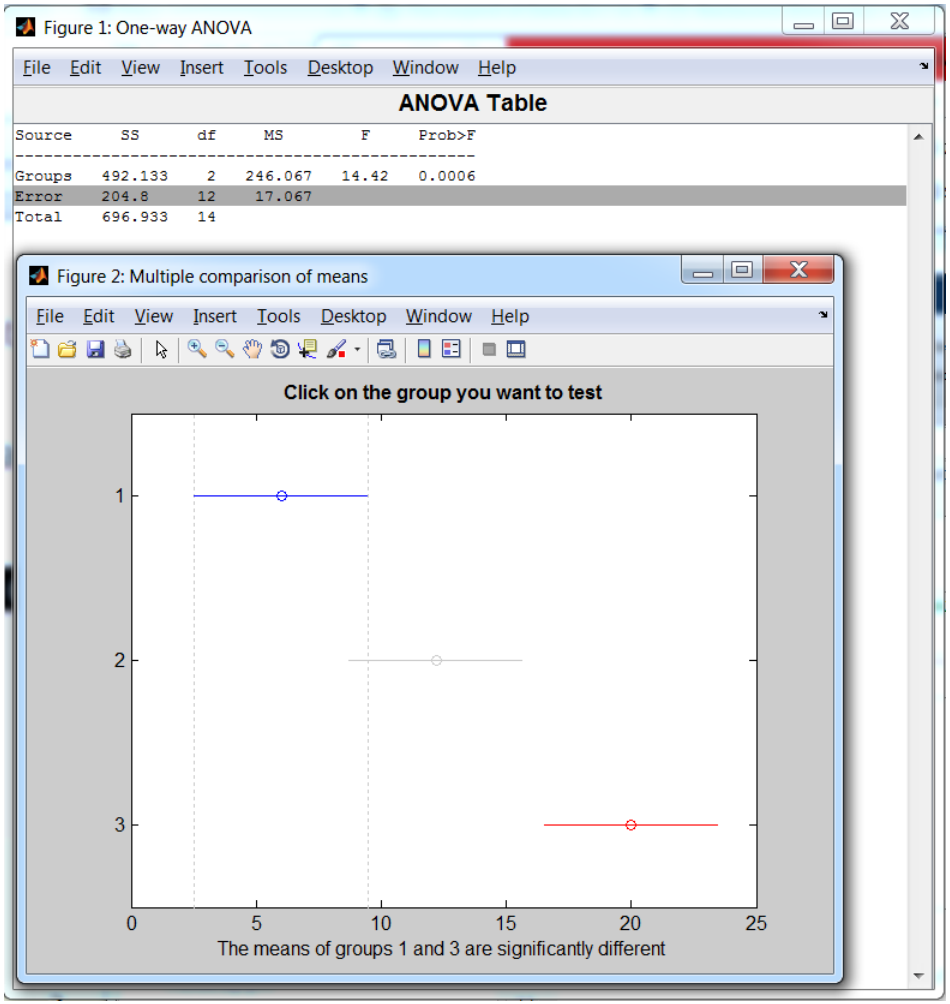
= 20-6 = 14 **OK**


vs.

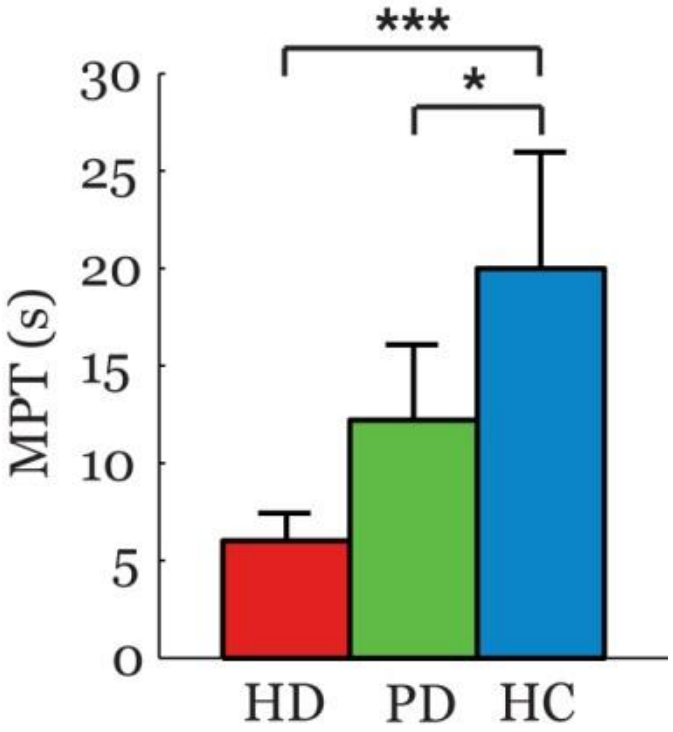
= 20-12 = 8 **OK**

Tukey's HSD: Matlab

```
[p,a,s] = anova1(Sample,Group)
[c,m,h,nms] = multcompare(s,'ctype','hsd','alpha',0.05);
```



Interpretation of investigated data



Using ANOVA, we found statistically significant Differences for MPT between all three groups [$F(2,12) = 14.4$, $p < 0.001$, $\eta^2 = 0.71$].

Post-hoc HSD comparison indicates statistically significant differences between HD and HC ($p < 0.001$) as well as PD and HC ($p < 0.05$).

Captions: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.