Co-financed by the Connecting Europe Facility of the European Union

ANOVA Analysis

Mod B Data Driven Healthcare
Prof. Paola Cerchiello

Chapter Goals

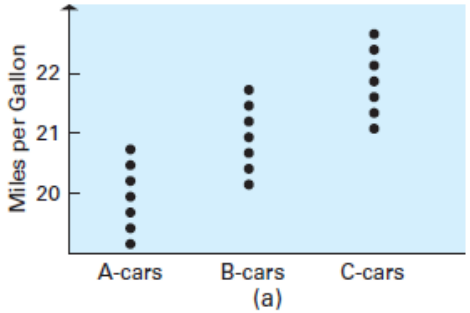
After completing this chapter, you should be able to:

- Recognize situations in which to use analysis of variance
- Understand different analysis of variance designs
- Perform a one-way and two-way analysis of variance and interpret the results

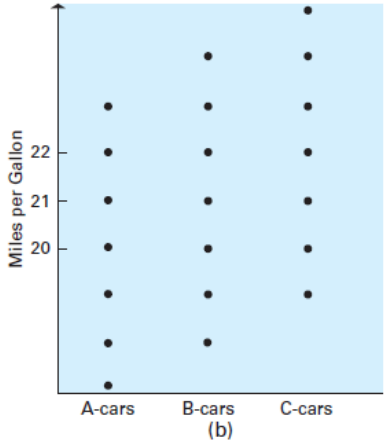
Comparison of Several Population Means (1 of 2)

- The null hypothesis is that the population means are all the same
- The critical factor is the variability involved in the data
  - If the variability around the sample means is small compared with the variability among the sample means, we reject the null hypothesis

Comparison of Several Population Means (2 of 2)



- Small variation around the sample means compared to the variation among the sample means



- Large variation around the sample means compared to the variation among the sample means

One-Way Analysis of Variance

- Evaluate the difference among the means of three or more groups

  Examples: Average production for 1st, 2nd, and 3rd shifts Expected mileage for five brands of tires

- Assumptions
  - Populations are normally distributed
  - Populations have equal variances
  - Samples are randomly and independently drawn

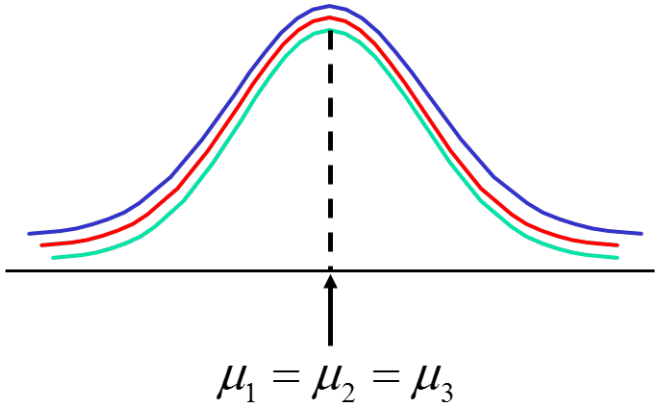
Hypotheses of One-Way ANOVA

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$
  - All population means are equal
  - i.e., no variation in means between groups
- $H_1 : \mu_i \neq \mu_j$ for at least one *i* , *j* pair.
  - At least one population mean is different
  - i.e., there is variation between groups
  - Does not mean that all population means are different (some pairs may be the same)

One-Way ANOVA (1 of 2)

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

H_1 : Not all μ_i are the same



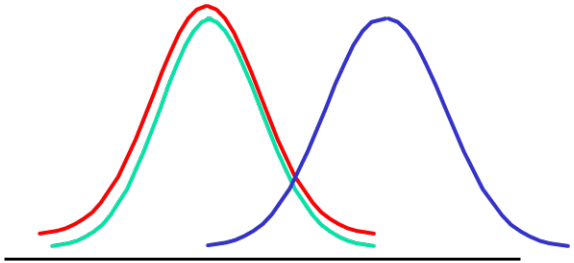
All Means are the same:
The Null Hypothesis is True
(No variation between groups)

One-Way ANOVA (2 of 2)

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

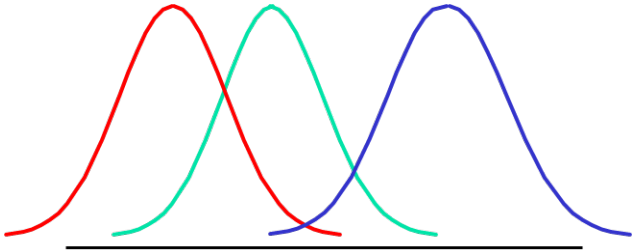
H_1 : Not all μ_i are the same.

At least one mean is different:
The Null Hypothesis is Not true
(Variation is present between groups)



$\mu_1 = \mu_2 \neq \mu_3$

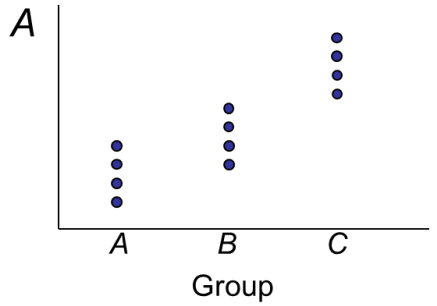
or



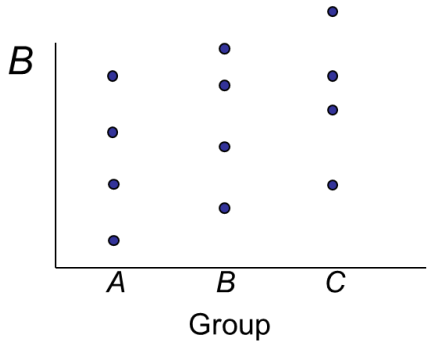
$\mu_1 \neq \mu_2 \neq \mu_3$

Variability

- The variability of the data is key factor to test the equality of means
- In each case below, the means may look different, but a large variation within groups in *B* makes the evidence that the means are different weak



Small variation within groups



Large variation within groups

Sum of Squares Decomposition (1 of 2)

- Total variation can be split into two parts:

$$\mathrm{SST} = \mathrm{SSW} + \mathrm{SSG}$$

SST = Total Sum of Squares

Total Variation = the aggregate dispersion of the individual data values across the various groups

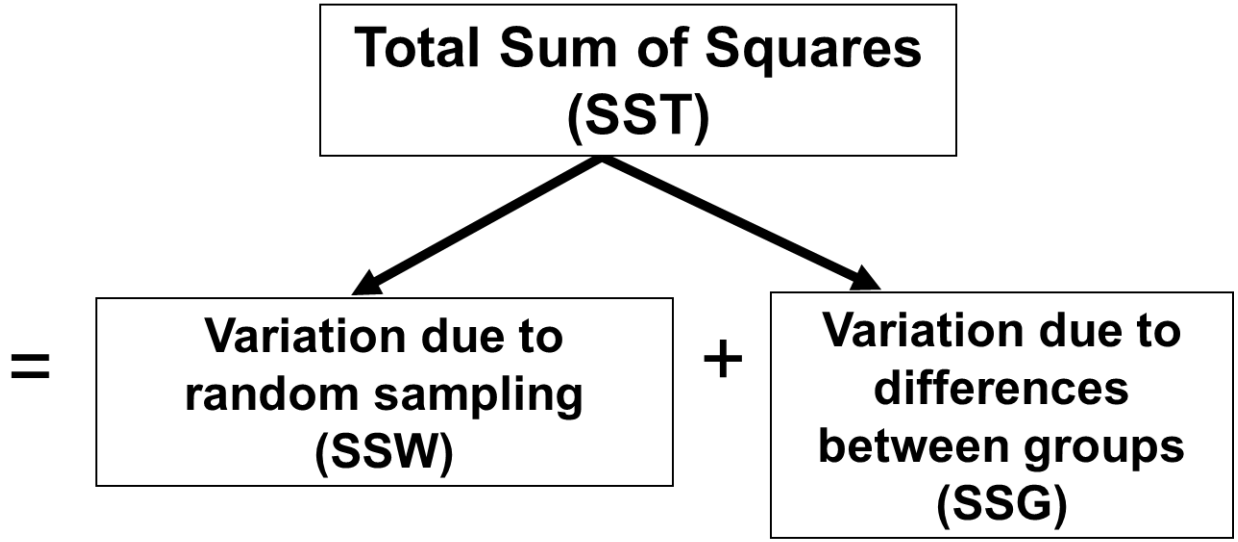
SSW = Sum of Squares Within Groups

Within-Group Variation = dispersion that exists among the data values within a particular group

SSG = Sum of Squares Between Groups

Between-Group Variation = dispersion between the group sample means

Sum of Squares Decomposition (2 of 2)



Total Sum of Squares (1 of 2)

$$SST = SSW + SSG$$

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{\overline{x}} \right)^2$$

Where:

SST = Total sum of squares

K = number of groups (levels or treatments)

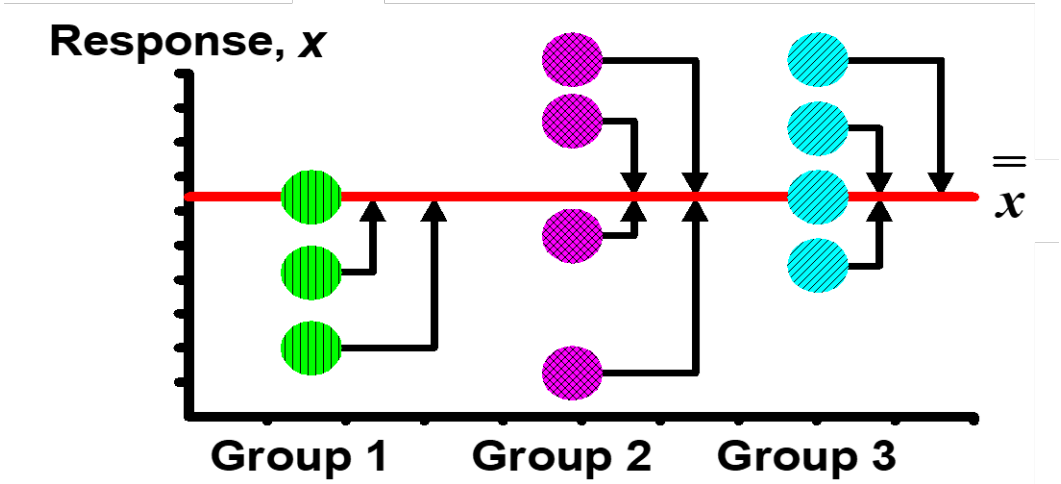
n_i = number of observations in group *i*

$x_{ij} = j^{th}$ observation from group *i*

$\overline{\overline{x}}$ = overall sample mean

Total Sum of Squares (2 of 2)

$$SST = \left(x_{11} - \bar{\bar{x}}\right)^2 + \left(x_{12} - \bar{\bar{x}}\right)^2 + \cdots + \left(x_{Kn_K} - \bar{\bar{x}}\right)^2$$



Within-Group Variation (1 of 3)

$$\text{SST} = \text{SSW} + \text{SSG}$$

$$\text{SSW} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}_i \right)^2$$

Where:

SSW = Sum of squares within groups

K = number of groups

n_i = sample size from group i

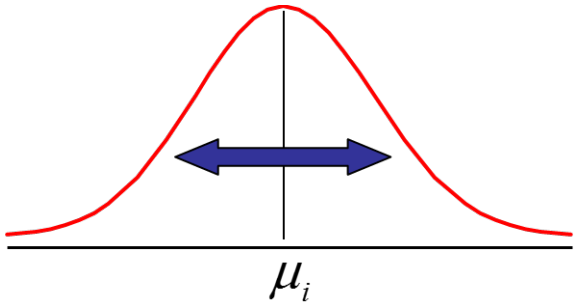
$\overline{x}_i$ = sample mean from group i

$x_{ij} = j^{\text{th}}$ observation in group i

Within-Group Variation (2 of 3)

$$\text{SSW} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i \right)^2$$

Summing the variation within each group and then adding over all groups

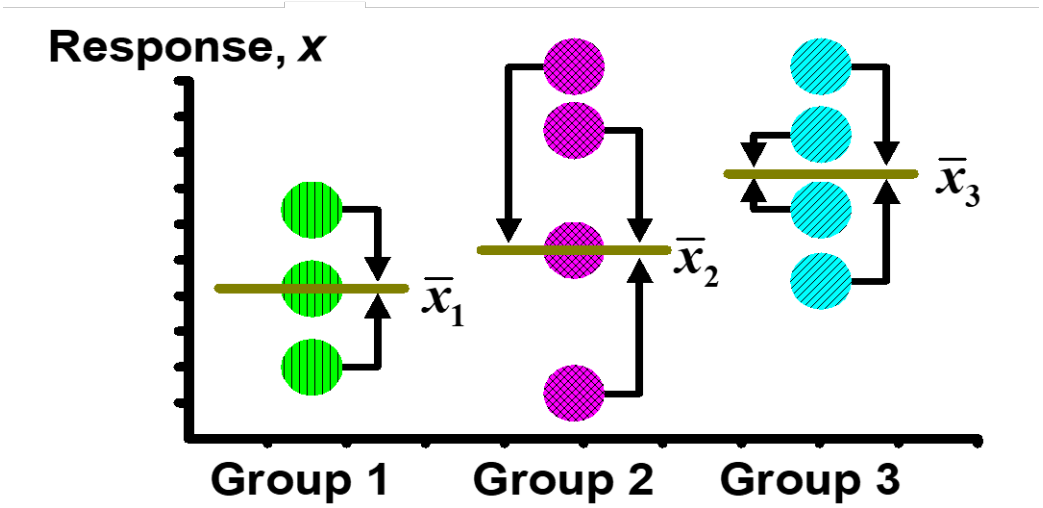


$$\text{MSW} = \frac{\text{SSW}}{n - K}$$

$$\text{Mean Square Within} = \frac{\text{SSW}}{\text{degrees of freedom}}$$

Within-Group Variation (3 of 3)

$$\text{SSW} = \left(x_{11} - \overline{x}_1\right)^2 + \left(x_{12} - \overline{x}_1\right)^2 + \ldots + \left(x_{Kn_K} - \overline{x}_K\right)^2$$



Between-Group Variation (1 of 3)

$$SST = SSW + SSG$$

$$SSG = \sum_{i=1}^{K} n_i \left(\overline{x}_i - \overline{\overline{x}} \right)^2$$

Where:

SSG = Sum of squares between groups

K = number of groups

n_i = sample size from group i

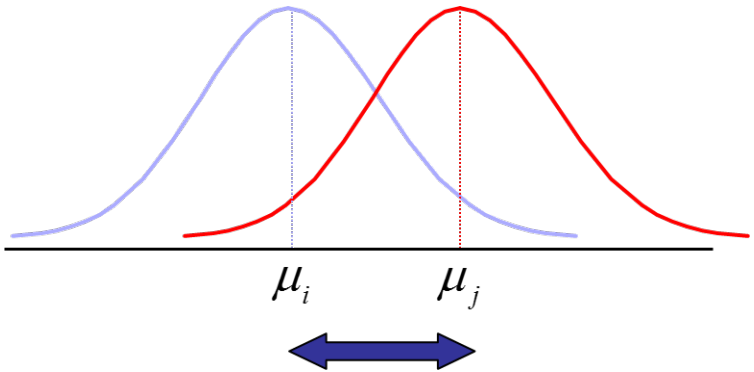
$\overline{x}_i$ = sample mean from group i

$\overline{\overline{x}}$ = grand mean (mean of all data values)

Between-Group Variation (2 of 3)

$$\text{SSG} = \sum_{i=1}^{K} n_i \left(\bar{x}_i - \bar{\bar{x}} \right)^2$$

Variation Due to Differences Between Groups



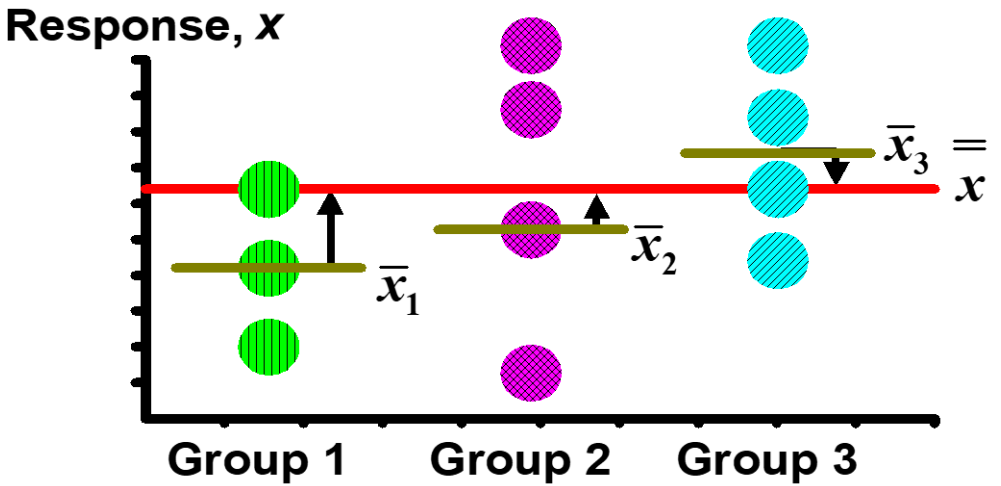
$$\text{MSG} = \frac{\text{SSG}}{K-1}$$

Mean Square Between Groups

$$= \frac{\text{SSG}}{\text{degrees of freedom}}$$

Between-Group Variation (3 of 3)

$$\text{SSG} = n_1\left(\bar{x}_1 - \bar{\bar{x}}\right)^2 + n_2\left(\bar{x}_2 - \bar{\bar{x}}\right)^2 + \ldots + n_K\left(\bar{x}_K - \bar{\bar{x}}\right)^2$$



Obtaining the Mean Squares

$$\text{MST} = \frac{\text{SST}}{n-1}$$

$$\text{MSW} = \frac{\text{SSW}}{n-K}$$

$$\text{MSG} = \frac{\text{SSG}}{K-1}$$

Where

n = sum of the sample sizes from all groups

K = number of populations

One-Way ANOVA Table

Source of Variation	SS	df	MS (Variance)	*F* ratio
Between Groups	SSG	$K - 1$	$\text{MSG} = \frac{\text{SSG}}{K - 1}$	$F = \frac{\text{MSG}}{\text{MSW}}$
Within Groups	SSW	$n - K$	$\text{MSW} = \frac{\text{SSW}}{n - K}$	
Total	$\text{SST} = \text{SSG} + \text{SSW}$	$n - 1$		

K = number of groups

n = sum of the sample sizes from all groups

df = degrees of freedom

One-Factor ANOVA *F* Test Statistic

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_K$$

H_1 : At least two population means are different

- Test statistic

$$F = \frac{\text{MSG}}{\text{MSW}}$$

MSG is mean squares between variances

MSW is mean squares within variances

- Degrees of freedom

$$\text{df}_1 = K - 1$$ (*K* = number of groups)

$$\text{df}_2 = n - K$$ (*n* = sum of sample sizes from all groups)

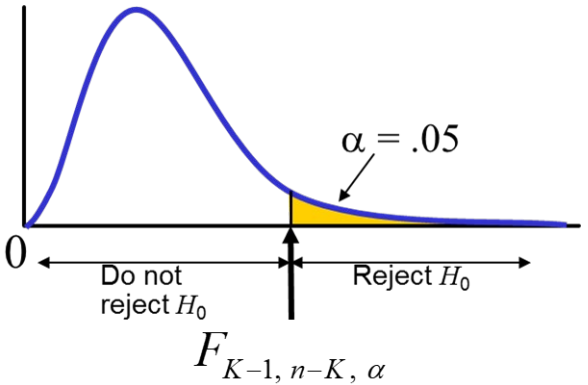
Interpreting the *F* Statistic

- The *F* statistic is the ratio of the between estimate of variance and the within estimate of variance
  - The ratio must always be positive
  - $df_1 = K - 1$ will typically be small
  - $df_2 = n - K$ will typically be large

Decision Rule:

- Reject H_0

$$F > F_{K-1,\, n-K,\, \alpha}$$



One-Factor ANOVA *F* Test Example

You want to see if three different patients yield different blood measurement. You randomly select five measurements from a laboratory.

At the .05 significance level, is there a difference in mean value?

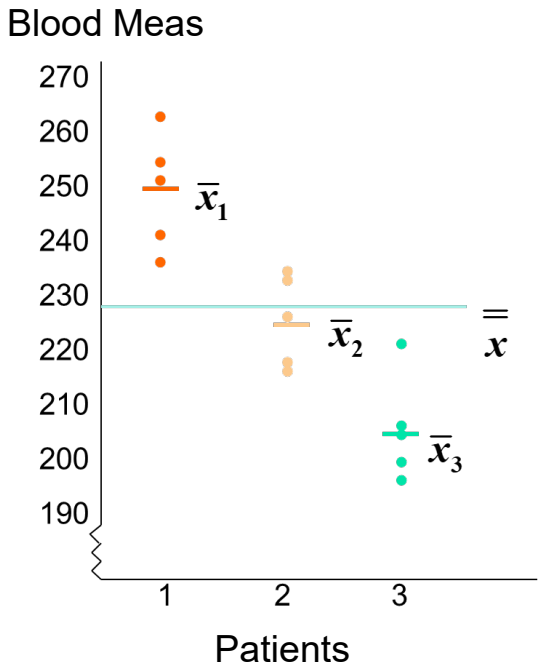
Patient 1	Patient 2	Patient 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

One-Factor ANOVA Example: Scatter Diagram

Patient 1	Patient 2	Patient 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2 \quad \bar{x}_2 = 226.0 \quad \bar{x}_3 = 205.8$

$\bar{\bar{x}} = 227.0$



One-Factor ANOVA Example Computations

Patient 1	Patient 2	Patient 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2 \qquad n_1 = 5$

$\bar{x}_2 = 226.0 \qquad n_2 = 5$

$\bar{x}_3 = 205.8 \qquad n_3 = 5$

$\bar{\bar{x}} = 227.0 \qquad n = 15$

$k = 3$

$$\text{SSG} = 5(249.2 - 227)^2 + 5(226 - 227)^2 + 5(205.8 - 227)^2 = 4716.4$$

$$\text{SSW} = (254 - 249.2)^2 + (263 - 249.2)^2 + \ldots + (204 - 205.8)^2 = 1119.6$$

$$\text{MSG} = \frac{4716.4}{(3-1)} = 2358.2$$

$$\text{MSW} = \frac{1119.6}{(15-3)} = 93.3$$

$$F = \frac{2358.2}{93.3} = 25.275$$

One-Factor ANOVA Example Solution

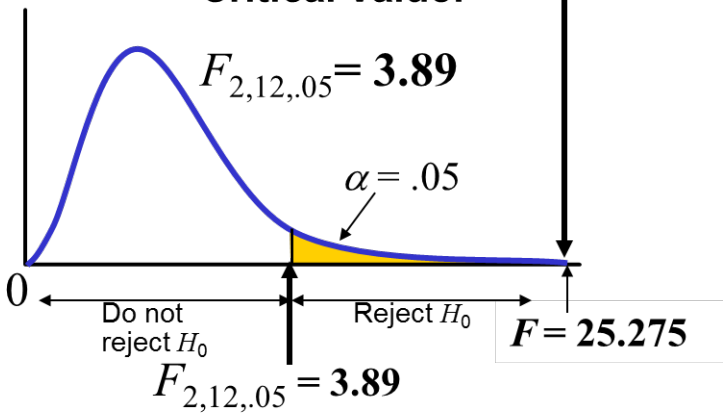
$H_0 : \mu_1 = \mu_2 = \mu_3$

$H_1 : \mu_i$ not all equal

$\alpha = .05$

$df_1 = 2 \qquad df_2 = 12$

Critical Value:



Test Statistic:

$$F = \frac{\text{MSA}}{\text{MSW}} = \frac{2358.2}{93.3} = 25.275$$

Decision:

Reject H_0 at $\alpha = 0.05$

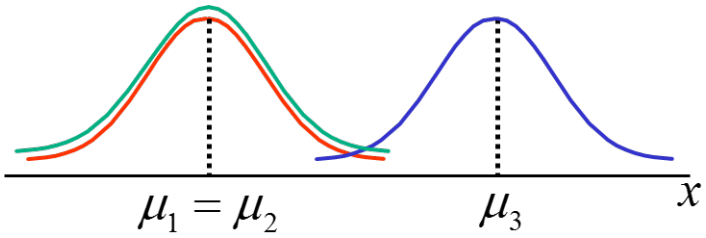
Conclusion:

There is evidence that at least one

μ_i differs from the rest

Multiple Comparisons Between Subgroup Means

- To test which population means are significantly different
  - e.g.: $\mu_1 = \mu_2 \neq \mu_3$
  - Done after rejection of equal means in single factor ANOVA design
- Allows pair-wise comparisons
  - Compare absolute mean differences with critical range



Two Subgroups

- When there are only two subgroups, compute the minimum significant difference (MSD)

$$\text{MSD} = t_{\frac{\alpha}{2}} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Where S_p is a pooled estimate of the variance

Multiple Subgroups (1 of 2) (procedure due to Tukey)

- The minimum significant difference between *k* subgroups is

$$\mathrm{MSD}(K) = Q\frac{s_p}{\sqrt{n}} \quad \text{where} \quad s_p = \sqrt{\mathrm{MSW}}$$

  - *Q* is a factor available in the relative tables for the chosen level of α
  - *K* = number of subgroups, and
    - MSW = Mean square within from ANOVA table

Multiple Subgroups (2 of 2)

$$\mathrm{MSD}(K) = Q\frac{s_p}{\sqrt{n}}$$

Compare:

$\left|\bar{x}_1 - \bar{x}_2\right|$

$\left|\bar{x}_1 - \bar{x}_3\right|$

$\left|\bar{x}_2 - \bar{x}_3\right|$

etc...

$$\left|\bar{x}_i - \bar{x}_j\right| > \mathrm{MSD}(K)?$$

If the absolute mean difference is greater than MSD then there is a significant difference between that pair of means at the chosen level of significance.

Multiple Subgroups: Example

$\bar{x}_1 = 249.2 \qquad n_1 = 5$

$\bar{x}_2 = 226.0 \qquad n_2 = 5$

$\bar{x}_3 = 205.8 \qquad n_3 = 5$

$$\text{MSD}(K) = Q\frac{s_p}{\sqrt{n}} = 3.77\frac{\sqrt{93.3}}{\sqrt{15}} = 9.387$$

(where $Q = 3.77$ is from Table 13 for $\alpha = .05$ and 12 df)

$|\bar{x}_1 - \bar{x}_2| = 23.2$

$|\bar{x}_1 - \bar{x}_3| = 43.4$

$|\bar{x}_2 - \bar{x}_3| = 20.2$

Since each difference is greater than 9.387, we conclude that all three means are different from one another at the .05 level of significance.

Alternatives to MSD Approach

- Another MSD for pairwise differences of means is based on the Bonferroni Inequality, where the MSD for groups k and l is

$$MSD_{kl} = t_{n-K,\frac{\alpha}{K(K-1)}} s_p \sqrt{\frac{1}{n_k} + \frac{1}{n_l}} \text{ with } s_p = \sqrt{MSW}.$$

Kruskal-Wallis Test

- Use when the normality assumption for one-way ANOVA is violated
- Assumptions:
  - The samples are random and independent
  - variables have a continuous distribution
  - the data can be ranked
  - populations have the same variability
  - populations have the same shape

Kruskal-Wallis Test Procedure (1 of 3)

- Obtain relative rankings for each value
  - In event of tie, each of the tied values gets the average rank
- Sum the rankings for data from each of the *K* groups
  - Compute the Kruskal-Wallis test statistic
  - Evaluate using the chi-square distribution with $K - 1$ degrees of freedom

Kruskal-Wallis Test Procedure (2 of 3)

- The Kruskal-Wallis test statistic:

(chi-square with K – 1 degrees of freedom)

$$W = \left[\frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} \right] - 3(n+1)$$

where:

n = sum of sample sizes in all groups

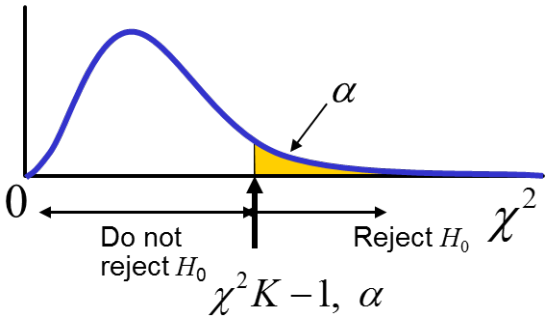
K = Number of samples

R_i = Sum of ranks in the i^{th} group

n_i = Size of the i^{th} group

Kruskal-Wallis Test Procedure (3 of 3)

- Complete the test by comparing the calculated *H* value to a critical χ^2 value from the chi-square distribution with $K - 1$ degrees of freedom



Decision rule

- Reject H_0 if $W > \chi^2_{K-1,\alpha}$
- Otherwise do not reject H_0

Kruskal-Wallis Example (1 of 4)

- Do different departments have different class sizes?

Class size (Math, M)	Class size (English, E)	Class size (Biology, B)
23	55	30
41	60	40
54	72	18
78	45	34
66	70	44

Size	Rank
18	1
23	2
30	3
34	4
40	5
41	6
44	7
45	8
54	9
55	10
60	11
66	12
70	13
72	14
78	15

Kruskal-Wallis Example (2 of 4)

- Do different departments have different class sizes?

Class size (Math, M)	Ranking	Class size (English, E)	Ranking	Class size (Biology, B)	Ranking
23	2	55	10	30	3
41	6	60	11	40	5
54	9	72	14	18	1
78	15	45	8	34	4
66	12	70	13	44	7
	$\Sigma = 44$		$\Sigma = 56$		$\Sigma = 20$

Kruskal-Wallis Example (3 of 4)

$H_0 : \text{Mean}_M = \text{Mean}_E = \text{Mean}_B$

$H_1 :$ Not all population means are equal

- The *W* statistic is

$$W = \left[\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right] - 3(n+1)$$

$$= \left[\frac{12}{15(15+1)} \left(\frac{44^2}{5} + \frac{56^2}{5} + \frac{20^2}{5} \right) \right] - 3(15+1) = 6.72$$

Kruskal-Wallis Example (4 of 4)

- Compare $W = 6.72$ to the critical value from the chi-square distribution for $3 - 1 = 2$ degrees of freedom and $\alpha = .05$:

$$\chi^2_{2,0.05} = 5.991$$

Since $H = 6.72 > \chi^2_{2,0.05} = 5.991,$

reject H_0

There is sufficient evidence to reject that the population means are all equal

Common pitfalls of ANOVA and alternative approaches

Despite its perceived simplicity, scientists frequently misuse ANOVA. A study by Wu et al. (2011) showed that from a survey of 10 leading Chinese medical journals in 2008, 446 articles used ANOVA, and of those articles, **59% of them used ANOVA incorrectly** (https://www.hindawi.com/journals/tswj/2011/139494/tab1/).

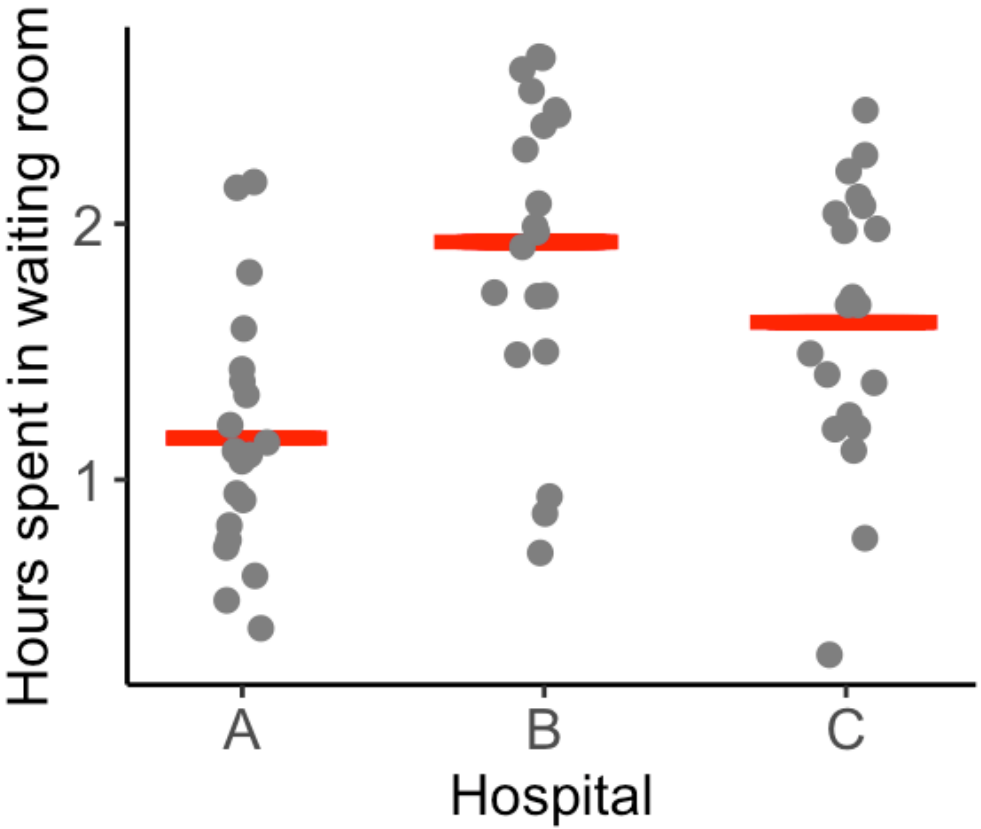
A simple toy example: hospital waiting times

Suppose that we have three hospitals, let's call them A, B and C (creative names, I know). We are interested in whether all three hospitals have the same the average waiting time for the emergency room.

We measured the waiting time for 20 unique individuals at each of these three hospitals (so there are 60 individuals in total). These waiting times (in hours) are recorded below.

Hospital A	Hospital B	Hospital C
1.8	0.9	1.4
1.4	0.7	2.1
0.7	2.6	1.4
0.8	1.7	1.2
0.5	2.5	2.1
2.1	2.4	2.3
0.9	2.4	1.7
2.2	2.3	1.2
1.2	2.0	1.1
1.3	1.7	1.3
1.1	2.1	0.3
1.1	0.9	1.7
0.4	2.7	1.5
1.4	1.5	1.7
0.8	2.0	2.0
1.1	1.9	0.8
0.6	2.6	2.0
1.1	2.4	2.4
1.6	1.5	2.2
0.9	1.7	2.0

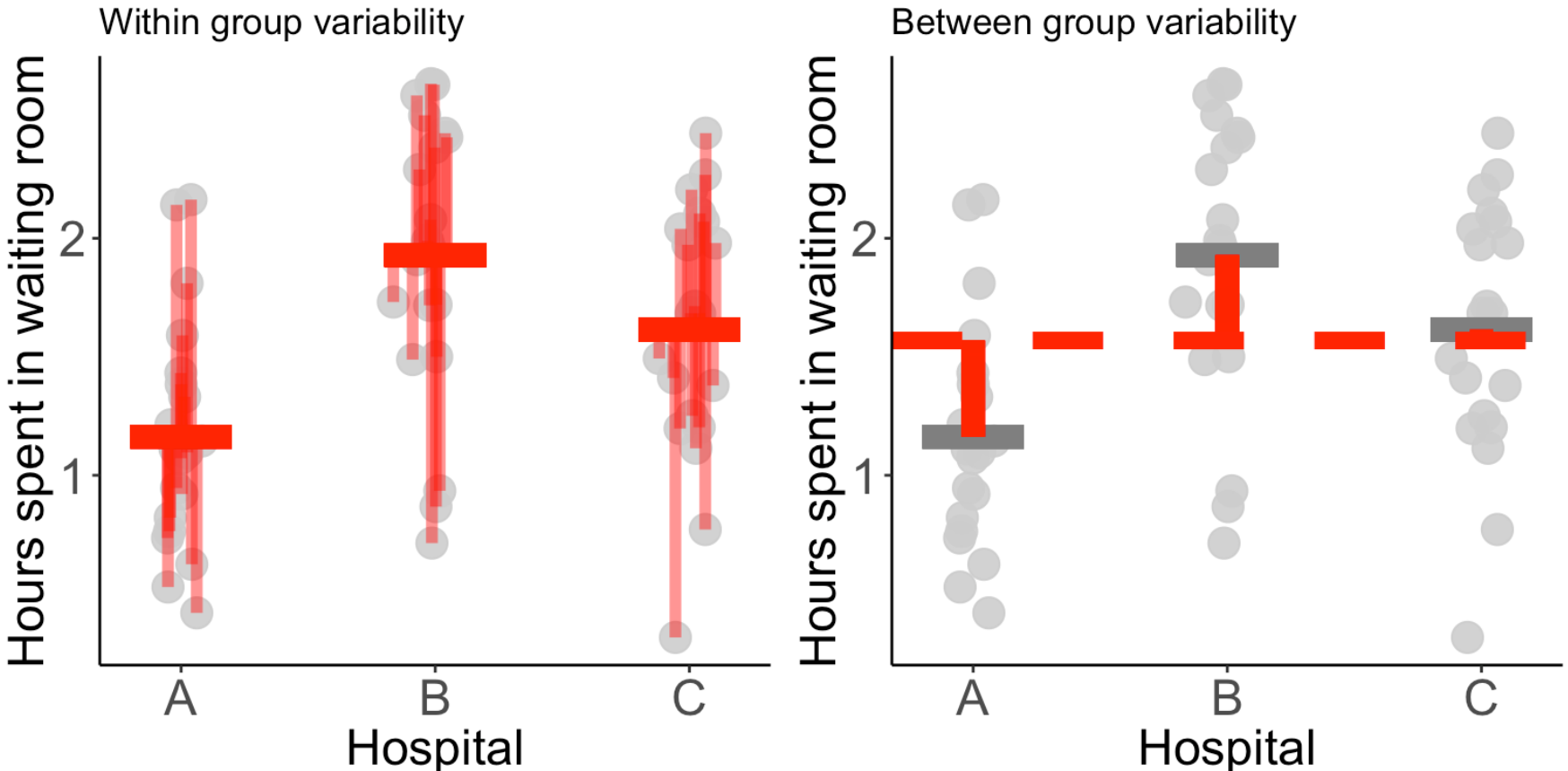
Let's inspect graphically. Most people seem to wait over an hour, with some unlucky individuals waiting for almost 3 hours. The mean waiting time for each hospital is highlighted by a red bar.



The question here is actually asking about equality between the average waiting times from the population of all patients who have ever, and will ever, wait in these waiting rooms, regardless of whether they fall in our sample.

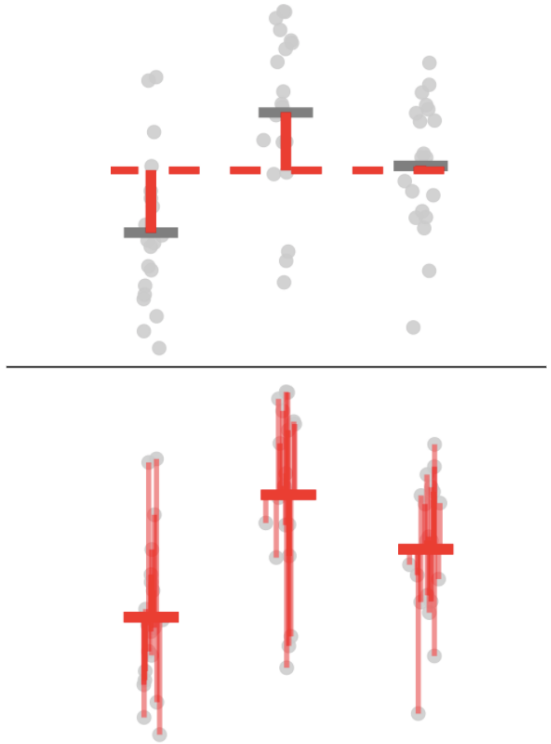
Although the sample means clearly aren't identical (the red bars are all at different heights), do we have enough evidence to show that the underlying population waiting time means are different, or are the differences that we observe are simply reflection of the inherent noise in the data?

The red bars in the left panel highlight the *within-group variance*, while the red bars in the right panel highlight the *between-group variance*.



ANOVA analysis works like this

$$\frac{\text{BSS} / (K - 1)}{\text{WSS} / (N - K)} =$$



Remember that there are assumptions…..

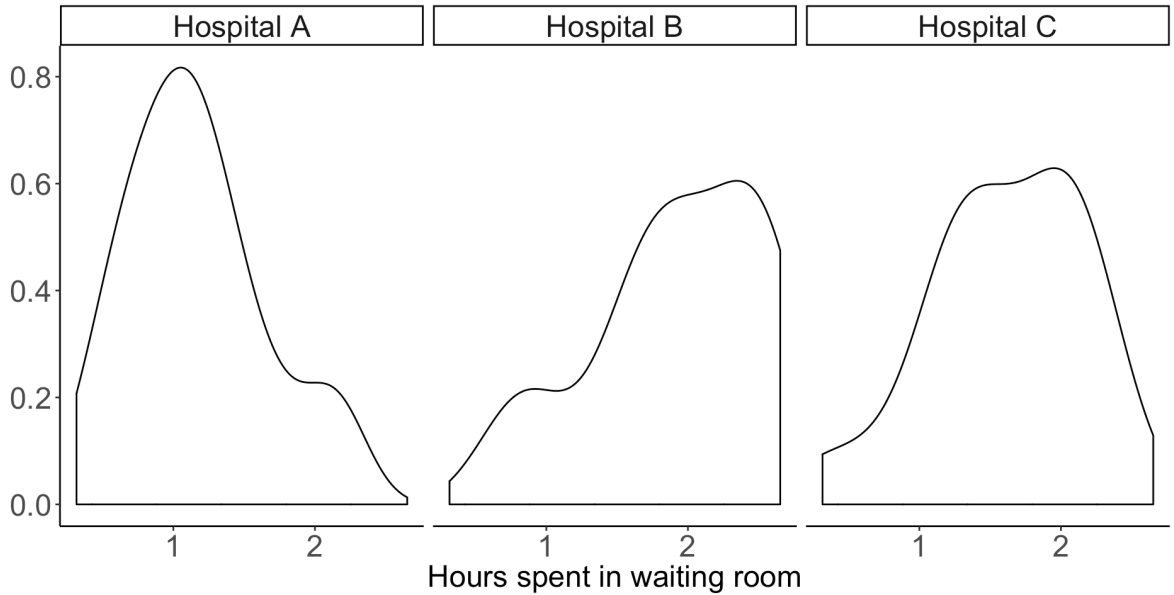
Assumption 1: The samples are independent.
Independence is an extremely common assumption that is hard to test in general.

Assumption 2: The data are normally distributed.
Not being a fan of such distributional assumptions myself, I am inclined to point the reader in the direction of non-parametric versions of ANOVA, including the Kruskal-Wallis test.
Those wishing to test the normality of their data can do so using a variety of methods such as plotting a QQ-plot, or using a normality test.

Assumption 3: Each group has the same variance.
The common variance assumption can be tested using common tests, such as the Bartlett test and the Fligner-Killeen test, which are easily implemented in R/Python.

Assumption 2:

The figure below plots the density estimation for the waiting times from each hospital. We know that if our data is normally distributed, it should look vaguely like a bell-curve (!!).
You can use this as a lesson on the difficult of drawing conclusions on normality from small samples (in this case, we have 20 observations in each group).



Assumption 2

A Shapiro-Wilk test for normality provides p-values of 0.39, 0.087, 0.52 for hospitals A, B and C, respectively.

Although none of these values are “significant” (even unadjusted for multiple testing), we have stumbled upon another lesson: small p-values (p=0.087 for hospital B) can certainly occur when
the null hypothesis is true (in this case, the null hypothesis is that the data are normally distributed)!

Assumption 3

Based on a visual assessment, *the common variance assumption* is probably fairly reasonable (and, again, since I simulated this data, I can confirm that the variance is the same for each hospital).

To test this formally, Bartlett's test for homogeneity of variances yields a p-value of 0.68, indicating that we do not have evidence that the variances are different.

We have now concluded that the assumptions for ANOVA are satisfied, and can proceed to do our calculations.

Calculating the between-sum-of-squares (BSS) and scaling by the degrees of freedom (the number of groups minus 1), and the within-sum-of-squares (WSS) and scaling by the degrees of freedom (the number of observations minus the number of groups), we get that

$$\mathbf{BSS/(K-1) = 5.94 / (3-1) = 2.97}$$

$$\mathbf{WSS/(N-K) = 16.96 / (60-3) = 0.30.}$$

Our test statistic turns out to be quite large indeed:

$$\mathbf{F = BSS / (K-1) / WSS / (N-k) = 2.97/ 0.3 = 9.98.}$$

Since we are confident that the ANOVA assumptions are satisfied, this F-statistic must follow an F distribution with suitable degrees of freedom.

Our p-value can thus be calculated as follows:

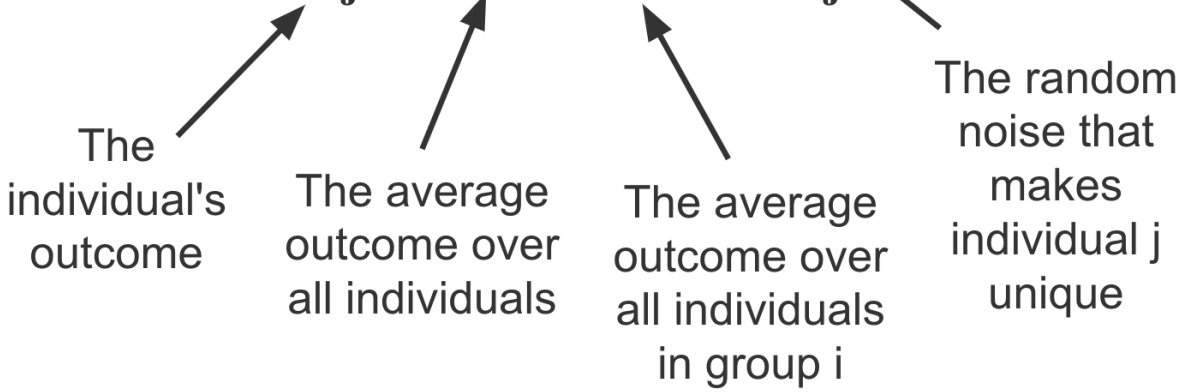
$$P(F_{2,53} \geq 9.98) = 0.000192$$

And we can claim to have evidence that the three group means are not all identical. Note that we can interpret this as the distances between the group means and the global mean is quite large relative to the distances between the individual observations and the group means.

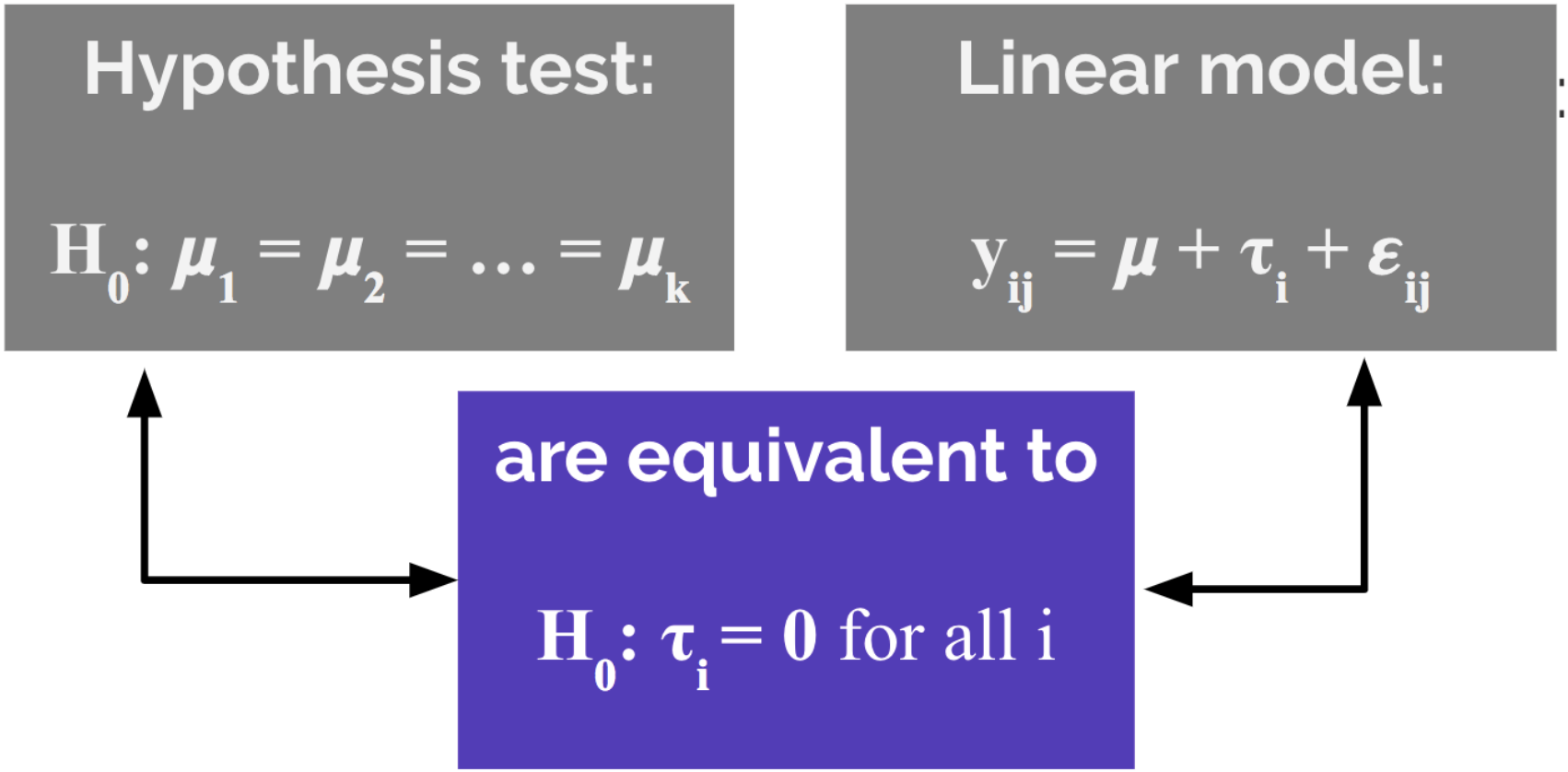
ANOVA as a linear model

It is more common to talk about ANOVA as a **linear model**.
The anova linear model can be written as follows:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$



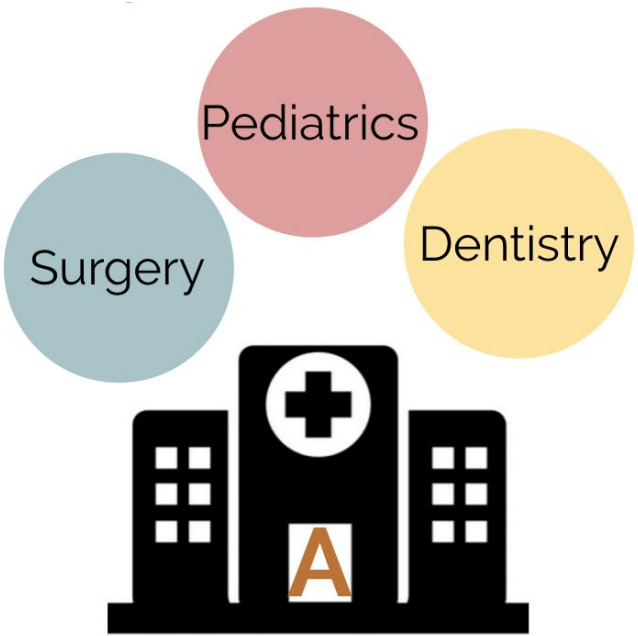
μ represents the **overall average** wait time across all hospitals, and τ_i represents the amount of time that is either added or subtracted from the overall average as a result of being at hospital i.
To get the average wait time for hospital i we can calculate $\mu_i := \mu + \tau_i$
ϵ_{ij} represents the “noise” term; the quantity that defines how the waiting time for individual j differs from the mean (within their group).



If the hospital-specific effects τ_A,τ_B and τ_C are all equal to zero, then the average effect across all groups is the same: $\mu_A=\mu_B=\mu_C=\mu$

COMMON PITFALLS OF ANOVA

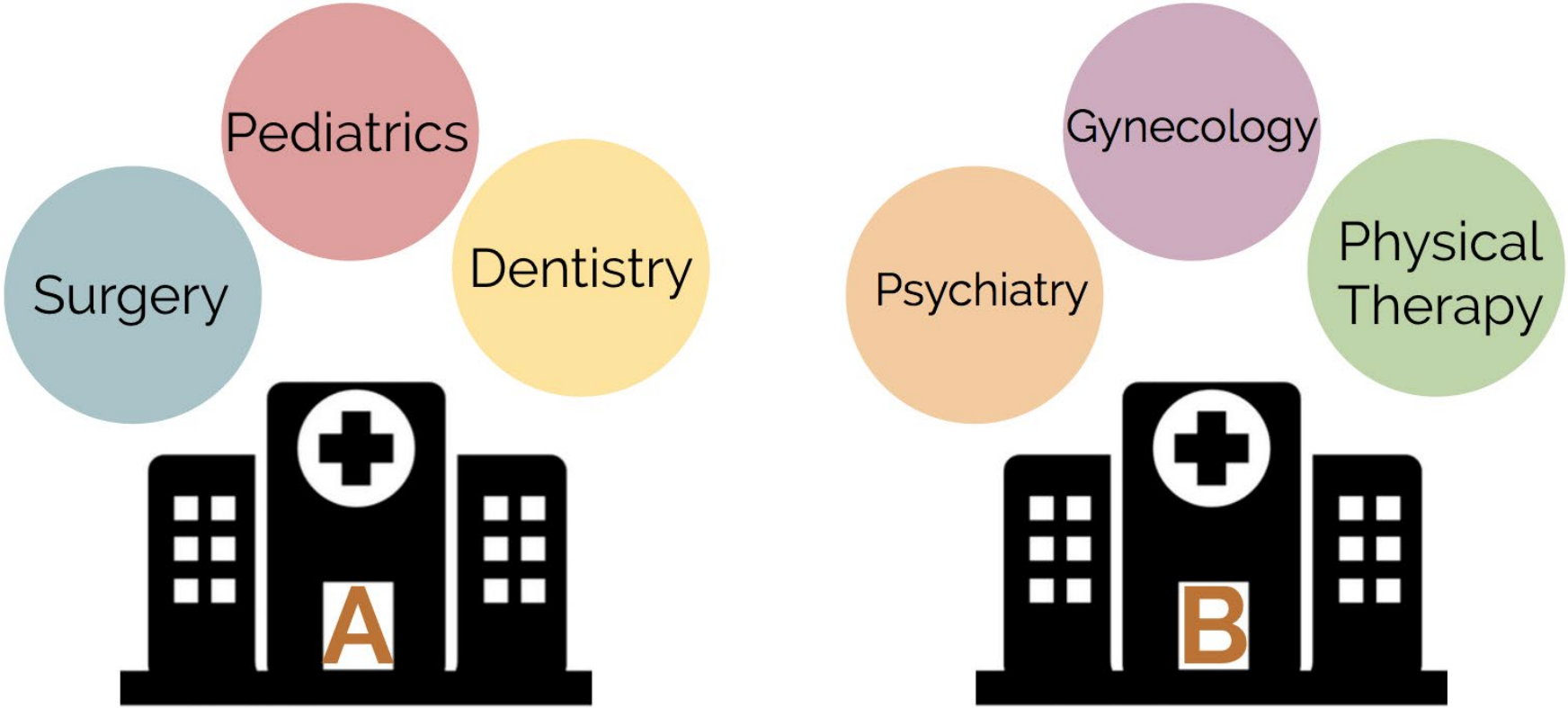
Using one-way ANOVA when there is more than one grouping variable



This is called two-way ANOVA

$$y_{ijk}=\mu+\tau_i+\gamma_j+\beta_{ij}+\epsilon_{ijk}$$

Using one-way ANOVA when there is more than one grouping variable and no groups in common (nested ANOVA)



Conducting ANOVA multiple times for multiple outcomes

Suppose that instead of simply being interested in whether there is a difference between waiting time for each hospital, we were also interested in differences in average *length of hospital stay* and *cost of visit*. Then the incorrect way to proceed would be to generate three separate ANOVA models and draw our conclusions separately for each model. This reeks of multiple testing issues and does not take into account any dependence between the different outcome variables.

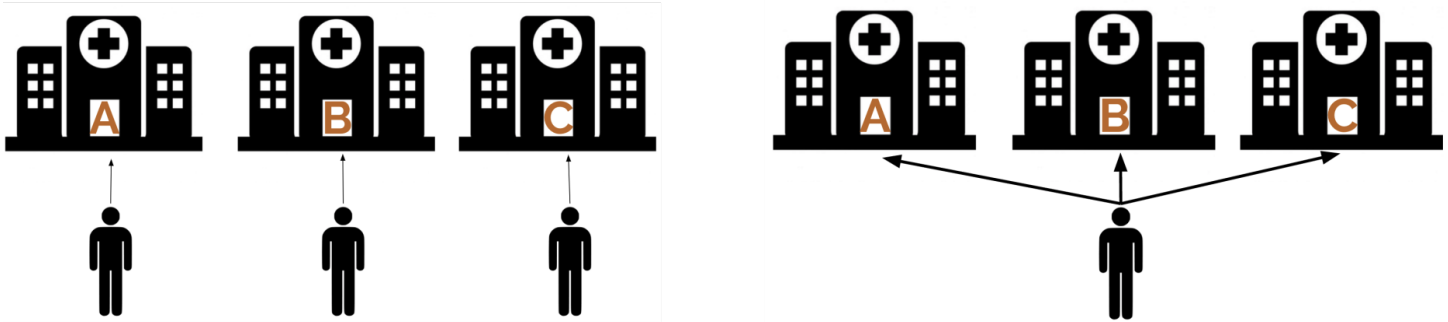
One should use Multivariate Analysis of Variance (**MANOVA**).

Incorrectly conducting multiple pair-wise comparisons following ANOVA

Upon obtaining a “significant” ANOVA p-value, a common mistake is to then go and test all of the pairwise differences to identify *which* of the populations had different means. This is another example of multiple hypothesis testing, and corrections on these p-values must be made.

Using ANOVA to analyse repeated-measures data

What if, instead of having measured the waiting room times on a different set of 20 people at each hospital (left-panel in fig below), we instead measured the waiting room times on the same set of 20 people at each hospital (right-panel in fig below)?



We have certainly violated the assumption that our observations are independent. Fortunately, **repeated measures ANOVA (rANOVA)** is a method for exactly this situation. https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php

Can I use ANOVA if my data violates the assumption of common variances?

if the sample size in each group is similar, and the difference between variance isn't too bad, you should be ok.

If my data are not normal, can I simply transform it and draw the conclusions as normal?

Yes, you can

How does the ANOVA for model comparison work?

It compares *nested* models wherein one model consists of a subset of the set of variables of the other model.

Note that the use of the word “nested” here has nothing to do with the nested anova discussed above in which the grouping variables themselves (rather than the models) were nested.

The comparison being made by ANOVA in this situation is whether the residual sum of squares (which is essentially the *within sum of squares* from one-way ANOVA) for model 1 (the larger model) is larger than the residual sum of squares for model 2 (the smaller model).