MA 214: Applied Statistics

Instructor: Ashis Gangopadhyay

ANOVA
Part I

Where We've Been

We were discussing….

- How to develop statistical models where the response variable Y is continuous and covariates are (mostly) continuous.

Where We're Going

- We want consider the scenarios where the response variable is still continuous, but <u>all</u> of the covariates are categorical.

Analysis of Variance

- Recall that a statistical model is selected on the basis of the characteristics of the dependent variable and the covariates as follows:

Dependent Variable

Covariates		*Continuous*	*Nominal*
	Continuous	Regression	Logistic Regression
	Nominal	ANOVA	Categorical Data Analysis

Analysis of Variance

- This suggests that an Analysis of Variance (ANOVA) model is appropriate when the dependent variable Y is continuous, but *all* covariates (or factors) in the model are nominal (categorical).

Dependent Variable

Covariates		*Continuous*	*Nominal*
	Continuous	Regression	Logistic Regression
	Nominal	$\boxed{\text{ANOVA}}$	Categorical Data Analysis

Analysis of Variance

- **Example 1:** A study was conducted to compare between three different techniques to lower blood pressure level of patients suffering from high blood pressure. 15 subjects were randomly assigned to one of three groups; and after four weeks, their reductions in blood pressure level were recorded.

Method	**Reduction in BP level**				
Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	8	8	4

Some definitions

- **Factor:** Covariates included in the experiment.

  In Example 1, the treatment method is the factor.

Method	Reduction in BP level				
Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	8	8	4

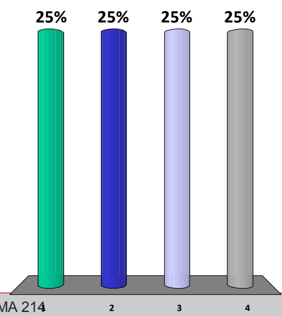
Clicker Question:

- Suppose in a market research, questionnaires are printed on papers of various colors (say, white, blue, red and pink) and the objective is to study the effect of the differences of the paper on the rate of response.

- **What is factor in this study?**

Factor in the study is

1. Survey respondents.
2. Questions listed in the questionnaire.
3. Color of the paper. ✓
4. Number of respondents.



Some definitions

- **Factor Level:** Specific forms of factor used in the experiment.

  In Example 1, it is the three forms of treatment methods; namely medication, exercise and diet, are the factor levels.

Method	Reduction in BP level				
Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	8	8	4

Some definitions

- **Factor Level:** Specific forms of factor used in the experiment.

  In the questionnaire example, the specific colors of the papers on which the questionnaires are printed are the factor levels.

Some definitions

- **Treatment:** Combinations of factor levels are called treatments. In one factor ANOVA, the factor levels are the treatments. In multifactor ANOVA, specific combinations of factor levels from various factors are the treatments.

Method	Reduction in BP level				
Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	8	8	4

Some definitions

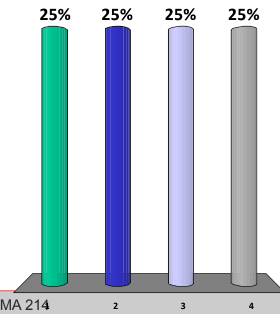
- In the BP example, suppose we wish to study the effect of two factors on the blood pressure level. The two factors are: method (with factor levels exercise, medication, and diet) and age (with factor levels young, middle-aged, and older).

  In this case, there are nine treatments:

  1. ***exercise, young***
  2. ***exercise, middle-aged***
  3. ***exercise, older***
  4. ***medication, young***
  5. ***medication, middle-aged***
  6. ***medication, older***
  7. ***diet, young***
  8. ***diet, middle-aged***
  9. ***diet, older***

Suppose a study was conducted to investigate the effects of education level (High School, College, Post-Graduate) and Gender of the person on his/her annual income. The number of treatments is:

1. 2
2. 3
3. 5
4. 6



Some definitions

- **Replications:** Number of observations within each treatment.

  In the BP example, the number of replications for all treatments is five.

Method	Reduction in BP level				
Medication	$\boxed{10 \quad 12 \quad 9 \quad 15 \quad 13}$				
Exercise	6	8	3	0	2
Diet	5	9	8	8	4

Some definitions

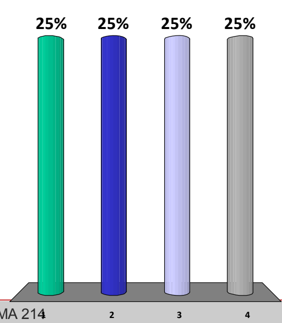
- An ANOVA is called *balanced*, if the number of replications in all treatments are equal, otherwise the ANOVA is called *unbalanced*.

Method	Reduction in BP level				
Medication	10	12	9	15	13
Exercise	6	8	3	0	2
Diet	5	9	8	8	4

} ***balanced***

If an ANOVA is balanced, then…

1. All factors have equal number of factor levels.
2. Number of factor levels must be even.
3. All treatments have equal number of observations.
4. All treatments have unequal number of observations.



One factor (one-way) ANOVA

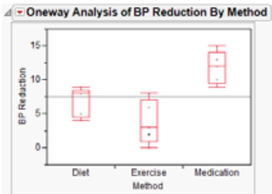
- Consider an experiment in which only one factor at k-levels is being studied. The general data structure is of the form:

Factor Level	Observations			
1	y_{11}	y_{12}	…	y_{1n_1}
2	y_{21}	y_{22}	…	y_{2n_2}
…	…	…	…	…
…	…	…	…	…
k	y_{k1}	y_{k2}	…	y_{kn_k}

Thus the factor has k-levels, and there are n_i observations in the ith factor level, i = 1, 2, …, k

One factor (one-way) ANOVA

- It is often useful to look at a graphical representation of the data, such as a side-by-side boxplot to get a better understanding of the differences between the observations in various factor levels.



One factor (one-way) ANOVA

- We assume the following statistical model that describes the data:

$$y_{ij} = \mu_i + \varepsilon_{ij},\ i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i$$

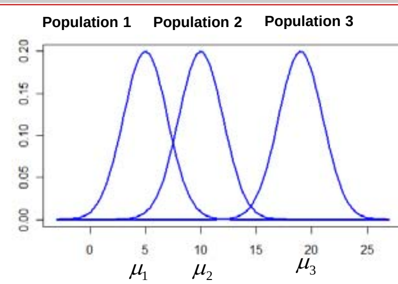
where the random error term ε_{ij} is assumed to be independent and normally distributed with mean 0 and variance σ^2.

One factor (one-way) ANOVA

- **Note:** the implication of the model is that the observations in the *i*th factor level describe a random sample of size n_i from a normal population with mean μ_i and variance σ^2, for i = 1, 2, …, k.

- In other words, we have k independent samples from k normal populations with mean μ_i and a common variance σ^2.

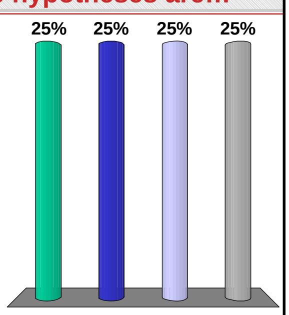
One factor ANOVA model



Normal distributions with different means but constant variance

Suppose we wish to test a hypothesis that all three populations are identical. Then the null and alternative hypotheses are…

1. $H_0: \mu_1 = \mu_2 = \mu_3$
   $H_a: \mu_1 \neq \mu_2 \neq \mu_3$
2. $H_0: \mu_1 \neq \mu_2 \neq \mu_3$
   $H_a: \mu_1 = \mu_2 = \mu_3$
3. $H_0: \mu_1 = \mu_2 = \mu_3$
   $H_a:$ Not all 3 μ's are equal
4. $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$
   $H_a:$ Not all 3 σ^2's are equal



One factor (one-way) ANOVA

- Thus, it is useful to note that one factor ANOVA model is a generalization of two-sample inference that we studied earlier in the course.

- One-factor ANOVA is often described as k-sample inference.

Objectives of one-factor ANOVA

- First objective is to verify whether the factor levels (treatments) have any effect on the response variable. This is done by testing the following hypothesis:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_a: \text{Not all } \mu_i\text{'s are equal.}$$

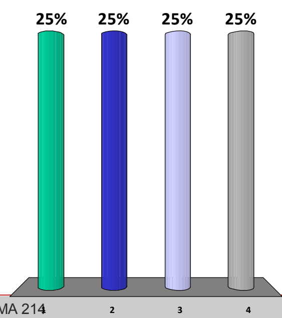
- Note that the null hypothesis states that the mean effect of all treatments are equal, and no treatment is better than the others; whereas the alternative hypothesis states that there is a statistically significant difference between the effects of treatments.

Objectives of one-factor ANOVA

- If we reject the null hypothesis, then we need to find the "best" treatment, or the best group of treatments.
- For this purpose, we use pairwise comparison techniques.

Which of the following is a necessary assumption to conduct a one-way ANOVA comparing *k* population means?

1. The *k* populations of values of the response variable associated with the treatments have equal variances.
2. The *k* populations of values of the response variable associated with the treatments all have normal distributions.
3. The samples of experimental units associated with the treatments are randomly selected, independent samples.
4. ✓ All of the above.



Some useful notations

y_{ij} : jth observation in ith factor level

$$n = \sum_{i=1}^{k} n_i \text{ (Total sample size)}$$

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$ (average of all observations in *i*th factor level)

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2$$ (sample variance of all observations in *i*th factor level)

ANOVA for one-factor model

- We use the analysis of variance approach that we developed in regression. So the basic idea is to consider the variability of the observations on the dependent variable *y*, and then figure out how much of that variability can be explained by the factor in the model.

ANOVA for one-factor model

- First, we ignore the fact that the observations are associated with different factor levels. The overall (baseline) variability of the observations on the response variable *Y* is given by:

$$SSTO = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2$$

ANOVA for one-factor model

- Now, we take into consideration the fact that observations are actually related to different factor levels, with possibly different mean effects. This leads to a revised measure of variability given by:

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2$$

ANOVA for one-factor model

- Thus, reduction in variability by the taking into effect of treatments is given by the Treatment Sum of Squares:

$$SSTR = SSTO - SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{y}_i - \bar{y}\right)^2$$

- As we have seen in regression, SSTR measures the part of the overall variability explained by the factor.

ANOVA for one-factor model

- Computational formulas:

$$SSTO = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2}{n}$$

$$SSE = \sum_{i=1}^{k}\left(n_i - 1\right)s_i^2$$

$$SSR = SSTO - SSE$$

ANOVA for one-factor model

- ANOVA for one-factor model

Source	*SS*	*DF*	*MS*	*F*
Treatment	*SSTR*	*k* – 1	*MSTR*	$F = \frac{MSTR}{MSE}$
Error	*SSE*	*n* – *k*	*MSE*	
Total	*SSTO*	*n* – 1		

Interpretation of ANOVA

- R^2 (coefficient of determination)

$$R^2 = \frac{SSTR}{SSTO}$$ (Proportion of the total variability explained by the factor)

- F-test

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_a : \text{Not all } \mu_i\text{'s are equal.}$$

Reject H_0 if $F > F_\alpha$ (F_α is based on ($k-1$, $n-k$) degrees of freedom).

Interpretation of ANOVA

- **Example:**

Method	**Reduction in BP level**					**Total**	**Average**	**Variance**
Medication	10	12	9	15	13	59	11.8	5.7
Exercise	6	8	3	0	2	19	3.8	10.2
Diet	5	9	8	8	4	38	6.8	4.7

Thus:

$$n_1 = n_2 = n_3 = 5, n = 15$$
$$s_1^2 = 5.7, s_2^2 = 10.2, s_3^2 = 4.7$$
$$\bar{y}_1 = 11.8, \bar{y}_2 = 3.8, \bar{y}_3 = 6.8$$

$$\sum_i^3\sum_{j=1}^5 y_{ij}^2 = 1082, \sum_i^3\sum_{j=1}^5 y_{ij} = 112$$
$$\bar{Y} = \frac{112}{15} = 7.47$$

Interpretation of ANOVA

- **Example:** $SSTO = 1082 - \frac{112^2}{15} = 245.73$

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2$$
$$SSE = 4\times 5.7 + 4\times 10.2 + 4\times 4.7 = 82.4$$
$$SSTR = 245.73 - 82.4 = 163.33$$

Source	*SS*	*DF*	*MS*	*F*
Treatment	*163.33*	*2*	*81.66*	*F = 11.88*
Error	*82.4*	*12*	*6.87*	
Total	*245.73*	*14*		

Interpretation of ANOVA

- Thus $R^2 = \frac{163.33}{245.73} = 0.66$, that is, 66% of the variability of the reduction in the BP level is explained by the treatment.

- Now, we wish to test the hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_a : \text{Not all } \mu_i\text{'s are equal.}$$

Reject H_0 if $F > F_{0.05}$ ($F_{0.05}$ = 3.89 using (*2, 12*) df).

Interpretation of ANOVA

- Since *F = 11.88,* we reject H_0 and conclude that treatments have statistically significant effect on the blood pressure level.



- Now that we have concluded that all treatments are not the same, how do we identify the "best" treatment(s)?

Pairwise comparison (multiple comparison) procedure

- If we have rejected the null hypothesis in ANOVA F-test, then we may wish to explore the relationship between specific subsets of treatments. In particular, one can consider a procedure in which all possible distinct pairs of treatments are compared, which can lead to identification of the "best" treatment, or the best group of treatments.

Pairwise comparison (multiple comparison) procedure

- Multiple comparison procedures are extremely useful in statistics, and there are several such techniques. Here is a list of some of the most common procedures:
  - Tukey multiple comparison
  - Bonferroni
  - Scheffe'
  - SNK

  … among many others.
- In this course we will discuss the Tukey procedure in detail.

Tukey multiple comparison procedure

- In Tukey multiple comparison procedure, we try to identify the best treatment (or group of treatments) by comparing all possible pairs of treatment means. Thus the objective is to test the hypothesis:

$$H_0 : \mu_i = \mu_j \quad \text{versus} \quad H_a : \mu_i \neq \mu_j$$

- Note that H_a states that there is a statistically significant difference between the *i*th and *j*th treatments. Thus the Tukey multiple comparison procedure for *k*-means involves ${}_kC_2 = k(k-1)/2$ many tests.

Tukey multiple comparison procedure

- However, the problem is that for each of these tests, there is a small probability α that the inference is incorrect (type I error), which can accumulate to a "large" probability of an incorrect decision when we combine the inferences based on these individual hypothesis testing procedures.

Suppose our inferences is based on *m* hypothesis testing procedures each with type I error *α*. Then

- P(at least one of the inferences is incorrect)
  = 1 – P(none of the inferences is incorrect)
  = $1 - (1 - \alpha)^m$

Tukey multiple comparison procedure

- **Example:** If our inference is based on 6 hypothesis testing procedures, each with probability of type I error 0.05, then the probability of making an incorrect inference is roughly $1 - 0.95^6 \approx 0.2649$, which is unacceptably high!

- This is why the hypothesis H_0: $\mu_i = \mu_j$ is evaluated by constructing simultaneous confidence intervals for the difference $(\mu_i - \mu_j)$ with *family confidence coefficient* $(1 - \alpha)$.

Tukey multiple comparison procedure

- **Family Confidence Coefficient:** A family confidence coefficient is defined as the proportion of correct families of estimates when repeated samples are selected and specified confidence intervals for the entire family are calculated for each sample.

1.2.3 Tukey multiple comparison procedure

Steps for Tukey multiple comparison:

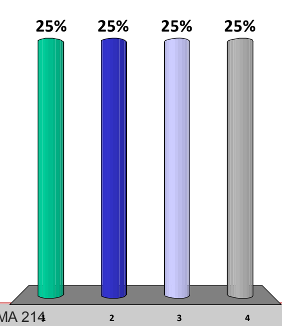
- **Step 1:** Construct Tukey confidence intervals for all distinct pairwise differences $(\mu_i - \mu_j)$ with family confidence coefficient $(1 - \alpha)$ as follows:

$$\left(\bar{y}_i - \bar{y}_j\right) \pm \frac{q_\alpha}{\sqrt{2}} \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $\hat{\sigma} = \sqrt{MSE}$, and q_α is the $(1 - \alpha)$ percentile of the studentized range distribution with df $(k, n - k)$. Note that the tables for percentiles of this distribution are available in your book.

Suppose we carry out pairwise comparisons between 5 treatment means. The numeber of Tukey CI is:

1. 5
2. 10
3. 15
4. None of the above



1.2.3 Tukey multiple comparison procedure

Steps for Tukey multiple comparison:

- **Step 2:** Accept H_0 if zero is included in the interval, and conclude that the difference between the means is not statistically significant. Otherwise, conclude that the difference is significant.

1.2.3 Tukey multiple comparison procedure

- **Example:** We wish to construct 95% Tukey confidence intervals for all pairwise comparisons.

Medication vs. Exercise $(H_0: \mu_1 - \mu_2 = 0)$

95% confidence interval for $(\mu_1 - \mu_2)$:

$$(\bar{y}_1 - \bar{y}_2) \pm \frac{q_{0.05}}{\sqrt{2}} \hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

From table: $q_{0.05}$ = 3.77 based on (k = 3, n – k = 12) df.

$\hat{\sigma} = \sqrt{6.87} = 2.62$

1.2.3 Tukey multiple comparison procedure

$$\frac{q_{0.05}}{\sqrt{2}} \hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{3.77}{\sqrt{2}} \times 2.62 \times \sqrt{\frac{1}{5} + \frac{1}{5}} = 4.42$$

$$(11.8 - 3.8) \pm 4.42 = (3.58, 12.42)$$

Thus the difference is statistically significant.

Tukey multiple comparison procedure

- **Example (cont'd):** We wish to construct 95% Tukey confidence intervals for all pairwise comparisons.

Medication vs. Diet $(H_0: \mu_1 - \mu_3 = 0)$

95% confidence interval for $(\mu_1 - \mu_3)$:

$$(\bar{y}_1 - \bar{y}_3) \pm \frac{q_{0.05}}{\sqrt{2}} \hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_3}}$$

From table: $q_{0.05}$ = 3.77 based on (k = 3, n – k = 12) df.

$\hat{\sigma} = \sqrt{6.87} = 2.62$

Tukey multiple comparison procedure

$$\frac{q_{0.05}}{\sqrt{2}} \hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_3}} = \frac{3.77}{\sqrt{2}} \times 2.62 \times \sqrt{\frac{1}{5} + \frac{1}{5}} = 4.42$$

$$(11.8 - 6.8) \pm 4.42 = (0.58, 9.42)$$

Thus the difference is statistically significant.

Tukey multiple comparison procedure

- **Example (cont'd):** We wish to construct 95% Tukey confidence intervals for all pairwise comparisons.

Exercise vs. Diet $(H_0: \mu_2 - \mu_3 = 0)$

95% confidence interval for $(\mu_2 - \mu_3)$:

$$(\bar{y}_2 - \bar{y}_3) \pm \frac{q_{0.05}}{\sqrt{2}} \hat{\sigma} \sqrt{\frac{1}{n_2} + \frac{1}{n_3}}$$

From table: $q_{0.05}$ = 3.77 based on (k = 3, n – k = 12) df.

$\hat{\sigma} = \sqrt{6.87} = 2.62$

Tukey multiple comparison procedure

$$\frac{q_{0.05}}{\sqrt{2}} \hat{\sigma} \sqrt{\frac{1}{n_2} + \frac{1}{n_3}} = \frac{3.77}{\sqrt{2}} \times 2.62 \times \sqrt{\frac{1}{5} + \frac{1}{5}} = 4.42$$

$$(3.8 - 6.8) \pm 4.42 = (-7.42, 1.42)$$

Thus the difference is NOT statistically significant.

Tukey multiple comparison procedure

- **Presentation of the results:** Arrange the treatments in the increasing order of their means, and identify pair of treatments that are not significantly different by joining them with lines below the treatments.

3.8	6.8	11.8
Exercise	**Diet**	**Medication**

Tukey multiple comparison procedure

- This leads to the Tukey groups {Exercise, Diet} and {Medication}. Thus in this context, since the higher average corresponds to better treatment effect, Medication is the best treatment, while Diet and Exercise are statistically equivalent.

3.8	6.8	11.8
Exercise	**Diet**	**Medication**

Tukey multiple comparison procedure

- It should be noted that the Tukey procedure controls the probability of the type I error well, but there are other tests (such as SNK, REGWQ, etc) that are more efficient than Tukey particularly when the sample sizes in each factor level are unequal, and should be considered while analyzing the data using a statistical software.