Topic 3 – Analysis of Several Population Means

Comparing Several Means

Def'n: ANalysis Of VAriance (ANOVA) is a procedure to test the equality of three or more population means. NOTE: the name of the test refers to comparing different sources of variability; it WILL test differences among means.

Test requires the following assumptions:

1. The populations are all normally distributed.
2. The populations all have the same standard deviation.
3. The samples, from different populations, are random and independent.

Robustness:

- Assumption #1 is not critical unless severe skewness or strong outliers exist
  - A higher n is always better.
- Assumption #2 is more critical but harder to assess. For Stat 252, consider the informal rule of the ratio of largest s being less than twice the smallest. If samples sizes (approx.) equal, the condition can be relaxed.

Notation:

- y_{ij} = observation for j^{th} subject in i^{th} group
- $i = 1, \ldots, I$ indexes groups
- $j = 1, \ldots, n_i$ indexes subjects within groups
- n_i = # of observations in i^{th} group; $N = \sum n_i$ = total # of observations
- $\bar{y}_i.$ and s_i^2 are sample mean and variance for the i^{th} group
- $\bar{y}..$ = grand mean = mean for combined sample:

$$\bar{y}.. = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{N}\sum_{i=1}^{I} n_i \bar{y}_i.$$

Statistical model, parameters, hypotheses:

Using a linear statistical model, each observation can be represented by

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, \ldots, I; j = 1, \ldots, n_i$$

where y_{ij} are independent random observations, μ is the *overall mean*, τ_i is a parameter associated with the i^{th} group called the i^{th} *treatment effect*, and ε_{ij} is a random error.

- random order of obs'ns & uniform environment → "completely randomized design"
- consider that only I groups exist → fixed-effects model
- parameters: σ, $\mu_1 = \mu + \tau_1, \ldots, \mu_I = \mu + \tau_I$; the τ_i are defined such that $\sum \tau_i = 0$.
- H_0: $\mu_1 = \ldots = \mu_I$
- H_a: the μ_i are not all equal (OR, at least 2 μ_i are different)

ANOVA test statistic:

For sources of variability in the model above, the *ANOVA identity* is

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}..)^2 = \sum_{i=1}^{I} n_i(\bar{y}_i. - \bar{y}..)^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i.)^2$$

$$TSS = SST + SSE$$

where *TSS* is the total sum of squares, *SST* is the (between-)treatment sum of squares and *SSE* is the residual sum of squares. Also, the presence of "squares" suggest a ratio test. Thus, for the above H_0, we have

$$F_0 = \frac{SST/(I-1)}{SSE/(N-I)} = \frac{MST}{MSE} \sim F_{I-1,\,N-I}$$

Recall that the *F* distribution has the following properties:

1. The units of an *F* distribution are nonnegative.
2. It is continuous and skewed to the right.
3. It has two parameters: *df* for the numerator and *df* for the denominator.

Reject H_0 if $F_0 > F_{\alpha,\,I-1,\,N-I}$

- Rationale: $E(SST) = (I-1)\sigma^2 + \sum_{i=1}^{I} n_i \tau_i^2$ and $E(SSE) = (N-I)\sigma^2$

→ If H_0 is true, then $E(MST) = \sigma^2 = E(MSE)$.

→ If H_a is true, then $E(MST) > \sigma^2 = E(MSE)$.

So, under H_0, we expect $F_0 \approx 1$, while under H_a, we expect $F_0 > 1$.

Using *p*-value, Table A.4 gives *F* critical values for several *p* (NOT *p*-values).

p-value = area under *F* distribution to *right* of observed *F*-value.

ANOVA Table:

Source	**df**	**SS**	**MS**	***F***	***p*-value**
Treatment	$I-1$	*SST*	*MST*	*MST*/*MSE*	$P(F > F_0)$
Error	$N-I$	*SSE*	*MSE*		
Total	$\boldsymbol{N-1}$	***TSS***			

Note that *TSS* = *SST* + *SSE*; df(total) = df(Group) + df(Error); MS = SS/df;

"Treatment" = "Between"; "Error" = "Within"

Calculating SS is tedious! More important to understand values and how they relate to other values in the ANOVA table. If you received an incomplete table, you should be able to fill it in.

e.g. Consider an *I*-mean problem. Five observations are gathered for each group. Use the given information in the table below to answer the questions. Assume all populations are normal with some common variance.

(a) What is *I*?

(b) What are the test statistic and *p*-value for the test to determine if any of the *I* groups are different?

Source	**df**	**SS**	**MS**	***F***	***p*-value**
Between					
Within			6		
Total	**54**	**360**			

Linear Combinations of Means

1. Parameter: $\gamma = C_1\mu_1 + C_2\mu_2 + ... + C_I\mu_I$

   If $\sum_{i=1}^{I} C_i = 0$, then γ is said to be a contrast.

2. Estimator: $g = C_1\overline{Y}_1. + C_2\overline{Y}_2. + ... + C_I\overline{Y}_I.$

3. Sampling Distribution of Estimator:

   Since *g* is a linear combination of independent sample means, which are normally distributed, it can be shown that:

$$g \sim N\left(\mu_g = \gamma, S.D.(g) = \sigma_g = \sqrt{\frac{C_1^2\sigma_1^2}{n_1} + ... + \frac{C_I^2\sigma_I^2}{n_I}}\right)$$

$$S.E.(g) = \sqrt{MSE}\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + ... + \frac{C_I^2}{n_I}}$$

   where, $\sqrt{MSE} = \sqrt{\frac{(n_1-1)s_1^2 + ... + (n_I-1)s_I^2}{N-I}}$ → (s_p for *I* groups)

4. i) Test statistic for testing H_0: $\gamma = 0$

$$t_0 = \frac{g}{S.E.(g)} \sim t_{N-I}$$

4. ii) A $(1-\alpha)100\%$ CI for γ is

$$g \pm t_{N-I,\,\alpha/2} \times S.E.(g)$$

Example (theoretical): Consider an $I = 5$ mean problem.

a) Suppose you wish to compare only μ_4 and μ_5.

This difference $(\mu_4 - \mu_5)$ is still a linear combination of all 5 means.

→ $\gamma = \mu_4 - \mu_5 = 0\mu_1 + 0\mu_2 + 0\mu_3 + 1\mu_4 - 1\mu_5$

$g = \bar{y}_4. - \bar{y}_5.$ and $S.E.(\bar{y}_4. - \bar{y}_5.) = \sqrt{MSE}\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + ... + \frac{C_5^2}{n_5}} = \sqrt{MSE}\sqrt{\frac{1}{n_4} + \frac{1}{n_5}}$

where, $\sqrt{MSE} = \sqrt{\frac{(n_1-1)s_1^2 + ... + (n_5-1)s_5^2}{N-5}}$ → (s_p for 5 groups)

- degrees of freedom = $N - 5$

b) Suppose you compare the average of the first 3 groups to the average of the last 2.

$$\gamma = \left(\frac{\mu_1 + \mu_2 + \mu_3}{3}\right) - \left(\frac{\mu_4 + \mu_5}{2}\right) = \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \frac{1}{2}\mu_4 - \frac{1}{2}\mu_5$$

Problem: In SPSS, coefficients (the C_i's) must be in decimal form and can have, at most, 3 decimal places. If this is not possible (as per above), multiply the linear combination by some constant to hold to this restriction. In this specific case,

$$6\gamma = 2(\mu_1 + \mu_2 + \mu_3) - 3(\mu_4 + \mu_5) = 2\mu_1 + 2\mu_2 + 2\mu_3 - 3\mu_4 - 3\mu_5$$

Enter these coefficients into SPSS, but note that

i) Value of contrast = $6g$ → Divide by 6 to get *g*.

ii) Standard error = $S.E.(6g)$

It can be shown (and verified from the S.E. formula) that
$S.E.(6g) = 6S.E.(g)$ → Divide by 6 to get $S.E.(g)$.

Additional Example: Suppose you were measuring the effect of 2 additives (A and B) in gasoline such that there are four blends: Control, Control + A, Control + B, Control + A and B. Define a linear combination of the four population means that describes the overall effect of additive B.

γ_B = effect of additive B
The effect of additive B when A is included is $\mu_4 - \mu_2$. The effect of additive B when A is not included is $\mu_3 - \mu_1$. The overall effect is the average of these two effects.

$$\gamma_B = \frac{(\mu_4 - \mu_2) + (\mu_3 - \mu_1)}{2} = -\frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 + \frac{1}{2}\mu_3 + \frac{1}{2}\mu_4$$

OR the "negative effect" of B: $\gamma_B = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3 - \frac{1}{2}\mu_4$

There are also linear combinations about *linear trends*. Please see p. 157, section 6.2.3.

Simultaneous Inferences or Multiple Comparisons (Unplanned Comparisons)

If some difference exists between I means, finding where it (they) exists requires two-by-two comparisons of all the means. (A.k.a. multiple comparisons.)

Def'n: The individual confidence level $(1 - \alpha_I)$ is the success rate of a single CI in capturing its true parameter. The individual error rate (α_I) is the probability of rejecting the null hypothesis when it is true (Type I error) for that one test.

The experiment-wise (or family-wise) confidence level $(1 - \alpha_E)$ is the overall success rate of several (m) CI all capturing their parameters. The experiment-wise (or family-wise) error rate (α_E) is the probability of rejecting the null hypothesis on at least one of several tests, when all null hypotheses are true.

Example: Consider a situation with three groups ($I = 3$) where we are interested in detecting any difference among their means. Their unknown means are μ_1, μ_2, and μ_3, respectively. Suppose we construct a 95% CI for each of

1. $\mu_1 - \mu_2$
2. $\mu_1 - \mu_3$
3. $\mu_2 - \mu_3$

Also, suppose it is known that $\mu_1 = \mu_2 = \mu_3$. If this were the case, all true differences would be zero. Therefore, in order for a CI to be considered successful, it should include zero.

What is the success rate that all intervals will contain zero? Or, what is the overall experiment-wise confidence level for the three intervals?

For any single 95% CI, there is a 95% success rate that the CI contains its true parameter. Conversely, there is a 0.05 error rate that the CI will *not* contain its true parameter (it will detect a difference when none exists). Thus, the overall success rate is about $0.95^3 = 0.857$ or 85.7%; and the experiment-wise error rate is roughly 0.143. Or, 14.3% of the time, we will see at least one difference among the 3 means, when in fact, no real differences exist. If $m = 10$, overall error rate increases to 0.401. This result is due to compound uncertainty, the chance of making *at least* one error. The experiment-wise confidence level, hence, needs to be fixed.

Methods to Fix Experiment-wise Confidence Level
Recall: A CI has the form

Estimate ± (Critical value) × *S.E.*(*Estimate*)

The critical value controls the level of confidence, so IT must be modified to fix the confidence level. For multiple comparisons, there are many procedures (defined by critical values or multipliers) that fix the experiment-wise confidence level. We will consider the following critical values (for m simultaneous confidence intervals with experiment-wise confidence level $(1 - \alpha_E)100\%$):

1. *Bonferroni's Procedure for Pairwise Comparisons*:
   Critical value: $t_{N-I,\alpha_E/2m}$
   This value gives an individual confidence level of $(1 - \alpha_E / m)$ to each interval. Thus, if $m = 10$, and $100(1 - \alpha_E)\% = 95\%$, then $\alpha_E = 0.05$ and this method would construct 10 simultaneous $100(1 - (0.05/10))\% = 99.5\%$ CIs. As a result, the critical value for each interval would be $t_{N-I,\,0.0025}$. The Bonferroni method ensures that the experiment-wise confidence level is *at least* $100(1 - \alpha_E)\%$.

2. *Tukey's W Procedure for Pairwise Comparisons*:
   Requires all group sizes to be equal.
   Critical value: $\dfrac{q_{\alpha_E}(I, N-I)}{\sqrt{2}}$
   where $q_{\alpha_E}(I, N-I)$ is a value off the Studentized Range distribution (Table A.5) with right tail probability α_E and degrees of freedom I and $N - I$.

3. *Scheffe's S Procedure for any linear contrast*:
   Critical value: $\sqrt{(I-1)F^{I-1}_{N-I,\alpha_E}}$

Other notes:
In SPSS, a '*' indicates a significant difference with family-wise confidence considered.
Planned Comparisons → Fix individual confidence, apply conventional methods (T-tools, CI)
Unplanned Comparisons → Fix experiment-wise confidence, modify the critical value in the CIs.

Extra-Sum-of-Squares *F*-Test

Using ANOVA to help introduce this general concept, the test compares two competing *models* for the population means. A *model* is just a structure or set of rules that defines how the *I* means are related to one another. Using ANOVA, what are the 2 models?

Null Hypothesis	Alternative Hypothesis
All means are equal	Not all means are equal → they are all *potentially* different
"One-mean" or "Equal-means" model	"*I*-mean" or "Separate-means" model

Abstractly again, the ESS *F*-test wishes to determine which "model" better describes the group structure of the means.

Def'n: Classify the two models as a reduced model (simpler, less complicated) and a full model (more complicated).

Full Model: A general model found to adequately describe the data. This is the model with the most parameters in the model for the means. The full model is considered the *alternative* hypothesis.

Reduced Model: A special case of the full model obtained by imposing some restriction. This is the model with the least parameters in the model for the means (as compared to the full model in question). The reduced model is considered the *null* hypothesis.

Thus, the ESS *F*-test is a test of

H_0: *reduced* model vs. H_a: *full* model

Similar to any hypothesis test, we assume the reduced model provides sufficient explanation for the data. Based on evidence gathered from samples, we will assess the strength of evidence against the reduced model in favour of the full model.

Residuals and SS Residuals

Comparing the models involves comparing their *sum-of-squared residuals* (or, a measure of unexplained variability). With each model, every observation is paired with an estimated group mean from that model. Thus,

Residual = observed value – estimated mean

Then, the *sum-of-squared residuals* is found by finding the residual of each observation, squaring it, then summing them all together. This concept leads to

SSR(*r*) – *sum-of-squared* residuals for the reduced model

SSR(*f*) – *sum-of-squared* residuals for the full model

Also, each model has its degrees of freedom, or *n* - # of unknown parameters in model. Denote *df*(*r*) and *df*(*f*) as the degrees of freedom for the reduced and full models, respectively.

Extra-Sum-of-Squares Principle
By adding more parameters to the reduced model, the SSR will always decrease. The more complicated full model will always have smaller SSR and "appear" to be a better fit; $SSR(r) > SSR(f)$. The ESS F-test determines if this reduction, $SSR(r) - SSR(f)$, is large enough to justify the full model. This difference is also known as the *Extra sum-of-squared residuals*, or *SSR*(Extra).
Note: If H_0 is true, then the 2 models should be equal in their ability to explain the data and thus, the SSRs should be about the same. If H_0 is incorrect, the $SSR(f)$ should be significantly smaller.
Converting *SSR*(Extra) to a known distribution to find p-values, it can be shown that

$$\frac{(SSR(r) - SSR(f))/(df(r) - df(f))}{SSR(f)/df(f)}$$

follows an F-distribution with numerator degrees of freedom of $df(r) - df(f)$ and denominator degrees of freedom of $df(f)$.

Summary
Again, we assume the reduced model is sufficient. Consider only the full model if there is significant evidence of providing a better fit. In other words, we assume the reduced is good enough unless the data show otherwise. Following the logic of hypothesis testing:

H_0: *reduced* model $\qquad$ H_a: *full* model

$$test - stat = F_0 = \frac{(SSR(r) - SSR(f))/(df(r) - df(f))}{SSR(f)/df(f)} \sim F_{df(f)}^{df(r)-df(f)}$$

$$p - value = P(F_{df(f)}^{df(r)-df(f)} > F_0)$$

Note: The difference in degrees of freedom, $df(r) - df(f)$, is also the number of additional parameters in the full model vs. the reduced model. In which case, the numerator can be thought of as the amount of reduction in variability ($SSR(r) - SSR(f)$) obtained by adding $df(r) - df(f)$ parameters. This is all compared to the variability in the full model.

ANOVA as an Extra-Sum-of-Squares *F*-Test
Going back to ANOVA for I means, the ANOVA F-test compares these 2 models:

The One-Mean Model (Reduced Model): This model says that all I population means are equal. In other words, there is one common mean to describe all of the I population means.
$\rightarrow \mu_1 = \mu_2 = \ldots = \mu_I = \mu$

The I-Mean Model (Full Model): This model says that all I means are potentially different. In other words, they are all not equal.
$\rightarrow \mu_1, \mu_2, \ldots, \mu_I$

Thus, H_0: (*reduced*) One-Mean model, $\mu_1 = \mu_2 = \ldots = \mu_I$
H_a: (*full*) I-Mean model, $\mu_1, \mu_2, \ldots, \mu_I$

Calculating SSR

1. Under H_0, all means are equal to a common mean, μ, so the estimated mean for any observation is the sample mean from all observations. The best estimate for μ is $\bar{y}..$, the overall sample mean. Let *SSR*(1) denote *SSR* for the One-Mean model under H_0. Hence,

$$SSR(r) = SSR(1) = TSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}..)^2 \qquad \text{(with } df = N - 1\text{)}$$

2. Under H_a, all means are *potentially* different, so the estimated mean for any observation is the sample mean of all observations within its group. The best estimate for μ_i is $\bar{y}_i.$, the sample mean for the i^{th} group. Let *SSR*(*I*) denote *SSR* for the *I*-Mean model under H_a. Hence,

$$SSR(f) = SSR(I) = SSE = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i.)^2 \qquad \text{(with } df = N - I\text{)}$$

Assessing evidence

Comparing the values (or, indirectly, models) requires testing the significance of the difference between them. An ESS *F*-test provides the process to do so.

$$test - stat = F_0 = \frac{(SSR(r) - SSR(f))/(df(r) - df(f))}{SSR(f)/df(f)}$$
$$= \frac{(SSR(1) - SSR(I))/(I-1)}{SSR(I)/(N-I)} \sim F_{N-I}^{I-1}$$

$$\rightarrow p - value = P(F_{N-I}^{I-1} > F_0)$$

Other notation

It can be shown that

$$SSR(\text{Extra}) = SSR(1) - SSR(I) = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}..)^2 - \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i.)^2$$
$$= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{y}_i. - \bar{y}..)^2 = \sum_{i=1}^{I} n_i(\bar{y}_i. - \bar{y}..)^2 = SST$$

If we define any SSR divided by its degrees of freedom as the mean square of residuals (MS), then

$$MS(I) = \frac{SSR(I)}{df(I)} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i.)^2}{N - I} = MSE$$

$$MS(\text{Extra}) = \frac{SSR(1) - SSR(I)}{df(1) - df(I)} = \frac{\sum_{i=1}^{I} n_i(\bar{y}_i. - \bar{y}..)^2}{I - 1} = MST$$

$$\rightarrow F_0 = \frac{MS(Extra)}{MS(I)} = \frac{MST}{MSE}$$

Thus, the text, approach, or software may change the notation, but the concept remains the same.