Lecture XI: Inference with Three or More Sample Means

X. Inference with Three or More Sample Means

1. Logic of ANOVA
2. Computation of ANOVA
3. Measures of association

Logic of ANOVA

- We have just encountered research problems in which we want to compare two sample means in order to make inferences about the two groups in the larger population. Now suppose that we want to compare means from three or more samples. In this case, we have a categorical independent with three or more categories, and a continuous dependent variable. When we have three or more sample means, the question we are interested in answering is this: Do the differences that we observe among the sample means indicate that there are significant differences across groups in the population? This is just a generalization of the two-sample case. There are several examples of such problems: sentence length as a function of offense type (violent, property, drug, other), fear of crime as a function of residential location (urban, suburban, rural), and offending as a function of race/ethnicity (white, black, Hispanic, Asian, other).

- One way to deal with this type of data is to carry out a series of two-sample hypothesis tests. An obvious disadvantage to this strategy is that the more groups that we have, the greater number of pairwise t-tests that we have to conduct. For example, say that we have an independent variable with three categories. This means that we would have to conduct (using the combination rule) $3!/2!(3-2)! = 3$ separate hypothesis tests. With four groups there are 6 hypothesis tests, with five groups there are 10 tests, with six groups there are 15 tests, with seven groups there are 21 tests, and so on. As you can see, the number of hypothesis tests quickly becomes unmanageable.

- A more substantive disadvantage to conducting multiple hypothesis tests with three or more samples is that, since each of the tests is conducted on the same data, they are not independent of one another. The practical implication of this is that over multiple tests, the probability of committing a type I error (i.e., the probability of falsely rejecting the null hypothesis) on any given test is greater than ??. As an example, suppose that we have three groups. We can determine using the binomial formula that the probability that we falsely reject the null hypothesis on at least one hypothesis test using a 5% criterion is $p(r > 0|n = 3) = 1 - \mathrm{C}_0^3(.05)^0(.95)^3 = .143$. If we increase the number of hypothesis tests, we find that $p(r > 0|n = 4) = .185$, $p(r > 0|n = 5) = .226$, $p(r > 0|n = 6) = .265$, $p(r > 0|n = 7) = .302$, and so on. Clearly, the probability that we would falsely reject in at least one hypothesis test becomes considerably greater than $\alpha = .05$.

- As you already know, we want a statistical test that will help us decide whether these observed differences are the result of sampling variation or (presumably) real differences in sentence length. Analysis of variance (ANOVA) is useful for determining the extent to which there are statistically significant differences between three or more sample means. We refer to ANOVA as a global test, which means that it tests the joint significance of several sample means, rather than differences among specific pairs. The advantage to conducting a global test is that the probability of committing a type I error is constant.

- Why is variance so important, when we are actually interested in comparing means? With ANOVA, we speak of two different kinds of variability: variability within and between groups. Let's consider each separately. Variability *within* groups refers to how tightly clustered individual scores are from their group mean. When this variability is small, each of the cases within a group cluster tightly around their respective group means, indicating that more of the cases are similar within a particular group than are different. Variability *between* groups (or across groups) refers to how tightly clustered the sample means are from each other, or from what we refer to as a *grand mean.* When the variability between groups is large, the group means are only loosely clustered around the grand mean, indicating that the group means are more different than they are similar.

- Now consider the ratio of between-group variability to within-group variability, and its implication for statistical inference. We refer to this as an F-ratio. When there is more variability within groups than between groups, the F-ratio will be less than one. This means that there is a great deal of overlap among the group distributions, and thus there is no relationship between group membership and the dependent variable. When there is more variability between groups than within groups, the F-ratio will be greater than one. This means that there is little or no overlap among the groups, and thus group membership is associated with the dependent variable.
- When we use ANOVA, we rely on a new probability distribution: the F-distribution. An F-ratio is simply the ratio of the variability between groups to the variability within groups. When we have a large F-ratio (i.e., one that is significantly greater than one), we will be led to reject the null hypothesis of no association between group membership and the outcome of interest.

Computation of ANOVA

- In ANOVA terminology, we are interested in a measure of variability called the sum of squared deviations about the mean, or simply the *sum of squares*. You might remember that the sum of squares is simply the numerator of the variance formula. A quick tutorial on notation: n refers to the sample size of a particular group, whereas N refers to the number of cases across all groups. There are three different sums of squares that we want to know in ANOVA.
  - **Total sum of squares**: This is the sum of the squared deviations of each case around the *grand mean.*
  $$SS_T = \sum (\mathrm{x}_{ik} - \overline{\mathrm{x}}_G)^2 = \sum \mathrm{x}_i^2 - N\overline{\mathrm{x}}_G^2$$
  - **Between-groups sum of squares**: This is the sum of the squared deviations of each group mean around the *grand mean.*
  $$SS_B = \sum n_k(\overline{\mathrm{x}}_k - \overline{\mathrm{x}}_G)^2 = \sum n_k\overline{\mathrm{x}}_k^2 - N\overline{\mathrm{x}}_G^2$$
  - **Within-groups sum of squares**. This is the sum of squared deviations of each case around its respective group mean.
  $$SS_W = \sum (\mathrm{x}_{ik} - \overline{\mathrm{x}}_k)^2 = \sum \mathrm{x}_i^2 - \sum n_k\overline{\mathrm{x}}_k^2 = SS_T - SS_B$$
  - The relationship among these three different sums of squares is straightforward.
  $$SS_T = SS_B + SS_W$$
- In addition to the sums of squares, we will also need to know our degrees of freedome:
  - **Total degrees of freedom**: $df_T = N - 1$
  - **Between-group degrees of freedom**: $df_B = k - 1$
  - **Within-group degrees of freedom**: $df_W = N - k$
- Using this information, we can partition the variance into the total variance, the between-group variance, and the within-group variance. The variance is computed as the sum of squares divided by the respective degrees of freedom. The F-statistic is calculated as the ratio of the between-group variance to the within-group variance.
- Let's carry out a full hypothesis test with the data from a sentence length example. Suppose that we have a sample of 40 offenders that have committed one of four types of offenses. Our independent variable is offense type, and our dependent variable is sentence length in months.
  - **Step 1: State hypotheses** - With ANOVA problerms, the alternatiove hypothesis is always $\mathrm{H}_1 : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$, and the null hypothesis is: $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. We make the assumption under the null hypothesis that the population means are equivalent. Under the alternative hypothesis, we make the assumption that at least one population mean is significantly different from at least one other population mean. ANOVA is what we call a *global test*, in that it can tell us if there are significant differences in the means, but it cannot tell us exactly which means are significantly different.
  - **Step 2: Obtain a probability distribution** - In order to do an ANOVA, we have to resort to the F-distribution. The F-distribution is one-tailed, because we square the deviations. The shape of the F-distribution is defined by two sets of degrees of freedom. The df in the numerator is $df_B = k - 1 = 3$ (top row of the F-table) and the df in the denominator is $df_W = N - k = 36$ (far left column of the F-table).
  - **Step 3: Make decision rules** - Let's use α=.01. The critical value of F is 4.31 (round up to 40 within-group degrees of freedome), thus we will reject the null hypothesis if F>4.31.

- **Step 4: Calculate the test statistic**.

Violent		Property		Drug		Other	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2
6	36	4	16	6	36	1	1
18	324	6	36	3	9	3	9
20	400	3	9	3	9	1	1
15	225	10	100	4	16	1	1
20	400	12	144	6	36	6	36
30	900	8	64	9	81	9	81
25	625	6	36	10	100	3	9
12	144	10	100	3	9	6	36
24	576	15	225	3	9	4	16
190	4030	82	794	49	309	36	194

- There are several pieces of information that we need to obtain to calculate the F-statistic. First, we need to compute the group means as well as the grand mean.
  - $\overline{x}_1 = \sum \mathrm{x}_{i1}/n_1 = 190/10 = 19$
  - $\overline{x}_2 = \sum \mathrm{x}_{i2}/n_2 = 82/10 = 8.2$
  - $\overline{x}_3 = \sum \mathrm{x}_{i3}/n_3 = 49/10 = 4.9$
  - $\overline{x}_4 = \sum \mathrm{x}_{i4}/n_4 = 36/10 = 3.6$
  - $\overline{x}_G = \sum \mathrm{x}_i/N = \dfrac{190+82+49+36}{10+10+10+10} = 8.925$
- Second, we need to compute the sums of squares. This step only requires us to find the total sum of squares and between-group sum of squares, which we can use to solve for the within-group sum of squares.
  - $SS_T = \sum \mathrm{x}_i^2 - N\overline{\mathrm{x}}_G^2 = (4030+794+309+194) - 40(8.9)^2 = 5327 - 3168.40 = 2158.6$
  - $SS_B = \sum \mathrm{n}_k\overline{\mathrm{x}}_k^2 - N\overline{\mathrm{x}}_G^2 = 10(19.0)^2 + 10(8.2)^2 + 10(4.9)^2 + 10(3.6)^2 - 40(8.9)^2 = 1483.7$
  - $SS_W = \sum \mathrm{x}_i^2 - \sum \mathrm{n}_k\overline{\mathrm{x}}_k^2 = SS_T - SS_B = 2158.6 - 1483.7 = 674.9$
- Third, we use this information to calculate the F-statistic. It is convenient to put ANOVA data into the form of a table.

Source	SS	df	$MS = SS/df$	$F = MS_B/MS_W$
Between groups	1483.7	$k-1=3$	494.57	$\dfrac{494.57}{18.75} = 26.38$
Within groups	674.9	$N-k=36$	18.75	
Total	2158.6	$N-1=39$	55.35	

- **Step 5: Make a decision about the null hypothesis** - Make a decision about the null hypothesis. Since $F = 26.39$, we reject the null hypothesis, and conclude that offense type is significantly associated with sentence length. Recall, however, that an F-test, since it is global, cannot tell us anything more substantive than this. We know that there are significant differences, but without conducting further tests (which we call post-hoc tests) we are unable to draw any further conclusions.

Let’s consider a couple more examples

- You collect data on fear of crime from a sample of 30 individuals, divided equally among urban, suburban, and rural areas. The research question is this: Is area of residence related to fear of crime?

Urban		Suburban		Rural	
x_U	x_U^2	x_S	x_S^2	x_R	x_R^2
22	484	23	529	19	361
29	841	22	484	24	576
31	961	26	676	24	576
28	784	25	625	19	361
30	900	24	576	20	400
32	1024	25	625	24	576
32	1024	24	576	21	441
31	961	24	576	17	289
28	784	27	729	23	529
30	900	23	529	19	361
293	8663	243	5925	210	4470

  - **Step 1: State hypotheses -** $\text{H}_1 : \mu_U \neq \mu_S \neq \mu_R$; $\text{H}_0 : \mu_U = \mu_S = \mu_R$
    - **Step 2: Obtain a probability distribution -** F-distribution, $df_B = 3 - 1 = 2$, $df_W = 30 - 3 = 27$
    - **Step 3: Make decision rules -** $\alpha = .05$’ $F_{crit} =$; reject H_0 if $F >$
    - **Step 4: Calculate the test statistic -**
      - $\overline{x}_1 = \sum \text{x}_{i1}/n_1 = 293/10 = 29.3$
      - $\overline{x}_2 = \sum \text{x}_{i2}/n_2 = 243/10 = 24.3$
      - $\overline{x}_3 = \sum \text{x}_{i3}/n_3 = 210/10 = 21.0$
      - $\overline{x}_G = \sum \text{x}_i/N = \dfrac{293 + 243 + 210}{10 + 10 + 10} = 24.87$
      - $SS_T = \sum \text{x}_i^2 - N\overline{\text{x}}_G^2 = (8663 + 5925 + 4470) - 30(24.87)^2 = 502.493$
      - $SS_B = \sum \text{n}_k\overline{\text{x}}_k^2 - N\overline{\text{x}}_G^2 = 10(29.3)^2 + 10(24.3)^2 + 10(21.0)^2 - 30(24.87)^2 = 344.293$
      - $SS_W = \sum \text{x}_i^2 - \sum \text{n}_k\overline{\text{x}}_k^2 = SS_T - SS_B = 502.493 - 344.293 = 158.2$

Source	SS	df	$MS = SS/df$	$F = MS_B/MS_W$
Between groups	344.293	$k - 1 = 2$	172.147	$\dfrac{172.147}{5.859} = 29.38$
Within groups	158.2	$N - k = 27$	5.859	
Total	502.493	$N - 1 = 29$	17.327	

Measures of Association

- While ANOVA can tell us whether there is a significant relationship between two variables, it cannot tell us anything about the strength of the relationship. We can, however, utilize what we know about variance to compute a measure of the strength of the association between the variables. We have available to us two measures of association called eta-square and epsilon-square, which are computed, respectively, as

  $$\eta^2 = \frac{SS_B}{SS_T}$$

  $$\epsilon^2 = 1 - \frac{MS_W}{MS_T} = 1 - \frac{SS_W/df_W}{SS_T/df_T}$$

  Both have an **explained variance** interpretation. They tell us what proportion of the total variance in the dependent variable is explained by the independent variable. Epsilon-square (ϵ^2) is a more conservative estimate, since it takes into account degrees of freedom.

- Let's consider our sentence length example. We compute the measures of association as

  $$\eta^2 = \frac{SS_B}{SS_T} = \frac{1483.7}{2158.6} = 0.687$$

  $$\epsilon^2 = 1 - \frac{MS_W}{MS_T} = 1 - \frac{SS_W/df_W}{SS_T/df_T} = 1 - \frac{18.75}{55.35} = 0.661$$

  We interpret these by saying that between 66.1% and 68.7% of the variance in sentence length is explained by offense type. Anything above 50% is generally a strong association.

- Let's compute these for the fear of crime example.

  $$\eta^2 = \frac{SS_B}{SS_T} = \frac{344.293}{502.493} = 0.685$$

  $$\epsilon^2 = 1 - \frac{MS_W}{MS_T} = 1 - \frac{SS_W/df_W}{SS_T/df_T} = 1 - \frac{5.859}{17.327} = 0.662$$

- Between 66.2% and 68.5% of the variance in fear of crime is explained by residential location.