STA 570 — Spring 2012 — Dr. Charnigo

Lecture 11

The two-way analysis of variance

Introduction. The one-way analysis of variance discussed in Lecture 10 allows us to compare populations characterized by a single factor. However, we may also wish to compare populations characterized by two factors.

For instance, we may wish to compare the following six populations with respect to mean full-scale IQ: males living extremely close to a lead smelter, males living fairly close to a lead smelter, males not living close to a lead smelter, females living extremely close to a lead smelter, females living fairly close to a lead smelter, and females not living close to a lead smelter. The two factors characterizing these six populations are proximity to a lead smelter and gender.

We could label the six population means as μ_1 through μ_6 and apply the one-way analysis of variance. However, detecting an interaction between proximity to a lead smelter and gender (i.e., attempting to discover that lead exposure affects males differently than females) through the one-way analysis of variance would be cumbersome. Instead, let us label the six population means as μ_{11}, μ_{21}, μ_{31}, μ_{12}, μ_{22}, and μ_{32}, where the first subscript indicates proximity to the lead smelter (1 = extremely close, 2 = fairly close, 3 = not close) and the second subscript indicates gender (1 = male, 2 = female).

Thinking of the six populations in this manner will allow us to perform the two-way analysis of variance, which can be used to detect an interaction between proximity to a lead smelter and gender. Moreover, the two-way analysis of variance will allow us to detect "main effects" for lead exposure as well as main effects for gender.

Notation and statistical model. Let a denote the number of levels of the first factor and b the number of levels of the second factor. Considering our example from the Introduction, if proximity to a lead smelter is the first factor and gender is the second factor, then we have $a = 3$ and $b = 2$ since there are three possible designations for proximity to a lead smelter and two possible genders.

Let n_{ij} denote the number of observations in our data set that are at both the i^{th} level of the first factor and the j^{th} level of the second factor. So, for instance, n_{32} is the number of observations in our data set for females not living near a lead smelter. If $n_{11} = \cdots = n_{ab}$, then we let n denote the common sample size and refer to the data as "balanced". Otherwise, we refer to the data as "unbalanced".

Let Y_{ijk} denote the random conceptualization of the measurement for the k^{th} person in our data set who is at the i^{th} level of the first factor and the j^{th} level of the second factor. Thus, for instance, Y_{325} pertains to the 5^{th} female not living near a lead smelter. Our statistical model is (Equation 12.22)

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where the ϵ_{ijk} are independent normal random variables with mean 0 and unknown variance σ^2. We identify $\mu + \alpha_i + \beta_j + \gamma_{ij}$ with μ_{ij}, the population mean at level i of the first factor and level j of the second factor. This prescription implies a belief that all populations are normal and share a common variance σ^2.

In addition, we must place constraints on the α_i, β_j, and γ_{ij}. For balanced data, these constraints are $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0$. Once the constraints are in place, we can address the following questions:

- Are any of the "interaction effects" $\gamma_{11}, \ldots, \gamma_{ab}$ nonzero? Nonzero interaction effects imply that the impact of the first factor depends on the level of the second factor.

• Are any of the "main effects" for the first factor $\alpha_1, \ldots, \alpha_a$ nonzero? Nonzero main effects for the first factor imply that the first factor has a nontrivial overall impact (overall in the sense of averaging over levels of the second factor or, in the absence of interaction, applying at all levels of the second factor).

• Are any of the main effects for the second factor $\beta_1, \ldots, \beta_b$ nonzero? Nonzero main effects for the second factor imply that the second factor has a nontrivial overall impact (overall in the sense of averaging over levels of the first factor or, in the absence of interaction, applying at all levels of the first factor).

Sums of squares. Let $\bar{y}_{..}$ denote the mean across all ab samples; let $\bar{y}_{i.}$ denote the mean for individuals sampled at level i of the first factor; let $\bar{y}_{.j}$ denote the mean for individuals sampled at level j of the second factor; and, let $\bar{y}_{ij}$ and s^2_{ij} denote the mean and variance for individuals sampled at level i of the first factor and level j of the second factor.

If we have balanced data, then we can decompose a gross measure of variability into pieces related to main effects for the first factor, main effects for the second factor, interaction effects, and dissimilarities within groups:

$$\begin{aligned} SST &:= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{..})^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{.j}-\bar{y}_{..})^2 + \\ &\quad \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij})^2 \\ &=: SSA + SSB + SSAB + SSE \end{aligned}$$

If we have unbalanced data, then SST and SSE remain the same. However, if we have unbalanced data, then there are several ways in which we can define SSA, SSB, and $SSAB$.

The "Type I" sums of squares preserve the equality $SST = SSA+SSB+SSAB+SSE$, but they are ambiguous in the sense that the order in which the two factors are specified matters. The "Type III" sums of squares do not preserve the equality $SST = SSA+SSB+SSAB+SSE$ with unbalanced data, but they are unambiguous. Your textbook author prefers the Type III sums of squares, and I agree with this preference.

Computational formulas for desk calculators, balanced data. If you have balanced data, then sums of squares may be computed as follows:

$$SST = \sum_{i=1}^{a} \sum_{j=1}^{b} \left[(n-1)s_{ij}^2 + n\bar{y}_{ij}^2 \right] - abn\bar{y}_{..}^2$$

$$SSA = bn \sum_{i=1}^{a} \bar{y}_{i.}^2 - abn\bar{y}_{..}^2$$

$$SSB = an \sum_{j=1}^{b} \bar{y}_{.j}^2 - abn\bar{y}_{..}^2$$

$$SSAB = n \sum_{i=1}^{a} \sum_{j=1}^{b} \bar{y}_{ij}^2 - abn\bar{y}_{..}^2 - (SSA + SSB)$$

$$SSE = SST - (SSA + SSB + SSAB)$$

Using SAS (or other software) to compute sums of squares is preferable for balanced data but essential for unbalanced data.

Performing the hypothesis tests. Let N denote the total number of subjects across all ab samples. For balanced data, we have $N = abn$.

Define $MSA := SSA/(a-1)$, $MSB := SSB/(b-1)$, $MSAB := SSAB/((a-1)(b-1))$, and $MSE := SSE/(N-ab)$.

We let

$$f_{AB} := MSAB/MSE$$

and reject $H_0 : \gamma_{ij} = 0$ for all (i, j) if $f_{AB} > f_{((a-1)(b-1)),(N-ab),1-\alpha}$.

We let

$$f_A := MSA/MSE$$

and reject $H_0 : \alpha_i = 0$ for all i if $f_A > f_{(a-1),(N-ab),1-\alpha}$.

We let

$$f_B := MSB/MSE$$

and reject $H_0 : \beta_j = 0$ for all j if $f_B > f_{(b-1),(N-ab),1-\alpha}$.

Example (performing the hypothesis tests). Refer to page 6 of {ANOVAExamples.pdf}. We have $N = 124$, $(a-1) = 2$, $(b-1) = 1$, and $(a-1)(b-1) = 2$. Also, we have $SSA = 526.57$, $SSB = 12.617$, $SSAB = 550.33$, $SSE = 24585.37$, $SST = 25519.19$.

Thus, we have $MSAB = 275.16$, $MSE = 208.35$, and $f_{AB} = 1.32$. The corresponding p-value is 0.2709, and we cannot reject $H_0 : \gamma_{ij} = 0$ for all (i, j) at level $\alpha = 0.05$. Similarly, we cannot reject the null hypotheses of zero main effects for proximity and zero main effects for gender.

In summary, the combination of proximity to a lead smelter and gender does not seem to affect mean full-scale IQ. Of course, we must be cognizant that non-significant findings may represent Type II errors.

The Kruskal-Wallis test

Introduction. The one-way analysis of variance in Lecture 10 requires us to assume normality for the populations characterized by the one factor of interest. However, this assumption is not always reasonable. Sometimes a transformation of the response variable can alleviate the problem (Cf. Lecture 8), but if not we would like to have a generalized version of the rank-sum test (Cf. Lecture 9) that could accommodate more than two populations. The Kruskal-Wallis test meets this need.

Scenario. We obtain measurements of an ordinal or a cardinal variable for n_1 subjects under one treatment, n_2 different subjects under a second treatment, n_3 different subjects under a third treatment, and so forth. Let k denote the total number of treatments and $N := n_1 + \cdots + n_k$ the total number of subjects across all k samples. Let $\Delta_1, \ldots, \Delta_k$ denote the population medians corresponding to the k treatments. We wish to test $H_0 : \Delta_1 = \cdots = \Delta_k$ against the complementary alternative.

Rank assignment. We begin by assigning ranks to the observations as illustrated in Tables 12.17 and 12.18. These data arise from the experiment described in Example 12.23 on page 600. Briefly, a study was conducted to compare the anti-inflammatory effects of four different treatments (indomethicin, aspirin, piroxicam, BW755C). Six rabbits were assigned to each treatment. Each rabbit received a score between -3 and $+3$, a higher score indicating greater effectiveness of the treatment.

There were no rabbits with a score of -3 or -2. There was one rabbit with a score of -1, so this rabbit receives the rank of 1. There were five rabbits with a score of 0. These rabbits could be ranked from 2 to 6, but there is no reason to rank one of the five rabbits ahead of the others; hence, all five rabbits get assigned the average rank of 4.0. Similarly, the five rabbits with a score of +1 get assigned the average rank of 9.0, the four with a score of +2 get assigned the average rank of 13.5, and the nine with a score of +3 get assigned the average rank of 20.0.

Performing the test. Let r_i denote the sum of ranks for sample i. The Kruskal-Wallis test statistic is

$$\frac{\frac{12}{N(N+1)} \times \sum_{i=1}^{k} \frac{r_i^2}{n_i} - 3(N+1)}{1 - \frac{\sum_j (t_j^3 - t_j)}{N^3 - N}},$$

where the t_j correspond to the entries in the "Frequency" column of Table 12.18. In the present example, the denominator summation runs from $j = -1$ to $j = +3$ with $t_{-1} = 1$, $t_0 = 5$, $t_{+1} = 5$, $t_{+2} = 4$, and $t_{+3} = 9$.

With not-too-small samples ($n_1, \ldots, n_k \geq 5$), we may reject H_0 if the test statistic exceeds $\chi^2_{k-1,1-\alpha}$ (Equation 12.26).

Example (performing the test). Example 12.24 on pages 601 through 603 shows how the Kruskal-Wallis test is carried out for the study on anti-inflammatory treatments. We obtain $r_1 = 97.5$ by adding the average ranks of 13.5, 20.0, 20.0, 20.0, 20.0, and 4.0 for the six rabbits receiving the first treatment (indomethicin). Similarly, we obtain $r_2 = 85.0$, $r_3 = 91.5$, and $r_4 = 26.0$. Noting that $N = 24$, we can now calculate the test statistic. Arithmetic details are shown on page 603. The end result is 11.804, which exceeds $\chi^2_{3,0.95} = 7.81$. Thus, we reject H_0 at level $\alpha = 0.05$. This decision is confirmed by the p-value 0.0081 in the "Kruskal-Wallis Test" box on page 4 of {NONPARExamples.pdf}. We conclude that not all of the anti-inflammatory treatments are equally effective.

Follow-up testing. If we reject $H_0 : \Delta_1 = \cdots = \Delta_k$, then we can perform follow-up tests concerning pairs of medians. This can be accomplished with the Dunn procedure (Equation 12.28). As with follow-up tests concerning pairs of means in the one-way analysis of variance, we may wish to employ a Bonferroni adjustment. You will not be asked to carry out the Dunn procedure in STA 570, but you are responsible for knowing its purpose.

Correlation

Introduction. So far this semester, we have not discussed methods for analyzing the relationship between two continuous variables. Today I will present a method that can be used if there is not a clear distinction between variables, in the sense that no obvious choice of response (dependent) variable and explanatory (independent) variable exists. Our goal will be simply to quantify the strength and direction of the (linear) relationship between variables. Another approach, useful if there is a clear distinction between variables (especially if we wish to predict one variable from the other), will be presented in Lecture 12.

Covariance. Let X and Y be two continuous variables. We define the (population) covariance between X and Y as (Definition 5.12)

$$Cov[X, Y] := E[(X - E[X])(Y - E[Y])].$$

If X and Y tend to be above their averages at the same time and below their averages at the same time, then $(X - E[X])(Y - E[Y])$ tends to be positive, so that the covariance is positive. [Here I am using the fact that a usually positive quantity has a positive expected value.] On the other hand, if X tends to be above its average when Y is below its average and Y tends to be above its average when X is below its average, then $(X - E[X])(Y - E[Y])$ tends to be negative, so that the covariance is negative.

We can estimate the population covariance from data. The sample covariance is

$$(n-1)^{-1} \sum_{i=1}^{n} \{(x_i - \bar{x})(y_i - \bar{y})\}.$$

Correlation. A difficulty with the covariance lies in the interpretation of its magnitude. A covariance of 9, for instance, may be either quite large (if the variances of X and Y are equal to, say, 10) or quite small (if the variances of X and Y are equal to, say, 1000). Thus, we would like to rescale the covariance so that its magnitude could be interpreted more easily.

The (population) correlation between X and Y is (Definition 5.13)

$$\rho := \frac{Cov[X, Y]}{SD[X] SD[Y]}.$$

The correlation must fall between -1 and $+1$. A correlation close to 1 in absolute value implies a strong (linear) relationship between X and Y, even if the covariance does not seem large; a correlation close to 0 implies a weak (linear) relationship, even if the covariance does not seem small.

We can estimate the population correlation from data. The sample or "Pearson" correlation is

$$r := \frac{(n-1)^{-1} \sum_{i=1}^{n} \{(x_i - \bar{x})(y_i - \bar{y})\}}{s_x s_y}$$

and can be computed via the formula $L_{xy}/\sqrt{L_{xx} L_{yy}}$, where

$$L_{xx} := \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 /n, \quad L_{xy} := \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right) /n,$$

and

$$L_{yy} := \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 /n.$$

Confidence interval. Assuming that X and Y are jointly normally distributed, a $100(1-\alpha)\%$ confidence interval for the population correlation ρ is (Equation 11.23, simplified)

$$\left[\frac{(1+r)\exp\left[-z_{1-\alpha/2}\sqrt{\frac{4}{n-3}}\right]-(1-r)}{(1+r)\exp\left[-z_{1-\alpha/2}\sqrt{\frac{4}{n-3}}\right]+(1-r)}, \frac{(1+r)\exp\left[z_{1-\alpha/2}\sqrt{\frac{4}{n-3}}\right]-(1-r)}{(1+r)\exp\left[z_{1-\alpha/2}\sqrt{\frac{4}{n-3}}\right]+(1-r)}\right].$$

Moreover, we can test $H_0 : \rho = \rho_0$ versus $H_1 : \rho \neq \rho_0$ at level α by rejecting H_0 if and only if ρ_0 does not appear in the $100(1-\alpha)\%$ confidence interval.

Example (confidence interval). Suppose that the Pearson correlation based on a sample of size $n = 500$ is only 0.10. Can we reject the null hypothesis that $\rho = 0$ at level $\alpha = 0.05$? To find out, we construct the 95% confidence interval

$$\left[\frac{(1+0.10)\exp\left[-1.96\sqrt{\frac{4}{497}}\right]-(1-0.10)}{(1+0.10)\exp\left[-1.96\sqrt{\frac{4}{497}}\right]+(1-0.10)}, \frac{(1+0.10)\exp\left[1.96\sqrt{\frac{4}{497}}\right]-(1-0.10)}{(1+0.10)\exp\left[1.96\sqrt{\frac{4}{497}}\right]+(1-0.10)}\right],$$

which simplifies to $[0.012, 0.186]$. Since the 95% confidence interval excludes 0, we reject the null hypothesis that $\rho = 0$ at level $\alpha = 0.05$.