Lecture 33

Nancy Pfenning Stats 1000

Chapter 16: Analysis of Variance

Last time, we wanted to test if the difference among 3 observed mean test scores—82, 66, and 60—could be easily enough attributed to chance variation:

$H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_a: not all the μ_i are equal. To express H_a with mathematical notation would be too awkward: one would have to write

$$H_a : \mu_1 = \mu_2 \neq \mu_3 \quad \text{or} \quad \mu_1 = \mu_3 \neq \mu_2 \quad \text{or} \quad \mu_2 = \mu_3 \neq \mu_1 \quad \text{or} \quad \mu_1 \neq \mu_2 \neq \mu_3$$

Our test statistic F is the ratio of variation *among* means MSG to variation *within* groups MSE. If this ratio is large, we have evidence that the means differ and we will reject H_0. An **ANOVA Table** organizes the calculations needed to perform the F test. For our exam problem, we have $I = 3$ groups, $N = 20$ observations. "Source" refers to source of variation and *df* are the degrees of freedom.

Source	*df*	Sum of Squares	Mean Sum of Squares	F	P-value
Group	$DFG = I - 1 = 2$	$SSG = 1720$	$MSG = \frac{1720}{2} = 860$	$\frac{MSG}{MSE} = 4.5$	in (.025, .05)
Error	$DFE = N - I = 17$	$SSE = 3245$	$MSE = \frac{3245}{17} = 191$		
Total	$N - 1 = 19$				

Under the null hypothesis of equal population means, the F statistic has a distribution with $I - 1$ *df* in the numerator and $N - I$ *df* in the denominator; in this example, it is $F(2, 17)$. The P-value is the probability that, assuming the null hypothesis is true, an $F(2, 17)$ R.V. would take a value at least as large as the one observed:

$$\text{P-value} = P(F \geq 4.5)$$

Consulting the F tables, page 586 shows F critical values for 2 *df* in the numerator, and 15 or 20 in the denominator. To be conservative, we will use 15 (slightly smaller critical values make it a little more difficult to reject the null hypothesis).

Since 4.5 is between 3.68 and 4.77, the P-value is between .025 and .050. This is, in general, small enough to reject H_0. We conclude that the difference in observed mean scores is unlikely to be a result of chance variation; rather, we have evidence that the three exams did not share the same level of difficulty.

In this course, we take the analysis no further. In practice, more detailed comparisons called *contrasts* can be made to pinpoint which means differ. For example, we may be able to show that there is only a significant difference between the first and the second two, not between those two—in other words, maybe the first exam was less difficult and the other two were comparable.

Example

Check if mean earnings could be equal for all 1st, 2nd, 3rd, 4th, and "other" year Pitt students.

Exercise: Compare values of a quantitative survey variable for more than two categorical groups by carrying out an ANOVA test in MINITAB. State your conclusions in terms of the particular variables chosen.

Lecture 34

Chapter 15: More About Categorical Variables

Example

Results of a labor survey in March 1988 for an SRS of 914 California men aged 35 to 44 are shown below:

⇓ Married?	Employed	Unemployed	Total	Proportion Employed
Currently	638	27	665	$\hat{p}_1 = \frac{638}{665} = .959$
Previously	133	8	141	$\hat{p}_2 = \frac{133}{141} = .943$
Never	102	6	108	$\hat{p}_3 = \frac{102}{108} = .944$
Total	873	41	914	overall $\hat{p} = \frac{873}{914} = .955$

We call this a 3×2 Table: 3 possibilities for the row variable (marital status) and 2 possibilities for the column variable (employment status). A natural assignment would be marital status as explanatory variable, employment as response. If we were only interested in deciding if there were a significant difference between the proportion of currently married men who were employed and the proportion of previously married men who were employed, a two-sample z test could be used on the difference in population proportions $p_1 - p_2$, as our textbook covered in Chapters 12 and 13. Here, since we want to compare more than 2 proportions at a time, we need to do a **chi-square** test.

Notice that the sample proportion employed is somewhat higher for currently married men than for previously or never married men. Could this just be chance variation when sampling from a population where the proportions are actually equal, or do we have evidence of a significant difference among proportions, indicating a relationship between marital status and employment?

We test $H_0 : p_1 = p_2 = p_3$ [equivalent to saying that there is no relationship between marital and employment status] against the "many-sided" alternative H_a : not all three proportions are equal [equivalent to saying that there *is* a relationship between marital and employment status]. Note: It would be incorrect to express H_a as $p_1 \neq p_2 \neq p_3$. In fact, to express it mathematically, we would need to write

$$H_a : p_1 = p_2 \neq p_3 \quad \text{or} \quad p_1 = p_3 \neq p_2 \quad \text{or} \quad p_2 = p_3 \neq p_1 \quad \text{or} \quad p_1 \neq p_2 \neq p_3$$

To test H_0, we check if the *observed* counts in each "cell" are reasonably close to what we'd *expect* under H_0, i.e., if there were no difference in underlying population proportions.

Altogether, $\frac{873}{914}$ of the men were employed, so we'd expect $\frac{873}{914}$ of the 665 currently married men to be employed: *expected* count for this cell is $\frac{873*665}{914} = 635$.

Similarly, we'd expect $\frac{873}{914}$ of the 141 previously married men to be employed: *expected* count is $\frac{873*141}{914} = 135$.

Likewise, we'd expect $\frac{873}{914}$ of the 108 never married men to be employed: *expected* count is $\frac{873*108}{914} = 103$.

Since $\frac{41}{914}$ of the men were unemployed, our expected counts for unemployed men in each marital category would be

$\frac{41*665}{914} = 30$ currently married men unemployed

$\frac{41*141}{914} = 6$ previously married men unemployed

$\frac{41*108}{914} = 5$ never married men unemployed

Note the pattern: each

$$\text{expected count} = \frac{\text{row total} * \text{column total}}{\text{overall total}}$$

To measure how far the observed counts as a group tend to be from the expected counts, we calculate the chi-square test statistic

$$X^2 = \Sigma_{\text{all cells}} \frac{(\text{observed minus expected})^2}{\text{expected}}$$

A chi-square R.V. is always positive. Its density curve is skewed more or less to the right, depending on how many cells are involved: We measure its degrees of freedom $df = (r-1)\times(c-1)$ where r is the number of rows, c is the number of columns. We look at tables of observed and expected counts in order to compute the value of X^2:

Observed:

638	27
133	8
102	6

Expected:

635	30
135	6
103	5

$$X^2 = \frac{(638-635)^2}{635} + \frac{(133-135)^2}{135} + \frac{(102-103)^2}{103} + \frac{(27-30)^2}{30} + \frac{(8-6)^2}{6} + \frac{(6-5)^2}{5}$$
$$= .014 + .030 + .010 + .300 + .667 + .200 = 1.221$$

Table A.5 shows the probability p that a chi-square R.V. with given df would take a value at least as large as the one observed: The P-value is the probability of a chi-square R.V. at least as large as the chi-square statistic. For $(3-1)*(2-1) = 2$ df, 1.221 is less than 1.39, so the P-value is greater than .50. The differences among these three sample proportions are easily attributed to chance alone: more than 50% of the time, random samples of this size taken from a population with equal proportions of employed men among currently, previously and never married men would result in sample proportions this different. In other words, there is no compelling evidence of a relationship between marital status and employment for the men studied.

Note that there are two interpretations of the chi-square hypotheses: the null hypothesis H_0 : $p_1 = p_2 = \cdots$ may be expressed as "the population proportions are equal" or as "there is no relationship between the row and column variables".

Also note: Another study by other researchers checked for a relationship between marital status and job grade. This other study rejected H_0 with a P-value less than .0005, concluding there *is* a relationship between job grade and marital status. Our example was *stratified* to control for age (results were given for men in the specific age bracket 35 to 44). The other study was not and so event though the p-value was very small, it did not prove causation—maybe it's because single men tend to be younger and less advanced. Studying different age groups separately in our own example helped control for the confounding variable age in this study.

Bar graphs can be used to display the information in the two-way table. Taking marital status as the explanatory variable, it should be graphed horizontally, and proportions employed—.959 vs. .041, .943 vs. .057, .944 vs. .056.—would be represented as bars of corresponding heights.

Taking employment status as the explanatory variable, it would be graphed horizontally, and proportions married, divorced, or single would be displayed with bars of heights .73 vs. .15 vs. .12 for employed, and .66 vs. .20 vs. .15 for unemployed.

Extra Credit: Pick two categorical variables from our survey. Decide which should be explanatory (row variable) and which response. Use MINITAB to compare conditional percentages in each row [explanatory variable must be entered before response] and carry out a chi-squared test for a relationship. (Survey data available on my website `http://www.pitt.edu/ nancyp/stat-0200/index.html` under `surveymmddyy.txt`.)

Lecture 35

Relationships Between Two Categorical Variables (More Examples)

Recall: To test $H_0 : p_1 = p_2 = \cdots = p_r$ for proportions falling in a particular category in an $r \times 2$ table, or more generally to test for a relationship between the row and column variables in an $r \times c$ table, our chi-square statistic X^2 measures the overall difference between observed counts and counts we would expect if population proportions were equal, or if no relationship existed between row and column variables. A test of significance is carried out with the following steps:

1. Set up the null hypothesis of equal proportions/no relationship and the alternative of unequal proportions/a relationship.
2. Calculate each
$$\text{expected count} = \frac{\text{column total} * \text{row total}}{\text{overall total}}$$
3. Calculate the chi-square statistic
$$X^2 = \Sigma_{\text{all cells}} \frac{(\text{observed minus expected})^2}{\text{expected}}$$
4. Find $df = (r-1) \times (c-1)$.
5. Use Table A.5 to check if the chi-square statistic is too large to permit us to attribute the differences to chance alone: Locate the P-value, which is the probability of a chi-square R.V. with the given df taking a value at least as large as the observed statistic. As usual, H_0 is rejected for "small" P-values. In other words, a small P-value is evidence of unequal proportions, or of a relationship between row and column variables.

Example

In Chapter 6 we constructed a two-way table (also called a contingency table) showing the relationship between gender (row variable) and lenswear (column variable) of statistics class members during a previous semester. We noticed that roughly equal proportions of males and females wore corrective lenses of some type. Males tended to wear glasses and females tended to wear contacts. Overall, do the data convince us that gender and lenswear are related?

Observed	None	Glasses	Contacts	Total
Male	65	36	37	138
Female	110	32	91	233
Total	175	68	128	371

1. H_0: gender and lenswear are not related.
   H_a: gender and lenswear are related.
2. Here is a table of counts expected if H_0 were true:

Expected	None	Glasses	Contacts
Male	$\frac{175*138}{371} = 65$	$\frac{68*138}{371} = 25$	$\frac{128*138}{371} = 48$
Female	$\frac{175*233}{371} = 110$	$\frac{68*233}{371} = 43$	$\frac{128*233}{371} = 80$

3. The chi-square statistic is
$$\frac{(65-65)^2}{65} + \frac{(110-110)^2}{110} + \frac{(36-25)^2}{25} + \frac{(32-43)^2}{43} + \frac{(37-48)^2}{48} + \frac{(91-80)^2}{80}$$
$$= 0 + 0 + 4.84 + 2.81 + 2.52 + 1.51 = 11.68$$

4. $df = (2-1)(3-1) = 2$
5. On Table A.5 in the 2 df row, 11.68 is between 10.60 and 13.82, so the P-value is between .005 and .001. Since it is so small, we have very strong evidence of a relationship between gender and lenswear. The P-value is small because the chi-square statistic is large; the chi-square statistic is large not because of the need for lenses (the observed counts not needing corrective lenses are identical to what we'd expect under H_0) but because of the type of lenses worn. Specifically, males wear glasses more and females wear contacts more.

Example

Is there a relationship between whether or not a student eats breakfast, and whether he/she lives on or off campus? Which should be the explanatory variable and which should be response? Do you have reason to expect a chi-square test to show statistical significance?

Living on or off campus would be the natural choice for explanatory variable, and eating breakfast would be the response. Some might expect on-campus students to be more likely to eat breakfast, because it is available in the cafeteria with no preparation required. On the other hand, off campus students may find it easier to eat breakfast just by walking into their kitchen, as opposed to making the trip to the cafeteria...

Here are data for a large group of students:

Observed	Breakfast	No Breakfast	Total
Off Campus	98	125	223
On Campus	101	121	222
Total	199	246	445

The proportions eating breakfast were $\frac{98}{223} = .439$ for off-campus students, $\frac{101}{222} = .455$ for on-campus students. The difference seems slight, but there could conceivably be a subtle but significant difference for a large sample like this, so we will carry out a formal test:

1. H_0: no relationship between living on or off campus and eating breakfast or not
   H_a: there is a relationship between living on or off campus and eating breakfast or not
2. Here is a table of counts expected if H_0 were true:

Expected	Breakfast	No Breakfast
Off Campus	$\frac{199*223}{445} = 100$	$\frac{246*223}{445} = 123$
On Campus	$\frac{199*222}{445} = 99$	$\frac{246*222}{445} = 123$

3. The chi-square statistic is

$$\frac{(98-100)^2}{100} + \frac{(101-99)^2}{99} + \frac{(125-123)^2}{123} + \frac{(121-123)^2}{123} = .04 + .04 + .03 + .03 = .14$$

4. $df = (2-1)(2-1) = 1$
5. On Table A.5 in the 1 df row, .14 is smaller than .45 so the P-value is larger than .50. There is no statistical evidence at all of a relationship. The data indicate that eating breakfast or not may well have nothing to do with living on or off campus.

Example

Is there a relationship between whether or not a student smokes, and whether he/she lives on or off campus? Which should be the explanatory variable and which should be response? Do you have reason to expect a chi-square test to show statistical significance?

In this case smoking would be the natural choice for explanatory variable, and living on or off campus would be the response. Because of restrictions on smoking in dormitories, it is reasonable to expect that smokers may tend to live off campus more than non-smokers.

Here are data for a large group of students:

Observed	Non-Smoker	Smoker	Total
Off Campus	162	61	223
On Campus	198	24	222
Total	360	85	445

The proportions living off campus were $\frac{162}{360} = .45$ for non-smokers, $\frac{61}{85} = .72$ for smokers. The difference seems substantial, but the only way to confirm statistical significance is to carry out a formal test:

1. H_0: no relationship between living on or off campus and smoking
   H_a: there is a relationship between living on or off campus and smoking
2. Here is a table of counts expected if H_0 were true:

Expected	Non-Smoker	Smoker
Off Campus	$\frac{360*223}{445} = 180$	$\frac{85*223}{445} = 43$
On Campus	$\frac{360*222}{445} = 180$	$\frac{85*222}{445} = 42$

3. The chi-square statistic is

$$\frac{(162-180)^2}{180} + \frac{(198-180)^2}{180} + \frac{(61-43)^2}{43} + \frac{(24-42)^2}{42} = 1.8 + 1.8 + 7.5 + 7.6 = 18.7$$

4. $df = (2-1)(2-1) = 1$
5. On Table A.5 in the 1 df row, 18.7 is larger than 10.83, so the P-value is smaller than .001. There is strong statistical evidence of a relationship. The data indicate that smokers are significantly more likely to live off campus.

Exercise: Pick two categorical variables from our survey. Decide which should be explanatory (row variable) and which response. Use MINITAB to compare conditional percentages in each row [explanatory variable must be entered before response] and carry out a chi-square test for a relationship. Use Table A.5 to give a range for the P-value.