QUAN 102 Statistics for Business Tutorial 11

Natt Krongkajonsook

11.1 Chi-Square (χ^2) Distribution

- $Z_1^2 + Z_2^2 + ... + Z_n^2 \sim \chi_n^2$, where $Z_1, Z_2, \ldots, Z_n$ are independent $N(0, 1)$ variables.
- The χ^2 distribution is non-negative and non-symmetrical.
- The degree of freedom measures the degree of variation. As the degree of freedom increases, the degree of variation and the asymmetry of the χ^2 distribution decreases.
- Since the χ^2 distribution is non-symmetrical, the method for looking up left-tail values is different from the method for looking up right tail values.
  - Area to the right - just use the area given and look up this area in the table.
  - Area to the left - the table provides the area to the right, so subtract the given area from 1 and look up the difference in the table.
  - Area in both tails - divide the area by two. Look up this area for the right critical value and 1 minus this area for the left critical value.

11.2 Goodness of Fit Test

- We test if the sample comes from the population with the claimed distribution. Another way of looking at this is to test if the frequency distribution fits a specific pattern. Since the frequency distribution and the probability distribution have the same properties, it is equivalent to test whether the probability distribution fits a specific pattern.
  - In Exercise 11.1, a die is thrown once, and X is the number that it will turn up. We test $H_0 : P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6$ against H_1 : at least one differs.
  - In Exercise 11.4, let X be the number of sites at which tuatara were found. We test $H_0 : X \sim Bin(3, .4)$ against $H_1 : X \nsim Bin(3, .4)$. Equivalently, we test $H_0 : P(X = 0) = .216, P(X = 1) = .432, P(X = 2) = .288, P(X = 3) = .064$ against H_1 : at least one differs.
- Two values are involved, an observed value, which is the frequency of a category from a sample, and the expected frequency, which is calculated based upon the claimed distribution. In particular, the expected frequency in category i (e_i) can be calculated as
$$e_i = n \times p_i^{H_0}$$
where $p_i^{H_0}$ is the probability for category i obtained from H_0, and n is the sum of all frequencies or the total number of observations in the sample. We do not need to calculate the observed frequency in category i (o_i) since it will be given.

  - In Exercise 11.1, $n = 100$, $o_i = \{19, 12, 11, 16, 20, 22\}$, and $e_i = 100 \times 1/6 = 16.67$.
  - In Exercise 11.4, $n = 24$, $o_i = \{7, 8, 6, 3\}$, and $e_i = \{5.184, 10.368, 6.912, 1.536\}$.

- If the claim is likely to be true, then the observed frequency should be really close to the claimed (expected) frequency, and therefore the square of the deviations will be small. The square of the deviation will be divided by the expected frequency to weight frequencies. A difference of 10 may be very significant if 12 was the expected frequency, but a difference of 10 will not be significant at all if the expected frequency was 1200.

- The test statistic is the sum of these weighted squared deviations

$$\sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

  where k is the total number of categories in the sample. If the test statistic is small, the observed frequencies are close to the expected frequencies and there would be no reason to reject the claim that it came from that distribution. The opposite occurs when the test statistic is too large. Therefore, the χ^2 goodness-of-fit test is always an upper tailed test.

  - In Exercise 11.1, the test statistic is $\frac{(19-16.67)^2}{16.67} + \cdots + \frac{(22-16.67)^2}{16.67} = 5.959$.
  - In Exercise 11.4, the test statistic is $\frac{(7-5.184)^2}{5.184} + \cdots + \frac{(3-1.536)^2}{1.536} = 2.693$.

- The test statistic has a χ^2 distribution with $k - 1$ degrees of freedom when the following assumptions are met.

  - The data is obtained from a random sample.
  - Each observation must fall into one and only one category.
  - At least 80% of the expected frequencies that we calculate must be at least 5. If the data fails this requirement, we combine some categories together so that this condition is met. The categories that we combine should be near to one another.

    * The data in Exercise 11.1 satisfies all assumptions, therefore the test statistic has the χ^2 distribution with 5 degrees of freedom. The critical value for this upper-tailed χ^2 test carried out at 5% level of significance is 11.07. We cannot reject H_0 since |test statistic| is less than |critical value|.
    * The data in Exercise 11.4 does not satisfy the last assumption. In fact only 75% (3 out of 4 categories) of the expected frequencies are at least 5. Thus the test statistic that we calculated earlier does not have the χ^2 distribution. We fix this problem by combining the last two categories together. The corresponding 0_i and e_i are $\{7, 8, 9\}$ and $\{5.184, 10.368, 8.448\}$ respectively. The new test statistic ($= 1.2131$) has the χ^2 distribution with 2 degrees of freedom. The critical value for this upper-tailed χ^2 test carried out at 5% level of significance is 5.991. We cannot reject H_0 since |test statistic| is less than |critical value|.

11.3 Test for Independence

- We test whether the row and column variables are independent of each other.
  - In Exercise 11.13, we test H_0 : the occupation and the opinion are independent against H_1 : they are not independent.
  - In Exercise 11.17, we test H_0 : the gender and the satisfaction are independent against H_1 : they are not.
  - In Exercise 11.25, we test H_0 : the patient improvement is independent of the use of cannabis or a placebo against H_1 : they are not independent.

- Recall that if A and B are independent events, then $P(A \cap B) = P(A) \times P(B)$. Thus another way of stating the two hypotheses is $H_0 : P(R_i \cap C_j) = P(R_i) \times P(C_j)$ and $H_1 : P(R_i \cap C_j) \neq P(R_i) \times P(C_j)$, where R_i = the event in the ith row, and C_i = the event in the jth column.
  - In Exercise 11.13, we test $H_0 : P$(the opinion in row ith and the occupation in column jth) $= P$(the opinion in row ith) $\times$ P(the occupation in column jth) against H_1 : at least one differs.
  - In Exercise 11.17, we test $H_0 : P$(the gender in row ith and the satisfaction in column jth) $= P$(the gender in row ith) $\times$ P(the satisfaction in column jth) against H_1 : at least one differs.
  - In Exercise 11.25, we test $H_0 : P$(the patient improvement in row ith and the treatment in column jth) $= P$(the patient improvement in row ith) $\times$ P(the treatment in column jth) against H_1 : at least one differs.

- Let r_i be the total number of observations for row i and c_j be the total number of observations for column j. Empirically, $P(R_i) = r_i/n$ and $P(C_j) = c_j/n$, and therefore $P(R_i \cap C_j) = r_i \times c_j/n^2$.
  - In Exercise 11.13, the probability distribution under H_0 can be seen in the following table:

	Student	Full Time	Carer	Unemployed
Yes	.148	.151	.103	.0413
No	.186	.189	.130	.0520

  - In Exercise 11.17, the probability distribution under H_0 can be seen in the following table:

	No Satisfied	Satisfied	Very satisfied
Male	.130	.261	.170
Female	.102	.204	.133

  - In Exercise 11.25, the probability distribution under H_0 can be seen in the following table:

	Placebo	Cannabis
Improved	.239	.239
Not improved	.261	.261

- The test statistic used is the same as the χ^2 goodness-of-fit test. The principle behind the test for independence and the principle behind the goodness-of-fit test are the same. (In fact, the test for independence is the goodness of fit test where the data is arranged into table form. This table is called the contingency table.) The observed frequency in each category will be given, so we only need to calculate the expected frequency in each category. This is simply the product of n and the probability for the category of interest obtained from H_0.
  - In Exercise 11.13, $n = 300$. The expected frequencies under H_0 can be seen in the following table:

	Student	Full Time	Carer	Unemployed
Yes	44.33	45.22	31.03	12.41
No	55.67	56.78	38.97	15.59

    The test statistic is $\frac{(32-44.33)^2}{44.33} + \cdots + \frac{(16-15.59)^2}{15.59} = 15.106$.
  - In Exercise 11.17, $n = 155$. The expected frequencies under H_0 can be seen in the following table:

	No Satisfied	Satisfied	Very satisfied
Male	20.21	40.41	26.38
Female	15.79	31.59	20.62

    The test statistic is $\frac{(13-20.21)^2}{20.21} + \cdots + \frac{(15-20.62)^2}{20.62} = 8.735$.
  - In Exercise 11.25, $n = 800$. The expected frequencies under H_0 can be seen in the following table:

	Placebo	Cannabis
Improved	191	191
Not improved	209	209

    The test statistic is $\frac{(238-191)^2}{191} + \cdots + \frac{(256-209)^2}{209} = 44.27$.
- **Tips**: Because the probability for the ijth category obtained from H_0 is $r_i \times c_j / n^2$, the expected frequency in the ijth category (e_{ij}) is given by

$$e_{ij} = n \times \frac{r_i \times c_j}{n^2} = \frac{r_i \times c_j}{n}.$$

  The test statistic is given by

$$\sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

  where o_{ij} is the observed frequency in the ijth category.
- If there are r rows and c columns in the sample, the test statistic will have the χ^2 distribution with $(r-1) \times (c-1)$ degrees of freedom when the following conditions are met.
  - The data is obtained from a random sample.
  - Each observation must fall into one and only one category.

- At least 80% of the expected frequencies that we calculate must be at least 5. If the data fails this requirement, we combine some of the row categories or some of the column categories together so that this condition is met. If there are 2 rows and more than 2 columns in the sample, we combine the column categories, and vice versa if there are 2 columns and more than 2 rows in the sample.
- For a 2×2 contingency table, the test statistic should be adjusted to

$$\sum \frac{(|o_{ij} - e_{ij}| - .5)^2}{e_{ij}}.$$

  * The data in Exercise 11.13 satisfies all conditions, therefore the test statistic has the χ^2 distribution with $(2-1) \times (4-1) = 3$ degrees of freedom. The critical value for this upper-tailed χ^2 test carried out at 5% level of significance is 7.815. We reject H_0 since |test statistic| is larger than |critical value|.
  * The data in Exercise 11.17 satisfies all conditions, therefore the test statistic has the χ^2 distribution with $(2-1) \times (3-1) = 2$ degrees of freedom. The critical value for this upper-tailed χ^2 test carried out at 5% level of significance is 5.991. We reject H_0 since |test statistic| is larger than |critical value|.
  * In Exercise 11.25, we need to adjust the test statistic to

$$\sum \frac{(|o_{ij} - e_{ij}| - .5)^2}{e_{ij}} = \frac{(|238 - 191| - .5)^2}{191} + \cdots + \frac{(|256 - 209| - .5)^2}{209} = 43.33.$$

    The adjusted test statistic has the χ^2 distribution with $(2-1) \times (2-1) = 1$ degrees of freedom. The critical value for this upper-tailed χ^2 test carried out at 5% level of significance is 3.841. We reject H_0 since |test statistic| is larger than |critical value|.

11.4 Testing The Slope of The Regression Line

- Suppose that the linear model that best explains the relationship between two variables X and Y is

$$Y_i = \alpha + \beta X_i + \epsilon_i.$$

  Where Y_i is the dependent variable, X_i is the independent variable, α is the intercept parameter, β is the slope parameter, and $\epsilon_i \sim N(0, \sigma^2)$. α, β, and σ^2 are unknown population parameters. In other words, α, β, and σ can only be calculated from the population.

- We can estimate α, β, and σ from the sample. The above equation becomes

$$\hat{Y}_i = a + bX_i.$$

  where a and b are point estimates for α and β respectively. Using the method of least squares

$$b = \frac{S_{XY}}{S_{XX}} = \frac{\sum XY - \frac{1}{n} \sum X \sum Y}{\sum X^2 - \frac{1}{n} \sum X^2}$$

  and $a = \bar{Y} - b\bar{X}$.

- Recall that $e_i = Y_i - \hat{Y}_i$ are the residuals. If $e_i \sim N(0, \hat{\sigma}^2)$, then $\hat{\sigma}^2$ can be used as the point estimate for σ^2. $\hat{\sigma}^2$ is just the sample variance of the residuals, i.e

$$\hat{\sigma}^2 = \frac{\sum(e_i - \bar{e})^2}{n-2} = \frac{S_{YY} - b \times S_{XY}}{n-2}$$

where $\bar{e} = 0$ and $S_{YY} = \sum Y^2 - \frac{1}{n}\sum Y^2$. We can check if e_i are normally distributed by plotting e_i against X_i and see if this plot exhibits any pattern. We can check for the symmetry by using a boxplot of e_i.

- If $e_i \sim N$, then it is not infeasible to say that $\epsilon_i \sim N$. If $\epsilon_i \sim N$, $b \sim N\left(\beta, se(b)^2\right)$, where $se(b) = \sigma/\sqrt{S_{XX}}$. Since σ in the formula for $se(b)$ is unknown, we will replace it by the value of $\hat{\sigma}$.

- We will be testing $H_0 : \beta = \beta^{H_0}$ against $H_1 : \beta < \beta^{H_0}$ or $H_1 : \beta > \beta^{H_0}$ or $H_1 : \beta \neq \beta^{H_0}$.
  - The test statistic has the same general pattern as before (observed minus claimed divided by standard error). Thus the test statistic in this case is $(b - \beta^{H_0})/se(b)$.
  - If $\epsilon_i \sim N$, then this test statistic has the t distribution with $n-2$ degrees of freedom. The decision to reject or not to reject H_0 can be made in the same way that it is made for other t tests.

- Exercise 12.4
  - We test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$.
  - $S_{XX} = 62477500$, $S_{YY} = 471.5244$, and $S_{XY} = 139971.2$, giving $\hat{\sigma} = 3.3588$, $b = .00224$, and $se(b) = 3.3588/\sqrt{62477500} = .000425$.
  - The test statistic = $(.00224 - 0)/.000425 = 5.2722$ has the t distribution with 14 degrees of freedom. Critical values for this two tailed test carried out at 5% level of significance are ± 2.977. We reject H_0 since |test statistic| is larger than |critical values|.

- Exercise 12.7
  - We test $H_0 : \beta = 20$ against $H_1 : \beta > 20$.
  - $S_{XX} = .18167$, $S_{YY} = 88.3825$, and $S_{XY} = 3.9$, giving $\hat{\sigma} = .6825$, $b = 21.46789$, and $se(b) = .6825/\sqrt{.18167} = 1.601$.
  - The test statistic = $(21.46789 - 20)/1.601 = .9167$ has the t distribution with 10 degrees of freedom. Critical values for this upper tailed test carried out at 5% level of significance are ± 1.812. We cannot reject H_0 since |test statistic| is less than |critical values|.