BAN 602: Quantitative Fundamentals

Lecture 5: Inferences about Population Variances and Multiple Proportions

Agenda

1. Big Picture of Chapter 11: Inferences about Population Variances
2. Big Picture of Chapter 12: Comparing Multiple Proportions
3. Example 1 (a population variance)
4. Example 2 (two population variances)
5. Example 3 (Test of Independence)
6. Example 4 (Goodness of Fit)

Big Picture of Chapter 11: Inferences about Population Variance(s)

Inference about a population variance	Inference about two population variances
● Sampling distribution of $(n-1)s^2/\sigma^2$ ○ Chi-square distribution with n-1 degrees of freedom ● Interval estimation of σ^2 ○ $(n-1)s^2/qchisq(1-\frac{\alpha}{2},n-1) \leq \sigma^2 \leq (n-1)s^2/qchisq(\frac{\alpha}{2},n-1)$ ● Hypothesis tests ○ Chi-square test statistic: $(n-1)s^2/\sigma_0^2$	● Sampling distribution of $s_1^2/\ s_2^2$ when $\sigma_1^2 =\ \sigma_2^2$ ○ F distribution with n_1-1 df for the numerator and n_2-1 df for the denominator ● Hypothesis tests ○ F test statistic = $s_1^2/\ s_2^2$

Big Picture of Ch. 12: Comparing Multiple Proportions

Testing the equality of 3 or more population proportions	Test of independence	Goodness of fit tests
● Chi-square test of the equality of 3 or more proportions (always upper-tailed test) ○ H_0: all proportions are equal ○ Chi-square test statistic ● A multiple comparison procedure (to identify where the differences are)	● Chi-square test of independence (always upper-tailed test) ○ H_0: independent (all proportions are equal) ○ Chi-square test statistic ● P-value vs. critical values	● To determine whether a population being samples has a specific probability distribution ○ Multinomial distribution ○ Normal distribution

Example 1: Inferences about a Population Variance

- The variance in drug weights is critical in the pharmaceutical industry. For a specific drug, with weights measured in grams, a sample of 18 units provided a sample variance of $s^2 = .36$.
- Construct a 90% confidence interval estimate of the population variance for the weight of this drug. ($\alpha = 10\%$)
- Construct a 90% confidence interval estimate of the population standard deviation.
- $(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$

Example 1: Inferences about a Population Variance

- The variance in drug weights is critical in the pharmaceutical industry. For a specific drug, with weights measured in grams, a sample of 18 units provided a sample variance of $s^2 = .36$.
- Construct a 90% confidence interval estimate of the population variance for the weight of this drug. ($\alpha = 10\%$)
- $qchisq(\frac{\alpha}{2}, n-1) \leq (n-1)s^2/\sigma^2 \leq qchisq(1-\frac{\alpha}{2}, n-1)$
- $\frac{(n-1)s^2}{qchisq(1-\frac{\alpha}{2}, n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{qchisq(\frac{\alpha}{2}, n-1)}, \frac{(18-1)0.36}{qchisq(0.95,17)} \leq \sigma^2 \leq \frac{(18-1)0.36}{qchisq(0.05,17)}$
- $qchisq(0.05,17) = 8.672, qchisq(0.95, 17) = 27.587$
- $0.222 \leq \sigma^2 \leq 0.706$

Example 1: Inferences about a Population Variance

- Construct a 90% confidence interval estimate of the population standard deviation.

Example 1: Inferences about a Population Variance

- Construct a 90% confidence interval estimate of the population standard deviation.
- $0.471 \leq \sigma \leq 0.840$

Example 2: Inferences about two population variances

- Investors commonly use the standard deviation of the monthly percentage return for a mutual fund as a measure of the risk for the fund; in such cases, a fund that has a larger standard deviation is considered more risky than a fund with a lower standard deviation. The standard deviation for the American Century Equity Growth fund (fund 1) and the standard deviation for the Fidelity Growth Discovery fund (fund 2) were recently reported to be 15.0% and 18.9%, respectively. Assume that each of these standard deviations is based on a sample of 60 months of returns. Do the sample results support the conclusion that the Fidelity fund (fund 2) has a larger population variance than the American Century fund? Which fund is more risky?

Example 2: Inferences about two population variances

- What are the hypotheses?

Example 2: Inferences about two population variances

- What are the hypotheses?

H_0: $\sigma_1^2 \geq \sigma_2^2$ or $\frac{\sigma_1^2}{\sigma_2^2} \geq 1.$

H_a: $\sigma_1^2 < \sigma_2^2$ or $\frac{\sigma_1^2}{\sigma_2^2} < 1.$

Example 2: Inferences about two population variances

- What is an appropriate test statistic? And what is the sampling distribution of this test statistic?

Example 2: Inferences about two population variances

- What is an appropriate test statistic? And what is the sampling distribution of this test statistic?

$$\frac{s_1^2}{s_2^2} \sim F(df_1 = n_1 - 1, df_2 = n_2 - 1)$$

Example 2: Inferences about two population variances

- Compute p-value for this test.

Example 2: Inferences about two population variances

- Compute p-value for this test.

$$p\ value = F\left(\frac{s_1^2}{s_2^2}, n_1 - 1, n_2 - 1\right) = F\left(\frac{0.15^2}{0.189^2}, 59{,}59\ \right)$$

$$= F\left(\frac{0.15^2}{0.189^2}, 59{,}59\ \right) = F(0.630{,}59{,}59\ ) = 3.92\%$$

Example 2: Inferences about two population variances

- Suppose α is 5%. Compute critical F value.

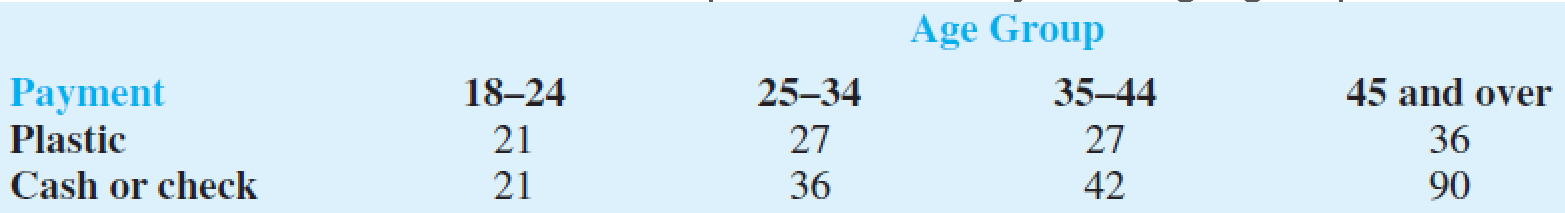
Example 2: Inferences about two population variances

- Suppose α is 5%. Compute critical F value.

$$F_c = F^{-1}(\alpha, n_1 - 1, n_2 - 1) = F^{-1}(0.05,59,59\ ) = 0.649$$

Example 3: Test of Independence

- Visa Card USA studied how frequently consumers of various age groups use plastic cards (debit and credit cards) when making purchases. Sample data for 300 customers shows the use of plastic cards by four age groups.



	Age Group			
Payment	18–24	25–34	35–44	45 and over
Plastic	21	27	27	36
Cash or check	21	36	42	90

- Test for the independence between method of payment and age group. What is the p-value? Using α = .05, what is your conclusion?
- If method of payment and age group are not independent, what observation can you make about how different age groups use plastic to make purchases?
- What implications does this study have for companies such as Visa, MasterCard, and Discover?

Example 3: Test of Independence

- What would be the hypotheses?

Example 3: Test of Independence

- What would be the hypotheses?

H_0: Age and method of payment are independent.
H_a: Age and method of payment are NOT independent.

Example 3: Test of Independence

- Let p_i be the proportion of people who use plastic payment in each of the four age groups. What would be the hypotheses?

Example 3: Test of Independence

- Let p_i be the proportion of people who use plastic payment in each of the four age groups. What would be the hypotheses?

  H_0: $p_1 = p_2 = p_3 = p_4$.
  H_a: At least one of the equation does NOT hold.

Example 3: Test of Independence

- What are the two random variables in question? What are their respective sample space? How can we interpret p_i?

Example 3: Test of Independence

- What are the two random variables in question? What are their respective sample space? How can we interpret p_i?

Method of payment $A \in \{A_1(plastic), A_2(cash)\}$.
Age group $B \in \{B_1, B_2, B_3, B_4\}$.
$p_i = prob(A_1|B_i), i = 1, 2, 3, or\ 4$.

Example 3: Test of Independence

- Why are age and method of payment independent if $p_1 = p_2 = p_3 = p_4$?

Example 3: Test of Independence

- Why are age and method of payment independent if $p_1 = p_2 = p_3 = p_4$?

If $p_1 = p_2 = p_3 = p_4$, then

- $prob(A_1|B_1) = prob(A_1|B_2) = prob(A_1|B_3) = prob(A_1|B_4)$, and
- $prob(A_2|B_1) = prob(A_2|B_2) = prob(A_2|B_3) = prob(A_2|B_4)$

Thus,

- $pr(A_1|B_1) = pr(A_1|B_2) = pr(A_1|B_3) = pr(A_1|B_4) = pr(A_1)$, and
- $pr(A_2|B_1) = pr(A_2|B_2) = pr(A_2|B_3) = pr(A_2|B_4) = pr(A_2)$.

Example 3: Test of Independence

- Why are age and method of payment independent if $p_1 = p_2 = p_3 = p_4$?

  - Alternatively, the two RVs are independent iff $pr(A_1 B_1) = pr(A_1)\ pr(B_1)$, $pr(A_1 B_2) = pr(A_1)\ pr(B_2)$, and $pr(A_1 B_3) = pr(A_1)\ pr(B_3)$.
  - More generally, as long as 3 joint probabilities from 3 different columns in the joint probability table are equal to the product of their respective marginal probabilities, Age and Method of payment are independent.
  - The degree of freedom of this example is 3. In general, df = (# or rows – 1)*(# of columns – 1).

Example 3: Test of Independence

- What is an appropriate test statistic and what is the sampling distribution of this test statistic?

Example 3: Test of Independence

- What is an appropriate test statistic and what is the sampling distribution of this test statistic?

$$\chi^2 statistic = \sum_{i=1}^{2}\sum_{j=1}^{4}\frac{\left(f_{ij}-e_{ij}\right)^2}{e_{ij}} \sim \chi^2(df = (2-1)(4-1)).$$

- f_{ij} is the observed frequency in row i and column j.
- e_{ij} is the expected frequency in row i and column j assuming independence. $e_{ij} = \frac{(row\ i\ total)(column\ j\ total)}{n}$.

Example 3: Test of Independence

- Why $e_{ij} = \frac{(row\ i\ total)(column\ j\ total)}{n}$ if we assume independence?

Example 3: Test of Independence

- Why $e_{ij} = \frac{(row\ i\ total)(column\ j\ total)}{n}$ if we assume independence?

  - If A and B are independent, then $pr\left(A_i B_j\right) = pr(A_i)\ pr\left(B_j\right)$, $\forall\ i, j$.
  - Therefore, $e_{ij} = n * pr(A_i)\ pr\left(B_j\right)$
  - $= \frac{(n * pr(A_i))(n * pr\left(B_j\right))}{n} = \frac{(row\ i\ total)(column\ j\ total)}{n}$.
  - e_{ij} is the expected frequency in row i and column j assuming independence. $e_{ij} = \frac{(row\ i\ total)(column\ j\ total)}{n}$.

Example 3: Test of Independence

- Another interpretation of degrees of freedom in test of independence.

  - In computing e_{ij} , we must know all the row totals and column totals.
  - In our 2×4 contingency table or frequency table, we only have freedom to randomly “choose” 3 values (from 3 different columns because we have more columns than rows). Then we can figure out the rest of 8-3 = 5 values.
  - In general, in test of independence with a $r \times c$ contingency table, the degree of freedom is $(r - 1) \times (c - 1)$.

Example 3: Test of Independence

- In test of independence, or more generally in test of equality of 3 or more population proportions, it is always going to be an upper-tailed test. Why?

Example 3: Test of Independence

- In test of independence, or more generally in test of equality of 3 or more population proportions, it is always going to be an upper-tailed test. Why?

  When two random variables tend to be independent, or when all the population proportions tend to be equal, the value of test statistic tends to be closer to zero. Otherwise, its value tends to be significantly GREATER than zero.

Example 3: Test of Independence

- Let α = 5% be the significant level. Write down R code for computing the p value and the critical value.

Example 3: Test of Independence

- Let α = 5% be the significant level. Write down R code for computing the p value and the critical value.

  - Compute p-value

    1 - $pchisq$(value of test statistic, df = (r-1)(c-1))

    1 - $pchisq$(7.947, 3) = 4.71%

  - Compute critical value

    $qchisq$(1 - α, df = (r-1)(c-1))

    $qchisq$(0.95, 3) = 7.815

  - Acceptance and rejection region?

Example 4: Goodness of Fit Test (Multinomial Dist.)

- Market study has shown that the market shares for certain product have stabilized at 30% for company A, 50% for company B, and 20% for company C. Assume that each customer buys from one of these three companies.
- Company C plans to introduce an improved product and wants to find out whether the improved product would change the market shares. A group of 200 customers were randomly selected and they indicated their product preference.

	A	B	C	total
Market share	0.3	0.5	0.2	1
observed	48	98	54	200
expected	60	100	40	200

Example 4: Goodness of Fit Test (Multinomial Dist.)

- What are the hypotheses?

Example 4: Goodness of Fit Test (Multinomial Dist.)

- What are the hypotheses?

  H_0: $p_A = 0.3, p_B = 0.5, p_C = 0.2.$

  H_a: Market shares are different from above.

Example 4: Goodness of Fit Test (Multinomial Dist.)

- What is an appropriate test statistic and what is its sampling distribution?

Example 4: Goodness of Fit Test (Multinomial Dist.)

- What is an appropriate test statistic and what is its sampling distribution?

$\chi^2 statistic = \sum_{i=A}^{C} \frac{(f_i - e_i)^2}{e_i} \sim \chi^2(df = 2)$.

- f_i is the observed frequency.
- e_i is the expected frequency. $e_i = p_i * n$.

Example 4: Goodness of Fit Test (Multinomial Dist.)

- Compute p value and critical value given α = 0.01.

Example 4: Goodness of Fit Test (Multinomial Dist.)

- Compute p value and critical value given α = 0.01.

  - Compute p-value

    1 - $pchisq$(value of test statistic, df = k-1)

    1 - $pchisq$(7.34, 2) = 2.55%
  - Compute critical value

    $qchisq$(1 - α, df = k-1)

    $qchisq$(0.99, 2) = 9.21
  - Acceptance and rejection region?

Quiz 5 Preparation

1. Understanding the sample statistics and their sampling distributions with respect to one and two variances, including their degrees of freedom.
2. Compute corresponding values of the sample statistics.
3. R code for computing confidence interval of one variance or sd.
4. Formulate hypotheses to test on two population variances.
5. R code for computing p-value and critical value.