Quantitative Methods – II

A.Y. 2021-22

Practice 12

Lorenzo Cavallo

For any clarification/meeting: cavallo@istat.it

THEME #1

Chi-Square Test

The Chi-Square Test

Goodness-of-Fit Test

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

where E is the expected frequency for a category (np) and O the observed frequency for a category.

Test of Independence or Homogeneity

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

where O and E are the observed and expected frequencies, respectively, for a cell.

Tests About the Population Variance

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

where s^2 is the sample variance and σ^2 is the hypothesized population variance.

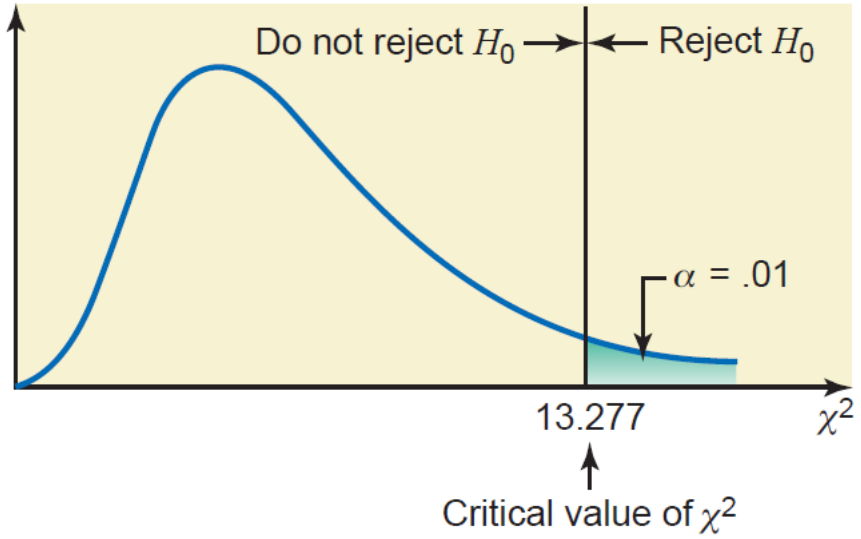
All the Chi-Square Test are Right-tailed Tests

The degrees of freedom in **Goodness-of-Fit Test** are ***df=k-1*** where k are the categories involved.

Tests of Independence or Homogeneity involve contingency table.
The degrees of freedom for tests of independence or homogeneity are ***df=(R-1)(C-1)*** where *R* and *C* are the number of rows and the number of columns in the contingency table.

Test of Indipendence: the null hypothesis is that two attributes of a population are not related.
Test of Homogeneity: the null hypothesis is that the proportions of elements with certain characteristics in two or more different populations are the same.

The degrees of freedom in **Tests About the Population Variance** are ***df=n-1*** where n is the sample size.



Population 2	Population 1 f	g	h	i	l	tot
a	5	3	5	6	7	26
b	4	1	3	3	8	19
c	3	2	5	8	9	27
d	2	4	7	2	1	16
e	1	3	4	1	3	12
tot	15	13	24	20	28	100

Goodness-of-Fit Test

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

Exercise 1

Number of bank transactions per month

Months	January	February	March	April	May	June	July	August	September	October	November	December	
Observed Frequency	35	33	23	28	52	54	34	28	72	33	76	32	n= 500

At the 95% level of significance, can we reject the null hypothesis that the number of transactions in the 12 months are the same?

Solution

n=500, categories k=12, 1-α=0.95

This is a goodness-of-fit test in which we have to test that the observed values (or proportions) of the sample are NOT different from to the expected frequencies (all the proportions/frequencies/values are the same for each category)

State the null hypothesis that the proportion of transaction per month are the same and the alternative that at least two are not equal, so:

H_0: $p_1=p_2=p_3=p_4=p_5=p_6=p_7=p_8=p_9=p_{10}=p_{11}=p_{12}$

H_1: At least two are not equal

Goodness-of-Fit Test

E=500/12=41.67

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

The expected frequencies are the frequencies in the case of equal distribution per month (all the months the same frequencies).

To calculate the expected frequencies we have to calculate the equal proportion per month (1/12 per month)

Then we have to multiplicate the proportion with the total frequencies

Category (Months)	Observed Frequency O	p	Expected Frequency E = np	(O - E)	(O - E)²	(O - E)²/E
January	35	0.0833=8.33%	500*(0.0833)=41.67	(35-41.67)=-6.67	44.49	1.07
February	33	0.0833=8.33%	500*(0.0833)=41.67	(33-41.67)=-8.67	75.17	1.8
March	23	0.0833=8.33%	500*(0.0833)=41.67	(23-41.67)=-18.67	348.57	8.36
April	28	0.0833=8.33%	500*(0.0833)=41.67	(28-41.67)=-13.67	186.87	4.48
May	52	0.0833=8.33%	500*(0.0833)=41.67	(52-41.67)=10.33	106.71	2.56
June	54	0.0833=8.33%	500*(0.0833)=41.67	(54-41.67)=12.33	152.03	3.65
July	34	0.0833=8.33%	500*(0.0833)=41.67	(34-41.67)=-7.67	58.83	1.41
August	28	0.0833=8.33%	500*(0.0833)=41.67	(28-41.67)=-13.67	186.87	4.48
September	72	0.0833=8.33%	500*(0.0833)=41.67	(72-41.67)=30.33	919.91	22.08
October	33	0.0833=8.33%	500*(0.0833)=41.67	(33-41.67)=-8.67	75.17	1.8
November	76	0.0833=8.33%	500*(0.0833)=41.67	(76-41.67)=34.33	1178.55	28.28
December	32	0.0833=8.33%	500*(0.0833)=41.67	(32-41.67)=-9.67	93.51	2.24
	n= 500					82.21

$\chi^2 = \frac{\sum(O-E)^2}{E} = 82.21$ and the critical value is $\chi^2_{k-1,\alpha} = \chi^2_{11,0.05} = 19.675$

$$\chi^2 = 82.21 > \chi^2_{11,0.05} = 19.675.$$

Then we reject H_o ; the bank transactions per month are not the same.

Exercise 2

Test the independence of age and salaries in a sample with n=115 (α=0.01).

		Salaries			
		<10,000	10,000-20,000	20,000-30,000	>30,000	tot
Age	18-28	1	3	5	1	10
	28-38	2	10	8	0	20
	38-48	3	9	6	2	20
	48-58	7	10	9	3	29
	58-68	8	12	12	4	36
	tot	21	44	40	10	115

Hypothesis testing:

H_0: Age and salaries are independent

H1: Age and salaries are dependent

$$Expected\ frequency = \frac{(Row\ Total)(Column\ Total)}{Sample\ size}$$

Solution

Observed

	<10,000	10,000-20,000	20,000-30,000	>30,000	tot
18-28	1	3	5	1	10
28-38	2	10	8	0	20
38-48	3	9	6	2	20
48-58	7	10	9	3	29
58-68	8	12	12	4	36
tot	21	44	40	10	115

Expected = $\frac{(Row\ Total)(Column\ Total)}{Sample\ size}$

	<10,000	10,000-20,000	20,000-30,000	>30,000	tot
18-28	10·21/115	10·44/115	10·40/115	10·10/115	10
28-38	20·21/115	20·44/115	20·40/115	20·10/115	20
38-48	20·21/115	20·44/115	20·40/115	20·10/115	20
48-58	29·21/115	29·44/115	29·40/115	20·10/115	29
58-68	36·21/115	36·44/115	36·40/115	36·10/115	36
tot	21	44	40	10	115

Expected

	<10,000	10,000-20,000	20,000-30,000	>30,000	tot
18-28	1.83	3.83	3.48	0.87	10
28-38	3.65	7.65	6.96	1.74	20
38-48	3.65	7.65	6.96	1.74	20
48-58	5.30	11.10	10.09	2.52	29
58-68	6.57	13.77	12.52	3.13	36
tot	21	44	40	10	115

(Observed- Expected)^2/Expected

	<10,000	10,000-20,000	20,000-30,000	>30,000	tot
18-28	0.3737	0.178	0.666	0.02	
28-38	0.7474	0.72	0.157	1.739	
38-48	0.1165	0.237	0.132	0.039	
48-58	0.5485	0.108	0.117	0.091	
58-68	0.3094	0.228	0.022	0.242	
tot					6.791

χ^2 is the Sum of all the E = (Observed value-Expected Value)^2/Expected Value = 6.791

Critical value with df=(R-1)(C-1)=4 · 3=12 and α=0.01 is: $\chi^2_{12,0.01}$=26.217

$\chi^2 < \chi^2_{12,0.01}$ We do not reject the Null Hypothesis, salaries and age are independent

Exercise 3

A shop sells t-shirts of different colors. Given a sample of 100 t-shirt sold in a month, can we reject the null hypothesis that the t-shirts sold are the same for each color? (at the 5% level of significance)

Blue	Black	White	Purple	Red	Grey	Other	n
11	18	21	15	15	12	8	100

Solution

Hypothesis testing:

H_0: $p_1=p_2=p_3=p_4=p_5=p_6=p_7$

H_1: At least two are not equal

The proportion for expected frequencies is p=1/7=0.1428 and the Expected frequency for each class is np=14.28

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

From the calculation of the Chi-square test we have that: $\chi^2=8.08$

Critical value with df=k-1=7-1=6 and $\alpha=0.05$ is: $\chi^2_{6,0.05}= 12.592$

The Chi-square test is lower of the critical value so, we do not reject the Null Hypothesis

Exercise 4

A shop sells t-shirts of different sizes. Given a sample of 150 t-shirt sold in a month, can we reject the null hypothesis that the t-shirts sold are the same for each size? (at the 1% level of significance)

XS	S	M	L	XL	XXL	n
33	16	25	37	24	15	150

Solution

Hypothesis testing:

H_0: $p_1=p_2=p_3=p_4=p_5=p_6$

H_1: At least two are not equal

The proportion for expected frequencies is p=1/6=0.1667 and the Expected frequency for each class is np=150 · 0.1667=25

$$\chi^2 = \frac{\sum(O-E)^2}{E}$$

From the calculation of the Chi-square test we have that: χ^2= 15.60

The critical value with df=k-1=6-1=5 and α=0.01 is: $\chi^2_{5,0.01}$= 15.086

The Chi-square test is higher of the critical value so, we reject the Null Hypothesis, at least two are not equal

Exercise 5

A random sample of 300 adults was selected, and these adults were asked if they favor giving more freedom to their children. The two-way classification of the responses of these adults is presented in the following table.

	In favor	Against	No opinion	tot
Men (M)	93	70	12	175
Women (W)	87	32	6	125
tot	180	102	18	300

Calculate the expected frequencies for this table, assuming that the two attributes, gender and opinions on the issue, are independent.

Does the sample provide sufficient evidence to conclude that the two attributes, gender and opinions of adults, are dependent? Use a 1% significance level.

H_0: Gender and opinions of adults are independent.

H_1: Gender and opinions of adults are dependent.

The observed frequencies are in the Table

	In favor	Against	No opinion	tot
Men (M)	93	70	12	175
Women (W)	87	32	6	125
tot	180	102	18	300

To calculate the expected frequencies we have to calculate the table of the expected frequencies by:

$$Expected\ frequency = \frac{(Row\ Total)(Column\ Total)}{Sample\ size}$$

Hence, the 6 values of the expected frequencies table are:

E for *Men* and *In Favor* cell $= (175)(180)/300 =$ **105.00**

E for *Men* and *Against* cell $= (175)(102)/300 =$ **59.50**

E for *Men* and *No Opinion* cell $= (175)(18)/300 =$ **10.50**

E for *Women* and *In Favor* cell $= (125)(180)/300 =$ **75.00**

E for *Women* and *Against* cell $= (125)(102)/300 =$ **42.50**

E for *Women* and *No Opinion* cell $= (125)(18)/300 =$ **7.50**

Observed frequencies

	In favor	Against	No opinion	tot
Men (M)	93	70	12	175
Women (W)	87	32	6	125
tot	180	102	18	300

Expected frequencies

	In favor	Against	No opinion	tot
Men (M)	105	59,5	10,5	175
Women (W)	75	42,5	7,5	125
tot	180	102	18	300

	In Favor (*F*)	**Against (*A*)**	**No Opinion (*N*)**	**Row Totals**
Men (*M*)	93 (105.00)	70 (59.50)	12 (10.50)	175
Women (*W*)	87 (75.00)	32 (42.50)	6 (7.50)	125
Column Totals	180	102	18	300

df=(R-1)(C-1)=(2-1)(3-1)=1 · 2=2 χ^2 for df=2 and α=0.01 is 9.210

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(93-105.00)^2}{105.00} + \frac{(70-59.50)^2}{59.50} + \frac{(12-10.50)^2}{10.50}$$
$$+ \frac{(87-75.00)^2}{75.00} + \frac{(32-42.50)^2}{42.50} + \frac{(6-7.50)^2}{7.50}$$
$$= 1.371 + 1.853 + .214 + 1.920 + 2.594 + .300 = 8.252$$

$\chi^2 = 8.252 < \chi^2_{2,0.01} = 9.210$ Hence falls in nonrejection region. We DO NOT reject H_0

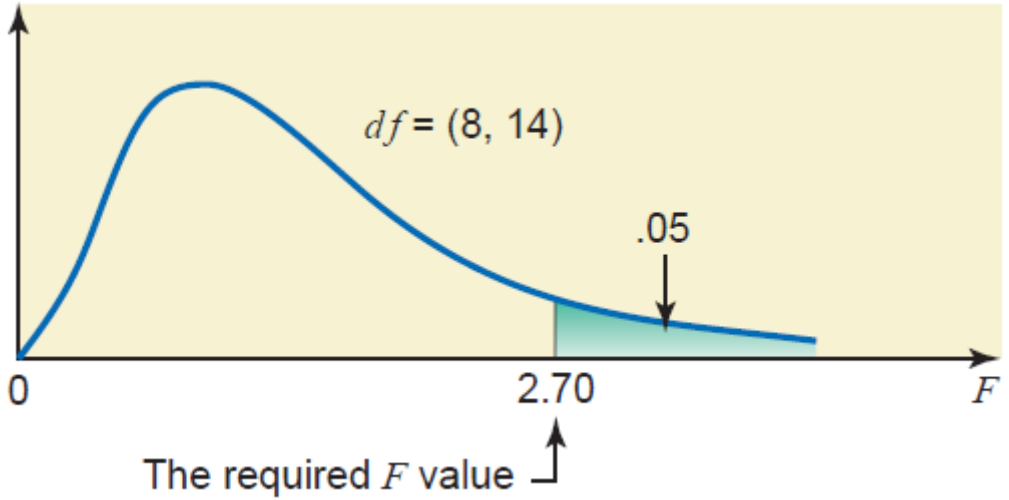
We can conclude that the null hyphotesis is true: gender and opinion are indipendent.

THEME #2

The analysis of variance procedure ANOVA

The *F* Distribution

1. The *F distribution* is continuous and skewed to the right.
2. The *F* distribution has two numbers of degrees of freedom: *df* for the numerator and *df* for the denominator.
3. The units of an *F* distribution, denoted by *F*, are nonnegative.



The **analysis of variance** procedure (**ANOVA**) is used to test the null hypothesis that the means of three or more populations are the same against the alternative hypothesis that not all population means are the same.

Test Statistic F for a One-Way ANOVA Test: The value of the test statistic F for an ANOVA test is calculated as:

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{MSB}{MSW}$$

One-Way ANOVA Test:

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{MSB}{MSW}$$

To calculate **MSB** and **MSW**, we first compute:

- the **between-samples sum of squares**, denoted by **SSB**,
- and the **within-samples sum of squares**, denoted by **SSW**.

The sum of **SSB** and **SSW** is called the **total sum of squares** and is denoted by **SST**:

$$SST = SSB + SSW$$

$$SST = SSB + SSW$$

The between-samples sum of squares, **denoted by SSB**, is calculated as

$$SSB = \sum \left(\frac{T_i^2}{n_i} \right) - \frac{(\sum x)^2}{n}$$

The within-samples sum of squares, denoted by SSW, is calculated as

$$SSW = \sum x^2 - \sum \left(\frac{T_i^2}{n_i} \right)$$

where:
T_i the sum of the values in sample i
n_i the size of sample i
n the number of values in all samples $(n_1+n_2+n_3+\ldots)$
Σx the sum of the values in all samples $(T_1+T_2+T_3+\ldots)$
Σx^2 the sum of the squares of the values in all samples

Exercise 6

Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (car types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use α = 5%.

Compact cars	Midsize cars	Full-size cars
1 15	4 5	7 12
2 10	5 8	8 10
3 30	6 26	9 18

Solution

Calculate the means of the 3 car types

Compact cars: $\bar{x}_1 = \frac{15+10+30}{3}=18.33$; Midsize cars: $\bar{x}_2 = \frac{5+8+26}{3}=13$; Full-size cars: $\bar{x}_3 = \frac{12+10+18}{3}= 13.33$

Then calculate the 3 sample standard deviations:

$s = \frac{\sum(x_i-\bar{x})^2}{n-1}$ then: $s_1=10.41$; $s_2=11.36$; $s_3=4.16$

The sample size n_i of the 3 samples is $n_1 =3$, $n_2 =3$, $n_3 =3$ and $n = n_1 + n_2 + n_3 = 9$

We have to calculate the SSB and the SSW.

Exercise 6

Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (car types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use $\alpha = 5\%$.

Compact cars	Midsize cars	Full-size cars
15	5	12
10	8	10
30	26	18

Solution

$$SSB = \sum \left(\frac{T_i^2}{n_i} \right) - \frac{(\sum x)^2}{n}$$

The T_i is the sum of the value per each sample: $T_1 = 15+10+30=55$; $T_2 = 5+8+26=39$; $T_3 = 12+10+18=40$

The T_i^2 are: $T_1^2 = 3025$; $T_2^2 = 1521$; $T_3^2 = 1600$ and T_i^2/n_i are: $T_1^2/n_1=1008.33$; $T_2^2/n_2=507$; $T_3^2/n_3=533.33$

$\Sigma(T_i^2/n_i)$ is $=T_1^2/n_1+T_2^2/n_2+T_3^2/n_3=1008.33+507+533.33=$**2048.7**

The total T is $=\Sigma x=T_1 + T_2 + T_3 = 55+39+40=134$ with $(\sum x)^2=T^2=134^2=17956$ and $(\Sigma x)^2/n = 17956/9=$**1995.1**

$$SSB = \sum \left(\frac{T_i^2}{n_i} \right) - \frac{(\sum x)^2}{n} = 2048.7 - 1995.1 = 53.6$$

Exercise 6

Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (car types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use $\alpha = 5\%$.

Compact cars	Midsize cars	Full-size cars
15	5	12
10	8	10
30	26	18

Solution

$$SSW = \sum x^2 - \sum \left(\frac{T_i^2}{n_i} \right)$$

We have to calculare Σx_i^2 that is: $15^2+10^2+30^2+5^2+8^2+26^2+12^2+10^2+18^2=2558$

$$SSW = \sum x^2 - \sum \left(\frac{T_i^2}{n_i} \right) = 2558 - 2048.7 = 509.3$$

The **total sum of squares is:**

$$SST = SSB + SSW = 53.6 + 509.3 = 562.9$$

One-Way ANOVA Test:

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{MSB}{MSW}$$

$$MSB = \frac{SSB}{k-1} \qquad MSW = \frac{SSW}{n-k}$$

where k is the number of different samples and k - *1* and n - k are, respectively, the *df* for the numerator and the *df* for the denominator for the *F distribution*.

ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Value of the Test Statistic
Between	$k-1$	SSB	MSB	$F = \frac{\text{MSB}}{\text{MSW}}$
Within	$n-k$	SSW	MSW	
Total	$n-1$	SST		

Exercise 6

Suppose the National Transportation Safety Board (NTSB) wants to examine the safety of compact cars, midsize cars, and full-size cars. It collects a sample of three for each of the treatments (cars types). Using the hypothetical data provided below, test whether the mean pressure applied to the driver's head during a crash test is equal for each types of car. Use $\alpha = 5\%$.

Solution

Source of Variation	DF	SUM OF SQUARE	MEAN SQUARE	
BEETWEEN	k-1	SSB	MSB = SSB/(k-1)	F=MSB/MSW
WITHIN	n-k	SSW	MSW = SSW/(n-k)	
TOTAL	n-1	SST = SSB + SSW		

Source of Variation	DF	SUM OF SQUARE	MEAN SQUARE	
BEETWEEN	3-1	SSB = 53.6	MSB = 53.6/2 = 26.8	F=26.8/84.9=0.32
WITHIN	9-3	SSW = 509.3	MSW=509.3/6 = 84.9	
TOTAL	9-1	SST = 562.9		

$F_{2,6}=5.14$ Hence we DO NOT reject the Null Hypothesis

Exercise 7

Calculate ANOVA of:

Sample 1	Sample 2	Sample 3
1	4	8
5	3	9
6	2	
	6	

Mean	4	4	8.5
n_i	3	4	2
T_i	12	15	17
T_i^2	144	225	289
T_i^2/n_i	48.00	56.25	144.50

n=	9
$T=\Sigma x=$	44
$\Sigma(T_i^2/n_i)=$	248.75
$\Sigma x_i^2=$	272
$\Sigma x^2=$	1936
$\Sigma x^2/n=$	215.1

SSB= $\Sigma(T_i^2/n_i)-\Sigma x^2/n=$ 33.6 MSB=16.82

SSW= $\Sigma x_i^2-\Sigma(T_i^2/n_i)=$ 23.3 MSW=3.88

SST= SSB + SSW = 56.9

$F=4.34 < F_{2,6}=5.14$ Hence we DO NOT reject the Null Hypothesis

Source of Variation	DF	SUM OF SQUARE	MEAN SQUARE	
BEETWEEN	2	33.6	16.82	
WITHIN	6	23.3	3.88	F=MSB/MSW=4.34
TOTAL	8	56.9		

Exercise 8

Calculate ANOVA of:

	Sample 1	Sample 2	Sample 3	Sample 4
	9	12	8	17
	6	16	8	15
	11	16	12	17
	14	12	7	16
	14	9	10	13
Mean	10.8	13.0	9.0	15.6

	Sample 1	Sample 2	Sample 3	Sample 4
n_i	5	5	5	5
T_i	54	65	45	78
T_i^2	2916	4225	2025	6084
T_i^2/n_i	583.20	845.00	405.00	1216.80

$n=$	20	SSB=	$\Sigma(T_i^2/n_i)-\Sigma x^2/n=$	121.8
$T=\Sigma x=$	242	SSW=	$\Sigma x_i^2-\Sigma(T_i^2/n_i)=$	110.0
$\Sigma x_i^2=$	3160	SST=	SSB + SSW =	231.8
$\Sigma(T_i^2/n_i)=$	3050.00			
$\Sigma x^2=$	58564			
$\Sigma x^2/n=$	2928.2			

Source of Variation	**DF**	**SUM OF SQUARE**	**MEAN SQUARE**	
BEETWEEN	3	121.8	40.6	
WITHIN	16	110	6.88	F=MSB/MSW= 5.91
TOTAL	**19**	**231.8**		

$F_{3,16}=3.24$ Hence we reject the Null Hypothesis at 95%