

# air pollution training institute



5.

## introduction to environmental statistics

self-instructional course

SI 473

MODULE V

CHI-SQUARE TECHNIQUES



Preceding modules have discussed the statistical techniques useful in comparing population parameters. The following ANOVA problem is an example of one of these techniques.

A researcher wants to test the following TSP data (24 hr.) from four cities to see if their means actually differ significantly at the 5% level.

<u>Day</u>	<u>N.Y.C.</u>	<u>Pittsburgh</u>	<u>Chicago</u>	<u>Los Angeles</u>
1	93	230	75	604
2	124	135	95	127
3	63	93	204	254
4	82	115	115	151
5	92	132	222	152

Knowing what we now know, we could quite easily answer the problem. However, let's now assume that we want to answer a slightly different question. We can see that the secondary standard of  $150 \mu\text{g}/\text{m}^3$  was exceeded on some of the above days, while on others it was not. How do you suppose we could answer this question: "Did the four cities differ significantly in the number of days on which the secondary standard was exceeded?"

The statisticians have developed a method to test just that. It's named the Chi-square test and we will discuss the application and computation of chi-square techniques in this module. Read pp. 129-133 in the text and then go to frame 2.

Answer:  $\chi^2$

24a

Step 3: Organize and arrange data.

Not necessary.

25

Be sure you have read pp. 129-133 (Chapter 12) in the text before continuing with this program.

Step 4: Compute the statistic.

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

We already know the O's.

	$f$
one	4
two	0
three	3
four	1
five	2
six	2

Assuming we have a fair die, the probability of obtaining any number is 1 in 6 (1/6). Since we are rolling the die twelve times, the expected frequency for each number is

$$1/6 \times 12$$

or

$$2$$

Calculate  $\chi^2$  for this problem.

When an experiment is performed, the experimenter will occasionally have an idea of the frequency with which certain data types will be observed. However, his expected frequencies are hardly ever observed. He can then use a \_\_\_\_\_ test to see if this difference between expected frequencies and observed frequencies is significant.

---

26a

Answer:

$$\begin{aligned} \chi^2 &= \frac{(4 - 2)^2}{2} + \frac{(0 - 2)^2}{2} + \frac{(3 - 2)^2}{2} + \frac{(1 - 2)^2}{2} + \frac{(2 - 2)^2}{2} + \frac{(2 - 2)^2}{2} \\ &= 5.0 \end{aligned}$$


---

27

Step 5: Determine critical values.

$$\alpha = .05$$

$df$  = the number of values we can arbitrarily set before the last one is determined. In this case, we have six categories for which we can set observed frequencies, but since the sum of these six must equal 12, we can only set 5 of these six arbitrarily.

$$df = \underline{5}$$

Consult Table I for the critical value.

$$\text{Critical Value} = \underline{5}$$



Answer: observed frequencies = expected frequencies

5

In  $\chi^2$  tests, the only alternative hypothesis we consider is that the observed frequencies and the expected frequencies differ.

$H_1$ : \_\_\_\_\_

28a

Answer: Accept

Since our obtained  $\chi^2$  was less than the tabled  $\chi^2$

We can then conclude that the die is a fair one.

Problem II: Given the following data on frequency of occasions in which emissions from four stacks were either above or below standards, can we conclude that the results are other than what we would expect if the stacks were the same? (use  $\alpha = .01$ ) 29

	A	B	C	D	Total
Higher	6	22	25	23	76
Lower	52	38	31	43	164
Total	58	60	56	66	240

Go to the next frame for the step-by-step solution.

Answer: observed frequencies  $\neq$  expected frequencies

---

6

To perform the  $\chi^2$  test we calculate

$$\frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

for each class and then sum all these values across all classes. The formula is:

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

where O = observed frequency

E = expected frequency

---

30

Step 1: Define the problem.

$H_0$ : \* \_\_\_\_\_

$H_1$ : \* \_\_\_\_\_

Let's try one. Suppose we toss a coin 100 times and obtain 70 heads and 30 tails. We now say to ourselves (statisticians are known to talk to themselves a lot), "Are those results different from what we expected?"

To answer this question, we decide to use the  $\chi^2$  test of significance:

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

Or in this case

$$\chi^2 = \frac{(\text{observed heads} - \text{expected heads})^2}{\text{expected heads}} + \frac{(\text{observed tails} - \text{expected tails})^2}{\text{expected tails}}$$

30a

Answer:  $H_0$ : observed frequencies = expected frequencies  
 $H_1$ : observed frequencies  $\neq$  expected frequencies

31

Step 2: Determine test and  $\alpha$  level.

test \*

$\alpha = f$

In this example we already know the observed frequency of heads and tails. What do you think the expected frequencies are? (Assuming we have an honest coin.)

Out of 100 tosses we expect # \_\_\_\_\_ heads and # \_\_\_\_\_ tails. Of course, it will rarely work out exactly this way; but this is our best estimate.

---

Answer:

$$\chi^2$$

$$\alpha = .01$$

31a

---

32

Step 3: Organize and arrange data.

Not necessary.

Answer: 50 heads

8a

50 tails

9

Fill in the values in this formula.

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

$$\chi^2 =$$

33

Step 4: Compute the statistic.

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

	A	B	C	D	Total
Higher	6	22	25	23	76
Lower	52	38	31	43	164
Total	58	60	56	66	240

Fill in expected frequencies.

$$E = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

Answer:

9a

$$\chi^2 = \frac{(70 - 50)^2}{50} + \frac{(30 - 50)^2}{50}$$

Carry out the calculation of  $\chi^2$ .

10

$$\chi^2 = \frac{(70 - 50)^2}{50} + \frac{(30 - 50)^2}{50}$$

$$\chi^2 = \# \underline{\hspace{2cm}}$$

Answer:

33a

	A	B	C	D	Total
Higher	6 <u>18.4</u>	22 <u>19.0</u>	25 <u>17.7</u>	23 <u>20.9</u>	76
Lower	52 <u>39.6</u>	38 <u>41.0</u>	31 <u>38.3</u>	43 <u>45.1</u>	164
Total	58	60	56	66	240

$$E \text{ for A Higher} = \frac{(58)(76)}{240} = 18.4$$

$$E \text{ for C Higher} = \frac{(56)(76)}{240} = 17.7$$

$$E \text{ for A Lower} = \frac{(58)(164)}{240} = 39.6$$

$$E \text{ for C Lower} = \frac{(56)(164)}{240} = 38.3$$

$$E \text{ for B Higher} = \frac{(60)(76)}{240} = 19.0$$

$$E \text{ for D Higher} = \frac{(66)(76)}{240} = 20.9$$

$$E \text{ for B Lower} = \frac{(60)(164)}{240} = 41.0$$

$$E \text{ for D Lower} = \frac{(66)(164)}{240} = 45.1$$

Compute  $\chi^2$ .

34

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

$$\chi^2 = \# \underline{\hspace{2cm}}$$

Answer:  $\chi^2 = 16$

11

To see whether this  $\chi^2$  value is significant, we need to use Table I in the Guide. You will notice that across the top of the table there are probabilities. Again, the .05 or .01 columns are of interest in this test. Also, down the left-hand side you see degrees of freedom again.

34a

Answer:

$$\begin{aligned} \chi^2 &= \frac{(6 - 18.4)^2}{18.4} + \frac{(22 - 19)^2}{19} + \frac{(25 - 17.7)^2}{17.7} + \frac{(23 - 20.9)^2}{20.9} + \\ &\quad \frac{(52 - 39.6)^2}{39.6} + \frac{(38 - 41)^2}{41} + \frac{(31 - 38.3)^2}{38.3} + \frac{(43 - 45.1)^2}{45.1} \\ &= 8.36 + .47 + 3.01 + .21 + 3.88 + .22 + 1.39 + .10 \\ &= 17.64 \end{aligned}$$

35

Step 5: Determine critical values.

$$\alpha = .01$$

$$df = (\# \text{ rows} - 1) (\# \text{ columns} - 1)$$

$$= \# \underline{\hspace{2cm}}$$

$$\text{Critical Value} = \# \underline{\hspace{2cm}}$$

The degrees of freedom for the  $\chi^2$  test equals the number of classes that can be arbitrarily assigned. In this case after arbitrarily assigning the number of heads, the number of tails is determined (100 tosses - 70 heads = number of tails). Therefore, our degrees of freedom equal 3.

---

Answer: 3

35a

Critical Value = 11.345

---

Step 6: Make the appropriate decision.

36

Accept or Reject  $H_0$ ? \* \_\_\_\_\_

Why? † \_\_\_\_\_

Answer: one

Using a .05 significance level and 1 *df*, we find # \_\_\_\_\_ at the intersection of the column and row. This is our critical value.

Answer: Reject

Obtained  $\chi^2$  exceeds table value

We can then conclude that the stacks do differ and this difference is what was responsible for the difference between the observed and expected frequencies.

Problem III: Three sites have been monitored for TSP concentrations. Previous experiments have shown that all three sites failed to meet standards one-fourth of the time. A new series of samples has been taken at each site with the following results:

	A	B	C	Total
Meets Standards	40	50	42	132
Does Not Meet Standards	20	10	18	48
Total	60	60	60	180

Can we conclude that these results are other than we expected? (use  $\alpha = .00$ )  
(Go to the next frame for the step-by-step solution.)

Answer: 3.841

---

Since our obtained  $\chi^2$  (16.0) exceeds the tabled value (3.841) we (accept/reject) the null hypothesis.

---

Step 1: Define the problem.

$H_0$ : \* \_\_\_\_\_

$H_1$ : \* \_\_\_\_\_

Answer: reject

---

15

Our conclusion then is that something other than chance was responsible for the results we obtained (at the 5% level of significance). In other words, we are 95% confident that our coin is "crooked".

---

38a

Answer:  $H_0$ : observed frequencies = expected frequencies

$H_1$ : observed frequencies  $\neq$  expected frequencies

---

39

Step 2: Determine test and  $\alpha$  level.

test = \_\_\_\_\_

$\alpha = \frac{1}{n}$  \_\_\_\_\_

Many times you won't be able to figure the expected frequencies that easily. But, fear not, the statisticians have devised a method. Suppose we sample 150 people (100 men, 50 women) and sort them according to political party. We find among the men 66 Republicans and 34 Democrats and among the women 27 Republicans and 23 Democrats. We can put these data into a *contingency table*:

	Republicans	Democrats	Total
Men	66	34	100
Women	27	23	50
Total	93	57	150

Studying this table, we see

# \_\_\_\_\_ men Republicans

# \_\_\_\_\_ women Democrats

# \_\_\_\_\_ total Democrats

# \_\_\_\_\_ total women

Answer:

$\chi^2$

$\alpha = .001$

39a

40

Step 3: Organize and arrange data.

Not necessary.

Go to:

Frame 41

Page V-18

Answer: 66  
 23  
 57  
 50

We still need to estimate the expected frequencies in order to calculate the  $\chi^2$ .

	Republicans	Democrats	Totals
Men	66	34	100
Women	27	23	50
Total	93	57	150

To estimate the expected frequency of men Republicans from the sample data, we take the total number of Republicans, divide it by the grand total of the sample, and multiply by the total number of men. In this way we create a proportion of men to total and Republicans to total, thereby giving us a value for men Republicans if everything else were equal.

$$\begin{aligned}
 E \text{ of men Republicans} &= \frac{\text{total Republicans}}{\text{total sample}} \quad (\text{total men}) \\
 &= \frac{93}{150} (100) \\
 &= 62
 \end{aligned}$$

Likewise,

$$\begin{aligned}
 E \text{ of women Democrats} &= \frac{\text{total Democrats}}{\text{total sample}} \quad (\text{total women}) \\
 &= \frac{57}{150} (50) \\
 &= 19
 \end{aligned}$$

Fill in the expected frequencies beside the obtained frequencies in the following table:

	Republicans	Democrats	Total
Men	66	34	100
Women	27	23	50
Total	93	57	150

	Republicans	Democrats	Total
Men	66 <u>62</u>	34 <u>38</u>	100
Women	27 <u>31</u>	23 <u>19</u>	50
Total	93	57	150

$$E \text{ men Republicans} = \frac{93}{150} (100)$$

$$E \text{ women Republicans} = \frac{93}{150} (50)$$

$$E \text{ men Democrats} = \frac{57}{150} (100)$$

$$E \text{ women Democrats} = \frac{57}{150} (50)$$

Step 4: Compute the statistic.

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

	A	B	C	Total
Meets Standards	40	50	42	132
Does Not Meet Standards	20	10	18	48
Total	60	60	60	180

We are given the observed frequencies and now need to determine the expected frequencies. In the problem we are told that past experience has shown that the concentrations have failed to meet standards one-fourth of the time. We, therefore, would expect to obtain this same proportion in the current sampling. Therefore, since 60 measures were taken at each site, we would expect that one-fourth of the 60, or 15, values would not meet the standards and three-fourths (45) would meet or exceed the standards. These are, then, our expected frequencies for all three sites. Plug these values into the table: Notice we are not using the method you have learned for getting expected frequencies. We are using past experience instead of current data.

	A	B	C	Total
Meets Standards	40	50	42	132
Does Not Meet Standards	20	10	18	48
Total	60	60	60	180

There is a more mechanical method for determining expected frequencies that is a little easier. You'll notice that each row (horizontal) and each column (vertical) has a total. Each cell (block) is particular to 1 row and 1 column. To find the expected frequency for each cell: multiply that cell's row total by that cell's column total, and divide by the grand total. While this is exactly what you did before, thinking about it this way makes it easier to remember.

	Republicans	Democrats	Total
Men	66	34	100
Women	27	23	50
Total	93	57	150

To find E men Republicans again:

$$E = \frac{\text{total of row (100)} \times \text{total of column (93)}}{\text{grand total (150)}}$$

$$E = 62$$

41a

Answer:

	A	B	C	Total
Meets Standards	40 <u>45</u>	50 <u>45</u>	42 <u>45</u>	132
Does Not Meet Standards	20 <u>15</u>	10 <u>15</u>	18 <u>15</u>	48
Total	60	60	60	180

42

Using these frequencies, calculate:

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

$$\chi^2 = \# \underline{\hspace{2cm}}$$

We now know the observed frequency and expected frequency for each cell.  
 All that remains is the computation of  $\chi^2$ , which is exactly as before:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

Or, in this example:

66	<u>62</u>	34	<u>38</u>
27	<u>31</u>	23	<u>19</u>

$$\begin{aligned} \chi^2 &= \frac{(66 - 62)^2}{62} + \frac{(27 - 31)^2}{31} + \frac{(34 - 38)^2}{38} + \frac{(23 - 19)^2}{19} \\ &= .258 + .516 + .421 + .842 \\ &= 2.037 \end{aligned}$$

Answer:

42a

$$\begin{aligned} \chi^2 &= \frac{(40 - 45)^2}{45} + \frac{(50 - 45)^2}{45} + \frac{(42 - 45)^2}{45} + \\ &\quad \frac{(20 - 15)^2}{15} + \frac{(10 - 15)^2}{15} + \frac{(18 - 15)^2}{15} \\ \chi^2 &= .56 + .56 + .2 + 1.67 + 1.67 + .6 \\ \chi^2 &= 5.26 \end{aligned}$$

43

Step 5: Determine critical value.

$$\alpha = .001$$

$$df = (\# \text{ rows} - 1) (\# \text{ columns} - 1)$$

$$= \# \underline{\hspace{2cm}}$$

$$\text{Critical Value} = \# \underline{\hspace{2cm}}$$

Our next consideration is to compare this value with a tabled value. In problems like this an easy way to determine degrees of freedom is:

$$df = (\# \text{ of rows} - 1) (\# \text{ of columns} - 1)$$

In this example:

$$\begin{aligned} df &= (2 - 1) (2 - 1) \\ &= 1 \end{aligned}$$

So in Table I at  $\alpha = .05$  and 1 *df*, we find # \_\_\_\_\_ as the critical value.

43a

Answer: 2

Critical Value = 13.815

44

Step 6: Make the appropriate decision.

Accept or Reject  $H_0$ ? \* \_\_\_\_\_

Why? † \_\_\_\_\_

**Answer:** 3.841

20a

---

21

Since our obtained  $\chi^2$  (2.037) is less than the critical value (3.841), we (accept/reject) that the observed and expected frequencies did not differ significantly.

---

44a

**Answer:** Accept

Obtained  $\chi^2$  less than critical value

Answer: accept

In other words, the observed cell values did not differ significantly from what we predicted, knowing the column and row totals.

---

Let's try three examples using the six-step statistical problem-solving procedure:

Problem I: In rolling a die (singular of dice) 12 times, an experimenter obtains the following frequencies:

	$f$
one	4
two	0
three	3
four	1
five	2
six	2

Can we conclude that these results are significantly different from what we would expect, if the die were fair. (Go on to the next frame.)

---

To review: you will occasionally need to determine whether observed measurements differ significantly from the results that were expected. To test this, you employ a  $\chi^2$  (chi-square) test. In  $\chi^2$ , the hypotheses being tested are:

$H_0$ : observed frequencies = expected frequencies

$H_1$ : observed frequencies  $\neq$  expected frequencies

The statistic is calculated by the formula:

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

where O = the observed frequency

E = the expected frequency

Step 1: Define the problem.

$H_0$ : \* \_\_\_\_\_

$H_1$ : \* \_\_\_\_\_

Frequently, however, the expected frequencies are not derivable from theory or past experience, and have to be estimated from contingency tables: 46

	A	B	C	Totals
K				
L				
M				
Total				Grand Total

E for the cell AK is determined by  $\frac{(\text{row K total}) (\text{column A total})}{\text{grand total}}$ .

Similarly, E for cell BK =  $\frac{(\text{row K total}) (\text{column B total})}{\text{grand total}}$ .

The statistic can then be calculated using the  $\chi^2$  formula:

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

Answer:  $H_0$ : observed frequencies = expected frequencies

$H_1$ : observed frequencies  $\neq$  expected frequencies

---

24

Step 2: Determine test and set  $\alpha$  level.

test is \* \_\_\_\_\_

$\alpha = .05$  (set arbitrarily since it wasn't given in problem)

Go to:

Frame 24e

Page V-1

---

47

The obtained  $\chi^2$  is then compared with the critical values found in Table I. Degrees of freedom equal (in cases when a contingency table is not used) the number of frequencies that can be arbitrarily set until the last one is determined; or, when a contingency table is used, degrees of freedom can be found by:

$$df = (\# \text{ of rows} - 1) (\# \text{ of columns} - 1)$$

If the obtained  $\chi^2$  exceeds the critical value, then the null hypothesis is rejected.

Proceed to Module VI.





