STATISTICS FOR DATA SCIENCE

HYPOTHESIS and INFERENCE

Dr. Deepa Nair
Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

UNIT-4 HYPOTHESIS and INFERENCE

Session-10

Chi-squared Test

Dr. Deepa Nair
Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Homogeneity:

- Sometimes several multinomial trials are conducted, each with the same set of possible outcomes.
- The null hypothesis is that the probabilities of the outcomes are the same for each experiment.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Homogeneity

In the previous example, we tested the null hypothesis that the probabilities of the outcomes for a multinomial trial were equal to a specified set of values. Sometimes several multinomial trials are conducted, each with the same set of possible outcomes. The null hypothesis is that the probabilities of the outcomes are the same for each experiment. We present an example. Four machines manufacture cylindrical steel pins. The pins are subject to a diameter specification. A pin may meet the specification, or it may be too thin or too thick. Pins are sampled from each machine, and the number of pins in each category is counted. Table 6.4 presents the results.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Observed numbers of pins in various categories with regard to a diameter specification

	TO Thin	ok	To thick	TOTAL
Machine 1	10	102	8	120
Machine 2	34	161	5	200
Machine 3	12	79	9	100
Machine 4	10	60	10	80
TOTAL	66	402	32	500

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Table above is an example of a contingency table. Each row specifies a category regarding one criterion (machine, in this case), and each column specifies a category regarding another criterion (thickness, in this case). Each intersection of row and column is called a cell, so there are 12 cells in Table 6.4.

The number in the cell at the intersection of row i and column j is the number of trials whose outcome was observed to fall into row category i and into column category j. This number is called the observed value for cell i j. Note that we have included the totals of the observed values for each row and column. These are called the marginal total

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The null hypothesis is that the proportion of pins that are too thin, OK, or too thick is the same for all machines. More generally, the null hypothesis says that no matter which row is chosen, the probabilities of the outcomes associated with the columns are the same. We will develop some notation with which to express H0 and to define the test statistic.

Let I denote the number of rows in the table, and let J denote the number of columns. Let pi j denote the probability that the outcome of a trial falls into column j given that it is in row i. Then the null hypothesis is

H_0 : For each column $j, P_{1j} = \cdots P_{Ij}$

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Let O_{ij} denote the observed value in cell $i\,j$. Let $O_{i}.$ denote the sum of the observed values in row i, let $O.j$ denote the sum of the observed values in column j, and let $O..$ denote the sum of the observed values in all the cells (see Table 6.5).

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Notation for observed values

	Column 1	Column 2	…….	Column J	Total
ROW1	O_{11}	O_{21}	…….	O_{1J}	$O_{1.}$
ROW2	O_{21}	O_{22}	……	O_{2J}	$O_{1.}$
.	.	.	…….	.	.
.	.	.	……	.	.
ROW I	O_{I1}	O_{I2}	…….	O_{IJ}	O_{I}
Total	$O_{.1}$	$O_{.2}$	…….	$O_{.J}$	$O_{..}$

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

To define a test statistic, we must compute an expected value for each cell in the table. Under H_0, the probability that the outcome of a trial falls into column j is the same for each row i. The best estimate of this probability is the proportion of trials whose outcome falls into column j. This proportion is $O.j\ /O$ … We need to compute the expected number of trials whose outcome falls into cell $i\ j$. We denote this expected value by E_{ij} . It is equal to the proportion of trials whose outcome falls into column j, multiplied by the number O_i. of trials in row i. That is,

$$E_{ij} = \frac{O_i O_j}{O_{..}}$$

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The test statistic is based on the differences between the observed and expected values:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Under H_0, this test statistic has a chi-square distribution with $(I-1)(J-1)$ degrees of freedom. Use of the chi-square distribution is appropriate whenever the expected values are all greater than or equal to 5.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Homogeneity:

Example:

	T00 thin	OK	Too thick	Total
Machine 1	10	102	8	120
Machine 2	34	161	5	200
Machine 3	12	79	9	100
Machine 4	10	60	10	80
Total	66	402	32	500

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Homogeneity:

Example:

Expected Values

	T00 thin	OK	Too thick	Total
Machine 1	15.84	96.48	7.68	120
Machine 2	26.40	160.8	12.8	200
Machine 3	13.2	80.4	6.40	100
Machine 4	10.58	64.32	5.12	80
Machine 5	66	402	32	500

Solution:

$$\chi 2 = (10 - 15.84)2/15.84 + \cdots + (10 - 5.12)2/5.12$$
$$= 34.1056/15.84 + \cdots + 23.8144/5.12$$
$$= 15.5844$$

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Solution:

- Since there are four rows and three columns, the number of degrees of freedom is $(4-1)(3-1) = 6$.
- To obtain the *P*-value, we consult the chi-square table Looking under six degrees of freedom, we find that the upper 2.5% point is 14.449, and the upper 1% point is 16.812.
- Therefore $0.01 < P < 0.025$. It is reasonable to conclude that the machines differ in the proportions of pins that are too thin, OK, or too thick.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Independence:

- In some cases, both row and column totals are random. In either case, we can test the null hypothesis that the probabilities of the column outcomes are the same for each row outcome, and the test is exactly the same in both cases.

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Independence:

Example:

- The cylindrical steel pins in previous example are subject to a length specification as well as a diameter specification.

- With respect to the length, a pin may meet the specification, or it may be too short or too long. A total of 1021 pins are sampled and categorized

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Example:

Observed Values

Length	T00 thin	OK	Too thick	Total
Too Short	13	117	4	134
OK	62	664	80	806
Too Long	5	68	8	81
Total	80	849	92	1021

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Example:

Expected Values

Length	T00 thin	OK	Too thick	Total
Too Short	10.50	111.43	12.07	134
OK	63.15	670.22	72.63	806
Too Long	6.35	67.36	7.30	81
Total	80	849	92	1021

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

$$\chi^2 = \frac{(13 - 10.50)^2}{10.50} + \cdots + \frac{(8 - 7.30)^2}{7.30}$$

$$= \frac{6.25}{10.50} + \cdots + \frac{0.49}{7.30} = 7.46$$

Source: Text Book Chapter6

STATISTICS FOR DATA SCIENCE

Chi-squared Test

- Since there are three rows and three columns, the number of degrees of freedom is $(3 - 1)(3 - 1) = 4$.
- To obtain the *P*-value, we consult the chi-square table Looking under four degrees of freedom, we find that the upper 10% point is 7.779. We conclude that $P > 0.10$.
- There is no evidence that the length and thickness are related.

Source: Text Book Chapter6

Dr. Deepa Nair

Department of Science and Humanities

deepanair@pes.edu