The Chi Square Test

Dr.Mamatha.H.R

Professor

Department of Computer Science and Engineering

PES University

Bangalore

Course material created using various Internet resources and text book

- Bernoulli trial, is a process that results in one of two possible outcomes, labelled “success” and “failure.”
- If a number of Bernoulli trials are conducted, and the number of successes is counted, we can test the null hypothesis that the success probability p is equal to a pre specified value p_0 .
- If two sets of Bernoulli trials are conducted, with success probability p_1 for the first set and p_2 for the second set, we can test the null hypothesis that $p_1 = p_2$.

- A generalization of the Bernoulli trial is the multinomial trial which is an experiment that can result in any one of k outcomes, where $k \geq 2$.
- The probabilities of the k outcomes are denoted p1 , . . . , pk .

- For example, the roll of a fair die is a multinomial trial with six outcomes 1, 2, 3, 4, 5, 6; and probabilities p1 =p2 = p3 = p4 = p5 = p6 = 1/6.

- we generalize the tests for a Bernoulli probability to multinomial trials.

- We begin with an example in which we test
- the null hypothesis that the multinomial probabilities p1 , p2 , . . . , pk are equal to a
- prespecified set of values p01 , p02 , . . . , p 0k , so that the null hypothesis has the form
- H 0 : p1 = p01 , p2 = p02 , . . . , pk = p0k .

- Imagine that a gambler wants to test a die to see whether it deviates from fairness.
- Let pi be the probability that the number i comes up.
- The null hypothesis will state that the die

is fair, so the probabilities specified under the null hypothesis are p01 = · · · = p06 = 1/6.

- The null hypothesis is H 0 : p1 = · · · = p6 = 1/6.

- The gambler rolls the die 600 times and obtains the results shown in Table

TABLE 6.3 Observed and expected values for 600 rolls of a die

Category	Observed	Expected
1	115	100
2	97	100
3	91	100
4	101	100
5	110	100
6	86	100
Total	600	600

- The expected value for a given outcome is the mean number of trials that would result in that outcome if *H*0 were true.

- To compute the expected values, let *N* be the total number of trials. (In the die example, *N* = 600.)
- When *H*0 is true, the probability that a trial results in outcome *i* is *p*0*i* , so the expected number of trials resulting in outcome *i* is *N*p0*i* .

In the die example, the expected number of trials for each outcome is 100.

- The idea behind the hypothesis test is that if *H*0 is true, then the observed and expected values are likely to be close to each other.

- Therefore we will construct a test statistic that measures the closeness of the observed to the expected values.

- The statistic is called the **chi-square statistic**

TABLE 6.3 Observed and expected values for 600 rolls of a die

Category	Observed	Expected
1	115	100
2	97	100
3	91	100
4	101	100
5	110	100
6	86	100
Total	600	600

Chi-Square as a Statistical Test

- ***Chi-square test:*** an **inferential statistics** technique designed to test for **significant relationships** between two variables organized in a bivariate table.

- Chi-square requires **no assumptions** about the shape of the population distribution from which a sample is drawn.

The Chi Square Test

- A statistical method used to determine goodness of fit
  - Goodness of fit refers to how close the observed data are to those predicted from a hypothesis

- Note:
  - The chi square test does not prove that a hypothesis is correct
    - It evaluates to what extent the data and the hypothesis have a good fit

Limitations of the Chi-Square Test

- The chi-square test does **not** give us much information about the ***strength*** of the relationship or its ***substantive significance*** in the population.
- The chi-square test is **sensitive** to ***sample size***. The size of the calculated chi-square is **directly proportional** to the size of the sample, independent of the strength of the relationship between the variables.
- The chi-square test is also **sensitive** to **small expected frequencies** in one or more of the cells in the table.

Statistical Independence

- ***Independence (statistical):*** the **absence of association** between two cross-tabulated variables. The percentage distributions of the dependent variable within each category of the independent variable are **identical**.

Hypothesis Testing with Chi-Square

Chi-square follows five steps:

1. Making assumptions (**random sampling**)
2. Stating the research and null hypotheses
3. Selecting the sampling distribution and specifying the test statistic
4. Computing the test statistic
5. Making a decision and interpreting the results

The Assumptions

- The chi-square test requires **no assumptions** about the **shape of the population distribution** from which the sample was drawn.

- However, like all inferential techniques it assumes **random sampling**.

Stating Research and Null Hypotheses

- The **research hypothesis** (H_1) proposes that the two variables are **related** in the population.

- The **null hypothesis** (H_0) states that **no association exists** between the two cross-tabulated variables in the population, and therefore the variables are **statistically independent**.

H_1: The two variables are **related** in the population.

Gender and fear of walking alone at night are ***statistically dependent***.

Afraid	Men	Women	Total
No	83.3%	57.2%	**71.1%**
Yes	16.7%	42.8%	**28.9%**
Total	**100%**	**100%**	**100%**

H_0: There is **no association** between the two variables.

Gender and fear of walking alone at night are ***statistically independent***.

Afraid	Men	Women	Total
No	71.1%	71.1%	**71.1%**
Yes	28.9%	28.9%	**28.9%**
Total	**100%**	**100%**	**100%**

The Concept of Expected Frequencies

Expected frequencies f_e : the cell frequencies that would be **expected** in a bivariate table **if** the two variables were **statistically independent**.

Observed frequencies f_o: the cell frequencies **actually observed** in a bivariate table.

Calculating Expected Frequencies

$$f_e = \frac{\text{(column marginal)(row marginal)}}{N}$$

To obtain the expected frequencies for any cell in any cross-tabulation in which the two variables are assumed independent, **multiply** the row and column totals for that cell and **divide** the product by the total number of cases in the table.

Chi-Square (obtained)

- The test statistic that **summarizes** the differences between the **observed** (*fo*) and the **expected** (*fe*) frequencies in a bivariate table.

Calculating the Obtained Chi-Square

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e}$$

f_e = expected frequencies

f_o = observed frequencies

- The larger the value of $\chi 2$, the stronger the evidence against $H0$.
- To determine the *P*-value for the test, we must know the null distribution of this test statistic.
- In general,we cannot determine the null distribution exactly.
- However, when the expected values are all sufficiently large, a good approximation is available.
- It is called the **chi-square distribution** with $k - 1$ degrees of freedom, denoted $\chi 2_{k-1}$.

- Note :number of degrees of freedom is one less than the number of categories.

- Use of the chi square distribution is appropriate whenever all the expected values are greater than or equal to 5.

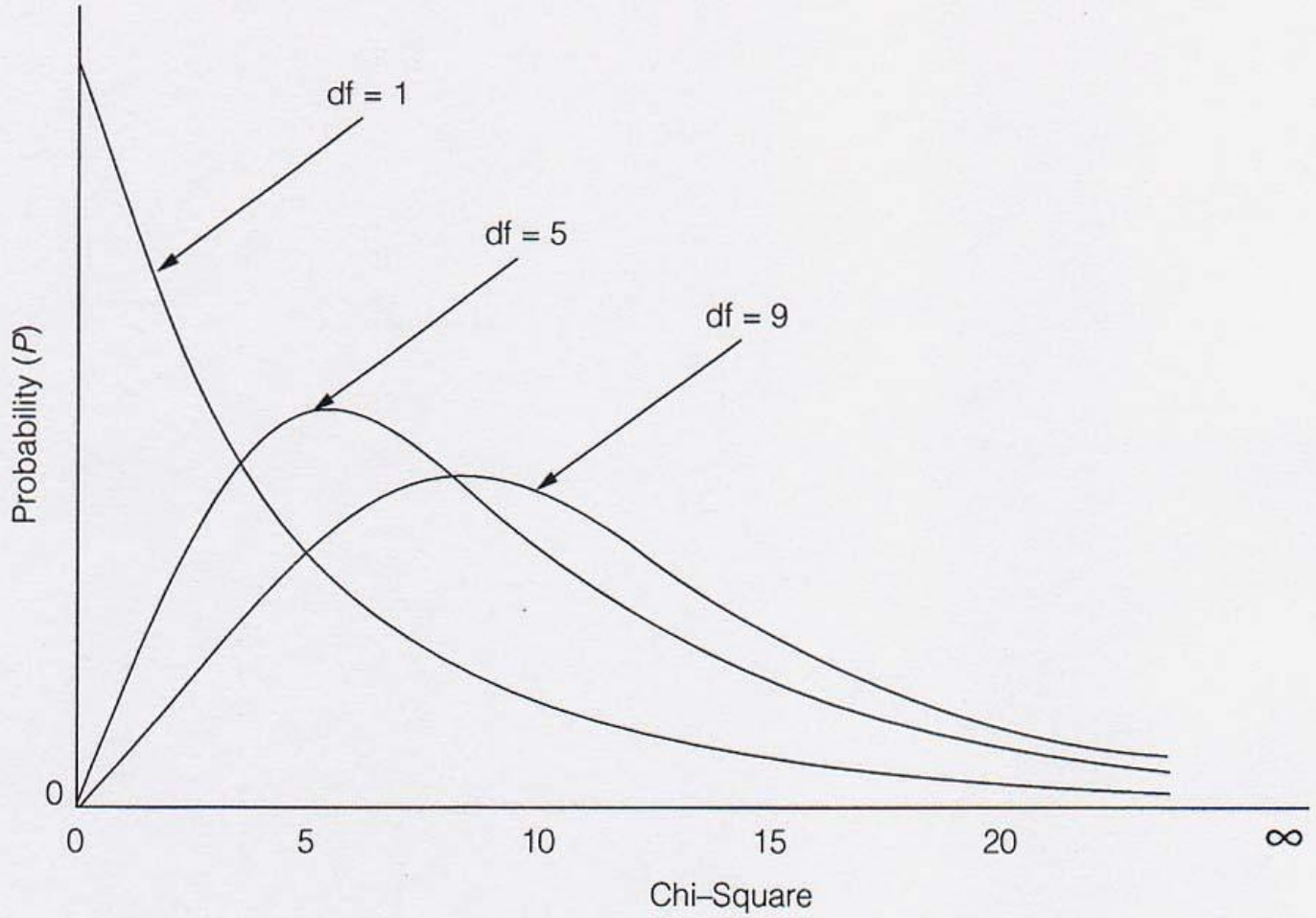
The Sampling Distribution of Chi-Square

- The sampling distribution of chi-square tells the **probability** of getting values of chi-square, **assuming no relationship** exists in the population.
- The chi-square sampling distributions depend on the **degrees of freedom**.
- The χ^2 sampling distribution is not one distribution, but is **a family of distributions**.

The Sampling Distribution of Chi-Square

- The distributions are **positively skewed**. The research hypothesis for the chi-square is **always** a **one-tailed test**.
- Chi-square values are **always positive**. The minimum possible value is zero, with **no upper limit** to its maximum value.
- As the number of degrees of freedom increases, the χ^2 distribution becomes **more symmetrical**.

Figure 14.1 **Chi-Square Distributions for 1, 5, and 9 Degrees of Freedom**



Determining the Degrees of Freedom

$$df = (r - 1)(c - 1)$$

where

r = the number of rows

c = the number of columns

Calculating Degrees of Freedom

How many degrees of freedom would a table with 3 rows and 2 columns have?

$(3 - 1)(2 - 1) = 2$

2 degrees of freedom

The Chi Square Test

- The general formula is

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

- where
  - O = observed data in each category
  - E = observed data in each category based on the experimenter's hypothesis
  - Σ = Sum of the calculations for each category

- Applying the chi square test
  - Step 1: Propose a null hypothesis (Ho) that allows us to calculate the expected values
  - Step 2: Calculate the expected values, based on the hypothesis
  - Step 3: Apply the chi square formula

- Step 4: Interpret the chi square value
  - The calculated chi square value can be used to obtain probabilities, or P values, from a chi square table
    - These probabilities allow us to determine the likelihood that the observed deviations are due to random chance alone
  - Low chi square values indicate a high probability that the observed deviations could be due to random chance alone
  - High chi square values indicate a low probability that the observed deviations are due to random chance alone
  - If the chi square value results in a probability that is less than 0.05 (ie: less than 5%) it is considered ***statistically significant***
    - The hypothesis is rejected

- Step 4: Interpret the chi square value

  - Before we can use the chi square table, we have to determine the degrees of freedom (*df*)
    - The *df* is a measure of the number of categories that are independent of each other
    - If you know the 3 of the 4 categories you can deduce the 4^{th} (total number– categories 1-3)
    - $df = n - 1$
      - where n = total number of categories

TABLE 2.1

Chi Square Values and Probability

Degrees of Freedom	$P = 0.99$	0.95	0.80	0.50	0.20	0.05	0.01
1	0.000157	0.00393	0.0642	0.455	1.642	**3.841**	**6.635**
2	0.020	0.103	0.446	1.386	3.219	**5.991**	**9.210**
3	0.115	0.352	1.005	2.366	4.642	**7.815**	**11.345**
4	0.297	0.711	1.649	3.357	5.989	**9.488**	**13.277**
5	0.554	1.145	2.343	4.351	7.289	**11.070**	**15.086**
6	0.872	1.635	3.070	5.348	8.558	**12.592**	**16.812**
7	1.239	2.167	3.822	6.346	9.803	**14.067**	**18.475**
8	1.646	2.733	4.594	7.344	11.030	**15.507**	**20.090**
9	2.088	3.325	5.380	8.343	12.242	**16.919**	**21.666**
10	2.558	3.940	6.179	9.342	13.442	**18.307**	**23.209**
15	5.229	7.261	10.307	14.339	19.311	**24.996**	**30.578**
20	8.260	10.851	14.578	19.337	25.038	**31.410**	**37.566**
25	11.524	14.611	18.940	24.337	30.675	**37.652**	**44.314**
30	14.953	18.493	23.364	29.336	36.250	**43.773**	**50.892**

1.06

From Fisher, R. A., and Yates, F. (1943) *Statistical Tables for Biological, Agricultural, and Medical Research.* Oliver and Boyd, London.

- Step 4: Interpret the chi square value
  - With *df* = *3,* the chi square value of 1.06 is slightly greater than 1.005 (which corresponds to P-value = 0.80)
  - P-value = 0.80 means that Chi-square values equal to or greater than 1.005 are expected to occur 80% of the time due to random chance alone; that is, when the null hypothesis is true.
  - Therefore, it is quite probable that the deviations between the observed and expected values in this experiment can be explained by random sampling error and *the null hypothesis is not rejected*.

TABLE 6.3 Observed and expected values for 600 rolls of a die

Category	Observed	Expected
1	115	100
2	97	100
3	91	100
4	101	100
5	110	100
6	86	100
Total	600	600

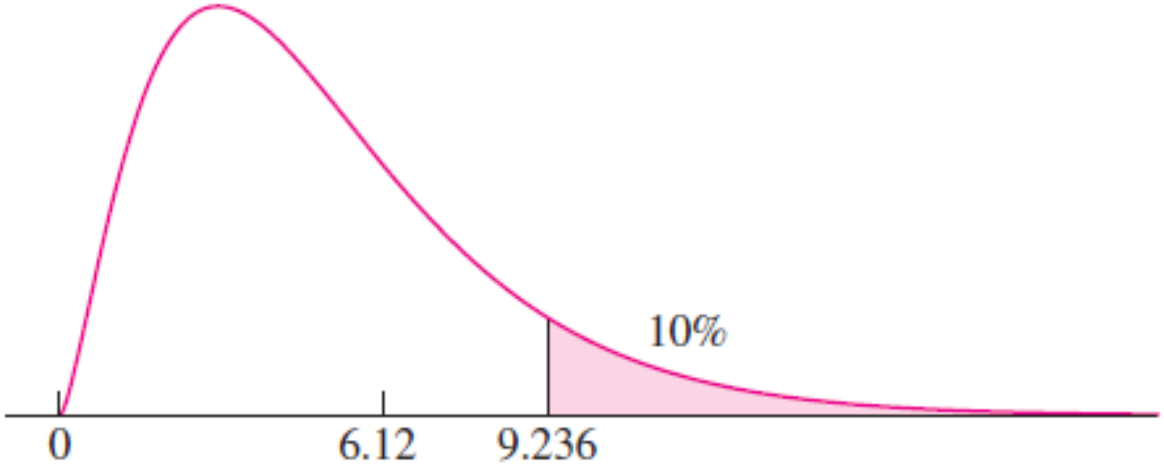


FIGURE 6.20 Probability density function of the χ^2_5 distribution. The observed value of the test statistic is 6.12. The upper 10% point is 9.236. Therefore the P-value is greater than 0.10.

- We conclude that $P > 0.10$.
- There is no evidence to suggest that the die is not fair.

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 pre 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.

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

	Too 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

- 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 *H*0 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*1 *j* = ⋯ = *pI j*

- Let *Oi j* denote the observed value in cell *i j*.
- Let *Oi.* 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

TABLE 6.5 Notation for observed values

	Column 1	Column 2	⋯	Column J	Total
Row 1	O_{11}	O_{12}	⋯	O_{1J}	$O_{1.}$
Row 2	O_{21}	O_{22}	⋯	O_{2J}	$O_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
Row I	O_{I1}	O_{I2}	⋯	O_{IJ}	$O_{I.}$
Total	$O_{.1}$	$O_{.2}$	⋯	$O_{.J}$	$O_{..}$

$$E_{ij} = \frac{O_{i.}O_{.j}}{O_{..}} \tag{6.8}$$

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

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{6.9}$$

Expected values for Table 6.4

	Too Thin	OK	Too Thick	Total
Machine 1	15.84	96.48	7.68	120.00
Machine 2	26.40	160.80	12.80	200.00
Machine 3	13.20	80.40	6.40	100.00
Machine 4	10.56	64.32	5.12	80.00
Total	66.00	402.00	32.00	500.00

Note that the observed row and column totals are identical to the expected row and column totals.

TABLE 6.4 Observed numbers of pins in various categories with regard to a diameter specification

	Too 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

Expected values for Table 6.4

	Too Thin	OK	Too Thick	Total
Machine 1	15.84	96.48	7.68	120.00
Machine 2	26.40	160.80	12.80	200.00
Machine 3	13.20	80.40	6.40	100.00
Machine 4	10.56	64.32	5.12	80.00
Total	66.00	402.00	32.00	500.00

- 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.

The Chi-Square Test for Independence

- In the previous Example the column totals were random, while the row totals were presumably fixed in advance, since they represented numbers of items sampled from various machines.
- 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

- 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 with respect to both length and diameter specification. The results are presented in the following table. Test the null hypothesis that the proportions of pins that are too thin, OK, or too thick with respect to the diameter specification do not depend on the classification with respect to the length specification.

Observed Values for 1021 Steel Pins

	Diameter			
Length	Too 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

Expected Values for 1021 Steel Pins

Length	Diameter: Too Thin	Diameter: OK	Diameter: Too Thick	Total
Too Short	10.50	111.43	12.07	134.0
OK	63.15	670.22	72.63	806.0
Too Long	6.35	67.36	7.30	81.0
Total	80.0	849.0	92.0	1021.0

- 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.