**Chi-Square Review**

|  |
| --- |
| What is Chi-Square Test?  A chi-square test is a hypothesis test that compares the observed distribution of your data to an expected distribution of data. |

Types of Chi-square test:

|  |
| --- |
| Chi-square goodness of fit test:  To test how well a sample of categorical data fits the theoretical distribution.  e.g. you can test whether a die is fair by rolling the die many times and using a chi-square goodness-of-fit test to determine whether the results follow a uniform distribution. In this case, the chi-square test quantifies how much the observed distribution of counts varies from the hypothesized distribution. |
| Chi-square test of association and independence:  The calculation part is the same but the question you are trying to answer is different.   * Test of association: use a test of association to determine whether a variable is associated with another different variable. E.g. determine whether the sales of different colors of cars depends on the city where they are sold.( appropriate a\when explanatory variable and response variable are clearly defined in the beginning of the study) * Test of independence: use a test of independence to determine whether the observed value of one variable depends on the observed value of a different variable. E.g. determine whether the candidate that a person votes for is independent of the voter’s gender.( appropriate when one sample is selected from one population and thus neither explanatory or response totals are known in advance) |

A two-way table is called contingency table, which is a useful tool for examining the relationship between categorical variables.

**E.g test of association:**

The data comes from the sinking of Titanic. Was there a relationship between the kind of ticket a passenger held on the Titanic and the passenger’s chance of making it into the lifeboat?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | First | Second | Third | Crew | **Total** |
| Alive | 202 | 118 | 178 | 212 | 710 |
| Dead | 123 | 167 | 528 | 673 | 1491 |
| **Total** | 325 | 285 | 706 | 885 | 2201 |

\*1. The way of representing categorical data using a two-way table is known as a “contingency table” representation. 2. We should and could also view this data set in a graphic form called “segmented bar chart”.

* Ho:Survival was independent of status on ship
* Ha:Survival was not independent of status on ship. The two variables are associated

Expected frequency = (Row Total)\*(Column Total)**/**(Grand Total)

E.g 104.84 = (710\*325)**/**(2201); 220.16=(1491\*325)/(2201)

“Expected Frequencies”

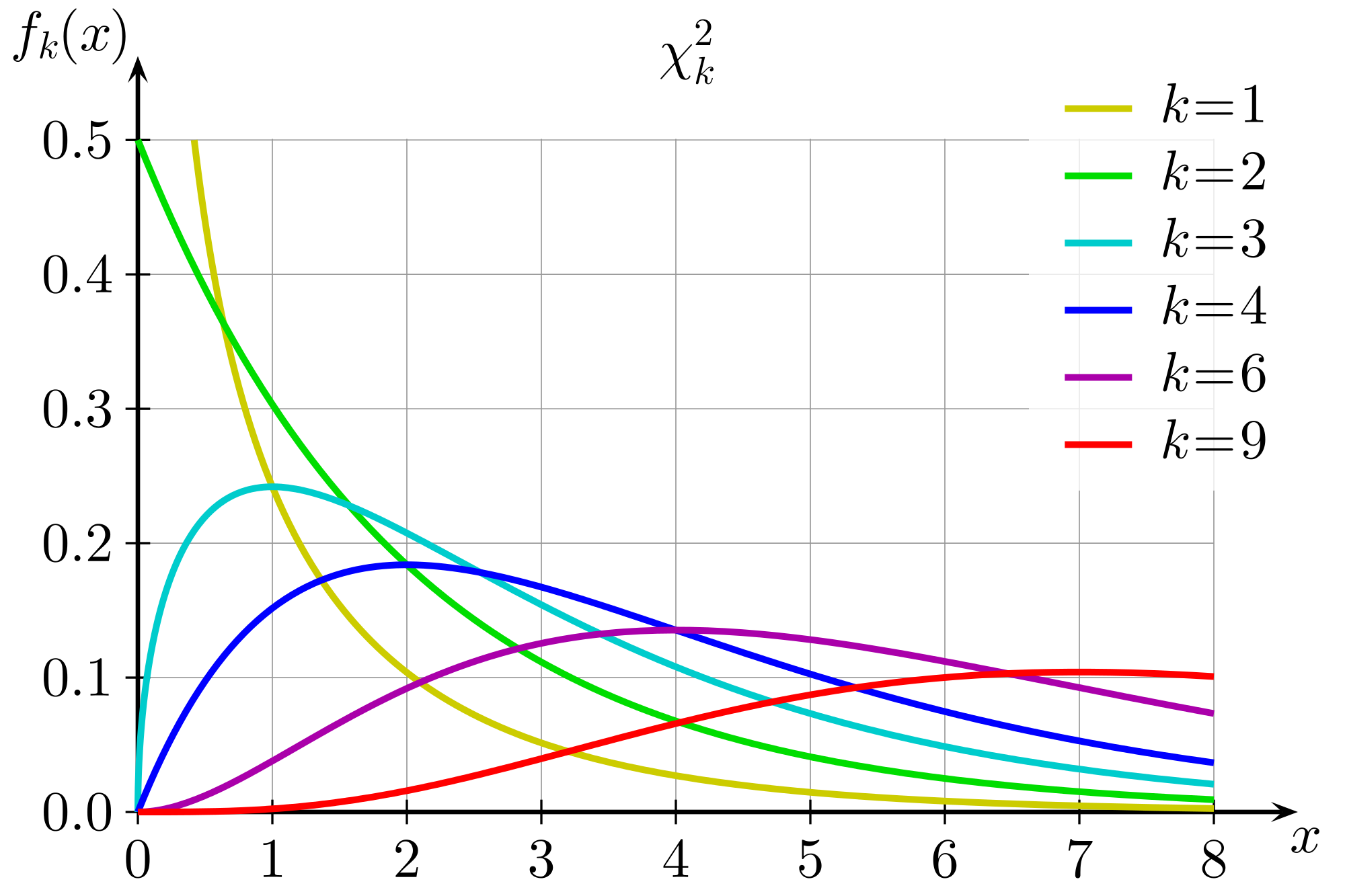
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | First | Second | Third | Crew | Total |
| Alive | 104.84 | 91.94 | 227.74 | 285.48 | 710 |
| Dead | 220.16 | 193.06 | 478.26 | 599.52 | 1491 |
| Total | 325 | 285 | 706 | 885 | 2201 |

\* Note that the expected frequencies(E) is different than the observed frequencies(O)

\* Test test statistic which we use to summarize the discrepancies between the observed and expected frequencies is given by 2 **/** E, where each cell has 2 **/** E

\* If the O’s and E’s agree well, this test statistic has a small value. Poor agreement gives large value

\* The test statistic has approximately a chi-square distribution if the null hypothesis of no association is true



Properties of Chi-square:

* Always > or = 0
* Skewed right
* Has integer df’s. df = (r-1)\*(c-1), r = number of rows in table, c = number of columns in table
* Has upper -a% critical points given in the Chi=square tables

In this test, df = (2-1)(4-1) = 3

The 5% critical value in chi-square table for df =3 is equal to 7.815

Our observed value is 187.793.

Since 197.793 > 7.815. We rejected the Ho.

|  |
| --- |
| Assumption and Condition required for a Chi-square test of association or independence |
| 1. Data must be in counts - check by examining the data |
| 2. Data in sample are independent - do we have a SRS and n < 10% of population? |
| 3. Sample size is sufficiently large - are all expected counts are greater than 5 |

**E.g test of goodness of fit:**

Table 1.1 Choice of correct responses of 400 AP multiple choice questions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Answer | A | B | C | D | E |
| Frequency | 85 | 90 | 79 | 78 | 68 |

The relative frequency of Table 1.1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Answer | A | B | C | D | E |
| Frequency | 85/400=0.2125 | 0.2250 | 0.1975 | 0.1950 | 0.1700 |

\*We can use a bar chart to examine the data like this

* Ho:pa=pb=pc=pd=pe=0.2
* Ha:at least one p =/= 0.20

Table 1.2 The expected counts

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Answer | A | B | C | D | E |
| Frequency | 400\*0.2=80 | 80 | 80 | 80 | 80 |

The test statistics : 2 **/** E, where each cell has 2 **/** E

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | R | Sum |
| Observed | 85 | 90 | 79 | 78 | 68 | 400 |
| Expected | 80 | 80 | 80 | 80 | 80 | 400 |
| 2 **/** E | 0.3125 | 1.25 | 0.0125 | 0.05 | 1.8 | 3.425 |

* Find degree of freedom: df = number of categories - 1
* Use statistical software to find P-value, the test should always be a right-tailed test. In this case, p = 0.4894. As the P-value is very large, there is no evidence to support the claim that the correct responses to the MC items are not all equally likely to occur.

|  |
| --- |
| Assumption and Condition required for a Chi-square goodness of fit test |
| 1. Data must be in counts - check by examining the data |
| 2. We have a random sample |
| 3. Sample size is large enough. (look at the expected counts, are each of the expected counts > 5? If so you can proceed some books suggest that no more than 20% of the expected counts be less than 5 |