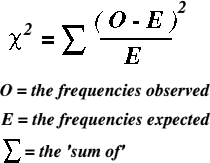
**A few notes on goodness-of-fit tests and the chi-square distribution**

The chi-square distribution can be used for goodness-of-fit tests, in which some observed frequencies are compared to expected frequencies (from a hypothesis or model).

The idea is the following. Let’s think of a die with FIVE sides. We would like to know whether it is well balanced (that is, if all sides are equally likely, this is the "null hypothesis") or not. To test our hypothesis, we roll the dice 100 times and count the number of ones, twos, threes, fours and fives. Let’s assume that the observed frequencies are the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| Observed\_freq | 30 | 10 | 20 | 15 | 25 |

Are these data compatible with our hypothesis that the die is balanced? To do that, we need to estimate the expected frequencies and see their similarities/differences. The expected frequencies can be deducted from the hypothesis that the die is balanced: out of 100 launches, as all the sides are assumed to be equally likely (probability=1/5 in each side), we expect a frequency of 20 for each side (100 \* 1/5):



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| Observed\_freq | 30 | 10 | 20 | 15 | 25 |
| Expected\_freq | 20 | 20 | 20 | 20 | 20 |

And now…how can we estimate the “goodness” of this fit? Well, we can do that using the "chi square" test. The formula is the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| Observed\_freq | 30 | 10 | 20 | 15 | 25 |
| Expected\_freq | 20 | 20 | 20 | 20 | 20 |
|  |  |  |  |  |  |
|  | 5 | 5 | 0 | 1.25 | 1.25 |
|  |  |  |  |  |  |
|  | SUM | 12.5 |  |  |  |

The sum of 12.5 (empirical chi-square), what does it mean? Is it high? The fit is not perfect, but is it still reasonable with the idea that the die is balanced?

The key idea here is that, if the dice were really balanced, the empirical "chi-square" would follow a well-known theoretical distribution from which we can deduce percentiles and so on: CHI-SQUARE WITH FOUR DEGREES OF FREEDOM (The 4 degrees of freedom is because actually knowing four of the frequencies of the cells, the fifth can be deducted, since the sum of frequencies of the cells in the example must be 100.)

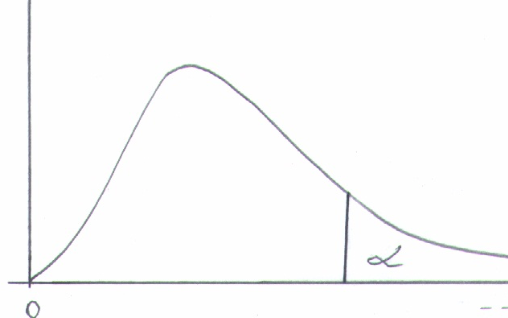
NOTE: if the dice were balanced and we do the test, say, 1000000000 times, and we plot a histogram, and we smooth it, then this smoothed histogram would fit perfectly with the theoretical chi-square distribution.

Ok, but now how do we decide whether to keep the "null hypothesis" (the die is balanced) or not? Obviously, a very small value in the chi-square test is an indication that the die is balanced, and a very high chi-square is an indication that the die is not. Where’s the cutoff point for the decision? In psychology, the boundary is traditionally the 95th percentile:

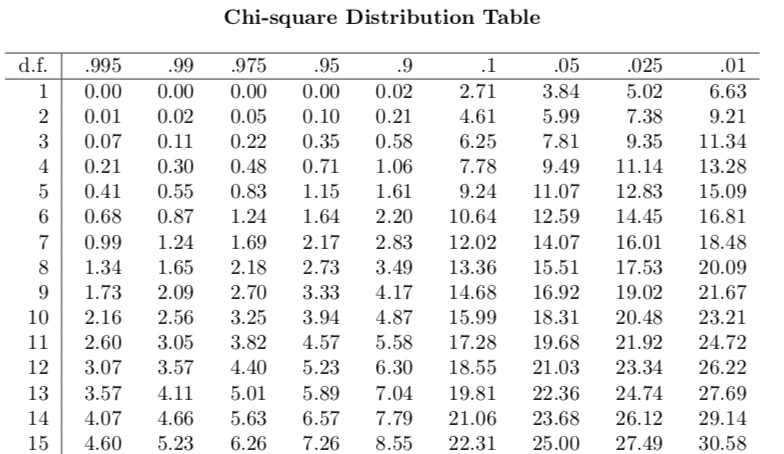
--If the result of the "chi-square" test exceeds the 95th percentile, then we say that the adjustment is too poor, so we reject the "null hypothesis" and conclude that the die is not balanced.

--If the result of the "chi-square" test is lower than the 95th percentile, then we say that the adjustment is ok, we keep the "null hypothesis" and conclude that the die is balanced.

What about the example?



The 95th percentile (in the table, the figures reflect the proportion of data to the right) is 9.49.



Since the result "chi-square" test (12.5) is greater than the 95th percentile (9.49), we conclude that the die is not balanced.

Another exercise:

A social psychologist claims that when two individuals sit down at a cafeteria table with 4 seats, they chose their seats at chance. To test this hypothesis, the social psychologist goes to a big cafeteria and counts the number of tables in which only two of two people are seated. Then she counts the number of cases in which the two individuals are facing each other (observed frequency=150) and the number of case in which the two individuals are next to each other (observed frequency=150). Are these data compatible with her hypothesis?

(Note: This is not as easy as it may seem at first glance. You need to pay close attention when deducing the expected frequencies.)