**AP Biology *χ*2 (chi square test for statistical significance)**

Pearson's chi-squared test is used to assess two types of comparison: tests of [goodness of fit](https://en.wikipedia.org/wiki/Goodness_of_fit) and tests of [independence](https://en.wikipedia.org/wiki/Independence_%28probability_theory%29)…and we will only study the latter.

* A test of **goodness of fit** establishes whether or not an observed [frequency distribution](https://en.wikipedia.org/wiki/Frequency_distribution) differs from a theoretical distribution. *This is the one we use in our genetics experiments!*
* A **test of independence** assesses whether paired observations on two variables, expressed in a [contingency table](https://en.wikipedia.org/wiki/Contingency_table), are independent of each other (e.g. polling responses from people of different nationalities to see if one's nationality is related to the response).

The procedure of the test of goodness of fit includes the following steps:

1. Calculate the chi-squared test [statistic](https://en.wikipedia.org/wiki/Statistic), , (see below on page 2).
2. Determine the [degrees of freedom](https://en.wikipedia.org/wiki/Degrees_of_freedom_%28statistics%29), *df*, of that statistic which is essentially the number of outcomes reduced by one.
3. Compare to the critical value from the [chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution) taking into account the *df* (degrees of freedom).

## Test for fit of a distribution

It should be noted that the degrees of freedom are not based on the number of observations. For example, if testing for a fair, six-sided die, there would be five degrees of freedom because there are six categories/parameters/outcomes (each number). The number of times the die is rolled will have absolutely no effect on the number of degrees of freedom. In our problem sets the *df* value will always be the number of possible outcomes measured minus one.

### Calculating the test-statistic

The value of the test-statistic is



where

= Pearson's cumulative test statistic, which asymptotically approaches a [distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution).

= an observed frequency;

= an expected (theoretical) frequency, asserted by the null hypothesis;

= the number of cells in the table.

The chi-squared statistic can then be used to calculate a [p-value](https://en.wikipedia.org/wiki/P-value) by [comparing the value of the statistic](https://en.wikipedia.org/wiki/Chi-squared_distribution#Table_of_.CF.872_value_vs_p-value) to a [chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution).

**Table of *χ*2 value vs p-value**

The [p-value](https://en.wikipedia.org/wiki/P-value) is the probability of observing a test statistic *at least* as extreme in a chi-squared distribution. Accordingly, since the distribution (called a CDF value…for cumulative distribution function) for the appropriate degrees of freedom *(df)* gives the probability of having obtained a value *less extreme* than this point, subtracting the CDF value from 1 gives the p-value. *Yesterday I gave you all the CDF value for 95%...which corresponds to a p-value of 5%!!!*

The table below gives a number of p-values matching to *χ*2 for the first 10 degrees of freedom.

A low p-value indicates greater [statistical significance](https://en.wikipedia.org/wiki/Statistical_significance), i.e. greater confidence that the observed deviation from the null hypothesis is significant.

*A p-value of 0.05 is often used as a bright-line cutoff between significant and not-significant results, because 95% of the time (CDF value) you will find values less extreme in an experiment and the null hypothesis can be accepted.*

|  |  |
| --- | --- |
| **Degrees of freedom (df)** | ***χ*2 value**[**[17]**](https://en.wikipedia.org/wiki/Chi-squared_distribution#cite_note-17) |
| 1 | 0.004 | 0.02 | 0.06 | 0.15 | 0.46 | 1.07 | 1.64 | 2.71 | 3.84 | 6.64 | 10.83 |
| 2 | 0.10 | 0.21 | 0.45 | 0.71 | 1.39 | 2.41 | 3.22 | 4.60 | 5.99 | 9.21 | 13.82 |
| 3 | 0.35 | 0.58 | 1.01 | 1.42 | 2.37 | 3.66 | 4.64 | 6.25 | 7.82 | 11.34 | 16.27 |
| 4 | 0.71 | 1.06 | 1.65 | 2.20 | 3.36 | 4.88 | 5.99 | 7.78 | 9.49 | 13.28 | 18.47 |
| 5 | 1.14 | 1.61 | 2.34 | 3.00 | 4.35 | 6.06 | 7.29 | 9.24 | 11.07 | 15.09 | 20.52 |
| 6 | 1.63 | 2.20 | 3.07 | 3.83 | 5.35 | 7.23 | 8.56 | 10.64 | 12.59 | 16.81 | 22.46 |
| 7 | 2.17 | 2.83 | 3.82 | 4.67 | 6.35 | 8.38 | 9.80 | 12.02 | 14.07 | 18.48 | 24.32 |
| 8 | 2.73 | 3.49 | 4.59 | 5.53 | 7.34 | 9.52 | 11.03 | 13.36 | 15.51 | 20.09 | 26.12 |
| 9 | 3.32 | 4.17 | 5.38 | 6.39 | 8.34 | 10.66 | 12.24 | 14.68 | 16.92 | 21.67 | 27.88 |
| 10 | 3.94 | 4.87 | 6.18 | 7.27 | 9.34 | 11.78 | 13.44 | 15.99 | 18.31 | 23.21 | 29.59 |
| **P value (Probability)** | 0.95 | 0.90 | 0.80 | 0.70 | 0.50 | 0.30 | 0.20 | 0.10 | 0.05 | 0.01 | 0.001 |

[Chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution), showing *X*2 on the x-axis and P-value on the y-axis.



### Goodness of fit

For example, to test the hypothesis that a random sample of 100 people has been drawn from a population in which men and women are equal in frequency, the observed number of men and women would be compared to the theoretical frequencies of 50 men and 50 women. If there were 44 men in the sample and 56 women, then



If the null hypothesis is true (i.e., men and women are chosen with equal probability), the test statistic will be drawn from a chi-squared distribution with one [degree of freedom](https://en.wikipedia.org/wiki/Degrees_of_freedom_%28statistics%29). If the male frequency is known, then the female frequency is determined.

Consultation of the [chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution) for 1 degree of freedom shows that the [probability](https://en.wikipedia.org/wiki/Probability) of observing this difference (or a more extreme difference than this) if men and women are equally numerous in the population is approximately 0.23 (in between p-values of .20 and .30). This probability is higher than conventional criteria for [statistical significance](https://en.wikipedia.org/wiki/Statistical_significance) (higher than the bright line cutoff for a p-value at 0.05), so normally we would not reject the null hypothesis that the number of men in the population is the same as the number of women (i.e., we would consider our sample within the range of what we'd expect for a 50/50 male/female ratio.)

For our problem 1a, our expected frequency was too low at 95% and the table does not apply.

The approximation to the chi-squared distribution breaks down if expected frequencies are too low. It will normally be acceptable so long as no more than 20% of the events have expected frequencies below 5. Where there is only 1 degree of freedom, the approximation is not reliable if expected frequencies are below 10 (ours was at 5% with 95% error expected??).

In this case, a better approximation can be obtained by reducing the absolute value of each difference between observed and expected frequencies by 0.5 before squaring; this is called [Yates's correction for continuity](https://en.wikipedia.org/wiki/Yates%27s_correction_for_continuity).