The Chi-Square Test

# Introduction

The chi-square test is a statistical test that can be used to determine whether observed frequencies are significantly different from expected frequencies. For example, after we calculated expected frequencies for different allozymes in the [HARDY-WEINBERG](http://www.tiem.utk.edu/~gross/bioed/bealsmodules/hardy-weinberg.html) module we would use a chi-square test to compare the observed and expected frequencies and determine whether there is a statistically significant difference between the two. As in other statistical tests, we begin by stating a null hypothesis (H0: there is no significant difference between observed and expected frequencies) and an alternative hypothesis (H1: there is a significant difference). Based on the outcome of the chi-square test we will either *reject* or *fail to reject* the null hypothesis.

# Importance

Chi-square tests enable us to compare observed and expected frequencies objectively, since it is not always possible to tell just by looking at them whether they are "different enough" to be considered statistically significant. Statistical significance in this case implies that the differences are not due to chance alone, but instead may be indicative of other processes at work.

# Questions

How is the chi-square test used to compare samples or populations? What does a comparison of observed and expected frequencies tell us about these samples?

# Variables

|  |  |
| --- | --- |
| χ2 | the chi-square test statistic |
| o | observed count or frequency |
| e | expected count or frequency |
| n | total number of observations |
| RT | row total |
| CT | column total |

# Methods

Shaklee *et al*. (1993) collected data to study genetic variation within a species of fish called the barramundi perch (*Lates calcarifer*). Many fish species are composed of breeding groups called stocks, which are populations that are genetically distinct from one another. One of the goals of Shaklee *et al*.'s study was to identify individual stocks of the barramundi perch on the basis of significant genetic differentiation. Of the 25 collections examined, those that were not significantly genetically distinct from one another were considered to be from the same stock; collections that were genetically distinct were considered to be from different stocks. Understanding species subdivision into stocks has important implications for conservation and fisheries management, since maintaining the genetic diversity of the species as a whole will require conservation of the different stocks.

We'll use some of their data here to illustrate the application of a simple chi-square test. Below are data showing allele frequencies at seven loci for eight collections of perch from different parts of the Australian coast (table adapted from Shaklee *et al*. 1993; all errors due to rounding are mine).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Locus & allele | # 1 | # 2 | # 14 | # 15 | # 18 | # 21 | # 22 | # 25 |
| *EST-2\** |   |   |   |   |   |   |   |   |
| *\*100+* | 249 | 78 | 97 | 115 | 101 | 242 | 128 | 116 |
| *\*98* | 26 | 4 | 0 | 1 | 2 | 0 | 2 | 30 |
| *\*95* | 126 | 41 | 60 | 60 | 52 | 226 | 125 | 70 |
| *ESTD\** |   |   |   |   |   |   |   |   |
| *\*100+* | 390 | 120 | 155 | 176 | 171 | 465 | 335 | 210 |
| *\*114* | 15 | 4 | 0 | 0 | 0 | 9 | 2 | 6 |
| *mIDHP\** |   |   |   |   |   |   |   |   |
| *\*100* | 387 | 123 | 152 | 167 | 152 | 474 | 333 | 216 |
| *\*78* | 0 | 0 | 5 | 10 | 4 | 1 | 0 | 0 |
| *sIDHP\** |   |   |   |   |   |   |   |   |
| *\*100* | 354 | 113 | 111 | 137 | 143 | 432 | 310 | 177 |
| *\*121+* | 37 | 7 | 44 | 33 | 27 | 39 | 18 | 28 |
| *\*83* | 9 | 3 | 0 | 0 | 0 | 1 | 1 | 3 |
| *LDH-C\** |   |   |   |   |   |   |   |   |
| *\*100* | 373 | 115 | 156 | 175 | 154 | 400 | 245 | 208 |
| *\*90+* | 29 | 9 | 1 | 1 | 1 | 75 | 25 | 5 |
| *PGDH\** |   |   |   |   |   |   |   |   |
| *\*100* | 382 | 122 | 130 | 145 | 153 | 378 | 240 | 199 |
| *\*88+* | 5 | 2 | 21 | 18 | 16 | 95 | 89 | 3 |
| *PROT\** |   |   |   |   |   |   |   |   |
| *\*100+* | 399 | 120 | 149 | 168 | 147 | 453 | 326 | 207 |
| *\*97* | 8 | 4 | 8 | 9 | 9 | 22 | 5 | 9 |

We can use the chi-square test to compare collections # 1 and # 25 at the *EST-2\**locus. The expected values are the allele frequencies we would expect if there were no difference between the two collections at this locus. We can calculate the expected allele frequencies using the row and column totals from a table of the observed frequencies for these two collections.

For the first cell (collection #1, allele*\*100+*) we begin by calculating the probability of an observation being in the first row, regardless of column. To do this, take the row total (365) and divide it by *n* (617) (note that *n* changes depending on which locus and which pair of populations is being compared). Based on these two collections, the probability of a barramundi perch having the *\*100+* allele at the *EST-2\** locus is 0.5916 (365/617). Next, we calculate the probability of an observation being in the first column, regardless of row, by taking the column total (401) and dividing it by *n* (617). The probability of an observation coming from collection #1 as opposed to collection #25 is 0.6499 (401/617).

We have now determined the probability of a perch having a given allele at this locus, and the probability of being in a given collection. But what is the probability that an individual observation will have the *\*100+* allele at the *EST-2\** locus *and* be from collection #1? The probability of two outcomes occurring together is called the joint probability and is calculated by multiplying the two separate probabilities: 0.5916 x 0.6499 = 0.3845. It follows that in a sample of 617 fish we would expect 617 x 0.3845 = 237 individuals to be from collection #1 and have the *\*100+* allele, and we have now calculated our expected value for the first cell in the table. This calculation can be simplified with the following formula:

e = (RT/n)(CT/n)\*n

Verify that the other expected frequencies have been calculated correctly.

|  |  |  |
| --- | --- | --- |
| **Observed frequencies** |  | **Expected frequencies** |
| **allele** | **# 1** | **# 25** | RT |  | **allele** | **# 1** | **# 25** | RT |
| *\*100+* | 249 | 116 | 365 |  | *\*100+* | 237 | 128 | 365 |
| *\*98* | 26 | 30 | 56 |  | *\*98* | 36 | 20 | 56 |
| *\*95* | 126 | 70 | 196 |  | *\*95* | 127 | 69 | 196 |
| CT | 401 | 216 | n=617 |  | CT | 401 | 216 | n=617 |

Note also that the row and column totals remain the same. Now we can use the chi-square test to compare the observed and expected frequencies. The chi-square test statistic is calculated with the following formula:

|  |  |
| --- | --- |
| $$χ^{2}=\sum\_{}^{}\frac{\left(o-e\right)^{2}}{e}$$ | LaTeX Code: \[ \chi^2 = \sum{\frac{(o - e)^2}{e}} \] |

For each cell, the expected frequency is subtracted from the observed frequency, the difference is squared, and the total is divided by the expected frequency. The values are then summed across all cells. This sum is the chi-square test statistic. For the example here,

|  |
| --- |
| $$χ^{2}=\frac{\left(249-237\right)^{2}}{237}+\frac{\left(26-36\right)^{2}}{36}+\frac{\left(126-127\right)^{2}}{127}+\frac{\left(116-128\right)^{2}}{128}+\frac{\left(30-20\right)^{2}}{20}+\frac{\left(70-69\right)^{2}}{69}=0.608+2.778+0.008+1.125+5.000+0.014=9.533$$ |
| LaTeX Code: \[ \frac{(249-237)^2}{237} + \frac{(26-36)^2}{36} + \frac{(126-127)^2}{127} + \frac{(116-128)^2}{128} + \frac{(30-20)^2}{20} + \frac{(70-69)^2}{69} = 0.608 + 2.778 + 0.008 + 1.125 + 5.000 + 0.014 = 9.533 \] |

# Interpretation

The critical value for the chi-square in this case (χ20.05,2) is 5.991; if the calculated chi-square value is equal to or greater than this critical value, we can conclude that the probability of the null hypothesis being correct is 0.05 or less -- a very small probability indeed! Our calculated value of 9.533 is greater than the critical value of 5.991. We therefore *reject* the null hypothesis and conclude that there is a significant difference between the observed and expected frequencies of alleles at the *EST-2\** locus for these two collections of barramundi perch. (Critical values for the chi-square are determined from a statistical table based on the significance level at which the test is being performed [0.05 in our case] and a number called *degrees of freedom* [2 in this example], but the details are beyond the scope of this module).

# Conclusion

Our rejection of the null hypothesis allows us to conclude that the two collections of barramundi perch compared here are genetically distinct at the *EST-2\** locus. In other words, the frequencies of the three alleles at this locus are significantly different between the two populations. Using somewhat more complicated applications of the chi-square test, the authors concluded that the 25 collections they analyzed came from seven genetically distinct stocks, or populations, from adjacent stretches of the northeastern Australian coast. One of the goals of conservation and/or management is the preservation of genetic diversity within a species. Management decisions based on the assumption that a species' genetic variation is distributed across populations could have disastrous consequences for the future of the species if the populations are indeed genetically distinct. Techniques for identifying amounts and patterns of genetic variation within a species are critical tools for biologists.

# Additional Questions

1) Are the allele frequencies at the other six loci also significantly different between collections #1 and #25? (\*\*For loci with two alleles instead of three, the critical value of the chi-square is 3.841, but otherwise the procedure is the same).

2) Use the chi-square test to compare allele frequencies for collections #14 and #15. Can you determine whether or not these two collections are from the same stock?

# Source

Rohlf, F. J. and R. R. Sokal. 1995. *Biometry, 3rd ed*. W. H. Freeman and Company, New York, NY.

Rohlf, F. J. and R. R. Sokal. 1995. *Statistical Tables, 3rd ed*. W. H. Freeman and Company, New York, NY.

Shaklee, J. B., J. Salini, and R. N. Garrett. 1993. Electrophoretic characterization of multiple genetic stocks of barramundi perch in Queensland, Australia. *Transactions of the American Fisheries Society 122*:685-701.

# About this Resource

This material was originally distributed as part of “Alternative Routes to Quantitative Literacy for the Life Sciences[[1]](#endnote-1)” - National Science Foundation Award DUE-9752339 to the University of Tennessee, Knoxville for August 1, 1998 - July 31, 2000. Principal Investigator: Louis J. Gross. Co-Principal Investigators: Beth C. Mullin and Susan E. Riechert.

This material is now being revised as part of the “Resources for Improving Quantitative Skills in Community College Biology[[2]](#endnote-2)” project. As part of that project is also aligned with the OpenStax Biology Textbook[[3]](#endnote-3).

It is published using the QUBES Open Education Resources publishing platform[[4]](#endnote-4).

1. http://www.tiem.utk.edu/~gross/bioed/ [↑](#endnote-ref-1)
2. https://qubeshub.org/community/groups/quantbioatcc/ [↑](#endnote-ref-2)
3. https://openstax.org/details/books/biology-2e [↑](#endnote-ref-3)
4. https://qubeshub.org/qubesresources/publications/1050/ [↑](#endnote-ref-4)