Chi-square (and contingency tables) are used to analyze two categorical data – counts or percents (relative frequencies).

Example: A 2-way table, with categorical variables (counts)

|  |
| --- |
| Years of School Completed by age, 1995 (Census Bureau)(Thousands of persons) |
| **Education** | **Age 25-34** | **Age 35-54** | **Age 55+** | **total** |
| **did not complete high school** |  5,325  |  9,152  |  16,035  |  30,512  |
| **completed high school** |  14,061  |  24,070  |  18,320  |  56,451  |
| **college 1-3 yrs** |  11,659  |  19,926  |  9,662  |  41,247  |
| **college 4+ yrs** |  10,342  |  19,878  |  8,005  |  38,225  |
| **total** |  41,387  |  73,026  |  52,022  |  166,435  |

Let’s take another example:

Fish in different habitats (sand, gravel, silt).

My scientific hypothesis is that I expect fish to prefer certain habitats. If fish were expected to be found in proportion to the habitat (regardless of their “preference”, how many would we expect?

What is my predictor variable?

response variable?

H0? The 2 variables are independent (habitat, #fish)

Ha:

|  |  |  |
| --- | --- | --- |
|   | # fish observed | # fish expected |
| sand (50%) | 8 |   |
| gravel (30%) | 18 |   |
| silt (20%) | 4 |   |

chi-square = Sum (observed count – expected count)2 / expected count



Intuitively, if the observed frequencies exactly matched the expected frequencies, what would this be? 0 – what does this say to you about what you want in order to reject H0? (is larger better? Or worse?)

Whether the chi-square statistic is significant varies with the Degrees of Freedom:

For a two way table, with 2 rows, 2 columns –

For our example, as soon as 1 cell value is specified, the others are fixed, so dfs = 1.

Dfs = (# rows – 1 ) \* (# columns – 1) -

Here – we have 3 rows, 2 columns (3-1) \* (2-1) = 2\*1 = 2

 THE NUMBER OF COUNTS DOES NOT MATTER, AS IT DOES WITH ANOVA OR REGRESSION…

WHY? Numbers have been transformed into percentages….

AGAIN, We use an approximation – the chi-square distribution



<http://sites.stat.psu.edu/~mga/401/tables/Chi-square-table.pdf>

(related to the approximation for binomial distributions)

Critical chi-square for p<.05, Dfs = 1, 5.991

Example in book: We want to know if protecting fragile species will have an effect on declining populations.

What do we do? Collect data on rare plant species, noting if protected or not….

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | Invasives Present? | Population declining? | Legally Protected? | Light level |
| Aristolochia | No | No | No | 2 |
| Hydrastis | No | Yes | No | 0 |
| Liatris | Yes | Yes | No | 4 |
| … | … | … | … | … |
| The 73rd species… |  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Population Status | Not protected | Protected | Row Total |
| Declining | 18 | 8 | 26 |
| Stable/Increasing | 15 | 32 | 47 |
|   | 33 | 40 | 73 |

What is the predictor variable?

What is the response variable?

What is the null hypothesis?

How do we calculate Expected Values for EACH cell (under H0)?

= (row probability \* column probability) / sample size…. Let’s do it!

.

.

.

My chi-square is 9.42

Degrees of freedom = 1

– we have 3 rows, 2 columns (3-1) \* (2-1) = 2\*1 = 2

Critical chi-square for p = .05 is 3.841…indeed for p=.005 it is 7.879 –

so we reject the null, and claim that protecting species does indeed increase populations (p<.005).

How do we run this in JMP?