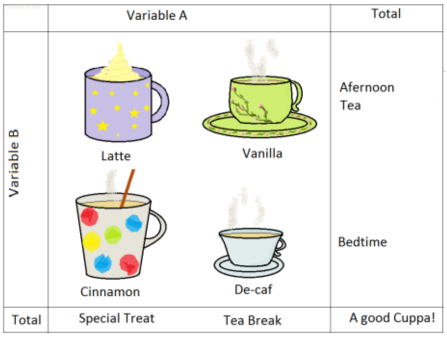
Analysis of frequencies – part I



The Chai-Squared Test

http://cartoonkirsty.blogspot.se/2013/01/chi-squared-test-statistics-cartoon.html

Lecture 10
Biological Statistics III
Ayco Tack

Outline

- *Analysis of frequencies*
  - ❖ *Goodness-of-fit to expected frequencies*
- *Contingency tables*
  - ❖ *Two-way contingency tables*
- *Analysis with* glm
  - ❖ *Generalized linear models with binomial response*

Analysis of frequencies

- Frequency data:
  - Originates from:
    - classification of observational units based on one or more qualitative variables
    - classification based on counts
    - Sometimes also from classification based on intervals of a quantitative variable (e.g. a histogram).
  - Counts of **independent** units in each classification group

- What do we do with frequency data?
  - Test for goodness of fit: compare the observed frequencies with some *a priori* expected frequencies
  - Test for covariation (contingency) between two or more qualitative variables
    - (For quantitative variables you would use a correlation test)
  - Test if a qualitative variable *y* depends on one or more qualitative x-variables
    - (For quantitative response variable you would use an ANOVA)
  - Test if a qualitative variable *y* depends on one or more qualitative and/or quantitative *x*-variables (generalized linear model), glm in R
    - (For a quantitative response variable you would use lm in R)

Goodness-of-fit

Scenario:

- We have n observations falling into a classes with observed frequency n_i and expected frequency f_i in class i (In statistics, frequency refers to the number of items occurring in a given category; www.dictionary.com)
- We want to compare the deviation between observed and expected frequencies in some natural way
- There are two common measures: the $\boldsymbol{X^2}$**-statistic** and the $\boldsymbol{G^2}$**-statistic**

$$X^2 = \sum_{i=1}^{a} \frac{(n_i - f_i)^2}{f_i}$$

$$G^2 = 2\sum_{i=1}^{a} \mathrm{n_i}\, ln\left(\frac{n_i}{f_i}\right)$$

When the expected frequencies are large, both these statistics are approximately $\boldsymbol{X^2}$-distributed with $df = a - 1$ given the null hypothesis H_0 that each observation has the probability

$$\hat{\pi} = \frac{f_i}{n}$$

of falling in class i, where n is the total number of occurrences. If we have estimated parameters from data to get the expected frequencies, we must deduct one *df* for each parameter. The G^2-statistic is also called the **log-likelihood ratio statistic** or the **deviance**

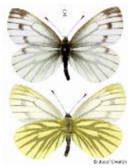
Goodness-of-fit test: *Pieris napi* sex ratio

Data: In a sample of 102 individuals there were 37 females and 65 males.

Aim: We want to test this observed distribution for goodness of fit to the expected distribution corresponding to an even sex ratio.

- Observed: $n_F = 37, n_M = 65$ (total $n = 102$)
- Expected: $f_F = 51, f_M = 51$ (total $n = 102$)
- Number of classes: $a = 2$

Sex	Female	Male	Total
Observed	37	65	102
Expected	51	51	



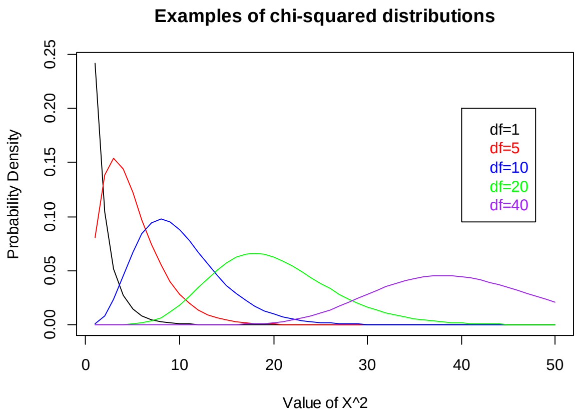
$$X^2 = \frac{(37-51)^2}{51} + \frac{(65-51)^2}{51} = 7.686$$

$$G^2 = 2\left[37log\left(\frac{37}{51}\right) + 65log\left(\frac{65}{51}\right)\right] = 7.786$$

With df $= 1$ for each test. Using the $\boldsymbol{X^2}$-distribution, we find $p = 0.0056$ and $p = 0.0053$. We **reject the null hypothesis** that there is an even sex ratio.

- Note that both statistics give very similar values; it does not really matter which one uses
- Some recommend that expected frequencies should be greater than five for these tests, otherwise one needs to pool classes

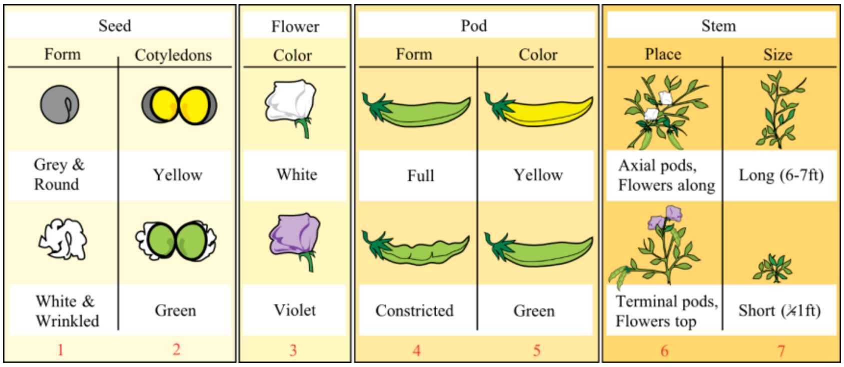
$\boldsymbol{X^2}$-distribution



- The mean of the distribution is equal to the number of degrees of freedom
- The variance equals two times the number of degrees of freedom
- When the degrees of freedom increase, the chi-square curve starts to approach a normal distribution

Mendel’s seven pea characters

Mendel studied the inheritance of seven traits in F2-crosses of pure-bred strains of *Pisum sativum*



For each trait, the alternative phenotypes are determined by the genotype at a single two-allele locus with dominance for one of the alleles

Mendel’s data on smooth and wrinkled peas

One of Mendel’s seven pea traits: smooth *versus* wrinkled seeds

- He first produced pure lines of each type (homozygotes)
- The F1 hybrids were all smooth (heterozygotes, smooth dominant)
- Out of a total of 7324 F2-hybrids, there were 5474 smooth and 1850 wrinkled
- From **Mendelian inheritance** and dominance, we expect a 3:1 ratio of smooth to wrinkled

- Observed: $n_s = 5474, n_W = 1850$ (total $n = 7324$)
- Expected: $f_S = 5493, f_W = 1831$ (total $n = 7324$)
- Number of classes: $a = 2$

$$X^2 = \frac{(5474 - 5493)^2}{5493} + \frac{(1850 - 1831)^2}{1850} = 0.263$$

With df $= 1$ and $p \; = \; 0.61$. We **accept the null hypothesis** of a 3:1 ratio

- The fit of observed to expected is very good. Is it too good?

Two-way contingency tables

Contingency: *the degree of association between theoretical and observed common frequencies of two graded or classified variables. It is measured by the chi-square test (thefreedictionary.com)*

We have two qualitative variables: the first has R values and the second has C values, giving a total of $R\ x\ C$ classes. We can present the frequencies in an $R\ x\ C$ table (having R rows and C columns).

Example of a 2 x 2 table:

- Out of 111 mice, 57 were injected with bacteria plus an antiserum and 54 were injected with only bacteria, and for each mouse it was noted whether it survived the injection

Observed table:

	Dead	*Alive*	**Total**
Bacterium + antiserum	13	44	57
Bacterium only	25	29	54
Total	38	73	111

Expected table:

	Dead	*Alive*	**Total**
Bacterium + antiserum	19.514	37.486	57
Bacterium only	18.486	35.514	54
Total	38	73	111

The expected table is computed from the marginal totals

Analysis of the antiserum $2\ x\ 2$ table

Calculation: We can analyse these observed and expected frequencies using our goodness-of-fit statistics

- Number of classes: $a = 4$
- Note that for the degrees of freedom, we must make an adjustment for the fact that we used the observed data to compute expected frequencies.

$$X^2 = \frac{(13 - 19.514)^2}{19.514} + \frac{(44 - 37.486)^2}{37.486} + \frac{(25 - 18.486)^2}{18.486} + \frac{(29 - 35.514)^2}{35.514} = 6.797$$

$$G^2 = 2\left[13log\left(\frac{13}{19.514}\right) + 44log\left(\frac{44}{37.486}\right) + 25log\left(\frac{25}{18.486}\right) + 29log\left(\frac{29}{35.514}\right)\right] = 6.879$$

Because we used observed data (the marginal totals) to calculate the expected values, we have $df\ = a - 1 - 2\ =\ 1$ for each test. Using the chi-square distribution, the *p*-values are $p\ =\ 0.0091$ and $p\ =\ 0.0087$ for the two tests. We **reject the null hypothesis** that the antiserum has no effect. There is an effect of antiserum on survival.

- Again it does not really matter which of the two statistics we use
- In general, for an $R\ x\ C$ table, the number of degrees of freedom of the chi-square test statistic is $(R - 1)(C - 1)$

Fisher’s exact test for a $2\ x\ 2$ table

Assumptions:

- Exact testing of contingency tables is based on the idea of enumerating all tables that have the same marginal totals as the observed table
- The p-value of the test is the proportion of these tables that are equally or more “extreme” compared with the observed table
- For a $2\ x\ 2$ table, Fisher developed a method of doing this
- For a $R\ x\ C$ table, one can use the multinomial probability of the table as a measure of how extreme it is
- It sometimes takes too long (even for a computer) to enumerate all tables, in which case simulation can be a good alternative

Example:

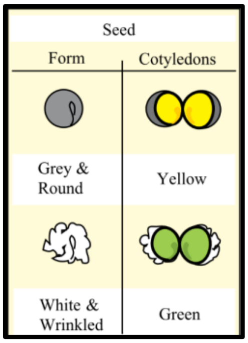
For the antiserum example, Fisher’s exact test gives $p\ =\ 0.0102$

Notes:

- Ideally, Fisher’s exact test should be applied to situations where all alternative hypotheses correspond to the same marginal totals.
- The antiserum example is not of this type, but Fisher’s exact test is often used in such situations anyway

Mendel’s second law

Mendel also published data on F2 phenotypes from a dihybrid cross (pure-bred lines differ in two traits)



Question: The purpose of the dihybrid cross is to test if the two traits are inherited independently (*Mendel’s second law*)

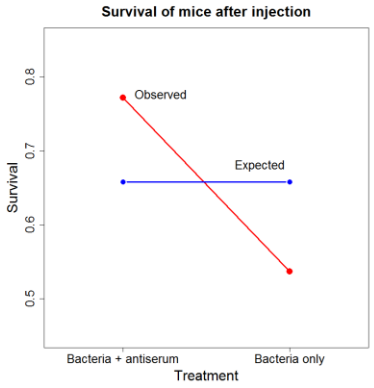
Data: A total of 556 F2 progeny:

	Round	*Wrinkled*	**Total**
Yellow	315	101	416
Green	108	32	140
Total	423	133	556

Null hypothesis: The null hypothesis is that the yellow/green proportion should be the same for both round and wrinkled seeds (or, to word it differently, that the round/wrinkled proportion would be the same for both yellow and green seeds)

- Fisher’s exact test gives p = 0.82
- The X^2-test gives the same p-value
- We have no reason to reject that the traits are inherited independently. Mendel’s second law worked very well.

Visualization of the antiserum example



We can think of the situation as an experiment to determine if antiserum treatment (*x*) has an effect on survival (*y*)

- *We can use this perspective to study model fitting with* glm

Analysis of the data with glm

Revisiting the antiserum data:

- We can view the data as consisting of two groups defined by the treatment (*x*), with 57 mice in the antiserum treatment and 54 mice in the bacteria only treatment. In each of these groups, the number of surviving mice (*y*) is binomially distributed.
- The null hypothesis is that the binomial proportions are the same in both groups
- With glm one tests this hypothesis against the alternative that binomial proportions differ between groups
- The test statistic used is the improvement in fit from a model with the same binomial proportions in each group to a model with binomial proportions equal to the observed proportions. As a measure of fit, the log-likelihood X^2-statistic (the deviance) is used.
- The deviance for the two fits (observed data and expected values) are:

$$G^2 = 2\left[13log\left(\frac{13}{19.514}\right) + 44log\left(\frac{44}{37.486}\right) + 25log\left(\frac{25}{18.486}\right) + 29log\left(\frac{29}{35.514}\right)\right] = 6.879$$

$$G^2 = 2\left[13log\left(\frac{13}{13}\right) + 44log\left(\frac{44}{44}\right) + 25log\left(\frac{25}{25}\right) + 29log\left(\frac{29}{29}\right)\right] = 0$$

The value of the test statistic for glm is the difference between 6.878 and 0, i.e. 6.878. With $df\ =\ 1$, we get $p\ =\ 0.0087$, assuming that the statistic follows a X^2 distribution. We **reject the null hypothesis** that the antiserum has no effect. There is an effect of antiserum on survival.

How to code the binomial relationship, and does it matter if you code it as binomial or binary?

https://aosmith.rbind.io/2019/10/04/expanding-binomial-to-binary/

Related reading and information

- **Quinn & Keough:** Sections 14.1 - 14.2
- **Crawley:** Sections 15.1 - 15.4, 17, 17.1 and 17.2

HAPPY HOLIDAYS!