

## Chi-square Tests

1

## Types of Chi-square Tests

- Tests of goodness of fit
  - e.g., does the frequency of education follow a normal distribution
- Tests of independence
  - e.g., is there a relationship between treatment and outcome
- Tests of homogeneity
  - e.g., is the relationship between treatment and outcome the same across gender

2

## Type of frequencies

- Observed frequencies
  - Frequencies of each combinations of data values in a sample
  - Frequencies tabulated and presented in a contingency table
- Expected frequencies
  - Frequencies that we would expect for each combination of data values in a sample
    - Calculated by multiplying the two marginals and dividing by the total

3

## Chi-Square Distribution

- Distribution of the sum of the differences between (observed and expected frequencies)<sup>2</sup> divided by the expected
- Equivalent to the square of the z-statistic
  - i.e.,  $z^2 = ((y - \mu) / \sigma)^2 \sim \chi^2$  with 1 d.f
- Chi-square statistic  $\chi^2 = \sum \left[ \frac{(O_i - E_i)^2}{E_i} \right]$  for each of i cells
- Reject for high values of chi-square only
- Degrees of freedom determined by (r-1)\*(c-1)

4

## Goodness of Fit Chi-Square Test

- Procedure: compare observed frequencies to the frequencies expected from a distribution
  - Only one sample test
- To some extent, however, all chi-square tests are goodness of fit tests since always testing the fit of the observed frequencies to the expected frequencies
- Once the expected frequencies are known, apply the usual chi-square test
- However, generating the expected frequencies can be challenging
- For the normal, standardize the values
  - After dividing the raw data into intervals, calculate the expected values from the standard normal distribution

5

## Chi-Square Goodness of Fit- contd

- **Basis**

To test the hypothesis  $H_0$  that a set of observations is consistent with a given **probability distribution (p.d.f.)**. For a set of categories, (distribution values), record the observed  $O_j$  and expected  $E_j$  number of observations that occur in each
- Under  $H_0$ ,
 
$$\text{Test Statistic} = \sum_{\text{all cells } j} \frac{(O_j - E_j)^2}{E_j} \sim \chi_{n-1}^2$$

distribution, where  $n$  is the number of categories.
- *E.g. A test of expected segregation ratio is a test of this kind. So, for Backcross mating, expected counts for the 2 genotypic classes in progeny can be calculated using  $0.5n$ , ( $B(n, 0.5)$ ). For F2 mating, expected counts two homozygous classes, one heterozygous class are  $0.25n, 0.25n, 0.5n$  respectively. For F2 with segregants for dominant gene, dominant/recessive exp. counts =  $0.75n$  and  $0.25n$  respectively.*

6

### Example.

**Example.** 40 dishes are counted to determine No. organisms as follows.  
Aim to test at the 0.05 level of significance if the results are consistent with hypothesis that outcomes across cultures randomly distributed.

No. organisms	1-25	26 - 50	51 - 75	76 - 100	Total
Observed No. dishes	6	12	14	8	40
Expected No. dishes	10	10	10	10	40

Test statistic =  $(6-10)^2/10 + (12-10)^2/10 + (14-10)^2/10 + (8-10)^2/10 = 4$ .

The 0.05 critical value of  $\chi^2_3 = 7.81$ , so the test is inconclusive.

**Note:** In general the chi square tests tend to be very conservative vis-a-vis other tests of hypothesis, (i.e. tend to give inconclusive results).

7

### Chi-Square Contingency Test

To test two random variables are **statistically independent**

**Under  $H_0$ ,** Expected number of observations for cell in row  $i$  and column  $j$  is the appropriate row total  $\times$  the column total divided by the grand total. The test statistic for table  $n$  rows,  $m$  columns

$$\sum_{\text{all cells } ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(n-1)(m-1)}$$

**D.o.f.**

Simply; - the chi-square distribution is the sum of  $k$  squares of independent random variables, i.e. defined in a  $k$ -dimensional space.

Constraints, e.g. forcing sum of observed and expected observations in a row or column to be equal, or e.g. estimating a parameter of the parent distribution from sample values, reduce dimensionality of the space by 1 **each time**, e.g. contingency table, with  $m$  rows,  $n$  columns has  $E_m, E_n$  predetermined, so d.o.f. of the test statistic is  $(m-1)(n-1)$ .

8

### Example

- In the following table, the figures in brackets are expected values.

Results	Method 1	Method 2	Method 3	Totals
High	100 (50)	70 (67)	30 (83)	200
Medium	130 (225)	320 (300)	450 (375)	900
Low	70 (25)	10 (33)	20 (42)	100
Totals	300	400	500	1200

- T.S. =  $(100-50)^2/50 + (70-67)^2/67 + (30-83)^2/83 + (130-225)^2/225 + (320-300)^2/300 + (450-375)^2/375 + (70-25)^2/25 + (10-33)^2/33 + (20-42)^2/42 = 248.976$
- The 0.05 critical value for  $\chi^2_{2 \times 2}$  is 9.49 so  $H_0$  **rejected** at the 0.05 level of significance.

9

### $\chi^2$ - Extensions

- Example:** Recall Mendel's data. The situation is one of **multiple populations**, i.e. round and wrinkled. Then

$$\chi^2_{Total} = \sum_{i=1}^m \sum_{j=1}^n \left[ \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

where subscript  $i$  indicates population,  $m$  is the total number of populations and  $n$  = No. plants, so calculate  $\chi^2$  for each cross and sum.

- Pooled  $\chi^2$**  estimated using marginal frequencies under assumption same S.R. all 10 plants

$$\chi^2_{Pooled} = \sum_{j=1}^n \left( \frac{\sum_{i=1}^m (O_{ij} - E_{ij})^2}{\sum_{i=1}^m E_{ij}} \right)$$

10

### $\chi^2$ -Extensions - contd.

So, a typical " $\chi^2$ -Table" for a single-locus segregation analysis, for  $n$  = No. genotypic classes and  $m$  = No. populations.

Source	dof	Chi-square
Total	$nm-1$	$\chi^2_{Total}$
Pooled	$n-1$	$\chi^2_{Pooled}$
Heterogeneity	$n(m-1)$	$\chi^2_{Total} - \chi^2_{Pooled}$

Thus for the Mendel experiment, testing separate null hypotheses:

- A single gene controls the seed character
- The F1 seed is round and heterozygous ( $Aa$ )
- Seeds with genotype  $aa$  are wrinkled
- The  $A$  allele (normal) is dominant to  $a$  allele (wrinkled)

11

### Fisher's Exact Test

- Used when there are small sample sizes in at least one cell
- Test for independence in a 2x2 table (extended to  $r \times c$  tables)
- Gives the exact p-value for the result (or more extreme) where the chi-square test is an approximation
- Today, can be used in virtually any situation, not just for small sample sizes
- Limitations on the chi-square test: not good when  $n < 20$  or when  $20 \leq n \leq 40$  and one cell size  $\leq 5$

12

### Fisher's Exact Test

- Computationally, Fisher's Exact Test is:

Status	Factor	No Factor	Total
Alive	a	b	a+b
Dead	c	d	c+d
Total	a+c	b+d	n

$$\frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!}$$

13

### Fisher's Exact Test

- Gives us the probability for only the observed table.
  - We need the probability of that table and all tables more extreme to be consistent with the approach to hypothesis testing
  - Use the hypergeometric distribution to test this

14