Keele University

Statistics & Experimental Design with R

Barbara Kitchenham

Keele University

Keele University

Proportions and Chi-squared

Keele University

Comparing *Independent* Probabilities

- Address questions such as
  - Is the failure rate of one set of projects greater than failure rate of another?
- General situation
  - We have one set of N_1 objects of which X have a characteristic
  - Another independent set of objects N_2 of which Y have the characteristic
  - Is $p_1=X/N_1$ significantly greater than $p_2=Y/N^2$
- There is an exact test based if X or Y are small based on the hyper geometric distribution
  - R function fisher.test

Large Sample Approximation Chi-Squared test of Homogeneity

	Success	Failures	Totals
Sample 1	O_{11}	O_{12}	$n_1.$
Sample 2	O_{21}	O_{22}	$n_2.$
Totals	$n._1$	$n._2$	$n..$

$$\hat{p} = \frac{n_{.1}}{n_{..}}$$

$$E_{11} = n_{1.}\hat{p} = \frac{n_{.1} \times n_{1.}}{n_{..}} \qquad E_{12} = n_{1.}(1-\hat{p}) = \frac{n_{.2} \times n_{1.}}{n_{..}}$$

$$E_{21} = n_{2.}\hat{p} = \frac{n_{.1} \times n_{2.}}{n_{..}} \qquad E_{22} = n_{2.}(1-\hat{p}) = \frac{n_{.2} \times n_{2.}}{n_{..}}$$

$$\chi^2 = \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \frac{(O_{21}-E_{21})^2}{E_{21}} + \frac{(O_{22}-E_{22})^2}{E_{22}}$$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{n_{..}(O_{11}O_{22} - O_{21}O_{12})^2}{n_{.1} \times n_{.2} \times n_{1.} \times n_{2.}}$$

Example

	Success	Failures	Totals
Sample 1	4	8	12
Sample 2	1	20	21
Totals	5	33	33

- R has the prop.test which accepts the data directly or via a matrix of the same format
  - prop.test(x=c(4,1),n=c(12,21) correct=F)
    - Chi-squared=4.849, df=1. p-value=0.0448
  - prop.test(x=c(4,1),n=c(12,21) correct=T)
    - Chi-squared=2.8812, p-value=0.8962
- Fisher test has p-value=0.0471

Another Classic solution

- Test a statistic of the form
  - (p_1-p_2)/(standard error)
  - Where the standard error is the square root variance of average effect i.e. $p_{ave}=(X+Y)/(N1+N2)$
  - This type of test is called a Wald test
- From Normal approximation
  - $Var(p_{ave})=p_{ave}(1-p_{ave})/(N1+N2)$
- Is there a potential problem?
  - If H0 is true p_1 and p_2 are both estimates of the probability estimated by p_{ave} and Var(p_{ave}) is best estimate of common variance
  - If H0 is false, the a "common" variance may be misleading particularly if
    - N1 and N2 are very different

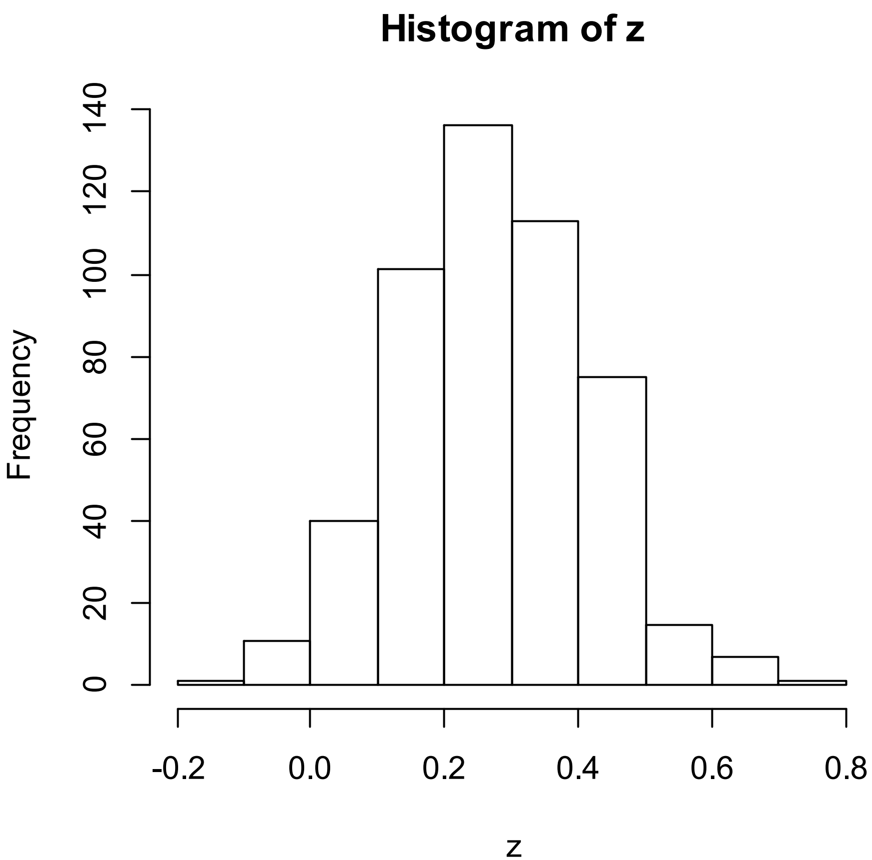
Alternative approach

- MonteCarlo simulation
  - Simulate two independent normal variables x and y
    - $M_i = p_i$ and $Var = p_i(1-p_i)/N_i$ (i=1, 2)
    - 500 of each
  - Calculate z=x-y to assess the distribution of the difference between the two parameters
  - Calculate the variance of z
  - Test statistic = $(p_1-p_2)/sqrt(var(z))$

Example

- Is p1=4/12 different to P2=1/21?
- Using classic approach
  - p1=0.333, p2=0.0476, p_{ave}=05/33=0.1515
  - Var(p_{ave})=0.1515*(1-0.1515)/33=. 2.0704
  - T=(0.3333-0.0476/sqrt(0.016796)=2.20
    - Critical-level one-sided =1.65 (based on standard normal distribution
- Using a simulation approach (based on a sample of 500 for x and y)
  - Var(diff)=0.01988, sd=.1410
  - T=0.2854/0.1410=2.02

Simulation results



Contingency Tables

- Items in a population are cross-classified in two dimensions
  - Are the characteristics independent?
- Confusion Matrix example
  - Is a predictor algorithm better at identifying faulty modules than chance?
    - Each module is classified according to its true status (faulty, Not faulty)
    - Also classified by predictor as faulty or not faulty
    - Are the correct classifications better than chance?
  - Also used for predicting failing projects

Confusion Matrix

Module predictions	Module Status		
	Faulty	Not Faulty	Totals
Faulty	O_{11}	O_{12}	$n_1.$
Not Faulty	O_{21}	O_{22}	$n_2.$
Totals	$n._1$	$n._2$	N=n..

- Let p_{ij}=probability of falling into cell i,j
- p_{11}=Prob(Prediction Faulty)×Prob(Module is faulty)
- H0: $p_{ij}=p_{i}. \times p._{j}$ for i,j
- Chi-squared approach is exactly the same

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{n_{..}(O_{11}O_{22} - O_{21}O_{12})^2}{n_{.1} \times n_{.2} \times n_{1.} \times n_{2.}}$$

SE Issues

- Being better than chance at predicting is a pretty weak criterion
  - Would like to assess the strength of the prediction model
  - Cramer Coefficient of Association (Mathews)
    - C=sqrt(chi-squared/N)
    - Exactly the same as Pearson correlation between all the individual pairs of 0 and 1
  - Would like to assess whether one model is better than another
    - Can compare the C values
      - Using correlation equality test

Keele University

Hypothetical Example

Estimated	Actual		Totals
	Failed	Succeeded	
Failed	26	15	41
Succeeded	7	37	44
Totals	33	52	85

- Chi-squared=20.166
- df=1
- p=7.099e-06
- Correlation=0.487

Keele University

Conclusions

- Handling proportions is relatively straightforward
- Chi-squared test works
  - For independent proportions
  - Contingency tables
- Contingency tables
  - Used frequently in SE to evaluated procedures for identifying failing projects/components
  - Chi-squared test identifies whether predictions better than chance
  - Correlation indicates strength of association