Stats fest 2007

Frequency analysis

murray.logan@sci.monash.edu.au

MONASH University

Analyzing frequencies

- **Not all dependent variables are normally distributed**
- **Frequencies (percentages) do not follow normal distributions**

Goodness of fit test

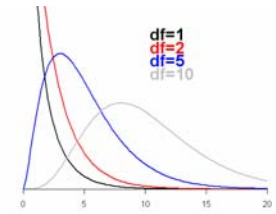
- **Aims**
  - Investigate whether observed ratios follow expected ratios
    - E.g. is the sex ratio likely to differ from 1:1
- **Data**
  - Counts (frequency) of units in each category
- **H_0:**
  - Observed data came from a population which has the specified expected frequencies
    - $O - E = 0$

Goodness of fit test

- **Chi square (χ^2) statistic**

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

  - o = observed frequencies
  - e = expected frequencies
  - *df* = *p* - 1



Goodness of fit test

- **Assumptions**
  - Observations must be classified independently
  - No more than 20% of categories have expected frequencies < 5

```
> chisq.test(c(obs),p=c(exp))$exp
```

    - Where obs are the observed counts and exp are the expected proportions

Goodness of fit test

- **Analysis sequence**
  - Design experiment/survey
  - Collect data
  - Test assumptions
  - Calculate χ^2 and compare to a χ^2 distribution

```
> chisq.test(c(obs),p=c(exp))
```

    - Where obs are the observed counts and exp are the expected proportions
  - Conclusions
    - Reject or not reject H_0

Contingency tables

- **Aims**
  - Cross-classification of two or more variables
  - Investigating the association of two categorical variables
- **Data**
  - Two or more categorical predictor variables
    - If have 3 or more variables – best to use log-linear models (*G* tests)
  - Dependent variable
    - Counts – number of observations

Contingency tables

- **H_0:**
  - The categorical variables are independent of one another
    - Equivalent to no interaction between categorical variables in ANOVA

		Categorical 1		
		A	**B**	**C**
Categorical 2	**1**			
	2			

Contingency tables

- **Assumptions**
  - Observations must be classified independently
  - No more than 20% of categories have expected frequencies < 5
    - Fishers exact test – for 2x2
- **Chi-square (χ^2)**
  - Simple to calculate
  - df = (rows-1)(cols-1)
- **Fishers exact test**
  - useful if have small sample sizes (problem with assumptions)
  - Computationally intense

Contingency tables

- **Analysis sequence**
  - Design experiment/survey
  - Collect data
  - Construct contingency table

```
> *.tab <- xtabs(response~cat1+cat2, dataset)
```

  - Test assumptions

```
> chisq.test(*.tab, correct=F)$exp
```

  - Perform test

```
> *.x2 <- chisq.test(*.tab, correct=F)
```

```
> *.x2 <- fisher.test(*.tab)
```

Generalized linear models

- **Aims**
  - Investigate the effects of one or more factors on a response variable
  - Accommodates a range of distributions
- **Data**
  - Response variable
    - Normally distribution – same as regression/ANOVA
    - **Poisson distribution – log-linear modeling**
    - Binomially distribution – logistic regression
  - Predictor variables
    - Categorical
    - Continuous – logistic regression

Generalized linear models

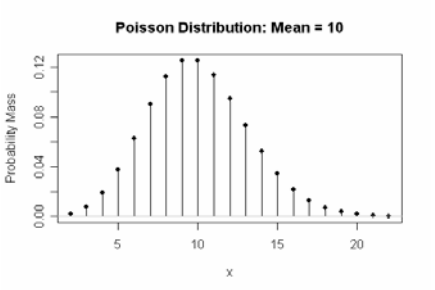
- **Link function**
  - **Poisson distribution – log-linear modeling**
    - **Log(μ)**
  - Binomially distribution – logistic regression
    - **logit**

Log-linear models

- **Aims**
  - Investigating the association of two or more categorical variables
    - Whether there is an interaction between two or more categorical variables
- **Data**
  - Response variable
    - **Poisson distribution – log-linear modelling**
  - Predictor variables
    - Categorical

Log-linear models

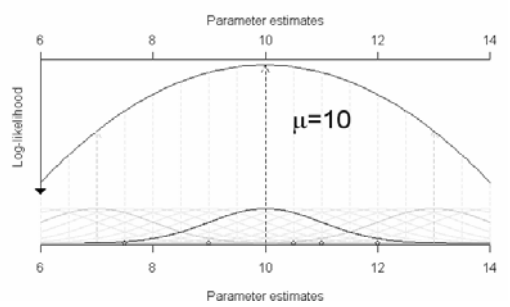
- **Poisson distribution**



Log-linear models

- **H_0:**
  - Predictor variable(s) independent
    - No interactions
- **Log-linear model**
  - Full model $\log f = \text{constant} + \lambda^{A} + \lambda^{B} + \lambda^{AB}$
  - Reduced model $\log f = \text{constant} + \lambda^{A} + \lambda^{B}$
- **Fit of model measured by log-likelihood (LL)**
- **Difference between full and reduced model (G^2) indicates importance of interaction term**

Log-likelihood



Log-linear models

- **Assumptions**
  - Observations must be classified independently
  - Response variable follows a Poisson distribution

Log-linear models

- **Analysis sequence**
  - Design experiment/survey
  - Collect data
  - Fit full generalized linear model

```
> *.glmF <- glm(RESPONSE~CAT1+CAT2+CAT1:CAT2, family=poisson, data)
```

  - Test H_0 (compare reduced model to full model)

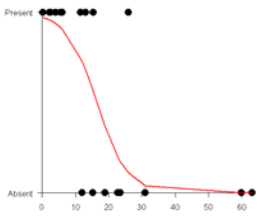
```
> anova(*.glmF, test="Chisq")
```

Generalized linear models

- **Aims**
  - Investigate the effects of one or more factors on a response variable
  - Accommodates a range of distributions
- **Data**
  - Response variable
    - Normally distribution – same as regression/ANOVA
    - Poisson distribution – log-linear modeling
    - **Binomially distribution – logistic regression**
  - Predictor variables
    - Categorical
    - Continuous – logistic regression

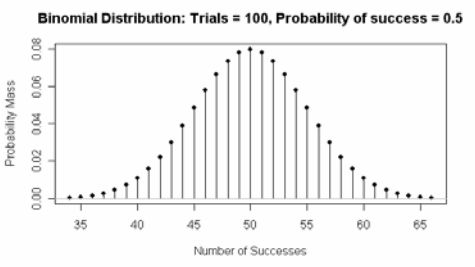
Logistic regression

- **Aims**
  - Investigate the relationship between a continuous predictor variable and a binary response variable
- **Data**
  - Response variable
    - **binary distribution**
    - 0 or 1, dead or alive,
    - yes or no,
    - present or absent
  - Predictor variables
    - Continuous



Logistic regression

- **Binomial distribution**



Logistic regression

- **Log-linear model**
  - Full model $g(x) = \beta_0 + \beta_1 x_1$
  - Reduced model $g(x) = \beta_0$

  Where g(x) is the probability of being either 1 or 0
- **H_0:**
  - Population slope (β_1) equals 0
- **Fit of model measured by log-likelihood (LL)**
- **Difference between full and reduced model (G^2) indicates importance of slope**

Logistic regression

- **Assumptions**
  - Response variable follows a binomial distribution
  - Absence of collinearity
    - Correlation matrix (SPLOM)

Log-linear models

- **Analysis sequence**
  - Design experiment/survey
  - Collect data
  - Fit generalized linear model

```
> *.glm <- glm(RESPONSE~PREDICTOR, family=binomial, data)
```

  - Test H_0

```
> summary(*.glm)
```

  - Calculate % variation explained

```
> 1-(*.glm$dev / *.glm$null)
```

  - Calculate LD50

```
> -*.glm$coef[1] / *.glm$coef[2]
```

Intercept Slope