Discrete data: counts and proportions

Biljana Jonoska Stojkova
Applied Statistics and Data Science Group (ASDa)
Department of Statistics, UBC

February 28, 2018

Resources for statistical assistance

Department of Statistics at UBC:

www.stat.ubc.ca/how-can-you-get-help-your-data

SOS Program - An hour of free consulting to UBC graduate students. Funded by the Provost and VP Research Office.

STAT 551 - Stat grad students taking this course offer free statistical advice. Fall semester every academic year.

Short Term Consulting Service - Advice from Stat grad students. Fee-for-service on small projects (less than 15 hours).

Hourly Projects - ASDa professional staff. Fee-for-service consulting.

Outline

Analysis of Count Data

Binary Data Analysis

Categorical Data Analysis

Generalized Linear Models

Types of Data

Continuous data: any value in a specified range is possible

- Normal distribution: entire real line
- Regression and ANOVA models

Count data: non-negative integer valued

- Poisson distribution
- Negative Binomial distribution

Categorical data: non numeric

- ordinal or nominal (binary is a special case)

Poisson Distribution

Number of events occuring at anytime in a fixed amount time or anywhere in a fixed amount of space (or some other index of size)

The rate at which events occur is constant

Occurrence of one event does not affect the probability that a second event will occur

Examples

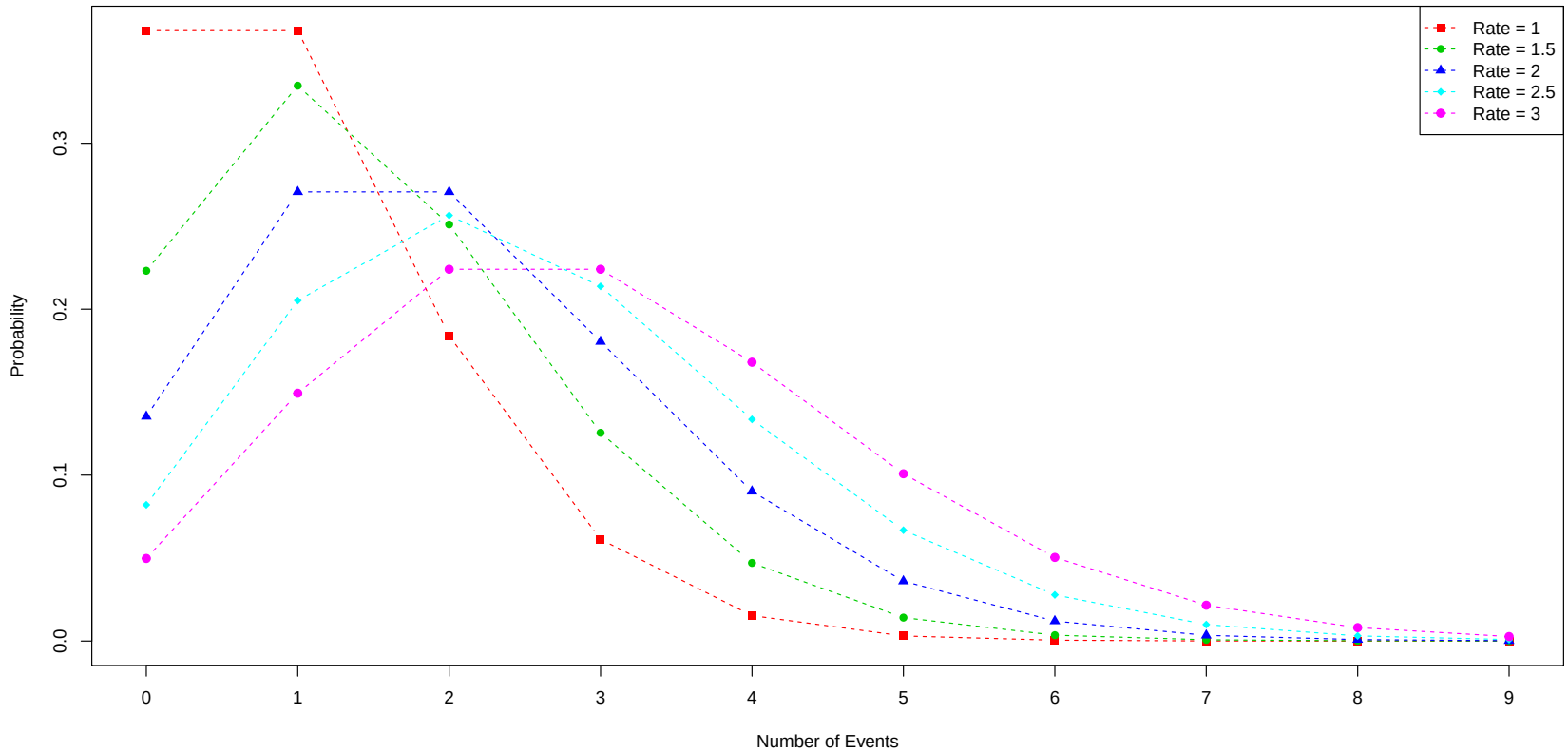
May follow a Poisson Distribution:

- number of phone calls received by a call center per hour
- decay events per second from a radioactive source

May violate the Poisson assumptions:

- number of students who arrive at the student union building per minute, rate not constant (low during class time, high between class time)
- number of magnitude 5 earthquakes per year in California, events not independent (one large earthquake increases the probability of aftershocks)

Poisson Distribution (continued)



Mean and variance both equal the rate (expected number of events)

The larger the rate the larger the spread in the data (a strong assumption)

If the rate is ≥ 20 then Poisson $\sim$ Normal

Number of colon polyps in 12 months

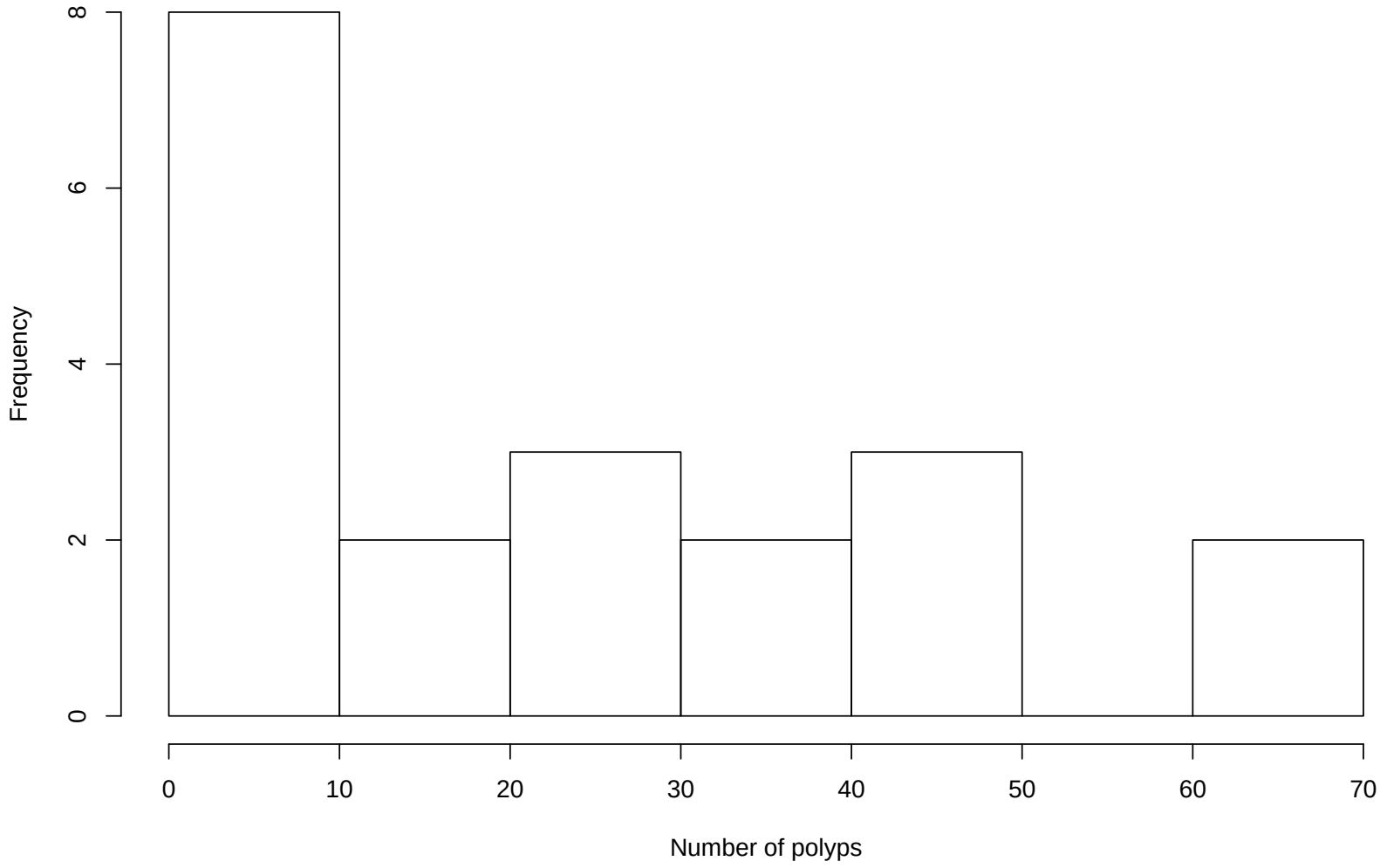
```
## Warning: package 'HSAUR3' was built under R version 3.4.

##  number   treat age
##      63 placebo  20
##       2    drug  16
##      28 placebo  18
##      17    drug  22
##      61 placebo  13
##       1    drug  23
##       7 placebo  34
##      15 placebo  50
##      44 placebo  19
##      25    drug  17
```

Estimated rate (mean number of polyps in 12 months) = 24.05

Can the normal distribution be used to describe this count data?

Number of colon polyps at 12 months



Mean = 24.05, Variance = 434.68 (Std. Dev = 20.85)

95% confidence interval for the estimated rate

If the rate is ≥ 20 then we can use a normal approximation:

Estimate the rate with $\bar{x} = \sum_1^n x_i/n$

Then a 95% CI is given by $\bar{x} \pm 1.96\sqrt{\bar{x}/n}$

If the rate is < 20 then the normal approximation may not be very good and a more exact method should be used

Comparing 2 Poisson rates

If the rate in each group is ≥ 20 then we do the following:

$$\bar{x} = \sum_1^n x_i/n \qquad \bar{y} = \sum_1^m y_j/m \qquad \bar{z} = \frac{\sum_1^n x_i + \sum_1^m y_j}{n+m}$$

Then under the null hypothesis

$$\frac{\bar{x}-\bar{y}}{\sqrt{\bar{z}}} / \sqrt{\frac{1}{n} + \frac{1}{m}} \sim N(0,1)$$

	sample size	total	rate
drug	9	89	9.89
placebo	11	392	35.64
	20	481	24.05

Statistic = -11.68

p-value = 1.57e-31

There are exact methods to compare the rate ratio of 2 groups.

Poisson Regression

Model the log of the rate (mean) as a function of other variables

$$\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...$$

```
##                 Estimate Std. Error z value   Pr(>|z|)
## (Intercept)        2.291     0.1060   21.62 1.225e-103
## treatplacebo       1.282     0.1174   10.92  9.446e-28
```

log(drug rate) $= 2.291$

drug rate $= e^{2.291} = 9.9$

log(placebo rate) $= 2.291 + 1.282$

placebo rate $= e^{2.291+1.282} = 35.6$

P/D relative risk $= 35.6/9.9 = 3.60$

P/D relative risk $= e^{2.291+1.282}/e^{2.291} = e^{1.282}$

log(P/D relative risk) $= 1.282$

Adjust for covariate (age)

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      3.1699    0.16821   18.85 3.22e-79
## treatplacebo     1.3591    0.11764   11.55 7.16e-31
## age             -0.0388    0.00596   -6.52 7.02e-11
```

For a fixed age:

log(P/D relative risk) = 1.359

P/D relative risk = $e^{1.359} = 3.89$

For a fixed treatment:

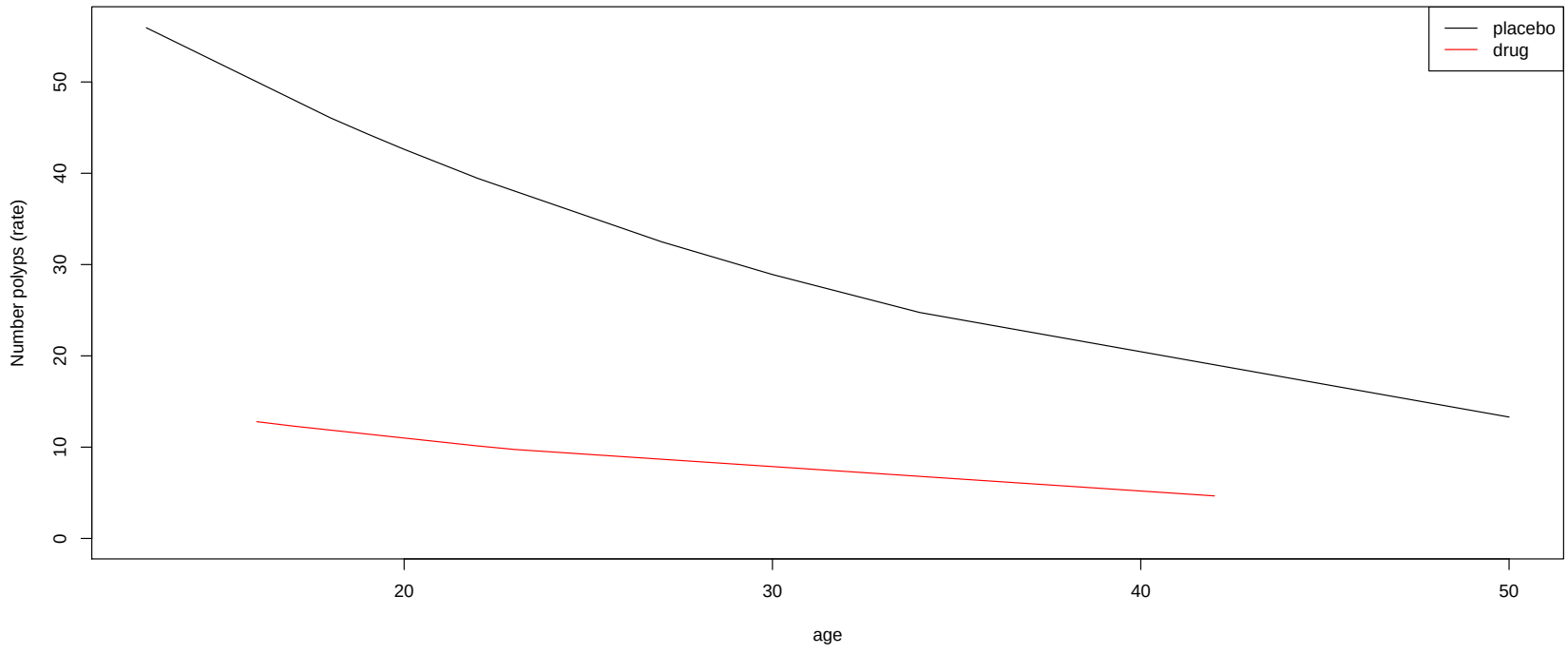
log(relative risk for one year increase in age) = -0.0388

Relative risk for one year increase in age = $e^{-0.0388} = 0.96$

Relative risk for 20 year increase in age = $e^{(-0.0388x20)} = 0.46$

If the Poisson assumption ($\sigma^2 = \mu$) is violated, the model can dramatically overstate the significance of the predictors.

Rate by age



At age=20:

placebo rate $= e^{3.170+1.359-(0.0388x20)} = 42.65$

drug rate $= e^{3.170-(0.0388x20)} = 10.96$

P/D relative risk $= 42.65/10.96 = 3.89$

Rate by age (continued)

At age=40:

placebo rate $= e^{3.170+1.359-(0.0388 \times 40)} = 19.63$

drug rate $= e^{3.170-(0.0388 \times 40)} = 5.04$

P/D relative risk $= 19.63/5.04 = 3.89$

Negative Binomial Distribution (NB)

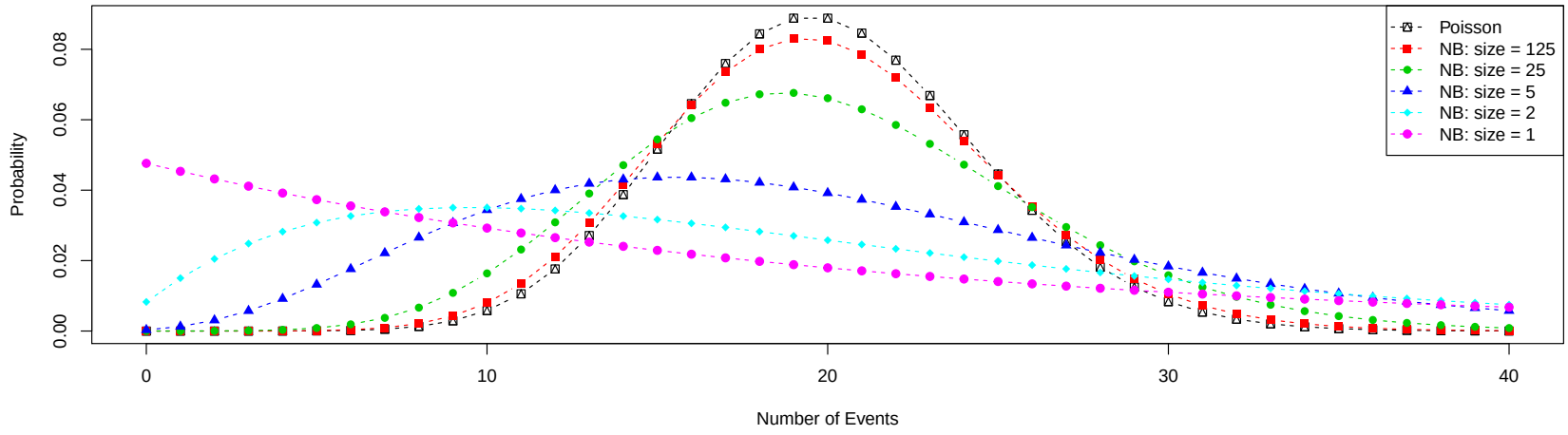
When your data show extra variation that is greater than the mean (overdispersion)

Has one parameter more than the Poisson distribution that adjusts the variance independently from the mean

$\sigma^2 = \mu + \mu^2/r$

Approaches the Poisson for large r, but has larger variance than the Poisson for small r

Negative Binomial (continued)



All of the above have a rate of 20. Note the high variation in the shapes of the curves.

Another parametrization typically used in Negative Binomial Regression:

$\sigma^2 = \tau^2 \mu$, where τ^2 is an overdispersion parameter

If $\tau = 1$, we have a Poisson

Estimating the Rate and 95% CI for NB data

Most count data we encounter in practice has $\sigma^2 > \mu$

Compute the sample mean and variance:

$\bar{x} = \sum_1^n x_i/n$ $\qquad$ $s^2 = \sum_1^n (x_i - \bar{x})^2/(n-1)$

We use the normal approximation for the 95% CI

$\bar{x} \pm 1.96 s/\sqrt{n}$

This approximation may not be very good

$\tau^2 = \sigma^2/\mu$ $\qquad$ is an estimate of the overdispersion

$r = \mu^2/(\sigma^2 - \mu)$ $\qquad$ is an estimate of the size parameter

Negative Binomial Regression

Like Poisson regression, the log of the rate (or mean) is modelled as a function of other variables

Poisson regression can include an overdispersion parameter. This is similar but not identical to negative binomial regression.

Negative Binomial regression will estimate either a single value for r or τ in addition to the rate. This allows σ^2 to be greater than the rate.

Number of colon polyps at 12 months

Overdispersed Poisson model:

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.1699     0.5510    5.75 2.34e-05
## treatplacebo     1.3591     0.3853    3.53 2.59e-03
## age             -0.0388     0.0195   -1.99 6.28e-02
```

```
## RR (P/D)    2.5 %   97.5 %
##    3.893    1.829    8.284
```

```
## Overdispersion (tau) = 10.56
```

Number of colon polyps at 12 months

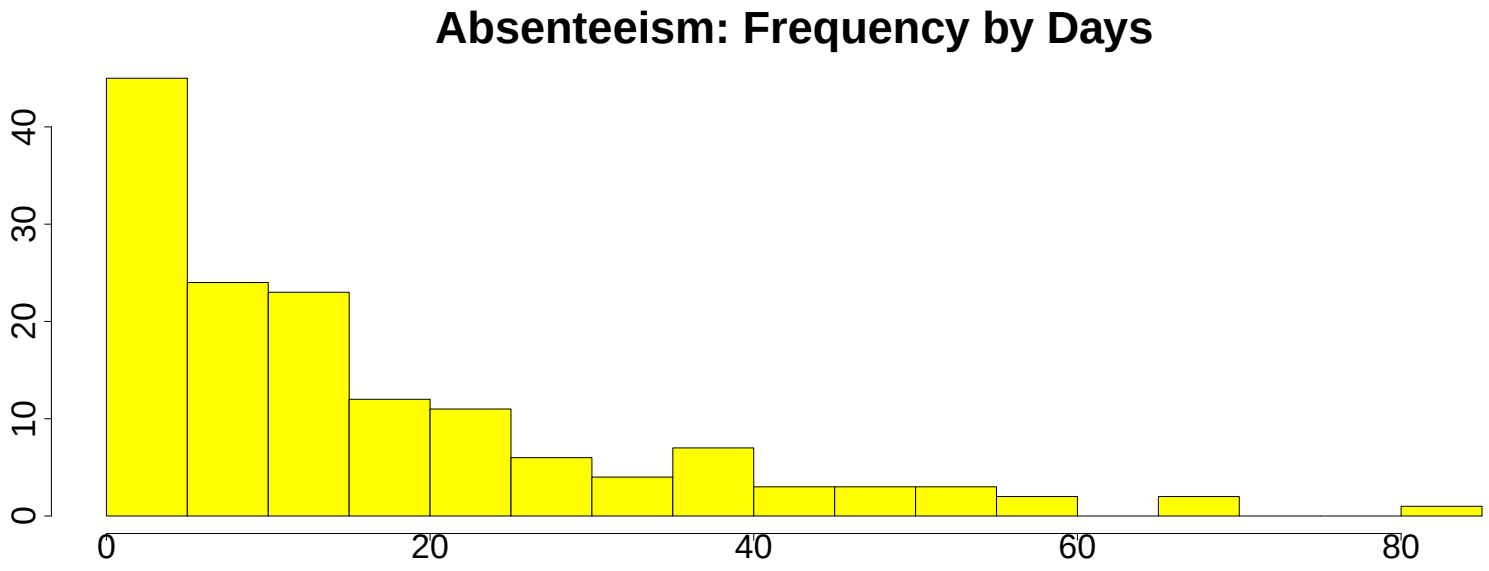
Negative Binomial model:

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      3.1579     0.5580    5.66 1.48e-08
## treatplacebo     1.3681     0.3690    3.71 2.09e-04
## age             -0.0386     0.0210   -1.84 6.58e-02
```

```
## RR (P/D)    2.5 %   97.5 %
##    3.928    1.906    8.096
```

```
## r = 1.719
```

Absenteeism from School in Rural New South Wales




```
##                 Eth  N  Mean    Var   V/M    r
## 1 (N)ot Aboriginal 77 12.18 183.89 15.10 0.86
## 2     (A)boriginal 69 21.23 313.95 14.79 1.54
```

Negative Binomial model

```
##                Estimate Std. Error z value   Pr(>|z|)
## (Intercept)      2.4999     0.1109  22.551 1.309e-112
## EthA             0.5556     0.1597   3.479  5.027e-04

## RR (A/N)     2.5 %   97.5 %
##    1.743     1.275    2.383

## r = 1.157
```

log(A/N relative rate) $= 0.5556$

A/N relative rate $= e^{0.5556} = 1.743$

Overdispersed Poisson model

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      2.4999     0.1262  19.803 6.066e-43
## EthA             0.5556     0.1617   3.436 7.716e-04

## RR (A/N)     2.5 %   97.5 %
##    1.743     1.270    2.393

## Overdispersion (tau) = 13.14
```

Overdispersion

Common in the modeling of counts

Not an issue in ordinary regression because the normal distribution has a separate variance parameter (not a function of the mean) to describe the variability

Does not address inadequacy due to an important term missing in the model

Example: Number of deaths due to vehicle accidents in a week. A Poisson model would assume each person has the same probability of dying. Factors such as amount of time spent driving, whether a person wears a seat belt and geographical location can cause fatality counts to display more variation than predicted by the Poisson model

Overview

We are interested in modelling the rate (count for fixed amount of time or space)

Main distributions are Poisson and Negative Binomial

When making comparisons we usually talk about the relative rate. For adverse events this is usually referred to as the relative risk.

If data are observed over varying time periods then we need to standardize the counts to make them comparable. Any analysis must adjust for these varying times.

Common to account for overdispersion. If it exists and isn't taken into account, this doesn't affect the model estimates but underestimates their standard errors.

Categorical Data

Two main types: Nominal and Ordinal

Nominal data is differentiated by label but otherwise there is no logical order (Gender, Ethnicity, Species)

Ordinal data is differentiated by a label that allows a logical order but the magnitude of the difference cannot be established (Likert Scales)

Binary data can be either ordinal or nominal. With only two possible outcomes, it is very easy to deal with. We can code the two outcomes as 0 or 1 but this is only an indicator that an outcome has occurred not an indication of order or a real number.

Binary Data 1/0 (special case of categorical data)

Binary data need not be coded as 1/0. It can be be coded as any binary indicator such as True/False, Success/Failure, etc.

We are interested with estimating the probability of each outcome. Although knowing one completely defines the other:

$P(X = 1) = p \qquad P(X = 0) = 1 - p$

Another parameter of interest is the odds (S/F) $= p/(1 - p)$

Or log odds (called the logit function):

$\eta = logit(p) = log(\frac{p}{1-p}) = log(p) - log(1 - p)$

PROBABILITY AND ODDS ARE NOT THE SAME

Suppose probability of success $= 0.9$

Odds S/F $= 0.9/0.1 = 9$

Odds F/S $= 0.1/0.9 = 0.11$

Binomial Distribution

Number of events (successes) for a fixed number of binary outcomes, n

A Binary outcome, x_i is 0 or 1 and $X = \sum x$ is the number of times our sample gave us a value of 1

$X \sim Binomial(n, p)$

Expected value of $X = np$

Variance of $X = np(1 - p)$

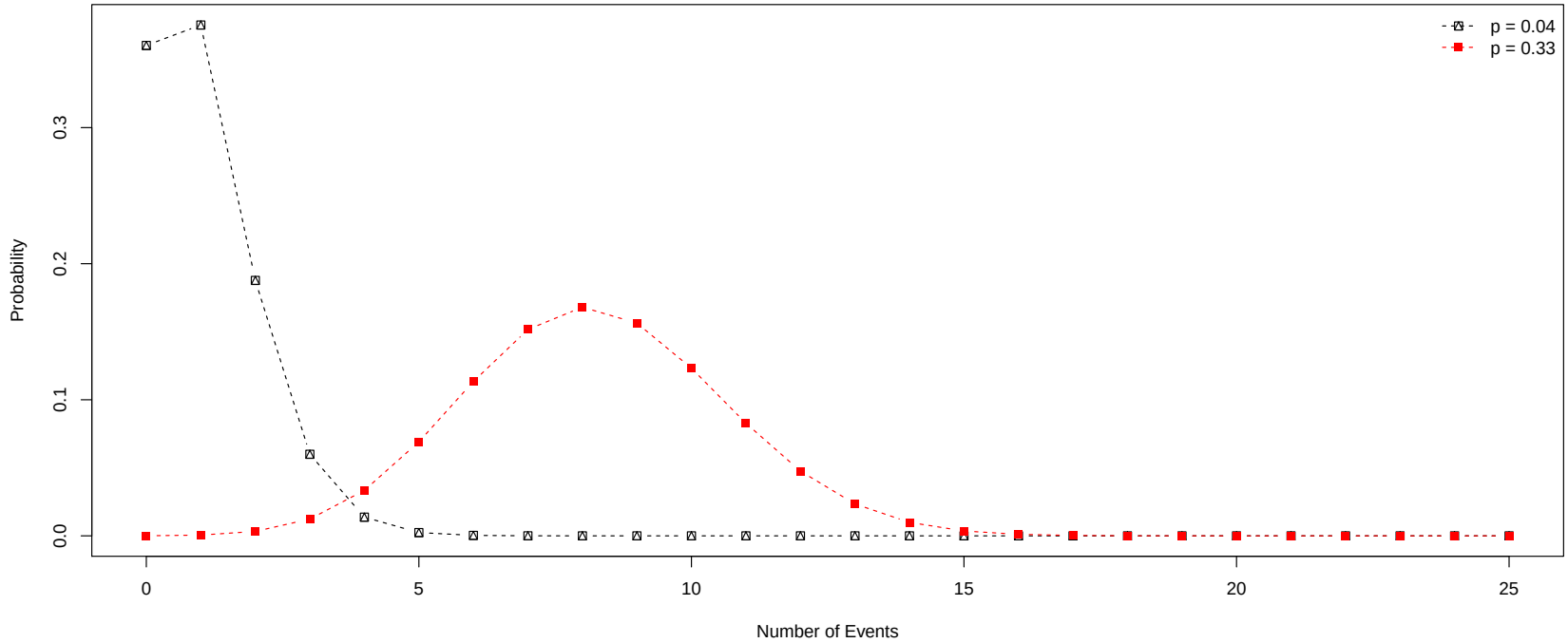
X can take any integer value between 0 and n

Can calculate the probability for any value of X

Example

Approximately 4.0% of Canadian adults were vegetarians as of 2003. A 2015 survey conducted by the Vancouver Humane Society and administered by polling company Environics "shows that 33 percent of Canadians, are either already vegetarian or are eating less meat."

A random sample of $n = 25$ students is selected from the GPS workshops. What is the distribution for the number who are vegetarian (or are eating less meat)?



Assuming p = 0.04:

mean = 25 x 0.04 = 1

sd = sqrt(25 x 0.04 x 0.96) = 0.98

Assuming p = 0.33:

mean = 25 x 0.33 = 8.25

sd = sqrt(25 x 0.33 x 0.67) = 2.35

Estimating p or $logit(p)$

Usually we are not interested in the number of successes but the probability or odds of a success

If we have n observations where n_1 are successes and n_0 are failures then we estimate the probability of a success by $\hat{p}$:

$$\hat{p} = \bar{x} = n_1/n \qquad \text{se}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

If we wish to estimate the log odds then:

$$\hat{\eta} = logit(\hat{p}) = log(n_1/n_0) = log(n_1) - log(n_0)$$

$$\text{se}_{\hat{\eta}} = \sqrt{1/n_1 + 1/n_0}$$

Testing the value of p

To test a specific value of p or $\eta = logit(p)$ we use a Wald test

Estimate the parameter from the data then plug the hypothesized value into the following:

$z_1 = (\hat{p} - p)/se_{\hat{p}}$ $\qquad$ $z_2 = (\hat{\eta} - \eta)/se_{\hat{\eta}}$

If $n_1 \geq 5$ and $n_0 \geq 5$ both z_1 and z_2 are approximately $N(0, 1)$

If the sample size is too small then exact methods based on binomial distributions are needed

Comparing a binary response between 2 groups

Create a 2x2 table of the data

	Group 1	Group 2	Total
False	n_{11}	n_{12}	n_{1+}
True	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n_{++}

The 4 numbers in the table are all that is needed

Example: UC Berkeley admissions by gender

Is there gender bias in admission practices at Berkeley?

	Male	Female	Total
Admitted	1198	557	1755
Rejected	1493	1278	2771
	2691	1835	4526

	Male	Female	Total
Admitted	0.45	0.30	0.39
Rejected	0.55	0.70	0.61
	0.59	0.41	1.00

Expected outcomes if hypothesis is true that there is no gender bias

	Male	Female	Total
Admitted	0.39	0.39	0.39
Rejected	0.61	0.61	0.61
	0.59	0.41	1.00

	Male	Female	Total
Admitted	1043	712	1755
Rejected	1648	1123	2771
	2691	1835	4526

Compare observed to expected

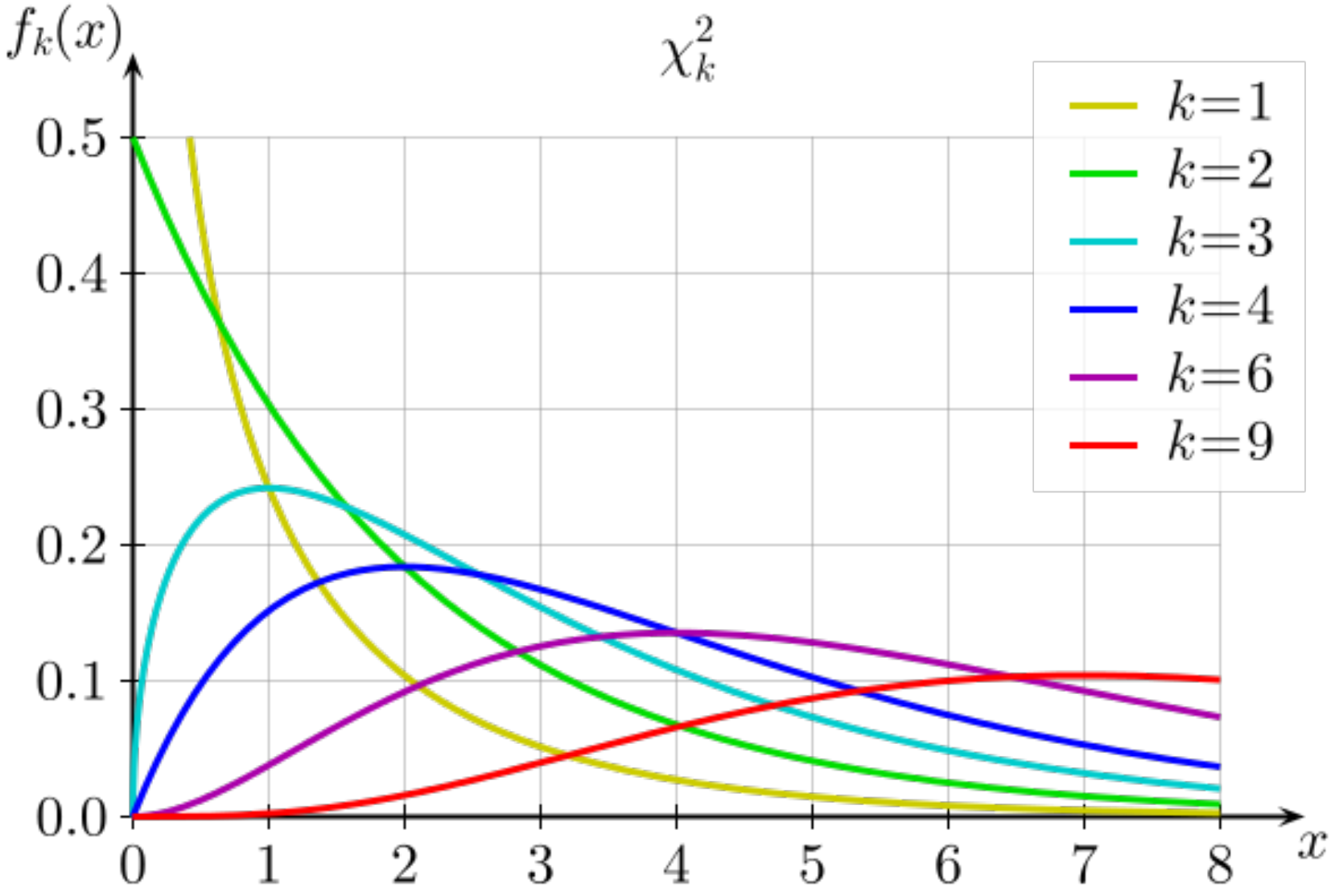
Pearson χ^2 test, uses a statistic that follows a χ^2 distribution approximately:

$(1198 - 1043)^2/1043 + (1493 - 1648)^2/1648 + (557 - 712)^2/712 + (1278 - 1123)^2/1123 = 92.7$

degrees of freedom = (#Columns - 1) x (#Rows - 1) = 1

p-value = 0

χ^2 distribution



Methods for comparing the numbers

Pearson χ^2 test

- should apply a continuity correction
- requires expected counts ≥ 5

Fisher's exact test

- Method is available in most software
- valid no matter what the counts in each cell are

z test for the log odds ratio

- good for estimating the size of the effect

Odds Ratio

```
##             Gender
## Admit       Male Female
##   Admitted 1198    557
##   Rejected 1493   1278
```

Odds of being admitted for Males $= 1198/1493 = 0.8024$

Odds of being admitted for Females $= 557/1278 = 0.4358$

M/F odds ratio (OR) of being admitted $= 0.8024/0.4358 = 1.8412$

The log of the odds ratio is approximately normally distributed:

Log OR $= log(1.8412) = 0.6104$

SE $= \sqrt{1/1198 + 1/1278 + 1/1493 + 1/557} = 0.06389$

p-value $= 1.25e\text{-}21$

Example: UC Berkeley admissions by 6 largest departments

Do admission rates differ by department?

```
##        Admitted Rejected %Admitted
## Dept
## A           601      332         64
## B           370      215         63
## C           322      596         35
## D           269      523         34
## E           147      437         25
## F            46      668          6
```

```
## Chisq = 778.9065   DF = 5    Pval = 0.0000
```

Overall %Admitted = 39

We can calculate individual odds ratios

There are 15 pairwise odds ratios to consider

Logistic Regression

Analyze the simultaneous effects of multiple variables, including mixtures of categorical and continuous variables and interaction terms

Odds of the response taking a particular value is modeled:

$$logit(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...$$

logit(p) = log(p/(1-p)) = log(odds)

The estimated effects in the model are log odds ratio for a unit change in the predictor

Example: Student Admissions at UC Berkeley

Dept	Gender	Admitted	Rejected	%Admitted
A	Male	512	313	62
	Female	89	19	82
B	Male	353	207	63
	Female	17	8	68
C	Male	120	205	37
	Female	202	391	34
D	Male	138	279	33
	Female	131	244	35
E	Male	53	138	28
	Female	94	299	24
F	Male	22	351	6
	Female	24	317	7
Sum	Male	1198	1493	45
	Female	557	1278	30

Analysis by Gender only

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(Admit, Reject)
##         LR Chisq Df Pr(>Chisq)
## Gender    93.45  1     <2e-16

##  OR (M/F)     2.5 %    97.5 %
##      1.84      1.62      2.09
```

There appears to be a gender bias

```
##                Estimate Std. Error
## (Intercept)    -0.8305     0.05077
## GenderM         0.6104     0.06389
```

For male: log(odds) = -0.8305 + 0.6104

odds = $e^{-0.8305+0.6104} = 0.802$

p = odds/(1 + odds) = 0.802/1.802 = 0.445

For female:

log(odds) = -0.8305

odds = $e^{-0.8305} = 0.436$

p = odds/(1 + odds) = 0.436/1.436 = 0.304

M/F odds ratio = 0.802/0.436 = 1.84

M/F odds ratio = $e^{-0.8305+0.6104}/e^{-0.8305} = e^{0.6104}$

log(M/F odds ratio) = 0.6104

Analysis by Gender and Department

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(Admit, Reject)
##         LR Chisq Df Pr(>Chisq)
## Gender      1.5  1      0.216
## Dept      763.4  5     <2e-16

##  OR (M/F)     2.5 %    97.5 %
##      0.90      0.77      1.06
```

There does not appear to be a gender bias

```
##               Estimate Std. Error
## (Intercept)  0.68192    0.09911
## GenderM     -0.09987    0.08085
## DeptB       -0.04340    0.10984
## DeptC       -1.26260    0.10663
## DeptD       -1.29461    0.10582
## DeptE       -1.73931    0.12611
## DeptF       -3.30648    0.16998
```

For male:

odds $= e^{0.68192-0.09987} = 1.790$

p $= 1.790/2.790 = 0.642$

For female:

odds $= e^{0.68192} = 1.978$

p $= 1.978/2.978 = 0.664$

M/F odds ratio $= e^{-0.09987} = 0.90$

Analysis of Gender within Department

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(Admit, Reject)
##              LR Chisq Df Pr(>Chisq)
## Gender           1.5  1    0.21593
## Dept           763.4  5    < 2e-16
## Gender:Dept     20.2  5    0.00114

##         OR (M/F) 2.5 % 97.5 %
## DeptA       0.35  0.21   0.58
## DeptB       0.80  0.34   1.89
## DeptC       1.13  0.85   1.50
## DeptD       0.92  0.69   1.24
## DeptE       1.22  0.83   1.81
## DeptF       0.83  0.46   1.51
```

```
##                 Estimate Std. Error
## (Intercept)       1.5442     0.2527
## GenderM          -1.0521     0.2627
## DeptB            -0.7904     0.4977
## DeptC            -2.2046     0.2672
## DeptD            -2.1662     0.2750
## DeptE            -2.7013     0.2790
## DeptF            -4.1250     0.3297
## GenderM:DeptB     0.8321     0.5104
## GenderM:DeptC     1.1770     0.2996
## GenderM:DeptD     0.9701     0.3026
## GenderM:DeptE     1.2523     0.3303
## GenderM:DeptF     0.8632     0.4027
```

DeptA OR (M/F):

$e^{1.5442-1.0521}/e^{1.5442} = e^{-1.0521} = 0.35$

DeptB OR (M/F):

$e^{1.5442-1.0521-0.7904+0.8321}/e^{1.5442-0.7904} = e^{-1.0521+0.8321} = 0.80$

Categorical variables with more than 2 levels

If our response has more than 2 levels then the models are more complicated

If we have a single categorical predictor we can do Pearson's χ^2 test or Fisher's exact test if some counts in the cross tabulation are small

```
##             Wine Rating
## Temperature  1  2  3  4  5
##        cold  5 16 13  2  0
##        warm  0  6 13 10  7
```

```
## Fisher's Exact test p-value = 7.366514e-05
```

Cumulative Logistic Regression Models

Can handle a response variable with k multiple categories as well as account for the ordering

The model indicates how each predictor variable uniquely affects the odds of being in category 2 or higher compared to category 1; being in category 3 or higher compared to being in category 2 or 1; ... up to being in category k compared to being in categories 1, 2, ..., k-1

Assumes that the predictors have the same effect on different levels of the response variable (proportional odds assumption)

Each comparison has its own intercept, but the same set of regression coefficient estimates

With nominal data no assumptions are made about structure so a more general model is fit

Copenhagen Housing Conditions Survey

Variables are

- Sat - Satisfaction with their present housing circumstances (Low, Medium, High)
- Infl - Perceived influence on the management of the property (Low, Medium, High)
- Type - (Tower, Atrium, Apartment, Terrace)

Satisfaction is the response (ordinal)

Predict Satisfaction by Influence and Type

The data

Infl	Type	Low	Medium	High
Low	Tower	21	21	28
	Apartment	61	23	17
	Atrium	13	9	10
	Terrace	18	6	7
Medium	Tower	34	22	36
	Apartment	43	35	40
	Atrium	8	8	12
	Terrace	15	13	13
High	Tower	10	11	36
	Apartment	26	18	54
	Atrium	6	7	9
	Terrace	7	5	11

Proportional odds logistic regession model

```
## Threshold Parameters for baseline (intercepts):

##  Low|Medium Medium|High
##  -0.3959673   0.6892151

## Shift parameters for predictors

##                    Estimate Std. Error Pr(>|z|)
## InflMedium        0.4901320  0.1655909   0.0031
## InflHigh          1.1934906  0.1865318   0.0000
## TypeApartment    -0.5277651  0.1667091   0.0015
## TypeAtrium       -0.2377054  0.2421692   0.3263
## TypeTerrace      -0.5632012  0.2316138   0.0150
```

Estimated logit equations

- estimated log odds (satisfaction falling into high category versus low and medium) = 0.6892 + 0.4901InflMedium + 1.1935InflHigh- 0.5278TypeApartment - 0.2377TypeAtrium - 0.5632TypeTerrace
- estimated log odds (satisfaction falling into medium or high category versus low) = -0.3960 + 0.4901InflMedium+1.1935InflHigh- 0.5278TypeApartment -0.2377TypeAtrium - 0.5632TypeTerrace

Predicted probabilities for proportional odds model

```
##                      Sat  Low Medium High
## Infl   Type
## Low    Tower            40.2   26.4 33.4
##        Apartment        53.3   23.9 22.8
##        Atrium           46.1   25.6 28.4
##        Terrace          54.2   23.6 22.2
## Medium Tower            29.2   25.8 45.0
##        Apartment        41.1   26.3 32.6
##        Atrium           34.3   26.4 39.3
##        Terrace          42.0   26.2 31.8
## High   Tower            16.9   20.7 62.3
##        Apartment        25.7   24.9 49.4
##        Atrium           20.6   22.8 56.6
##        Terrace          26.4   25.1 48.5
```

Nominal logistic regression model

```
## Estimated parameters


##                                        Est   PVal
## Low|Medium.(Intercept)          -0.3967147 0.0299
## Medium|High.(Intercept)          0.7000645 0.0001
## Low|Medium.InflMedium           -0.5188023 0.0044
## Medium|High.InflMedium          -0.4707628 0.0160
## Low|Medium.InflHigh             -1.0728042 0.0000
## Medium|High.InflHigh            -1.2567069 0.0000
## Low|Medium.TypeApartment         0.5347458 0.0050
## Medium|High.TypeApartment        0.5172442 0.0051
## Low|Medium.TypeAtrium            0.1309072 0.6436
## Medium|High.TypeAtrium           0.3230569 0.2343
## Low|Medium.TypeTerrace           0.5538032 0.0327
## Medium|High.TypeTerrace          0.5696173 0.0305
```

Predicted probabilities for Nominal model

```
##                      Sat  Low Medium High
## Infl   Type
## Low    Tower          40.2   26.6 33.2
##        Apartment      53.4   23.7 22.8
##        Atrium         43.4   30.2 26.4
##        Terrace        53.9   24.1 21.9
## Medium Tower          28.6   27.1 44.3
##        Apartment      40.6   27.2 32.2
##        Atrium         31.3   32.1 36.5
##        Terrace        41.1   27.9 31.0
## High   Tower          18.7   17.7 63.6
##        Apartment      28.2   20.8 51.0
##        Atrium         20.8   23.4 55.8
##        Terrace        28.6   21.7 49.7
```

Generalized Linear Models

All the models we have fit are generalized linear models

There is the link function which converts the mean into a linear function of the model parameters

$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ...$

Dist	Link	Variance	Dist	Link	Variance
Normal	μ	σ^2	Gamma	$1/\mu$	μ^2
Poisson	$\log(\mu)$	μ	NB	$\log(\mu)$	$\mu + \mu^2/r$
Binomial	$\text{logit}(\mu)$	$\mu(1-\mu)$	NB	$\log(\mu)$	$\tau^2\mu$

Resources for statistical assistance

Department of Statistics at UBC:

www.stat.ubc.ca/how-can-you-get-help-your-data

SOS Program - An hour of free consulting to UBC graduate students. Funded by the Provost and VP Research Office.

STAT 551 - Stat grad students taking this course offer free statistical advice. Fall semester every academic year.

Short Term Consulting Service - Advice from Stat grad students. Fee-for-service on small projects (less than 15 hours).

Hourly Projects - ASDa professional staff. Fee-for-service consulting.