Further Statistical Methods

Contingency Tables, Hilary Term, 2016

Robin Evans

(based on slides by Marco Scutari)

evans@stats.ox.ac.uk
Department of Statistics
University of Oxford

May 19, 2016

Course Information

Lectures

- Week 3: Friday 11am
- Week 4: Thursday 12pm, Friday 11am
- Week 5: Thursday 12pm

Practical Week 6 (unassessed)

Reference Books (further references in the next slide)

AA Agresti A (2013). Categorical Data Analysis. Wiley, 3rd edition.

DE Edwards D (2000). Introduction to Graphical Modelling. Springer, 2nd edition.

SF Fienberg SE (2007). The Analysis of Cross-Classified Categorical Data. Springer, 2nd edition.

JP Pearl J (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.

Other Useful Books on Contingency Tables

- Agresti A (2010). Analysis of Ordinal Categorical Data. Wiley, 2nd edition.
- Bishop YMM, Fienberg SE, Holland PW (2007). Discrete Multivariate Analysis: Theory and Practice. Springer.
- Koller D, Friedman N (2009). Probabilistic Graphical Models. MIT ress.
- Lauritzen S (1996). Graphical Models. Oxford University Press.
- Pesarin F, Salmaso L (2010). Permutation Tests for Complex Data: Theory, Applications and Software. Wiley.
- Whittaker J (1990). Graphical Models in Applied Multivariate Statistics. Wiley.

Overview

1. Models and Probability Distributions

   [AA 2, 3 & 8; DE 2]

2. Hypothesis Testing

   [AA; DE 5; SF 3.8]

3. Graphical Models

   [JP 3; DE 2]

Models and Probability Distributions

Example: Twin Study

Drton and Richardson (2008)* use data from a study of 597 pairs of identical twins.
For each pair we have (A_1, A_2, D_1, D_2):

- A_i indicates alcoholism in twin i ($1 =$ present, $0 =$ absent);
- D_i indicates depression in twin i.

A_1	A_2	$D_1 = 0$, $D_2 = 0$	$D_1 = 0$, $D_2 = 1$	$D_1 = 1$, $D_2 = 0$	$D_1 = 1$, $D_2 = 1$
0	0	288	80	92	51
	1	15	9	7	10
1	0	8	4	8	9
	1	3	2	4	7

* J. Roy. Statist. Soc. Ser. B, 70(2), pp. 287–309.

Example: General Social Survey

In the same paper data from the US General Social Survey are analysed. The data are answers to the following seven questions, asked at various times from 1975–1994.

- Can people be trusted?
- Are people helpful?
- How much confidence do you have in organised religion?
- How much confidence do you have in congress?
- How much confidence do you have in business?
- Are you a member of a union?
- Are you a member of a church?

Each question has multiple levels. We might be interested in modelling how these variables are related, and how that relationship changes over time and in different portions of the population.

Sufficient Statistics

For any collection of i.i.d. data $Y_1, \ldots, Y_n$, the unordered collection of data is sufficient. (So if I permute the data it won't change my inference.)
If $Y_i \in \mathcal{J} = \{1, \ldots, J\}$, a finite collection of values, then we can summarise the data by the counts or frequencies

$$n_j \equiv \sum_{i=1}^{n} \mathbb{1}_{\{Y_i = j\}}.$$

That is n_j is how many times we observed j.
Letting $\pi_j = P(Y_i = j)$ the log-likelihood is always

$$l(\boldsymbol{\pi}; \boldsymbol{n}) = \sum_{j \in \mathcal{J}} n_j \log \pi_j,$$

This means that the sufficient statistics are always at most J-dimensional, no matter how large n and how complicated the model.

What is a Contingency Table?

A contingency table represents the frequencies n_{ijk} of two or more discrete variables (X, Y, Z, ...), where each element of a finite set correponds to a different combination (configuration) of their values.

The corresponding probability table is denoted by $\pi_{ijk} = P(X = i, Y = j, Z = k)$.

Note that this is just a special case of the one dimensional table, with extra structure added.

Each of those discrete variables can be:

1. a categorical random variable, defined on an unordered set of values (i.e. the `level()`s of the `factor`);
2. an ordinal random variable, defined on an ordered set of values (e.g. small/large; $0 - 10, 11 - 20, > 20$).

The main difference is that in the latter case the CDF is defined, as is the concept of trend.

Notation for Cells and Totals

Standard notation is:

- We observe discrete i.i.d. random variables (X, Y, Z) over n individuals. We denote by n_{ijk} the number of observations with $X = i$, $Y = j$ and $Z = k$; each entry in the table is called a cell.
- Margins of the table are denoted (for e.g.)

$$n_{i++} = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{ijk}, \qquad n_{i+k} = \sum_{j=1}^{J} n_{ijk}, \qquad n_{+jk} = \sum_{i=1}^{I} n_{ijk}.$$

- n (or n_{+++}) is the sample size, i.e. the overall total of the table; in two-dimensional tables it is also denoted as n_{++}.

The notation for the probabilities follows the same scheme (e.g. π_{i++} is the probability associated with n_{i++}).

A Three-Dimensional Contingency Table: Lizards

This small data set is from [SF] and is also used extensively in [DE].

Species	Perch Diameter	Perch Height	
		> 4.75ft	$\leqslant 4.75$ft
Sagrei	$\leqslant 4$in	32	86
	> 4in	11	35
Distichus	$\leqslant 4$in	61	73
	> 4in	41	70

For a sample of 409 lizards, the following variables were recorded:

- the species, which can be either "Sagrei" or "Distichus";
- the diameter of the branch they were perched on, discretised in two categories `narrow` ($\leqslant 4$in) and `wide` (> 4in);
- the height of that same branch, discretised in two categories `high` (> 4.75ft) and `low` ($\leqslant 4.75$ft).

A Three-Dimensional Contingency Table

```
> lizards <- read.table("lizards.txt", header = TRUE)
> head(lizards, 3)

  Species Diameter Height
1  Sagrei   narrow    low
2  Sagrei   narrow    low
3  Sagrei   narrow    low
```

```
> table(lizards)

, , Height = high

           Diameter
Species     narrow wide
  Distichus     73   70
  Sagrei        86   35

, , Height = low

           Diameter
Species     narrow wide
  Distichus     61   41
  Sagrei        32   11
```

`margin.table()`: Totals and Marginals

We can compute marginals with `margin.table()`, which has a `margin` argument to specify which dimensions of the table to retain. For a single variable, it produces n_{i++}, n_{+j+} and n_{++k}.

```
> margin.table(table(lizards), margin = 1)

Species
Distichus    Sagrei
      245       164
```

For two variables, it produces n_{ij+}, n_{i+k} and n_{+jk}.

```
> margin.table(table(lizards), margin = 2:3)

        Height
Diameter high low
  narrow  159  93
  wide    105  52
```

Combining `margin.table()` with subsetting we can produce all sub-tables and marginals.

plyr: Expanding a Contingency Table

It is sometimes convenient to expand a contingency table into a data frame with one row for each observation; many functions in R can handle the latter but not the former.

```
> library(plyr)
> liz2 <- adply(table(lizards), .margins=1:3)
> head(liz2, 3)

    Species Diameter Height V1
1 Distichus   narrow   high 73
2    Sagrei   narrow   high 86
3 Distichus     wide   high 70

> # if you need separate rows for each observation:
> liz3 <- ddply(liz2, .variables=c(1:3),
+               .fun= function(x) x[rep(1,each=x$V1), 1:3])
> head(liz3, 3)

    Species Diameter Height
1 Distichus   narrow   high
2 Distichus   narrow   high
3 Distichus   narrow   high
```

Probabilistic Assumptions for Contingency Tables

The right distribution for the frequencies in a contingency tables depends on the underlying sampling distribution.

- Multinomial sampling treats counts n_{ijk} as the outcomes of a multinomial with probabilities π_{ijk} that sum up one. n is considered fixed.
- Independent multinomial sampling one or more sets of marginal counts are fixed, and each of the resulting sub-tables (e.g. n_{ijk} for fixed k) has an independent multinomial distribution with probabilities $\pi_{ij|k}$ such that $\sum_{ij} \pi_{ij|k} = 1$. As a side effect, n is also fixed as a result.
- Poisson sampling treats the counts n_{ijk} as independent Poissons with parameters μ_{ijk}, which means that the overall total n is not considered fixed.

The most common assumption is multinomial sampling.

Different Sampling Schemes: Are They Related?

Poisson sampling is simply

$$n_{ijk} \sim \text{Pois}(\mu_{ijk}) \qquad \text{independently for all } i, j, k.$$

From probability theory then we know that

$$n = \sum_{ijk} n_{ijk} \sim \sum_{ijk} \text{Pois}(\mu_{ijk}) = \text{Pois}\left(\sum_{ijk} \mu_{ijk}\right).$$

But (**exercise**) one can show that *conditional on the total* n, the distribution of n_{ijk} is jointly multinomial:

$$\{n_{ijk} \mid n\} \sim \text{Multinom}(n, \{\pi_{ijk}\}) \tag{1}$$

with probabilities $\pi_{ijk} = \mu_{ijk}/\mu$.
Moving to independent multinomial sampling involves conditioning on individual totals:

$$\{n_{ijk} \mid n_k\} \sim \text{Multinom}(n_k, \{\pi_{ij|k}\}) \quad \text{with} \quad \pi_{ij|k} = \pi_{ijk}/\pi_{++k}.$$

Roles of the Variables in a Contingency Table

Modelling a contingency table differs substantially depending on which roles we assign to the variables, which in turn depends on the aim of the analysis.

- If there is one clear variable of interest, we can use regression (such as with a GLM) and use the other variables as covariates. The resulting GLM will then be Binomial or Multinomial depending on how many levels the response has.
- We may also be interested in the joint structure of the variables: are they (conditionally) independent? is there symmetry? A common multivariate approach uses graphical models.
- We may interested in explaining the cell counts as a function of the variables, using (for example) a Poisson GLM.

A Two-Dimensional Contingency Table: Seat-Belts

This example from [AA 3] shows fatality results for children under 18 who were passengers in car accidents in Florida in 2008, according to whether the child was wearing a seat belt.

	Injury Outcome		
Seat Belt Use	Fatal	Nonfatal	Total
No	54	10325	10379
Yes	25	51790	51815
Total	79	62115	62194

The data is observational (it does not arise from a designed experiment) so none of the totals are fixed.

A Two-Dimensional Contingency Table (R Code)

```
> # data in table form.
> belt = matrix(c(54, 25, 10325, 51790), nrow = 2,
+           dimnames = list(Seatbelt = c("No", "Yes"),
+                           Injury = c("Fatal", "Nonfatal")))
> belt = as.table(belt)
> belt

        Injury
Seatbelt Fatal Nonfatal
     No     54    10325
     Yes    25    51790

> # data in data frame form.
> as.data.frame(belt)

  Seatbelt   Injury  Freq
1       No    Fatal    54
2      Yes    Fatal    25
3       No Nonfatal 10325
4      Yes Nonfatal 51790
```

Seatbelts Example

Are seatbelts protective in accidents? If not, we would expect that $\pi_{ij} = \pi_{i+}\pi_{+j}$ suggesting statistical independence between the wearing of seatbelts and fatalities.

	Fatal	Non-fatal
No Seatbelt	54	10325
Seatbelt	25	51790

Treating these counts as Poisson random variables we can fit a saturated model $n_{ij} \sim \text{Pois}(\mu_{ij})$ with

$$\log \mu_{ij} = \log(n\pi_{ij}) = \theta + \alpha_i + \beta_j + \gamma_{ij}$$

with identifiability constraints ($\alpha_1 = \beta_1 = \gamma_{11} = \gamma_{21} = \gamma_{12} = 0$).

Exercise: show that $\pi_{ij} = \pi_{i+}\pi_{+j}$ corresponds to $H_0 : \gamma_{22} = 0$.

Contingency Table as a Poisson GLM

```
> summary(glm(Freq ~ Seatbelt*Injury, data = as.data.frame(belt),
+   family = poisson))

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                 3.9890     0.1361  29.313  < 2e-16 ***
SeatbeltYes                -0.7701     0.2419  -3.184  0.00146 **
InjuryNonfatal              5.2533     0.1364  38.503  < 2e-16 ***
SeatbeltYes:InjuryNonfatal  2.3827     0.2421   9.840  < 2e-16 ***
---
    Null deviance: 1.1524e+05  on 3  degrees of freedom
Residual deviance: 5.6248e-12  on 0  degrees of freedom
AIC: 42.666
```

- We have $e^{\gamma_{22}} = 10.8\ (6.7, 17.4)$; that is, the odds of a non-fatal injury are 10.8 times higher for someone wearing a seatbelt.
- The residual deviance is zero, because the model is saturated.

Contingency Table as a Poisson GLM

```
> summary(glm(Freq ~ Seatbelt + Injury, data = as.data.frame(belt),
+   family = poisson))

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.57897    0.11286   22.85   <2e-16 ***
SeatbeltYes     1.60790    0.01075  149.52   <2e-16 ***
InjuryNonfatal  6.66729    0.11257   59.23   <2e-16 ***
---
    Null deviance: 115243.62  on 3  degrees of freedom
Residual deviance:    104.07  on 1  degrees of freedom
AIC: 144.74
```

- $e^{\alpha_2} = 5.0\ (4.9, 5.1)$ are the odds of a seatbelt having been worn (about 83%).
- Similarly the odds of fatal injury are $e^{-\beta_2} \approx 0.13\%$ (0.10%, 0.16%).
- The deviance is large $(104 \gg 1)$ because the data strongly suggest that $\gamma_{ij} \neq 0$.

Contingency Table as a Binomial GLM

We can model the rows of two column matrix (table) as binomial.

```
> belt <- belt[,2:1]
> sb <- factor(dimnames(belt)$Seatbelt)
> summary(glm(belt ~ sb, family = binomial))
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.253      0.136   38.50   <2e-16 ***
sbYes          2.383      0.242    9.84   <2e-16 ***
---

    Null deviance: 1.0407e+02  on 1  degrees of freedom
Residual deviance: 9.5479e-15  on 0  degrees of freedom
AIC: 14.89
```

$$\operatorname{logit} P(Y = \text{non-fatal} \,|\, X = i) = \beta_2 + \gamma_{i2}.$$

Compare to the Poisson model.

General Case for Log-Linear Models

Let's look at the general case of a Poisson model with

$$\log \mu_{ijk} = \log n\pi_{ijk} = \lambda^{\emptyset} + \lambda_j^X + \lambda_j^Y + \lambda_{ij}^{XY} + \cdots + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

These parameters λ are the log-linear parameters. Some restriction is needed for identifiability.

By default, for unordered factors, R sets parameters to 0 if *any* of the indices are baseline. e.g. $\lambda_1^X = \lambda_{1j}^{XY} = \lambda_{i1k}^{XYZ} = 0$ for all i, j, k (the corner point constraint).

Using this identifiability constraint, for each fixed i, k we find that

$$\log \frac{\pi_{j|ik}}{\pi_{1|ik}} = \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

so fitting this multinomial response model will give the same parameter estimates and standard errors.

GLMs for Multinomial Responses

Multinomial GLMs are an extension of Binomial GLMs. After choosing one level of the response as the baseline (say, 1), the model fits the following set of simultaneous equations:

$$\begin{aligned}\text{logit}(Y = j \mid Y \in \{1, j\}) &\equiv \log\left[\frac{P(Y = j \mid Y \in \{1, j\})}{1 - P(Y = j \mid Y \in \{1, j\})}\right] \\ &= \beta_{0(j)} + x_{i1}\beta_{1(j)} + \ldots + x_{ip}\beta_{p(j)} \qquad (2)\end{aligned}$$

for $j = 2, \ldots, J$. Only $J - 1$ simultaneous equations are needed, the logit values for other pairs of levels (j, j') can be derived as

$$\begin{aligned}&\text{logit}(Y = j \mid Y \in \{j, j'\}) = \\ &\qquad = \text{logit}(Y = j \mid Y \in \{1, j\}) - \text{logit}(Y = j' \mid Y \in \{1, j'\}) \qquad (3)\end{aligned}$$

and from there the parameters of the regression models.

Note that these parameters are all variation independent.

Multinomial GLMs in R

```
> library(MASS)
> head(housing, 3)
> library(nnet)
> multinom(Sat ~ Infl, data=housing, weights=Freq)
```

```
Coefficients:
       (Intercept) InflMedium InflHigh
Medium       -0.51       0.42      0.6
High         -0.48       0.73      1.5

Std. Errors:
       (Intercept) InflMedium InflHigh
Medium       0.097       0.14     0.18
High         0.096       0.13     0.16

Residual Deviance: 3543
AIC: 3555
```

Polynomial Contrasts

For ordered factors, it is useful to reparameterize in terms of quantities that take this ordering into account.
We can separate out effects of the covariate into orthogonal components: constant, linear, quadratic, etc. using polynomials. Let

$$b_{ij} = \text{logit}(Y = j \mid Y \in \{1, j\}, X = i),$$

so the model is described by $\boldsymbol{B}$.
Let $\boldsymbol{C}$ be a non-singular matrix with orthogonal rows: can equivalently look at $\boldsymbol{CB}$.

```
> # polynomials used in R
> t(contr.poly(4))

     [,1]  [,2]  [,3] [,4]
.L -0.67 -0.22  0.22 0.67
.Q  0.50 -0.50 -0.50 0.50
.C -0.22  0.67 -0.67 0.22
```

A Contingency Table with Ordinal Variables: Income

This small example from [DE 5] describes a survey on job satisfaction as a function of income in the United States. The sample size can considered to be fixed, as the number of questionnaires is fixed in advance.

```
> job = read.table("job.satisfaction.txt", header = TRUE)
> job$Income = ordered(job$Income,
+    levels = c("< 6000", "6000-15000", "15000-25000", "> 25000"))
> job$Satisfaction = ordered(job$Satisfaction,
+    levels = c("Very Dissatisfied", "Little Dissatisfied",
+               "Moderately Satisfied", "Very Satisfied"))
> table(job)

             Satisfaction
Income        Very Dissatisfied Little Dissatisfied Moderately Satisfied
  < 6000                     20                   24                   80
  6000-15000                 22                   38                  104
  15000-25000                13                   28                   81
  > 25000                     7                   18                   54
             Satisfaction
Income        Very Satisfied
  < 6000                  82
  6000-15000             125
  15000-25000            113
  > 25000                 92
```

Contingency Tables as Multinomial GLMs

By default, R uses polynomial constrasts for ordered factors. These give a parameter for the linear, quadratic, cubic, etc. component of the covariate.

```
> library(nnet)
> summary(multinom(Satisfaction ~ Income, data = job))
```

```
Coefficients:
                     (Intercept) Income.L Income.Q Income.C
Little Dissatisfied         0.61     0.56   -0.094    0.022
Moderately Satisfied        1.70     0.50    0.023   -0.038
Very Satisfied              1.97     0.88    0.044   -0.025

Std. Errors:
                     (Intercept) Income.L Income.Q Income.C
Little Dissatisfied         0.17     0.37     0.34     0.31
Moderately Satisfied        0.15     0.33     0.30     0.28
Very Satisfied              0.15     0.32     0.30     0.27

Residual Deviance: 2085
AIC: 2109
```

Ordered Factors: R Code

And we can get the same answer using the contrast matrix:

```
> tab <- table(job)
> t(tab)

                    Income
Satisfaction         < 6000 6000-15000 15000-25000 > 25000
  Very Dissatisfied      20         22          13       7
  Little Dissatisfied    24         38          28      18
  Moderately Satisfied   80        104          81      54
  Very Satisfied         82        125         113      92

>  t(log(tab) - log(tab[,1])) %*% contr.poly(4)


Satisfaction              .L     .Q     .C
  Very Dissatisfied    0.00  0.000  0.000
  Little Dissatisfied  0.56 -0.094  0.022
  Moderately Satisfied 0.50  0.023 -0.038
  Very Satisfied       0.88  0.044 -0.025
```

From Multinomial to Ordered Responses

If the response is an ordinal random variable, we can also model the cumulative logit with a GLM, i.e. a logit link on the cumulative distribution function

$$\begin{aligned}\operatorname{logit}(Y_i \leqslant j) &= \log\left(\frac{P(Y_i \leqslant j)}{P(Y_i > j)}\right) \\ &= \log\left(\frac{F_{Y_i}(j)}{1 - F_{Y_i}(j)}\right) = \beta_{0(j)} + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \quad (4)\end{aligned}$$

with a different intercept for each level but the same regression coefficients across levels. Intercepts $\beta_{0(j)}$ are constrained to be increasing in j so that $P(Y_i \leqslant j \mid \mathbf{X})$ increases in j for any fixed set of explanatory variables $\mathbf{X}$.

This is called a cumulative or proportional odds (ratio) model.

Contingency Tables as Ordinal Regressions

The `polr()` function in `MASS` does ordinal regression.

```
> library(MASS)
> summary(polr(Satisfaction ~ Income, data = job))
```

```
Coefficients:
           Value Std. Error t value
Income.L  0.4163      0.136   3.052
Income.Q  0.0538      0.128   0.422
Income.C -0.0150      0.119  -0.126

Intercepts:
                                                Value   Std. Error t value
Very Dissatisfied|Little Dissatisfied          -2.641    0.133    -19.881
Little Dissatisfied|Moderately Satisfied       -1.490    0.087    -17.173
Moderately Satisfied|Very Satisfied             0.151    0.068      2.215

Residual Deviance: 2087.63
AIC: 2099.63
```

Estimating Parameters in Contingency Tables

Under the multinomial sampling assumption, estimating the parameters of a contingency table means estimating the probabilities π_j associated with the cells.

- The usual frequentist estimator is the relative frequency

$$\hat{\pi}_j = \frac{n_j}{n}$$

  which is also the maximum likelihood estimator without further structure. The standard error is $\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)/n}$.
- Some careful considerations are required when dealing with sparse tables, i.e. tables with low counts and/or many zero cells.

prop.table(): Computing Cell Probabilities

The frequentist estimator $\hat{\pi}_{ijk}$ and marginal and conditional probabilities like $\hat{\pi}_{i++}$ and $\hat{\pi}_{ij|k}$ can all be computed with `prop.table()`. The syntax is similar to that of `margin.table()`, and the two functions can be combined.

```
> # get proportions within each level of Height:
> prop.table(table(lizards), margin = 3)

, , Height = high

           Diameter
Species     narrow  wide
  Distichus  0.277 0.265
  Sagrei     0.326 0.133

, , Height = low

           Diameter
Species     narrow  wide
  Distichus  0.421 0.283
  Sagrei     0.221 0.076
```

Sparse Contingency Tables: Small Cell Counts

The frequentist estimator $\hat{\pi}_j$ is problematic for sparse contingency tables, that is, when n is not large compared to the number of cells J because:

- some cells will have zero counts ($n_j = 0$); we may not know whether it is impossible to observe that configuration of the variables (a structural zero) or it is just rare enough that we do not have it in the sample (sampling zero);
- some estimated probabilities will be $\hat{\pi}_j = 0$, which places them right at the boundary of their domain and thus breaks the assumptions of most asymptotic results.

In such cases we have options:

- applying a continuity correction to the n_j or collapsing levels;
- using a smaller model that doesn't approach the boundary of the parameter space;
- using a Bayesian approach so that posterior mass is moved away from the boundary;
- use a shrinkage approach.

Collapsing Levels of Cell

In some situations the easiest solution to small n_j is to collapse levels for one or more variables, e.g. merging adjacent age brackets. The new cell counts are larger as a result, which makes the applicability of large sample properties more plausible.

Theoretical properties are generally preserved (though not conditional independence, as we see later).

You can recode a factor in R easily:

```
> x

 [1] B B B C A B B B C A B B
Levels: A B C

> levels(x) <- c("A", "A", "C")
> x

 [1] A A A C A A A A C A A A
Levels: A C
```

Bayesian Models

A common choice of prior for $\boldsymbol{\pi}$ is the Dirichlet distribution with parameter $\boldsymbol{\alpha} = \{\alpha_i\}$, $\alpha_j > 0$, with density

$$p(\boldsymbol{\pi}; \boldsymbol{\alpha}) \propto \prod_j \pi_j^{\alpha_j - 1}, \qquad \text{for } \pi_j \geq 0, \quad \textstyle\sum_j \pi_j = 1.$$

This is because it's conjugate to the multinomial:

$$p(\boldsymbol{n} \mid \boldsymbol{\pi}) \propto \prod_j \pi_j^n$$

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}, \boldsymbol{n}) \propto \prod_j \pi_j^n \times \prod_j \pi_j^{\alpha_j - 1} = \prod_j \pi_j^{\alpha_j + n_j - 1}, \quad \text{for } \pi_j \geq 0, \quad \textstyle\sum_j \pi_j = 1.$$

So the posterior distribution is a Dirichlet with parameters $\alpha_j + n_j - 1$. The α_js are sometimes called the imaginary sample size.
Letting $\alpha = \sum_j \alpha_j$, the posterior means and variances are then

$$\frac{\alpha_j + n_j}{\alpha + n} \qquad \frac{(\alpha_j + n_j)(\alpha + n - \alpha_j - n_j)}{(\alpha + n)^2(\alpha + n + 1)}$$

which are (almost) the same as the frequentist ones in the limit $\alpha \to 0$.

The Dirichlet-Multinomial Posterior Distribution

The Dirichlet prior for the π_j is

$$f(\{\pi_j\};\{\alpha_j\}) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_j \pi_j^{\alpha_j - 1},$$

$$\alpha_j > 0, \pi_j \geq 0, \sum_j \pi_j = 1;$$

and the multinomial density is

$$f(\{n_j\}; n, \{\pi_j\}) = \frac{n!}{\prod_j n_j!} \prod_j \pi_j^{n_j},$$

$$\pi_j \geq 0, n_j \in \mathbb{N}^+, \sum_j \pi_j = 1;$$

so the Dirichlet posterior is

$$f(\{\pi_j\};\{n\pi_j + \alpha_j\}) = \frac{\Gamma(n + \sum_j \alpha_j)}{\prod_j \Gamma(n\pi_j + \alpha_j)} \prod_j \pi_j^{n\pi_j + \alpha_j - 1}.$$

Posterior for the Independent Multinomial Sampling

In the case of independent multinomial sampling, we typically have a collection

$$f_k(\{\pi_{j|k}\};\{\alpha_{j|k}\}), \qquad\qquad k = 1,\ldots,K$$

of independent priors that result in K independent Dirichlet posteriors

$$f_k(\{\pi_{j|k}\};\{n\pi_{j|k}+\alpha_{j|k}\}) = \frac{\Gamma(n_{+k}+\sum_{jk}\alpha_{j|k})}{\prod_{jk}\Gamma(n\pi_{j|k}+\alpha_{j|k})}\prod_{jk}\pi_{jk}^{n\pi_{jk}+\alpha_{jk}-1},$$

which are then combined to give the overall posterior for the contingency table:

$$f(\{\pi_{j|k}\};\{n\pi_{j|k}+\alpha_{j|k}\}) = \prod_{k=1}^{L} f_k(\{\pi_{j|k}\};\{n\pi_{j|k}+\alpha_{j|k}\}).$$

Parameters in the Prior and the Posterior

Using this interpretation, the estimated probability for each cell in the prior is

$$\tau_j = \frac{\alpha_j}{N} \qquad \text{with} \qquad N = \sum\nolimits_j \alpha_j$$

and the corresponding estimate in the posterior is

$$\tilde{\pi}_j = \frac{\alpha_j + n\hat{\pi}_j}{n + \sum_j \alpha_j},$$

which can be rewritten as a convex combination of the prior and the observed cell probabilities

$$\frac{\alpha_j + n\pi_j}{n + \sum_j \alpha_j} = \frac{N\tau_j + n\pi_j}{n + N} = \frac{N}{N+n}\tau_j + \frac{n}{N+n}\pi_j.$$

The Imaginary Sample Size

The quantity $N = \sum_j \alpha_j$ is called the imaginary sample size, and controls the "weight" of the prior compared to the observed data:

- if $N \gg n$ then the prior dominates the likelihood;
- if $n \gg N$ then the likelihood dominates the prior.

We prefer the latter because when we are using a simple prior, such as the uniform

$$\alpha_j = \frac{N}{J} \qquad \text{for all } j \tag{5}$$

the ratio N/n acts as a smoothing or regularisation parameter for the posterior.

Note the uniform prior is often called non-informative, and indeed we know from information theory is has the highest possible entropy. This does not mean that it is completely uninformative!

Shrinkage: the James-Stein Estimator

A shrinkage estimator $\tilde{\pi}_j$ is defined as the convex combination of the observed distribution and a target distribution τ_j, which in the case of contingency tables means

$$\tilde{\pi}_j = \lambda \tau_j + (1 - \lambda)\hat{\pi}_j, \qquad \lambda \in [0, 1].$$

For the Bayesian posterior estimator we have

$$\hat{\lambda} = \frac{N}{N + n}.$$

The James-Stein estimator chooses λ to minimise the mean squared error, so it compromises between the biased, low variance target and the unbiased, high variance MLE. A closed-form estimate for the optimal shrinkage coefficient λ is

$$\hat{\lambda} = \frac{1 - \sum_j \hat{\pi}_j^2}{(n-1)\sum_j (\tau_j - \hat{\pi}_j)^2} \tag{6}$$

as derived in Hausser & Strimmer (JMLR 10:1469–1484, 2009) from James & Stein (1961) and Ledoit & Wolf (2003).

Shrinkage Estimators and Bayesian Posteriors

It is clear from the respective definitions that there is a one-to-one correspondence between shrinkage and posterior estimators:

- the target distribution plays the role of the prior;
- and the shrinkage coefficient is determined by the sample size and the imaginary sample size.

Both have a few properties in common:

- they provide regularised estimates for small samples;
- as $n \to \infty$ they converge to the maximum likelihood estimates, i.e. $\tilde{\pi}_j \to \hat{\pi}_j$;
- for small n they smooth estimated probabilities and provide non-zero estimated probabilities for cells with zero counts, i.e. $\tilde{\pi}_j > 0$.

The shrinkage estimator is an empirical Bayes estimator since $\hat{\lambda}$ is chosen from the data, whereas the posterior estimator is fully Bayesian.

Hypothesis Testing

Tables with Structure

Sometimes the structure of the data is best modelled by a model tailored to the underlying sampling mechanism. The Bradley-Terry model for pairwise comparisons assumes an individual skill level for participants in games. It is based on symmetric logit functions for each pair (i, j),

$$\log\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = \beta_i - \beta_j, \tag{7}$$

and level i is 'better than' j if $\beta_i > \beta_j$. The estimated probability of person i winning is then

$$\hat{\pi}_{ij} = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}$$

with a confidence interval based on the covariance matrix of the maximum likelihood estimates through

$$\mathrm{Var}(\hat{\beta}_i - \hat{\beta}_j) = \mathrm{Var}(\hat{\beta}_i) + \mathrm{Var}(\hat{\beta}_j) - 2\,\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j).$$

See AA Chapter 11.

Common Hypotheses of Interest

A large part of the analysis of contingency tables is testing different hypotheses. Some examples:

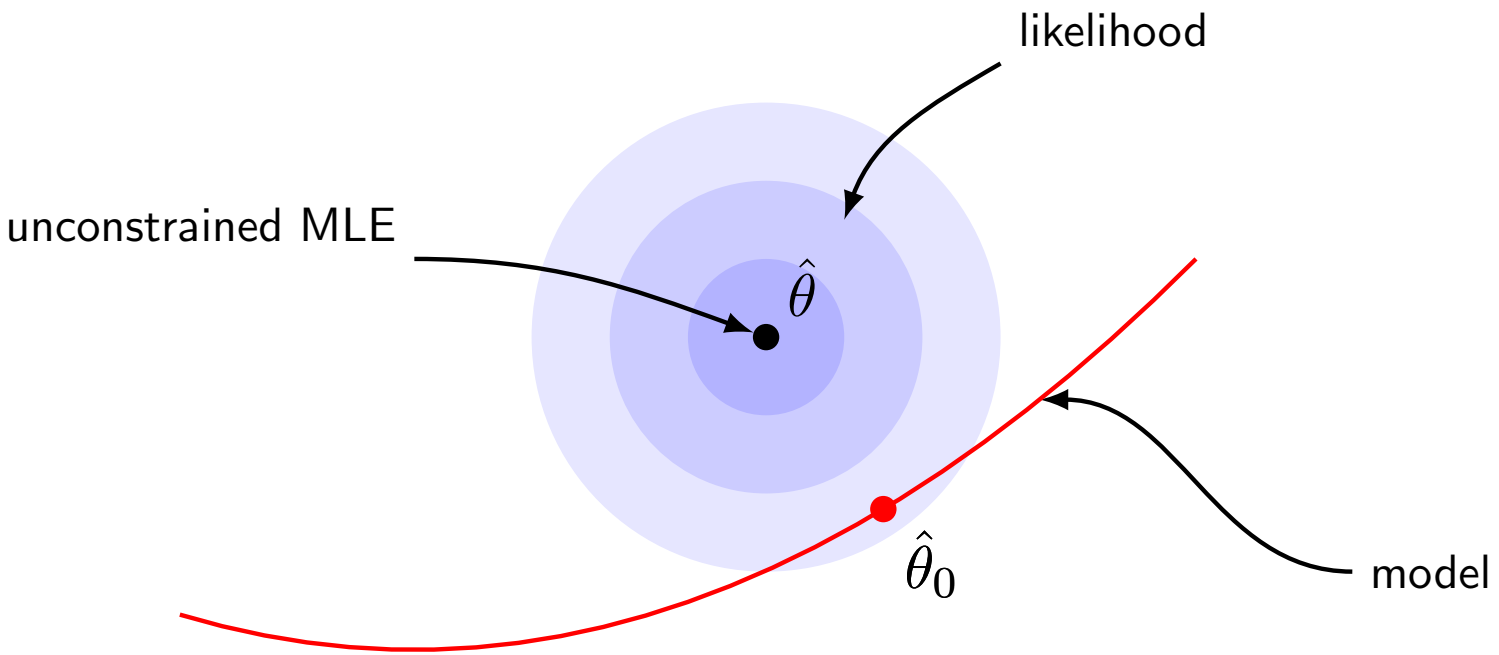
- whether two variables are marginally or conditionally independent;
- whether one ordinal variable shows a trend (increasing or decreasing) as a function of a second ordinal variable;
- whether one or more categorical variables have the same distribution for all the levels of a separate set of variables (a homogeneity test);
- testing paired observations for a statistically significant difference between the two measures.

We generally approach this using likelihood ratio tests, i.e. by looking at the change in deviance of a model and comparing it to the degrees of freedom.

Deviance

What is deviance?

Suppose we have a p dimensional model inside the $q > p$ dimensional space of distributions.



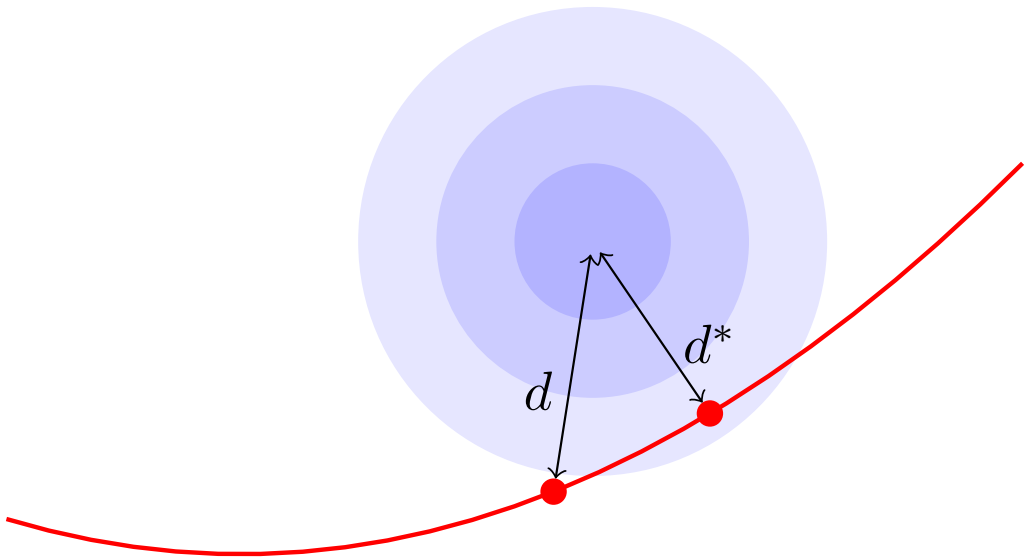
Deviance just measures the reduction in likelihood from the global MLE to the highest point in your model.

Deviance

Why is it χ^2 distributed? Well, remember that MLEs are asymptotically normal with mean θ.

We expect that the unconstrained estimator is a distance $d = (\sum_{i=1}^{q} Z_i^2)^{1/2}$ from the truth, where Z_i are i.i.d. standard normals. Then d^2 has a χ^2_q distribution.

BUT any distance which is within the p dimensional span of the model won't add to the distance between the MLE and the model, so only the other $q - p$ components will contribute to d^*.



So Just Count the Dimension

Some special cases:

- **Independence of rows and columns**: in an $I \times J$ table: full model has $q = IJ - 1$ parameters, independence assumes $\pi_{ij} = \pi_{i+}\pi_{+j}$ so only $p = (I - 1) + (J - 1)$ parameters. So deviance should use $q - p = (I - 1)(J - 1)$ degrees of freedom.
- **Symmetry**: in an $I \times I$ table: full model has $p = I^2 - 1$ parameters, symmetry assumes $\pi_{ij} = \pi_{ji}$ for $i \neq j$, which is $q - p = I(I - 1)/2$ constraints. So symmetry model has dimension $p = I(I + 1)/2 - 1$.
- **Bradley-Terry**: in an $I \times I$ table with no diagonal (competitors do not play themselves) the full model has $(I^2 - I)/2$ parameters, the model has $p = I - 1$ parameters (one skill level for each player, with one identifiability constraint). So there are $I^2 - 2I$ degrees of freedom.

χ^2-tests

The easiest way to compare nested models is via a likelihood ratio test. Recall that, if we have models $\mathcal{M}_0 \subset \mathcal{M}_1$ of dimensions p and $q > p$, with maximum likelihood estimates $\hat{\boldsymbol{\pi}}^0$ and $\hat{\boldsymbol{\pi}}^1$, then

$$2\left\{l(\hat{\boldsymbol{\pi}}^1; \boldsymbol{n}) - l(\hat{\boldsymbol{\pi}}^0; \boldsymbol{n})\right\} \to \chi^2_{q-p}$$

$$2\sum_j n_j \log\left(\frac{\hat{\pi}^1_j}{\hat{\pi}^0_j}\right).$$

In particular, if $\mathcal{M}_1$ is the saturated model with dimension $q = J - 1$, then the MLEs are the normalised empirical frequencies: $\hat{\pi}^1_j = n_j/n$. So this becomes

$$G^2 \equiv 2\sum_j n_j \log\left(\frac{n_j}{n\hat{\pi}^0_j}\right) \to \chi^2_{J-1-p}$$

$J - 1 - p$ is the number of independent constraints imposed by the model on $\boldsymbol{\pi}$.

Likelihood Ratio Tests

The likelihood ratio test (also called the G^2 test or mutual information test) rejects $H_0 : \boldsymbol{\pi} \in \mathcal{M}_0$ if G^2 is larger than the upper $(1-\alpha)$ point of the χ^2_{J-1-p}.
The G^2 statistic is often written as

$$G^2 \equiv 2\sum_j O_j \log\left(\frac{O_j}{E_j}\right) = 2\sum_j n_j \log\frac{n_j}{n\hat{\pi}_j}$$

where $O_j = n_j$ is the observed count and $E_j = n\hat{\pi}_j$ is the expected count (that is, our estimate of $\mathbb{E}n_j$ under the model).
A related test uses Pearson's X^2 statistic:

$$X^2 \equiv \sum_j \frac{(O_j - E_j)^2}{E_j} = \sum_j \frac{(n_j - n\hat{\pi}_j)^2}{n\hat{\pi}_j}.$$

You may recognise them both as estimates of the deviance parameters in Poisson GLMs.

The Relationship Between the X^2 and G^2 Tests

It can be shown that X^2 and G^2 are approximately the same:

$$\log(1+x) = x - \frac{1}{2}x^2 + O(x^3),$$

Under the null $(O_j - E_j)/E_j = \mathcal{O}_p(n^{-1/2})$, so

$$\log \frac{O_j}{E_j} = \log\left(1 + \frac{O_j - E_j}{E_j}\right) = \frac{O_j - E_j}{E_j} - \frac{(O_j - E_j)^2}{2E_j^2} + \mathcal{O}_p(n^{-3/2}),$$

then

$$\begin{aligned} O_j \log \frac{O_j}{E_j} &= \frac{O_j}{E_j}\left((O_j - E_j) - \frac{(O_j - E_j)^2}{2E_j}\right) + \mathcal{O}_p(n^{-1/2}) \\ &= \left((O_j - E_j) - \frac{(O_j - E_j)^2}{2E_j}\right) + \mathcal{O}_p(n^{-1/2}) \end{aligned}$$

$$\implies \qquad 2\sum_j O_j \log \frac{O_j}{E_j} \simeq \sum_j \frac{(O_j - E_j)^2}{E_j}$$

since $\sum_j (O_j - E_j) = 0$. So G^2 and X^2 are asymptotically equivalent.

Example: Marginal Independence

Marginal independence between rows and columns in a contingency table correponds to $\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$ for each i, j. There are $(I-1)+(J-1)$ free parameters, corresponding to the two marginal distributions.
The log-likelihood is

$$\begin{aligned}\sum_{i,j} n_{ij} \log \pi_{ij} &= \sum_{i,j} n_{ij} \log \pi_{i+} + \sum_{i,j} n_{ij} \log \pi_{+j} \\ &= \sum_{i} n_{i+} \log \pi_{i+} + \sum_{j} n_{+j} \log \pi_{+j}\end{aligned}$$

so we just maximise the terms separately, and find $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$. Then $\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n$ and the LR statistic is

$$G^2(X,Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log \frac{n_{ij}n}{n_{i+}n_{+j}}. \tag{8}$$

and is asymptotically distributed under the null as a $\chi^2_{(I-1)(J-1)}$. The degrees of freedom are computed as the difference between the number of free parameters in the observed table $(I \times J - 1)$ and the number of free parameters under the null $(I - 1 + J - 1)$.

Degrees of Freedom and Sparse Contingency Tables

In sparse contingency tables, some of the n_{ij} may be zero, as well as some of the $\{n_{i+}\}$ and $\{n_{+j}\}$. Some n_{ij} may be zero because the underlying π_{ij} is small compared to the sample size and that configuration of variables has not been observed; we call this a sampling zero. On the other hand, it may be that $\pi_{ij} = 0$ so it is impossible to observe configuration of variables; we call the cell a structural zero and the contingency table an incomplete table.
In the general case, the adjusted degrees of freedom for the χ^2 are

$$\nu = (T_e - z_e) - (T_p - z_p) \tag{9}$$

where (from [DF 3.8]):

- T_e is the total number of cells;
- T_p is the number of parameters fitted;
- z_e is the number of cells with $\hat{\pi}_{ij} = 0$ (i.e. the sampling zeros);
- z_p is the number of parameters $\hat{\pi}_{ij}$ cannot be estimated (because either $\hat{\pi}_{i+} = 0$ or $\hat{\pi}_{+j} = 0$ or both, i.e. the structural zeros).

Parametric, Nonparametric and Semiparametric Tests

Problems can arise if we don't have enough data for these asymptotics to kick in.

In order to determine a threshold for the test, we need a null distribution; there is more than one way to go about this, and we classify tests as follows.

- Parametric tests: the full distribution is completely specified by the null hypothesis. They can be:
  - asymptotic tests (e.g. χ^2 log-likelihood ratio tests);
  - exact tests (e.g. F tests in linear models).
- Nonparametric tests: no distributional assumption is made, and an empirical null distribution is built using either bootstrap resampling or permutations.
- Semiparametric tests: the null distribution is specified up to one or more parameters, which are estimated from the empirical null distribution through bootstrap resampling or permutations.

Pros and Cons of Different Types of Tests

- Parametric tests can be biased when assumptions are violated or sample size is not large enough for the test statistic to converge to the asymptotic distribution.
- Nonparametric tests are slower than parametric tests due to the need of generating the permutations or the bootstrap samples and to evaluate the test statistic on each of them. But we can use a nonparametric test even when a closed-form null distribution is not available.
- Permutation tests are always unbiased by construction, so they always reject the null hypothesis $\alpha \times 100\%$ of the time.
- Semiparametric tests are a compromise that require much less resampling (typically $10\times$ less for the same precision) while still being reasonably robust.
- Nonparametric tests condition on the observed data set, whereas parametric tests are defined on the general population the sample is drawn from. This affects the interpretation of inference results.

Using Permutation Tests

We can 'unravel' our counts as a vector of observations:

$$\begin{array}{cccccc} X_1 & X_2 & X_3 & \cdots & X_{n-1} & X_n \\ Y_1 & Y_2 & Y_3 & \cdots & Y_{n-1} & Y_n \end{array}$$

The first n_{11} pairs are $(X_i = 1, Y_i = 1)$, and so on.

If the null hypothesis that $X_i \perp\!\!\!\perp Y_i$ is true, then the probability of observing this is the same if we take any permutation of the X_is but keep the Y_is the same (or vice versa).

This gives a new table, with the same margins n_{i+}, n_{+j}.

Examples:

4	0
0	4

3	1
1	3

2	2
2	2

1	3
3	1

0	4
4	0

`chisq.test()`: Asymptotic and Permutation X^2

- Asymptotic χ^2 test

```
> chisq.test(belt, correct = FALSE)


Pearson's Chi-squared test

data:  belt
X-squared = 200, df = 1, p-value <2e-16
```

- Monte Carlo permutation test, with $B = 5000$ permutations.

```
> chisq.test(belt, simulate.p.value = TRUE, B = 5000)


Pearson's Chi-squared test with simulated p-value (based on 5000 replica

data:  belt
X-squared = 200, df = NA, p-value = 2e-04
```

Monte Carlo Permutation Test for Independence

A Monte Carlo implementation of such a permutation test then is as follows:

1. Compute the marginals $\{n_{i+}\}$ and $\{n_{+j}\}$ from $\boldsymbol{n}$.
2. Compute the value of the test statistic T (X^2 or G^2 or whatever) for $\boldsymbol{n}$.
3. Generate a large enough number B of random contingency tables $\boldsymbol{n}^*$ with fixed marginals $\{n_{i+}\}$ and $\{n_{+j}\}$ by (e.g.) permuting the X's.
4. Estimate the empirical distribution of Pearson's G^2 under H_0 as $\{T(\boldsymbol{n}_1^*), \ldots, T(\boldsymbol{n}_B^*)\}$.
5. Compute the p-value for the test statistic as

$$P(T(\boldsymbol{n}^*) \geqslant T(\boldsymbol{n})) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(T(\boldsymbol{n}_b^*) \geqslant T(\boldsymbol{n})) \tag{10}$$

   using the right tail of the empirical distribution under H_0.

Fisher's Exact Test

It is actually possible to calculate the probability of obtaining a particular 2×2 table under the permutation test using a hypergeometric distribution. If you can evaluate this probability for all possible tables, then one can calculate exact p-values.

Proceeding in this manner leads to Fisher's exact test. The p-value of the test is (again) the probability of a generated contingency table $\boldsymbol{n}^*$ having a test statistic at least as large as $\boldsymbol{n}$: $T(\boldsymbol{n}^*) \geq T(\boldsymbol{n})$.

This test is not computationally feasible to use on tables with very large numbers of cells because there are too many possible tables to enumerate. Permutation tests or Markov chain approximations are used instead.

If the table is not sparse, then it is better just to use the χ^2-approximation.

Fisher's test: `fisher.test()`

```
> tab <- matrix(c(6,2,3,7),2,2)
> tab
     [,1] [,2]
[1,]    6    3
[2,]    2    7
> ## fisher.test(tab)

Fisher's Exact Test for Count Data
p-value = 0.1534
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.6218933 100.0509462
sample estimates:
odds ratio
  6.176771
```

For larger tables, adding the option `simulate.p.value=TRUE` uses a Markov chain to obtain the p-value.

Packages for Hypothesis Testing: coin and bnlearn

Two other packages that implement permutation tests are bnlearn and coin; bnlearn implements all of parametric, semiparametric and nonparametric tests, coin just nonparametric tests. Both packages provide both marginal and conditional tests.

```
> library(vcdExtra)
> belt.df = expand.dft(belt)
> library(bnlearn)
> ci.test("Seatbelt", "Injury", data = belt.df, test = "x2")

Pearson's X^2

data:  Seatbelt ~ Injury
x2 = 200, df = 1, p-value <2e-16
alternative hypothesis: true value is greater than 0
```

```
> ci.test("Seatbelt", "Injury", data = belt.df, test = "mc-x2")

Pearson's X^2 (MC)

data:  Seatbelt ~ Injury
mc-x2 = 200, Monte Carlo samples = 5000, p-value <2e-16
alternative hypothesis: true value is greater than 0
```

From Marginal to Conditional Independence

In contingency tables with more than two dimensions, we may also want to test the more general hypothesis of conditional independence.
Two variables X and Y are conditionally independent given Z if

$$f_{X|YZ}(x \,|\, y, z) = f_{X|Z}(x \,|\, z).$$

That is, the conditional distribution (density) of X given Y and Z only depends upon Z. We write this as $X \perp\!\!\!\perp Y \mid Z$.

Conditional independence is equivalent to the factorization of the density or mass function:

$$\begin{aligned} f_{XYZ}(x, y, z) = \frac{f_{XZ}(x, z) f_{YZ}(y, z)}{f_Z(z)} &= f_{X|Z}(x \mid z) f_{YZ}(y, z) \\ &= g(x, z)\, h(y, z). \end{aligned}$$

(This shows the definition is symmetric in X and Y.) Consequently, the likelihood for X, Y, Z also factorizes, which is very useful as it makes computation simpler.

Conditional Independence Testing

Note that the likelihood for the three way table can be factorized as:

$$\begin{aligned} l(\boldsymbol{\pi}; \boldsymbol{n}) &= \sum_{i,j,k} n_{ijk} \log \pi_{ijk} = \sum_{i,j,k} n_{ijk} \log(\pi_{ij|k}\pi_{++k}) \\ &= \sum_{i,j,k} n_{ijk} \log \pi_{ij|k} + \sum_{i,j,k} n_{ijk} \log \pi_{++k} \\ &= \sum_{i,j,k} n_{ijk} \log \pi_{ij|k} + \sum_{k} n_{++k} \log \pi_{++k}. \end{aligned}$$

Now conditional independence is equivalent to $\pi_{ij|k} = \pi_{+j|k}\pi_{i+|k}$, so likelihood-based inference just considers the first term.

$$l(\boldsymbol{\pi}; \boldsymbol{n}) = \sum_{i,j,k} n_{ijk} \log \pi_{ij|k}.$$

But note that this is just the same as the sum of K log-likelihoods for K separate (independent) tables $\boldsymbol{n}_{\bullet\bullet k}$.
So, it follows from our derivation for marginal independence that the MLEs are

$$\hat{\pi}_{ijk} = \hat{\pi}_{i|k}\hat{\pi}_{j|k}\hat{\pi}_{k} = \frac{n_{i+k}}{n_{++k}} \frac{n_{+jk}}{n_{++k}} \frac{n_{++k}}{n} = \frac{n_{i+k} \cdot n_{+jk}}{n_{++k} n}$$

Null Distribution of Conditional Independence Tests

The model with $X \perp\!\!\!\perp Y \mid Z$ has $(I-1)(J-1)K$ restrictions (an independence for each of the K levels of Z).

Hence the asymptotic null distribution of the X^2 and G^2 statistics is a χ^2-distribution with $(I-1)(J-1)K$ degrees of freedom.

$$\mathrm{X}^2 = \sum_{ijk} \frac{(\hat{n}_{ijk} - n_{ijk})^2}{\hat{n}_{ijk}} \longrightarrow \chi^2_{(I-1)(J-1)K}$$

Note that this is the same as what we would conclude if we took separate MLEs independence for each two-way table $\boldsymbol{n}_{\bullet\bullet k}$, and added the X^2 statistics together.

Conditional Pearson's X^2 in `bnlearn`

The syntax is the same as before, but in addition to `"Diameter"` (`x` argument) and `"Height"` (`y` argument) we also specify the conditioning variable(s) `"Species"` (`z` argument).

```
> ci.test("Diameter", "Height", "Species", data = lizards, test = "x2")


Pearson's X^2

data:  Diameter ~ Height | Species
x2 = 2, df = 2, p-value = 0.4
alternative hypothesis: true value is greater than 0
```

As an alternative, we can perform the corresponding permutation test by using `test = "mc-x2"`.

Conditional Independence as a Log-Linear Model

Recall the Poisson perspective:

$$\log(n\pi_{ijk}) = \lambda^{\emptyset} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

Suppose that $\lambda_{ij}^{XY} = \lambda_{ijk}^{XYZ} = 0$ for all i, j, k. Then

$$\begin{aligned} \log \pi_{ijk} &= \lambda^{\emptyset} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \\ &= \log r_{ik} + \log s_{jk} \\ \pi_{ijk} &= r_{ik} \cdot s_{jk}; \end{aligned}$$

i.e. $X \perp\!\!\!\perp Y \mid Z$.

In fact the converse is true, so that $X \perp\!\!\!\perp Y \mid Z$ if and only if $\lambda_{ij}^{XY} = \lambda_{ijk}^{XYZ} = 0$ for all i, j, k.

Conditional Independence: Example

Recall the lizards data.

```
> mod1 <- glm(V1 ~ Species*(Diameter+Height),
+                    family=poisson, data=liz2)
> summary(mod1)

Residual deviance:  2.0256  on 2  degrees of freedom
```

```
> 1 - pchisq(2.0256, df=2)

[1] 0.36
```

Alternatively,

```
> library(bnlearn)
> ci.test("Height", "Diameter", "Species", data=lizards)


Mutual Information (disc.)

data:  Height ~ Diameter | Species
mi = 2, df = 2, p-value = 0.4
alternative hypothesis: true value is greater than 0
```

How Are Permutations Done in Conditional Tests?

In the presence of a set of conditioning variables $\mathbf{Z}$, the conditional test is constructed as a collection of marginal tests. As a result, the sufficient statistics under the null hypothesis are the sufficient statistics for each of the sub-tables the marginal test statistics are computed on.

Therefore to permute the data and obtain the empirical null distribution:

1. We fix the marginal counts $\{n_{i+k}\}$ and $\{n_{+jk}\}$ (and thus the n_{++k} subtotal) for all the K configurations.
2. For each configuration in turn, we permute the corresponding sub-table to get $\boldsymbol{n}^*_{b(k)}$, $b = 1, \ldots, B$ and $k = 1, \ldots, K$;
3. We construct the overall permuted table as $\boldsymbol{n}^*_b = \{\boldsymbol{n}^*_{b(k)}, k = 1, \ldots, K\}$.
4. We compute $T(\boldsymbol{n}^*_b)$ a large number of times to obtain the empirical null distribution.

This may be computationally infeasible for larger K.

A Compromise: Semiparametric Tests

The semiparametric versions of G^2 and Pearson's X^2 use the asymptotic χ^2-distribution but estimate the degrees of freedom from the data as

$$df = \frac{1}{B}\sum_{b=1}^{B} \mathrm{X}^2(\boldsymbol{n}_b^*) \qquad \text{or} \qquad df = \frac{1}{B}\sum_{b=1}^{B} G^2(\boldsymbol{n}_b^*) \tag{11}$$

because if $Z \sim \chi^2_d$ then $\mathbb{E}Z = d$ and therefore d can be approximated by the mean of the test statistics obtained from the permutations.

This is a much easier estimation problem than that of a nonparametric test, because we are computing a point estimate of the mean instead of an empirical estimate of the whole distribution. Fewer permutations are required, and the degrees of freedom are self-adjusting in the presence of zero cell counts.

All G^2 Tests

```
> ci.test("Diameter", "Height", "Species", data = lizards,
+         test = "sp-mi")


Mutual Information (disc., semipar.)

data:  Diameter ~ Height | Species
sp-mi = 2, df = 1.8, Monte Carlo samples = 100.0, p-value = 0.3
alternative hypothesis: true value is greater than 0
```

```
> ci.test("Diameter", "Height", "Species", data = lizards,
+         test = "mc-mi")    # Monte Carlo


Mutual Information (disc., MC)

data:  Diameter ~ Height | Species
mc-mi = 2, Monte Carlo samples = 5000, p-value = 0.4
alternative hypothesis: true value is greater than 0
```

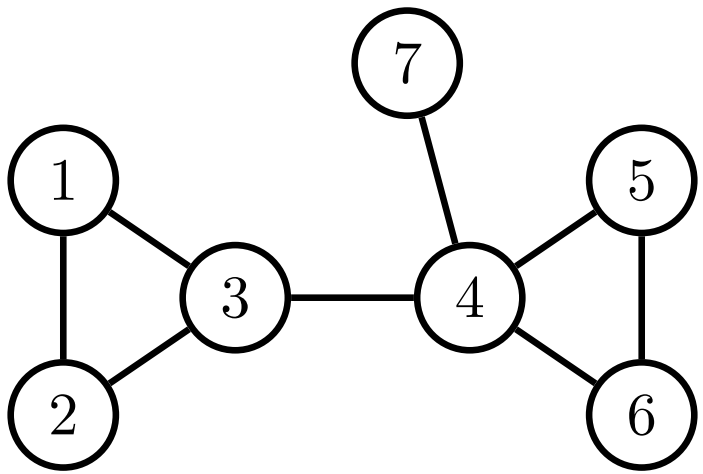
Graphical Models

Undirected Graphical Models

Graphical models relates the structure of a graph to structural restrictions on probability distributions. There are various types; we will study two: directed and undirected graphical models.

An undirected graph $\mathcal{G}$ is a pair (V, E), where V is a finite set of vertices, and E is a collection of unordered pairs of elements of V called edges.

We represent graphs by drawing the vertices (also called nodes) and joining them with a line if the corresponding edge is present.



Example: $E = \{\{1,2\},\{1,3\},\{2,3\},\{3,4\},\ldots,\{4,7\}\}$

Markov Properties (pairwise)

Let $\mathcal{G} = (V, E)$ be an undirected graph, and with each $v \in V$ associate a random variable X_v. We will use the graph to define a model on the joint distribution of $X_V \equiv \{X_v, v \in V\}$.

Say that a distribution P obeys the pairwise Markov property if:

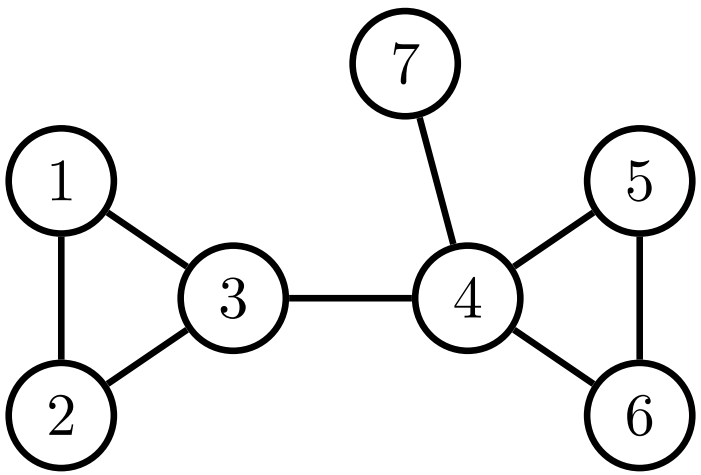
$$X_a \perp\!\!\!\perp X_b \mid X_{V \setminus \{a,b\}} \, [P] \qquad \text{whenever } (a, b) \notin E.$$

This corresponds to log-linear parameters $\lambda^A = 0$ whenever A contains both a and b.

(We abuse notation slightly and write λ^A for λ^{X_A}.)

Separation

Say that sets A and B are separated by C in $\mathcal{G}$ if every path from any $a \in A$ to any $b \in B$ goes through some $c \in C$.



For example: $\{1, 3\}$ and $\{5, 6\}$ are separated by $\{4\}$.

Global Markov Property

Say that a distribution P obeys the global Markov property if:

$$X_A \perp\!\!\!\perp X_B \mid X_C \, [P] \qquad \text{whenever } A \text{ and } B \text{ are separated by } C.$$

This is clearly stronger than the pairwise property. In fact: if P is positive, then the global Markov property for $\mathcal{G}$ holds if and only if the pairwise property holds.

Log-Linear Models for Contingency Tables (again)

Consider a multinomial model π_{ijk} for a contingency table n_{ijk}. If $\pi_{ijk} > 0$ (positivity) then we can write

$$\log(\pi_{ijk}) = \lambda^{\emptyset} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

Suppose that $\lambda_{ij}^{XY} = \lambda_{ijk}^{XYZ} = 0$ for all i, j, k. Then

$$\begin{aligned} \log \pi_{ijk} &= \lambda^{\emptyset} + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \\ &= \log r_{ik} + \log s_{jk} \\ \pi_{ijk} &= r_{ik} \cdot s_{jk}. \end{aligned}$$

In fact the converse is true, so that $X \perp\!\!\!\perp Y \mid Z$ if and only if $\lambda_{ij}^{XY} = \lambda_{ijk}^{XYZ} = 0$ for all i, j, k.

In fact we can say the following: let $\mathcal{C}$ be the set of complete sets in the graph. That is, the set of subsets $C \subseteq V$ such that $\{c, d\} \in E$ for every $c, d \in C$.

Then $P > 0$ satsifies the (global or pairwise) Markov property for $\mathcal{G}$ if and only if

$$\lambda^D = 0 \qquad \text{whenever } D \notin \mathcal{C}.$$

I.e. if edge is missing in D, the associated parameter is 0.

The maximal complete sets are called the cliques, denoted $\bar{\mathcal{C}}$.

Cliques: Example 1

In this example the cliques are:

$$\{1,2\} \qquad \{2,3\}$$

and the complete sets are these plus $\{1\},\{2\},\{3\}$.

So the model corresponds to:

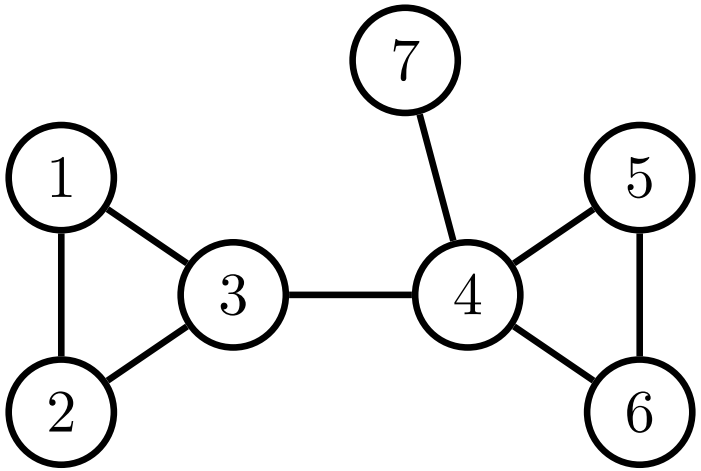
$$\log \pi(i_1,i_2,i_3) = \lambda^{\emptyset} + \lambda^1(i_1) + \lambda^2(i_2) + \lambda^{12}(i_1,i_2) + \lambda^3(i_3) + \lambda^{23}(i_2,i_3).$$

Fitting an Undirected Model

```
> mod1 <- glm(V1 ~ Species*Diameter + Species*Height,
+                  family=poisson, data=liz2)
> summary(mod1)
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                 4.3594     0.1019  42.801  < 2e-16 ***
SpeciesSagrei               0.1072     0.1450   0.739 0.459691
Diameterwide               -0.1883     0.1283  -1.467 0.142309
Heightlow                  -0.3379     0.1296  -2.607 0.009135 **
SpeciesSagrei:Diameterwide -0.7537     0.2161  -3.488 0.000486 ***
SpeciesSagrei:Heightlow    -0.6967     0.2198  -3.170 0.001526 **
---

    Null deviance: 98.5830  on 7  degrees of freedom
Residual deviance:  2.0256  on 2  degrees of freedom
AIC: 59.004
```

Cliques: Example 2



In this example the cliques are:

$$\{1,2,3\} \qquad \{3,4\} \qquad \{4,5,6\} \qquad \{4,7\}$$

and the complete sets are any subsets of these.
So the model corresponds to:

$$\log \pi(i_V) = \lambda^{123}(i_{123}) + \lambda^{34}(i_{34}) + \lambda^{456}(i_{456}) + \lambda^{47}(i_{47}) + \cdots .$$

Hierarchical Models

Graphical models are hierarchical: a hierarchical model satisfies

$$\lambda^C = 0 \implies \lambda^D = 0 \qquad \text{whenever } D \supset C.$$

As a general principle in modelling, we don't include interaction effects without including all the main effects as well.

Not all hierarchical models are graphical: for example

$$\log \pi_{ijk} = \lambda^{\emptyset} + \lambda^1_i + \lambda^2_j + \lambda^{12}_{ij} + \lambda^3_{ik} + \lambda^{13}_{ik} + \lambda^{23}_{jk}.$$

Should include λ^{123}_{ijk} to be graphical.

Factorization

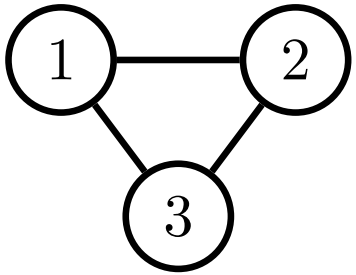
We can see from the hierarchical model form, that $\pi(i_V) > 0$ satisfies the Markov properties if and only if

$$\pi(i_V) = \prod_{C \in \bar{\mathcal{C}}} \psi_C(i_C)$$

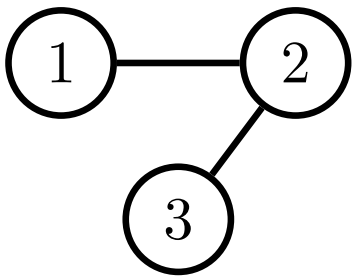
for some functions $\psi_C > 0$. This is the factorization property.

For positive distributions the factorization property, the pairwise Markov property and the global Markov property are all equivalent (see Lauritzen, 1996, for details).

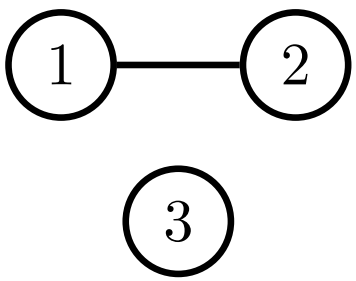
Models



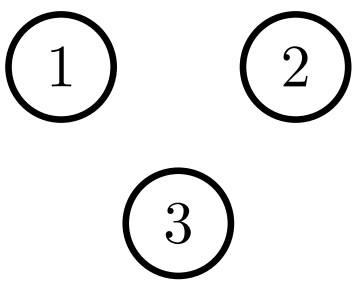
saturated model $\pi_{ijk} = \psi_{123}(i, j, k)$



$X_1 \perp\!\!\!\perp X_3 \mid X_2$ $\pi_{ijk} = \psi_{12}(i, j)\psi_{23}(j, k)$



$X_1, X_2 \perp\!\!\!\perp X_3$ $\pi_{ijk} = \psi_{12}(i, j)\psi_3(k)$



full independence $\pi_{ijk} = \psi_1(i)\psi_2(j)\psi_3(k)$

Hierarchical Non-Graphical Models

These data are found in AA Chapter 9; they consist of answers of high schoolers to a Dayton, Ohio survey on substance use.

Alcohol	Tobacco	Marijuana	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

```
> library(vcdExtra)
> library(plyr)
> data(DaytonSurvey)
> subdf <- ddply(DaytonSurvey, .variables = 1:3,
+                .fun = function(x) c(Freq = sum(x$Freq)))
```

```
> summary(glm(Freq ~ alcohol*cigarette*marijuana,
+             family=poisson, data=subdf))
```

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        5.63121    0.05987  94.060  < 2e-16
alcoholYes                         0.49128    0.07601   6.464 1.02e-10
cigaretteYes                      -1.87001    0.16383 -11.414  < 2e-16
marijuanaYes                      -4.93806    0.70964  -6.959 3.44e-12
alcoholYes:cigaretteYes            2.03538    0.17576  11.580  < 2e-16
alcoholYes:marijuanaYes            2.59976    0.72698   3.576 0.000349
cigaretteYes:marijuanaYes          2.27548    0.92746   2.453 0.014149
alcoholYes:cigaretteYes:marijuanaYes  0.58951    0.94236   0.626 0.531600
---
    Null deviance:  2.8515e+03  on 7  degrees of freedom
Residual deviance: -2.9821e-13  on 0  degrees of freedom
```

No Three-Way Interaction Model

Fitting the model without the higher order parameter gives a good fit:

```
> summary(glm(Freq ~ (alcohol + cigarette + marijuana)^2,
+                 family=poisson, data=subdf))
```

```
Residual deviance:    0.37399  on 1  degrees of freedom
AIC: 63.42
```

How do we interpret this? Well, it's equivalent to say that the conditional odds ratio between X and Y given Z is the same for each level k of Z.

So the association between (e.g.) smoking cigarettes and smoking marijuana is the same whether you drink or not. (Don't forget that odds ratios are not collapsible though!)

Decomposability

Let $\mathcal{G}$ be a graph with vertices V. A decomposition of $\mathcal{G}$ is a pair of sets $A, B \subseteq V$ such that:

- $V = A \cup B$;
- $S = A \cap B$ is complete;
- there are no edges between $A \setminus S$ and $B \setminus S$.

We say that a graph is decomposable if either (i) it is complete, or (ii) there is a decomposition A, B such that the subgraphs over A and B are decomposable.

Decomposable Models

Decomposable models (i.e. graphical models where the graph is decomposable) are particularly easy to work with.

$$\begin{aligned} P(X_V = i_V) &= P(X_A = i_A) \cdot P(X_{B \setminus S} = i_{B \setminus S} \mid X_A = i_A) \\ &= P(X_A = i_A) \cdot P(X_{B \setminus S} = i_{B \setminus S} \mid X_S = i_S) \\ &= \frac{P(X_A = i_A) \cdot P(X_B = i_B)}{P(X_S = i_S)}. \end{aligned}$$

Looking at the likelihood:

$$\begin{aligned} &\sum_{i_V} n(i_V) \log \pi(i_V) \\ &= \sum_{i_A} n(i_A) \log \pi(i_A) + \sum_{i_B} n(i_B) \log \pi(i_B) - \sum_{i_S} n(i_S) \log \pi(i_S) \\ &= \sum_{i_A} n(i_A) \log \pi(i_A) + \sum_{i_B} n(i_B) \log \pi(i_{B \setminus S} \mid i_S) \end{aligned}$$

Decomposable Models

It follows that the MLE of a decomposable model is:

$$\hat{\pi}(i_V) = \frac{\hat{\pi}(i_A) \cdot \hat{\pi}(i_B)}{\hat{\pi}(i_S)}.$$

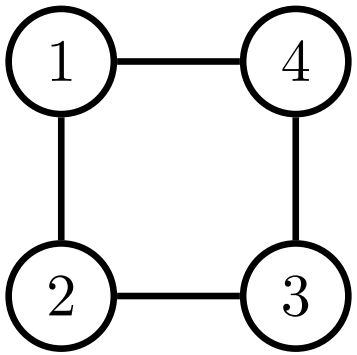
Similarly, if we put a prior distribution over $\pi(i_V)$ which factorizes into pieces for $\pi(i_A)$ and $\pi(i_B)$, then so will the posterior distribution.

```
> liz_sub <- adply(table(lizards[,1:2]), 1:2)
> summary(glm(V1 ~ Diameter*Species, family=poisson, data=liz_sub))
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                4.20469    0.08639  48.673  < 2e-16 ***
SpeciesSagrei             -0.12716    0.12624  -1.007 0.313824
Diameterwide              -0.18831    0.12834  -1.467 0.142309
SpeciesSagrei:Diameterwide -0.75373    0.21607  -3.488 0.000486 ***
```

Example: 4-cycle

The simplest non-decomposable model is the following graph on four variables:



The cliques are $\{1,2\}$, $\{2,3\}$, $\{3,4\}$ and $\{1,4\}$. The model is defined by distributions of the form

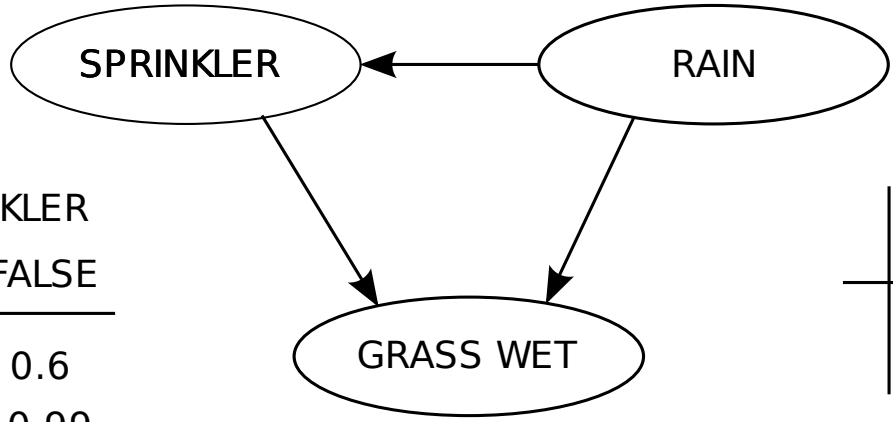
$$\log \pi_{ijkl} = \lambda^{\emptyset} + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_l^4 + \lambda_{ij}^{12} + \lambda_{jk}^{23} + \lambda_{kl}^{34} + \lambda_{il}^{14}$$

Equivalently:

$$X_1 \perp\!\!\!\perp X_3 \mid X_2, X_4 \qquad\qquad X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3.$$

A Simple Bayesian Network: Watson's Lawn

There are many instances in which it's easier to specify conditional distributions than joint ones.



RAIN	SPRINKLER	
	TRUE	FALSE
FALSE	0.4	0.6
TRUE	0.01	0.99

RAIN	
TRUE	FALSE
0.2	0.8

SPRINKLER	RAIN	GRASS WET	
		TRUE	FALSE
FALSE	FALSE	0.0	1.0
FALSE	TRUE	0.8	0.2
TRUE	FALSE	0.9	0.1
TRUE	TRUE	0.99	0.01

Sequential Regressions

Suppose we have variables $X_V = (X_1, \ldots, X_k)$. We can always write:

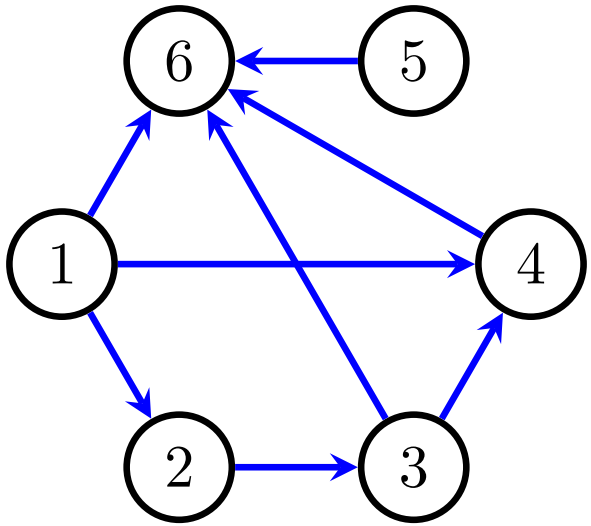
$$P(X_V = i_V) = \prod_{v \in V} P(X_v = i_v \mid X_1 = i_1, \ldots, X_{v-1} = i_{v-1}). \quad (12)$$

A conditional independence constraint would mean that, for example, $P(X_v = i_v \mid X_1 = i_1, \ldots, X_{v-1} = i_{v-1})$ only depends upon a subset of $i_1, \ldots, i_{v-1}$.

It's convenient to represent this dependence by drawing a graph.

Directed Acyclic Graphs

A directed acyclic graph $\mathcal{G}$ is a set of vertices V, and ordered pairs of edges E.



Directed cycles e.g. $1 \to 2 \to 3 \to 1$ are not allowed.

The vertices which have edges into v are called its parents, $\mathrm{pa}(v)$.

$$\mathrm{pa}(4) = \{1, 3\}$$

Bayesian Networks

So starting from our previous factorization:

$$P(X_V = i_V) = \prod_{v \in V} P(X_v = i_v \mid X_1 = i_1, \ldots, X_{v-1} = i_{v-1}).$$

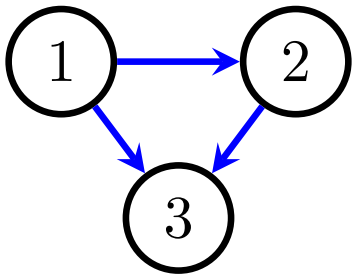
We say that a distribution P factorizes according to a directed acyclic graph $\mathcal{G}$ if

$$P(X_V = i_V) = \prod_{v \in V} P(X_v = i_v \mid X_{\mathrm{pa}(v)} = i_{\mathrm{pa}(v)}). \tag{13}$$

The model defined by this restriction is called the Bayesian networka associated with $\mathcal{G}$.

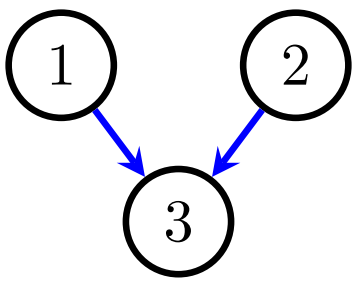
Bayesian Networks

Start with a graph over our variables, but with a directed edge into each variable from its predecessors.



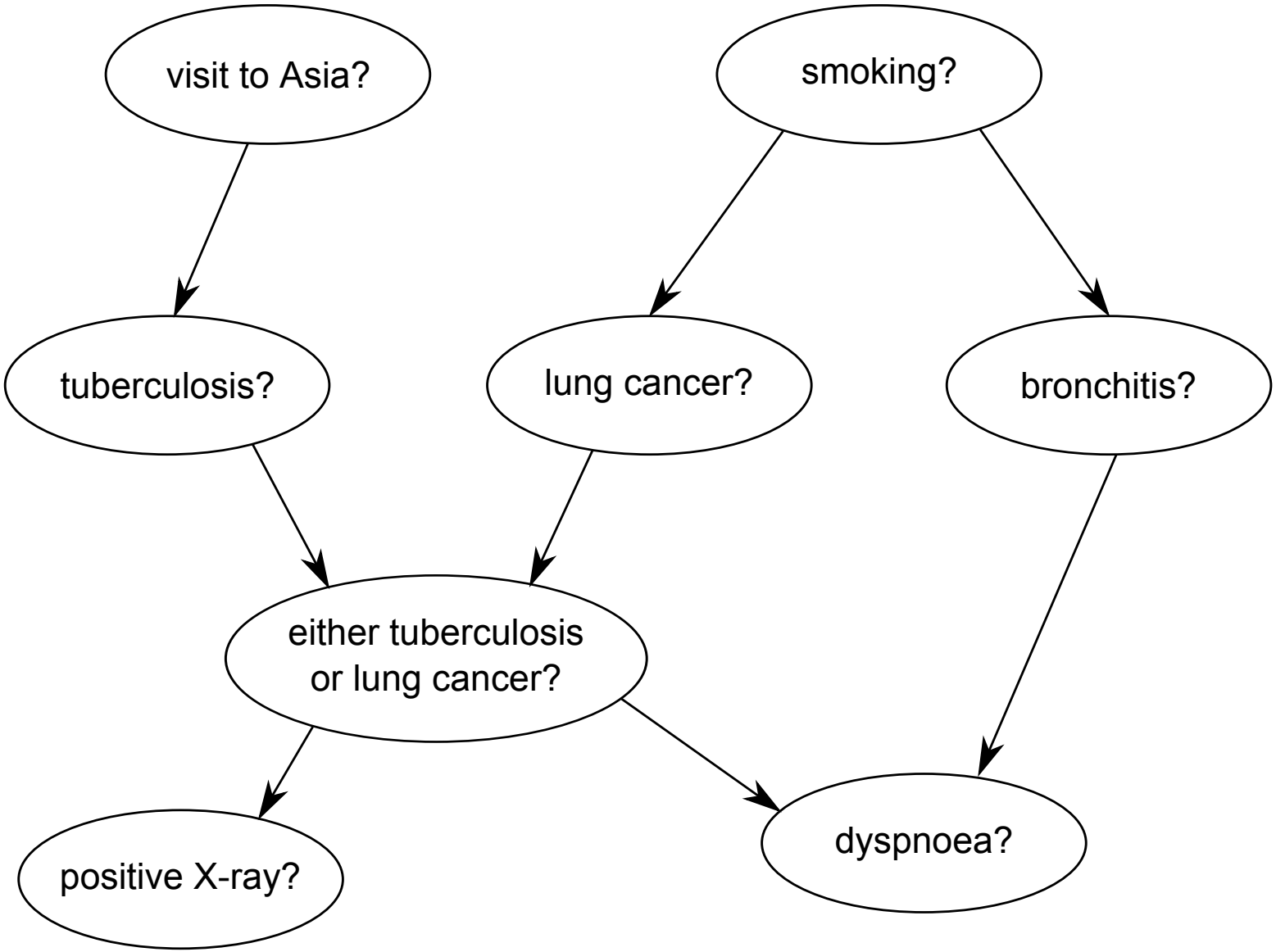
$$P(X_1 = i_1) \cdot P(X_2 = i_2 \mid X_1 = i_1) \cdot P(X_3 = i_3 \mid X_1 = i_1, X_2 = i_2).$$

Then, whenever the regression of X_v on $X_1, \ldots, X_{v-1}$ doesn't depend upon X_w, remove the $w \to v$ edge.



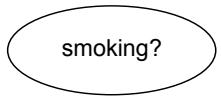
$$P(X_1 = i_1) \cdot P(X_2 = i_2) \cdot P(X_3 = i_3 \mid X_1 = i_1, X_2 = i_2).$$

A Discrete Bayesian Network

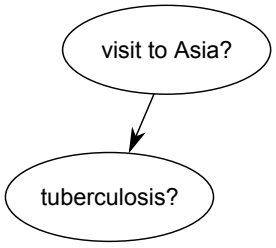


A Discrete Bayesian Network

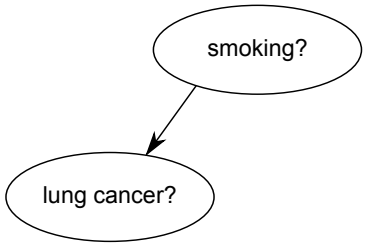
yes	no
0.01	0.99



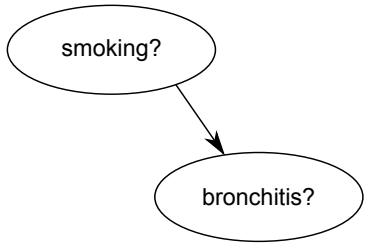
yes	no
0.50	0.50



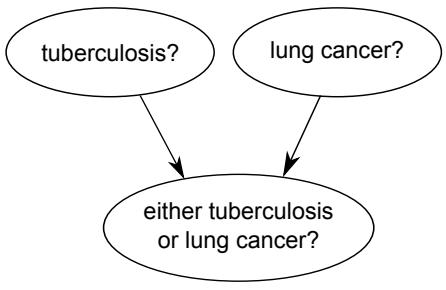
	yes	no
yes	0.05	0.95
no	0.01	0.99



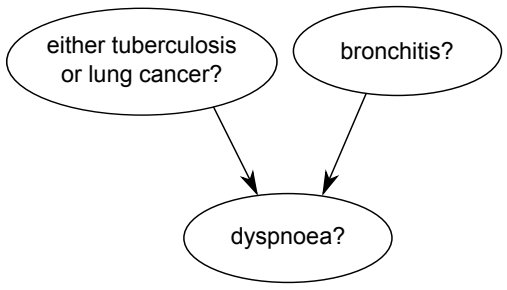
	yes	no
yes	0.10	0.90
no	0.01	0.99



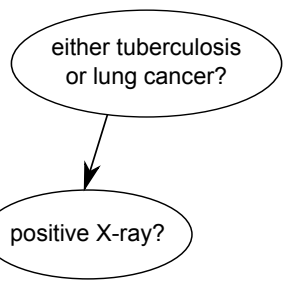
	yes	no
yes	0.60	0.40
no	0.30	0.70



	yes	no
yes:yes	1	0
yes:no	1	0
no:yes	1	0
no:no	0	1



	yes	no
yes:yes	0.90	0.10
yes:no	0.70	0.30
no:yes	0.80	0.20
no:no	0.10	0.90



	yes	no
yes	0.98	0.02
no	0.05	0.95

Fitting to Data

The `bnlearn` package can be used to fit and score models.

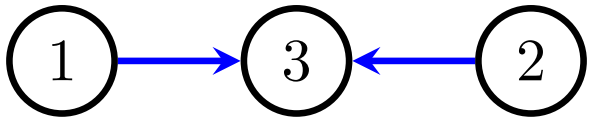
```
> library(bnlearn)
> my_bn <- as.bn("[Species][Height|Species][Diameter|Species]")
> my_bn2 <- as.bn("[Height][Diameter][Species|Height:Diameter]")
> score(my_bn, data = lizards, "loglik")

[1] -802.2

> score(my_bn2, data = lizards, "loglik")

[1] -801.5
```

Other options include `"aic"` and `"bic"`.

Relationship Between Directed and Undirected Models

Some directed models are also undirected models:

But some are not.



In fact, it is precisely decomposable models which can be represented by both undirected and directed graphical models.

Graphical Model Selection

Besides prior knowledge, model selection can be based on any of the usual methods as well as some new ones:

- score-based methods (e.g. AIC, BIC, posterior mass)
- constraint based methods (sequentially testing conditional independences or likelihood ratios);
- hybrid methods.

Typically there are too many models to search through exhaustively, so some sort of clever search method is used. Much of this is implemented in the `bnlearn` package.

Finding the best-fitting Bayesian network is an NP hard problem.

That's It!