last change: oct.25^{th} 2012@20.00

2 Discrete Multivariate Analysis

Summary 1 *joint,marginal,conditional,multinomial,binomial,poisson, likelihood estimation, likelihood ratio test, multinomial point hypothesis, independence test, chi-square tests, simpsons paradox, joint independence from, conditional independence on, contitional independence test, collapsibility, mutlitivariate conditional poisson, Appendix: Fisher test, multi hypergeometric, Mantel-Haenszel test,*

2.1 Distributions of contingency tables in 2 dimensions

also: define risk, relative risk, odds, odds ratio, life tables, rates, likelihood

joint distribution $p_{ij} = \mathbb{P}(X = x_i, Y = y_j)$

$$\begin{array}{c|c|c|c|c|c|}
\multicolumn{6}{c}{Y} \\
\cline{2-6}
 & & y_1 & \dots & y_J & \\
\cline{2-6}
 & x_1 & p_{11} & \dots & p_{1J} & p_{1\cdot} \\
\cline{2-6}
X & \vdots & \vdots & & \vdots & \vdots \\
\cline{2-6}
 & x_I & p_{I1} & \dots & p_{IJ} & p_{p\cdot} \\
\cline{2-6}
 & & p_{\cdot 1} & \dots & p_{\cdot q} & 1 \\
\cline{2-6}
\end{array}$$

marginals $p_{i\cdot} = \mathbb{P}(X = x_i),\ p_{\cdot j} = \mathbb{P}(Y = y_j)$

conditional distributions: $\acute{p}_{j\ |i} = \mathbb{P}(Y = y_j|\ X = x_i)$

$$\begin{array}{c|c|c|c|c|c|}
\multicolumn{6}{c}{Y} \\
\cline{2-6}
 & & y_1 & \dots & y_J & \\
\cline{2-6}
 & x_1 & p_{1|1} & \dots & p_{J|1} & 1 \\
\cline{2-6}
X & \vdots & \vdots & & \vdots & \vdots \\
\cline{2-6}
 & x_I & p_{1|I} & \dots & p_{J|I} & 1 \\
\cline{2-6}
\end{array}$$

odds

case: binary response variable

$$\text{Odds}=\frac{\mathbb{P}\left(Y = y_1 \mid X = x_i\right)}{\mathbb{P}\left(Y = y_2 \mid X = x_i\right)} = \frac{\mathbb{P}\left(Y = y_1 \mid X = x_i\right)}{1 - \mathbb{P}\left(Y = y_1 \mid X = x_i\right)} = \frac{p_{1|i}}{1 - p_{1|i}}$$

define odds ratio (**OR**)

$$\textbf{OR}= \frac{\mathbb{P}\left(Y = y_1 \mid X = x_{i_1}\right)}{\mathbb{P}\left(Y = y_2 \mid X = x_{i_1}\right)} \Bigg/ \frac{\mathbb{P}\left(Y = y_1 \mid X = x_{i_2}\right)}{\mathbb{P}\left(Y = y_2 \mid X = x_{i_2}\right)}$$

Example: Chances of survival to dying for members of group i_1 to memebers of group i_2.

general case: consider the odds w.resp. to a **baseline** group

$$\text{Odds} = \frac{\mathbb{P}\left(Y = y_j \mid X = x_i\right)}{\mathbb{P}\left(Y = y_1 \mid X = x_i\right)} = \frac{p_{j|i}}{p_{1|i}}$$

resp

$$\text{Odds} = \frac{\mathbb{P}\left(Y = y_j \mid X = x_{i_1}\right)}{\mathbb{P}\left(Y = y_j \mid X = x_1\right)} = \frac{p_{j|i}}{p_{j|1}}$$

Example: Chances of survival for exposed to non-exposed

Data:

survival

smoke		no	yes
	yes	139	443
	no	230	502

replace counts by estimated cell probabilities

OR: survival chances

$$\Longrightarrow \begin{cases} \text{smoke:} & \text{odds(live/dead)}=\frac{443/(139+443)}{139/(139+443)} = \frac{443}{139} =: 3.19 \\ \text{no smoke:} & \text{odds(live/dead)}=\frac{502/(230+502)}{230(230+502)} = \frac{502}{230} =: 2.18 \end{cases}$$

$\Longrightarrow$ OR$=\frac{3.19}{2.18} = 1.46 \Longrightarrow$ higher odds of live/dead for smoker 46% higher than for non-smokers $\Longrightarrow$ smoking is good for your health !?!

RR **relative risk:**

$$\text{RR} = \frac{\mathbb{P}\left(Y = y_j \mid X = x_{i_1}\right)}{\mathbb{P}\left(Y = y_j \mid X = x_{i_2}\right)} = \frac{p_{j|i}}{p_{j|i_2}}$$

Example: risk of dying for expose and non-exposed

risk for smokers:	$\mathbb{P}(\text{dead/smoke})=\frac{139}{139+443}$	$= \frac{139}{582} = 0.24$
risk for non smokers:	$\mathbb{P}(\text{dead/no smoke})=\frac{230}{230+502}$	$= \frac{230}{732} = 0.31$

$\Longrightarrow$ RR(dead/nosmoke : dead/smoke)=0.31/0.24 $= 1.2917$

Notes:

1. this unplausible result is to be explained later ("**confounder**")
2. Note: the relative risk is most widely used for extreme probabilities (e.g. epidemiology):

"Since relative risk is a more intuitive measure of effectiveness, the distinction is important especially in cases of medium to high probabilities. If action A carries a risk of 99.9% and action B a risk of 99.0% then the relative risk is just over 1, while the odds associated with action A are almost 10 times higher than the odds with B."

Sampling distributions

multinomial sampling: the underlying assumptions are similar to those of **Bernoulli trials** for the binomial distribution

1. the observations per cell are independent
2. the cell probabilities are constant
3. the total number of trials is fixed in advance

Note: this is a very important assumption. not every cross tabulated data set can be modelled in this way.

Data (contingency table) $n_{ij} = \#\text{obs}(X = x_i, Y = y_j)$

Y

		y_1	$\cdots$	y_J	
	x_1	n_{11}	$\cdots$	n_{1J}	$n_{1\cdot}$
X	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_I	n_{I1}	$\cdots$	n_{1J}	$n_{I\cdot}$
		$n_{\cdot 1}$	$\cdots$	$n_{\cdot J}$	$n_{\cdot\cdot}$

joint: (X, Y) (fixed total sample size)

$\mathbb{P}(n_{11}, ..., n_{IJ}| \ n_{\cdot\cdot}) = \binom{n_{\cdot\cdot}}{n_{11},...,n_{IJ}} p_{11}^{n_{11}} \cdots p_{IJ}^{n_{IJ}}$ with $p_{ij} = \mathbb{P}(X = i \ Y = j)$ $(\Longrightarrow \sum_{i,j} p_{ij} = 1)$

marginals: Y

$\mathbb{P}(n_{\cdot 1}, ..., n_{\cdot J} \mid n_{\cdot\cdot}) = \binom{n_{\cdot\cdot}}{n_{\cdot 1},...,n_{\cdot J}} p_{\cdot 1}^{n_{\cdot 1}} \cdots p_{\cdot J}^{n_{\cdot J}}$ with $p_{\cdot j} = \mathbb{P}(Y = i)$ $(\Longrightarrow \sum_j p_{\cdot j} = 1)$

conditionals: $(Y \mid X)$

$\mathbb{P}(n_{i1}, ..., n_{iq}| \ n_{i\cdot}) = \binom{n_{i\cdot}}{n_{i1},...,n_{iJ}} p_{1/i}^{n_{i1}} \cdots p_{J/i}^{n_{iJ}}$ with $p_{j/i} = \mathbb{P}(Y = j| \ X = i)$ $(\Longrightarrow \sum_j p_{j/i} = 1)$

joint unconditional: (random sample size**)**

$\mathbb{P}(n_{11}, ..., n_{IJ}) = \prod_{i,j} \mathbb{P}(n_{ij})$ with $\mathbb{P}(n_{ij}) = \frac{\lambda_{ij}^{n_{ij}}}{n_{ij}!} e^{-\lambda_{ij}}$, $\lambda_{ij} = np_{ij}$

note: in the case of **fixed** sample size $\sum n_{ij} = n_{\cdot\cdot}$ the Poisson turns into a multinomial:

$$\mathbb{P}(n_{11}, ..., n_{IJ} \mid n_{\cdot\cdot}) = \text{Multinom}\,(n_{11}, ..., n_{IJ}, p_{11}, ..., p_{IJ})$$

with

$$p_{ij} = \frac{\lambda_{ij}}{\sum \lambda_{ij}}$$

(a proof is given in the appendix)

2.2 Maximum Likelihood Estimation

likelihood of a Binomial $Y \backsim$Bnom(p,n)

$$L(p \mid y) = \binom{n}{y} p^y(1-p)^{n-y} \backsim p^y(1-p)^{n-y}$$

skip the factor $\binom{n}{y}$ since it does not contain information on the unknown parameter ("ancillary")

ml estimator

$$l(p \mid y) = \log L(p \mid y) = y \log p + (n-y)\log(1-p)$$

MLE[1]

$$\widehat{p} = \arg\max_p l(p| \ x) = \frac{y}{n}$$

note: strictly, maximization has to take place under the restrictions $0 \leq p \leq 1$.

properties:

- unbiasedness: $\mathbb{E}\widehat{p} = p$
- variance: $\mathbb{V}\widehat{p} = p(1-p)/n$
- consistency: $\mathbb{P}(\ |\widehat{p}-p| > \epsilon) \to 0$ as $n \to \infty$ (law of large numbers)
- asymptotic distribution: (CLT)

$$\frac{\widehat{p}-p}{\sqrt{p(1-p)/\,n}} \to \mathcal{N}(0,1)$$

  as $n \to \infty$

these properties are also the properties of the mle (see corresponding section)

likelihood of a multinomial: $(X_1, ..., X_q) \backsim$Mnom$(p_1, ..., p_q, n)$

$$L(\mathbf{p} \mid \mathbf{x}) = \prod_i^q p_i^{x_i} \Longrightarrow l(\mathbf{p} \mid \mathbf{x}) = \sum_i^q x_i \log p_i$$

[1] By introducing the **reparametrization**

$$p = \frac{e^\beta}{1+e^\beta}$$

this maximization in β is without restriction. this will be used in the logit model. Note that if

$$\widehat{\beta} \text{ mle of } \beta \ \Longrightarrow \widehat{p} = \frac{\exp\widehat{\beta}}{1+\exp\widehat{\beta}} \text{ mle of } p = \frac{\exp\beta}{1+\exp\beta}$$

generalization: $p_i = \frac{\exp\beta_i}{1+\sum_i^{j-1}\exp\beta_i} \ \ i = 1, \ldots, j-1$

(the multinomial coefficients are uninformative).

MLE

$$\widehat{\mathbf{p}} = \arg\max_{\mathbf{p}} \sum_i^q x_i \log p_i \quad \text{with} \quad \left\{ \begin{array}{c} p_i \geq 0 \\ \sum_i^q p_i = 1 \end{array} \right.$$

where $\sum_i^q x_i = n$

strictly, the restricted **ml estimator** needs to be found**:**

Lagrange function:

$$\begin{array}{c} \mathcal{L} = \sum_i^q x_i \log p_i - \lambda(1 - \sum_i^q p_i) \\ \left. \begin{array}{r} \frac{\partial \mathcal{L}}{\partial p_j} = \frac{x_j}{p_j} - \lambda = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \sum_i^q p_i = 0 \end{array} \right\} \Longrightarrow \widehat{p}_j = \frac{x_j}{\lambda} = \frac{x_j}{n} \end{array}$$

Poisson Sampling:

Data

X	x_1	$\cdots$	x_I
n	n_1	$\cdots$	n_I

Distribution

	X		
N	x_1	...	x_I
	p_1	...	p_I

$n \geq 1$

Model $n_1, ..., n_p$ independent, $n_i \sim$Pois(λ_i)

likelihood

$$L(\lambda_1, ..., \lambda_p \mid n_1, ..., n_p) = \prod_i^p \frac{\lambda_i^{n_i}}{n_i!} \exp(-\lambda_i)$$

log likelihood ($n_1!, ..., n_p!$ uninformative)

$$l(\lambda_1, ..., \lambda_p \mid n_1, ..., n_p) = \sum_i^p (n_i \log \lambda_i - \lambda_i)$$
$$\Longrightarrow \widehat{\lambda}_i = n_i$$

since every cell probability is modelled individually this is called a **saturated model**

2.3 gof-tests for multinomial

MLE gof=**g**oodness-**o**f-**f**it: which model to choose? take the one with the (substantially) higher likelihood.

2.3.1 Likelihood-ratio tests:

mle $\widehat{\mathbf{p}}$

$$\widehat{\mathbf{p}} = \arg\sup_{\mathbf{p}\in\Theta} L(\,\mathbf{p}\mid n_1,...,n_k)$$

Null Hypothesis

$$H_0: \ \mathbf{p} \in\subset \Theta_0 \quad H_1: \ \mathbf{p} \in \Theta - \Theta_0$$

with $\Theta_0 \subset \Theta$ (**nested** hypothesis)

LR-test statistic

$$LR = \frac{\sup_{\mathbf{p}\in\Theta_0} L(\,\mathbf{p}\mid n_1,...,n_k)}{\sup_{\mathbf{p}\in\Theta} L(\,\mathbf{p}\mid n_1,...,n_k)}$$

rule: reject the null if this statistic too small in absolute value. Distribution:

Theorem: as $n \to \infty$ and there are no empty cells

$$-2\log LR \approx \chi^2(\mathrm{df}_1 - \mathrm{df}_0)$$

where

df_1	degrees of freedom model with $\mathbf{p} \in \Theta$	$\mathrm{df}_1 = \dim(\Theta)$
df_0	degrees of freedom model with $\mathbf{p} \in \Theta_0$	$\mathrm{df}_0 = \dim(\Theta_0)$

2.3.2 multinomial gof-tests

multinomial $\mathbf{N} = (N_1,...,N_q)' \sim$Mnom$(\mathbf{p}, n)$

$$\mathbb{P}(\,n_1,...,n_k\mid\mathbf{p}) = \binom{n}{n_1,\ldots,n_k} p_1^{n_1}\cdots p_k^{n_k}$$

point hypothesis: $\quad (N_1,...,N_q) \sim$Mnom$(n, p_1,...,p_q)$

$$\begin{aligned} &\mathrm{H}_0: \ p_i = p_{i0} \ , \ \ i = 1,...,q \\ &\mathrm{H}_1: \ \text{no constraints} \end{aligned}$$

degrees of freedom

df_1	$q-1$ free parameters: $p_1,...,p_{q-1}$
df_0	0 free parameters

Log-Lik

$$l(\mathbf{p}) = \sum_i n_i \log p_i$$

ML-est

$$\nabla l(\mathbf{p}) = 0 \Longrightarrow \widehat{p}_i = \frac{n_i}{n}$$

LR-test

$$-2\log LR = -2\left(l(\mathbf{p}_0) - l(\widehat{\mathbf{p}})\right)$$
$$LR = \frac{L(\mathbf{p}_0)}{L(\widehat{\mathbf{p}})}$$

LR-test statistic[2]**:**

$$-2\log LR = -2\sum_i n_i\left(\log p_{i0} - \log \widehat{p}_i\right) \approx \sum_i^q \frac{\left(n_i - n\ p_{i0}\right)^2}{n\widehat{p}_i}$$
$$= \sum_i^q \frac{\left(\text{observed cell count} - \text{expected cell count}\right)^2}{n\widehat{p}_i} \backsim \chi^2(q-1)$$

NB: There are other (asymptotically equivalent tests). Alternatively,

Pearson's χ^2

$$\chi^2 = \sum_{i=1}^q \frac{\left(n_i - n\ p_{i0}\right)^2}{np_{i0}} \sim \chi^2(q-1)$$

("observed minus expected under the null"). Both tests are asymptotically equivalent (see appendix)

Independence test for 2-dim contingency table multinomial

$$\mathbb{P}\left(n_{11}, ..., n_{IJ} \mid n_{..}\right) = \binom{n_{..}}{n_{11}, ..., n_{IJ}}\ p_{11}^{n_{11}} \cdots p_{IJ}^{n_{IJ}}$$

with $p_{ij} = \mathbb{P}(X = i,\ Y = j),\ i = 1, \ldots, I,\ j = 1, \ldots, J$

log likelihood

$$l\left(\mathbf{p} \mid n_1, ..., n_k\right) = \sum\nolimits_{i,j} n_{ij} \log p_{ij}$$

Independence hypothesis:

$$\mathrm{H}_0:\ p_{ij} = p_{i\cdot}p_{\cdot j}\ ,\ \ i = 1, ..., I,\ \ j = 1, ..., J$$

[2] see appendix

degrees of freedom:

$$\begin{array}{ll} \text{df}_1 & IJ-1 \text{ free parameters: } p_1, ..., p_{IJ-1} \\ \text{df}_0 & I-1+J-1 \text{ free parameters} \end{array}$$

$$\Longrightarrow \text{df}_1 - \text{df}_0 = IJ - 1 - (I - 1 + J - 1) = (I-1)(J-1)$$

LR-test: $\Theta_0 = \{p_{1\cdot}, p_{2\cdot}, \dots, p_{\cdot 1}, p_{\cdot 2}, \dots : p_{ij} = p_{i\cdot} p_{\cdot j}\}$

$$\begin{aligned} -2 \log \frac{\sup_{\mathbf{p} \in \Theta_0} L(\mathbf{p}|\ n_1, ..., n_k)}{\sup_{\mathbf{p} \in \Theta} L(\mathbf{p}|\ n_1, ..., n_k)} &= -2 \sum\nolimits_{i,j} n_{ij} \left(\log \widehat{p}_{i\cdot} \widehat{p}_{\cdot j} - \log \widehat{p}_{ij} \right) \\ &\sim \chi^2 (I-1)(J-1) \end{aligned}$$

Notes:

1. this test can be written in a simplified, more intuitive form (this is not relevant for the present purpose).
2. widely used alternatives to the LR-test for independence are a (version of) Pearson's (asymptotic) χ^2 goodness-of-fit test and Fisher's **exact** test (i.e., no asymptotic χ^2-approximation) based on the hypergeometric distribution

2.4 Higher dimensional tables

The above tests carry over to higher dimensions. A higher dimension may be crucial as in the following case.

Simpsons paradox

Kidney stone treatment: This is a real-life example from a medical study comparing the success rates of two treatments for kidney stones.

The first table shows the overall success rates and numbers of treatments for both treatments (where Treatment A includes all open procedures and Treatment B is percutaneous nephrolithotomy):

Treatment A	Treatment B
78% (273/350)	83% (289/350)

This seems to show treatment B is more effective. If we include data about kidney stone size, however, the same set of treatments reveals a different answer:

	Treatment A	Treatment B
Small Stones	Group 1 93% (81/87)	Group 2 87% (234/270)
Large Stones	Group 3 73% (192/263)	Group 4 69% (55/80)
Both	78% (273/350)	83% (289/350)

Reference for this and other highly instructive examples see *http://en.wikipedia.org/wiki/Simpson%27s_paradox*

Conclusions:

- Analysis: doctors give weaker treatment to benign cases and doctors give stronger treatment to malign cases
- Consequence: Variables: X treatment, Y outcome, Z severity of cases. Z *hidden variable* ("**confounder**"). ignoring the variable Z leads to wrong inference (in this case).
- confounding with observed or unobserved variables is avoided by assigning treatments to individuals at random (**randomization**)
- similar problems can arise if higher dimensional tables are **collapsed** ("marginalized") into a lower-dimensional (marginal) table

Confounding not a problem e.g. under:

2.4.1 Joint independence (from), resp marginal independence

X and Y are **jointly independent** from Z :

$$\mathbb{P}(X, Y \mid Z) = \mathbb{P}(X, Y)$$

for all values of X, Y, Z.

Notation:

$$(X, Y) \perp Z$$

Proposition: X, Y are jointly independent from Z implies X and Z are marginally independent (likewise, Y and Z are marginally independent)

Proof:

$$\begin{aligned}\mathbb{P}(X, Y \mid Z) = \mathbb{P}(X, Y) \implies \sum\nolimits_y \mathbb{P}(X, Y \mid Z) = \sum\nolimits_y \mathbb{P}(X, Y) \implies \mathbb{P}(X \mid Z) = \mathbb{P}(X) \\ \implies \mathbb{P}(X, Z) \ / \mathbb{P}(Z) = \mathbb{P}(X \mid Z) = \mathbb{P}(X) \\ \implies \mathbb{P}(X, Z) = \mathbb{P}(X)\ \mathbb{P}(Z) \implies X \perp Z\end{aligned}$$

$\boxed{\text{qed}}$

Notes on joint independence from Z :

- if X and Y are jointly independent of Z the table is certainly **collapsible** w.r. to Z for the study of the relation between X and Y (though conditional independence is a weaker condition, see later)

- Remember that X and Y are **marginally** independent (within the joint marginal dist of X, Y) if $\mathbb{P}(X,Y) = \mathbb{P}(X)\ \mathbb{P}(Y)$. But: X, Y jointly indep from $Z \not\Leftrightarrow X, Y$ marginally indep.
- equivalently to calling X, Y jointly independent from Z one could say that the joint distribution of X, Y, Z factors into the (joint) marginal distributions of X, Y and the marginal of Z. *This way it could be considered a more general form of marginal independence.*
- Joint independence of X, Y (from Z) is different from $\mathbb{P}(X,Y,Z) = \mathbb{P}(X)\,\mathbb{P}(Y)\,\mathbb{P}(Z)$ called independence, or joint independence, or total independence etc.

Testing for joint independence (from Z): multinomial

$$\mathbb{P}(n_{111}, ..., n_{IJK} \mid n_{...}) = \binom{n_{...}}{n_{111}, ..., n_{IJK}} p_{111}^{n_{111}} \cdots p_{IJK}^{n_{IJK}}$$

with $p_{ijk} = \mathbb{P}(X = i,\ Y = j,\ Z = k)$

Hypothesis of joint independence from Z:

$$\begin{aligned}
&\mathrm{H}_0: \ p_{ijk} = p_{ij\cdot}p_{\cdot\cdot k}\ ,\ \ i = 1, ..., I,\ \ j = 1, ..., J,\ k = 1, ..., K \\
&\mathrm{H}_1: \text{no restriction}
\end{aligned}$$

because:

$$\begin{aligned}
\mathbb{P}\,(X = i,\ Y = j,\ Z = k) \quad &= \mathbb{P}\,(\,X = i,\ Y = j \mid Z = k)\ \mathbb{P}\,(Z = k)) && \text{always} \\
&= \mathbb{P}(X = i,\ Y = j)\ \mathbb{P}(Z = k) && \text{under } \mathrm{H}_0
\end{aligned}$$

degrees of freedom:

$$\begin{aligned}
&\mathrm{df}_1 \quad IJK - 1 \text{ free parameters: } p_1, ..., p_{IJK-1} \\
&\mathrm{df}_0 \quad (IJ - 1) + (K - 1) \text{ free parameters}
\end{aligned}$$

$$\Longrightarrow \mathrm{df}_1 - \mathrm{df}_0 = IJK - 1 - (IJ - 1 + K - 1) = IJK - IJ - K + 1 = (IJ - 1)(K - 1)$$

LR-test:

$$-2\sum\nolimits_{i,j,k} n_{ijk}\,(\log \widehat{p}_{ij\cdot}\widehat{p}_{\cdot\cdot k} - \log \widehat{p}_{ijk}) \sim \chi^2(IJ - 1)(K - 1)$$

2.4.2 Conditional Independence

A higher dimensional table is also collapsible under the weaker relation of conditional independence

Definition: X and Y are conditionally independent Given the information on X provided by Z if there is no additional information (to that provided by Z) that Y can provide on X.

$$\mathbb{P}(X \mid Y, Z) = \mathbb{P}(X \mid Z)$$

for all values of X, Y, Z.

Notation:

$$(X \perp Y) \mid Z$$

Therefore, for the study of the dependence of the distribution of X on Z the table is collapsible w.r.to Y.

Proposition: An equivalent condition to conditional independence is given in the **symmetric form**:

$$\mathbb{P}(X, Y \mid Z) = \mathbb{P}(X \mid Z)\ \mathbb{P}(Y \mid Z)$$

Proof (only one side): under cond ind

$$\begin{aligned} \mathbb{P}(X, Y \mid Z) &= \mathbb{P}(X \mid Y, Z)\ \mathbb{P}(Y \mid Z) && \text{always} \\ &= \mathbb{P}(X \mid Z)\ \mathbb{P}(Y \mid Z) && \text{by hypothesis} \end{aligned}$$

$\boxed{\text{qed}}$

That "joint independence from" is stronger than "conditional independence on" is shown next:

Proposition:

$$(X, Z) \perp Y \implies (X \perp Y) \mid Z$$

Proof: remember $(X, Z) \perp Y \Leftrightarrow \mathbb{P}(X, Z \mid Y) = \mathbb{P}(X, Z)$

always (by rules of cond prob)

$$\mathbb{P}(X, Z \mid Y) = \mathbb{P}(X \mid Z, Y)\mathbb{P}(Z \mid Y)$$

remember that **joint independence from ... implies marginal independence from...**:

$$\mathbb{P}(X, Z \mid Y) \text{ indep of } Y \implies \mathbb{P}(Z \mid Y) \text{ indep of } Y$$

hence

$$\underbrace{\mathbb{P}(X, Z \mid Y)}_{\text{indep of } Y \text{ by joint indep from } Y} = \mathbb{P}(X \mid Z, Y) \underbrace{\mathbb{P}(Z \mid Y)}_{\text{indep of } Y \text{ bymarginal indep from } Y}$$

i.e.

$$
\begin{array}{ll}
\mathbb{P}(X, Z \mid Y) = \mathbb{P}(X, Z) & \text{, i.e. does not depend on (the values of) } Y \\
\mathbb{P}(Z \mid Y) = \mathbb{P}(Z) & \text{, i.e. does not depend on (the values of) } Y
\end{array}
$$

$\Longrightarrow$

$$\mathbb{P}(X \mid Z, Y)\text{, i.e. does not depend on (the values of) } Y$$

$\Longrightarrow$

$$\mathbb{P}(X \mid Z, Y) = \mathbb{P}(X \mid Z) \implies (X \perp Y) \mid Z$$

$\boxed{\text{qed}}$

Notes: There are 4 notions of (in)dependence among variables X, Y, Z.

1. (joint) independence among X, Y, Z
2. Marginal (pairwise) independence among X, Y or X, Z or Y, Z.
3. joint independence of X, Y from Z, of X, Z from Y, of Y, Z from X
4. Conditional independence of X, Y given Z, of X, Z given Y, of Y, Z given X

the wrong independence notion can lead to wrong results

Example: Causality for newborns

$$
\left.\begin{array}{l}
X: \#\text{storks} \\
Y: \#\text{babies} \\
Z: \#\text{families} \\
\text{sampling units: villages}
\end{array}\right\} \Longrightarrow \left\{\begin{array}{l}
X, Y \text{ marginally dependent} \\
X, Z \text{ marginally dependent} \\
Y, Z \text{ marginally dependent} \\
X, Y \text{ conditionally independent given } Z
\end{array}\right.
$$

Also: X, Y are *conditional independent from Z but not jointly independent* from Z : already the ranges of X, Y depend on Z (many families implies many storks and babies and few families implies fewer storks and babies). Hence condl independence from $Z \not\Rightarrow$ joint independence from Z.

On omission variable bias in regression analysis see Greene (Econometric Analysis): if income is omitted then the regression of demand for gasoline on prices for gasoline comes out positive. this is counterintuitive since higer prices should lead to a decrease in demand. the omitted variable in the model is income. high income leads to high demand in gasoline and this effect overrides the negative effect that high prices have on gasoline.

Collapsibility condition

Agresti (2007). For three-way tables, (X, Y, Z) marginal odds ratios (in the conditional $X \mid Y-$table) and conditional odds ratios (in the $(X \mid Z, Y)-$table) are identical if either Z and X are contionally independent given Y or if Z and Y are conditionally independent given X.i.e.,

$$\boxed{\text{collapsible in } Z \text{ if } (X \perp Z) \mid Y \quad \text{or} \quad (Y \perp Z) \mid X}$$

such that in neither case the variable Z has any influence on the conditional distribution of $\mathbb{P}(X \mid Z, Y)$ resp $\mathbb{P}(Z \mid Z, X)$ from with odds ratios are determined, i.e. are identical to the odds ratios of $\mathbb{P}(X \mid Y)$ resp $\mathbb{P}(Z \mid X)$

Notes:

1. This means only that marginal and conditional conditional odds are equal for the true distributions $\mathbb{P}(X = x_i \mid Y = y_j)$, they will not be equal for the actual data, i.e., the estimates are different.

2. This means only that marginal and conditional conditional odds are equal for the true distributions $\mathbb{P}(X = x_i \mid Y = y_j)$. It does not mean that we can draw inference on the joint distribution of $\mathbb{P}(X = x_i, Y = y_j)$ which requires knowledge of $\mathbb{P}(Y)$ in addition to $\mathbb{P}(X \,/ Y)$

3. $(Z \perp X) \mid Y \Leftrightarrow \mathbb{P}(X \mid Z, Y) = \mathbb{P}(X \mid Y)$ means that all conditional odds ratios in the conditional tables for given $Z = z$ equal the unconditional, i.e., marginal odds ratios in the collapsed tables

$$\mathrm{OR}(X \,/ Y) = \mathrm{OR}(X \,/ Y \mid Z = z)$$

   For example, conditional independence (OR=1|z) between X, Y given Z is equivalent to marginal independence (OR=1)in this case. similarly, $(Z \perp Y) \mid X \Leftrightarrow \mathbb{P}(Y \mid Z, X) = \mathbb{P}(Y \mid X)$ means that all conditional odds ratios in the conditional tables for given $Z = z$ equal the unconditional, i.e., marginal odds ratios in the collapsed tables

$$\mathrm{OR}(Y \,/ X) = \mathrm{OR}(Y \,/ X \mid Z = z)$$

   i.e., the odds ratios between X, Y can be computed from the collapsed (i.e., marginal tables).

4. as example consider the chance of survival (X) in different age groups (Y) for people with or without university degrees (Z). Suppose it has been decided that $(X \perp Z) | Y$. the odds can now safely be estimated from $X | Y$

5. There is a corresponding result for linear regression models (see Appendix)

$$y = \beta_1 x + \underbrace{\gamma z + \varepsilon}_{\varepsilon^*}$$
$$y = \beta_2 x + \varepsilon^*$$

   (misspecified model)

$$\beta_2 = \beta_1 + \gamma \frac{cov(x, z)}{var(x)}$$

$\Longrightarrow$

$$\beta_2 = \beta_1 \text{ if } cov(x, z) = 0 \text{ or } \gamma = 0$$

i.e., under multinormality, margininal independence of X, Z or cond independence of Z, Y given X will lead to the same regression slopes.

Testing a 3-dim contingency table for conditional independence The LR-test can be derived. This test is based on the second (symmetric) formulation of the cond indep condition. A popular alternative is the Mantel-Haenszel-test (see below) as another version of Pearson's g-o-f.

general case: **multinomial**

$$\mathbb{P}\left(n_{111}, ..., n_{IJK} \mid n_{...}\right) = \binom{n_{...}}{n_{111}, ..., n_{IJK}} p_{111}^{n_{111}} \cdots p_{IJK}^{n_{IJK}}$$

with $p_{ijk} = \mathbb{P}(X = i,\ Y = j,\ Z = k)$
introduce $p_{ij|k} = \mathbb{P}(X = i,\ Y = j|\ Z = k) = p_{ijk}\ / p_{\cdot\cdot k}$

Hypothesis of conditional independence:

$$\mathbb{P}(X = i,\ Y = j|\ Z = k) = \mathbb{P}(X = i \mid Z = k)\ \mathbb{P}(Y = j \mid Z = k)$$
$$\Leftrightarrow p_{ijk} = p_{ij|k}\ p_{\cdot\cdot k} = p_{i\cdot|k}\ p_{\cdot j|k}\ p_{\cdot\cdot k}$$

test by maximum likelihood. **likelihood reduced model:**

$$p_{ijk} = p_{i\cdot|k}\ p_{\cdot j|k}\ p_{\cdot\cdot k}$$

$$L_0(\mathbf{p} \mid \mathbf{n}) = \prod_{i,j,k} (p_{i\cdot/k} p_{\cdot j/k} p_{\cdot\cdot k})^{n_{ijk}}$$

likelihood full model: with

$$p_{ijk} = p_{ij/k} p_{\cdot\cdot k}$$

$$L_1(\mathbf{p} \mid \mathbf{n}) = \prod_{i,j,k} p_{ijk}^{n_{ijk}} = \prod_{i,j,k} (p_{ij/k} p_{\cdot\cdot k})^{n_{ijk}}$$

$\mathrm{H}_0:\ p_{ij/k} = p_{i\cdot/k} p_{\cdot j/k}$
$\Longrightarrow$ **LR-test:**

$$\begin{aligned} -2\log LR &= -2\sum_{i,j,k} n_{ijk}\left(\log(\widehat{p}_{i\cdot/k}\widehat{p}_{\cdot j/k}\widehat{p}_{\cdot\cdot k}) - \log(\widehat{p}_{ij/k}\widehat{p}_{\cdot\cdot k})\right) \\ &= -2\sum_{i,j,k} n_{ijk}\left(\log \widehat{p}_{i\cdot/k}\widehat{p}_{\cdot j/k} - \log \widehat{p}_{ij/k}\right) \end{aligned}$$

where $\widehat{p}_{i\cdot/k} = \dfrac{n_{i\cdot k}}{n_{\cdot\cdot k}},\ \widehat{p}_{\cdot j/k} = \dfrac{n_{\cdot jk}}{n_{\cdot\cdot k}},\ \widehat{p}_{ij/k} = \dfrac{n_{ijk}}{n_{\cdot\cdot k}}$

computation of df:

fullmod: p_{ijk} $\Longrightarrow$ df(full)$= IJK-1$

redmod: $p_{ijk} = p_{i\cdot|k}\, p_{\cdot j|k}\, p_{\cdot\cdot k}$ $\Longrightarrow$ $\begin{aligned}\text{df(red)} &= (I-1)\,K + (J-1)\,K + (K-1)\\ &= IK + JK - K - 1\end{aligned}$

$\Longrightarrow$ df(full)-df(red) $= IJK - 1 - (IK + JK - K) + 1 = (I-1)\,(J-1)\,K$

Another way to understand the df: There are K decoupled $I \times J$ tables now. On each $I \times J$ table perform independence test and sum over k. This results in a test which under H_0 is distributed $\chi^2\left((I-1)\,(J-1)\,K\right)$

Example:

`Agresti (2002), p. 287f and p. 297. Job Satisfaction example.`

`V_D: very dissatisfied, L_D: low satisfied, M_S: medium satisfied, V_S: very satisfied`

Job Satisfaction		V_D	L_S	M_S	V_S
Gender	Income				
Female	<5000	1	3	11	2
	5000-15000	2	3	17	3
	15000-25000	0	1	8	5
	>25000	0	2	4	2
Male	<5000	1	1	2	1
	5000-15000	0	3	5	1
	15000-25000	0	0	7	3
	>25000	0	1	9	6

testing for independence of satisfaction and income

$$\mathrm{H}_0 : \mathbb{P}(\texttt{Job Satisfaction} \mid \texttt{Income}) = \mathbb{P}(\texttt{Job Satisfaction})$$

`Pearson's Chi-squared test`

`X-squared = 11.5243, df = 9`$(= (I-1)\,(J-1))$`, p-value = 0.2415`

Job Satisfaction				
Income	V_D	L_S	M_S	V_S
<5000	2	4	13	3
5000-15000	2	6	22	4
15000-25000	0	1	15	8
25000	0	3	13	8

testing for conditional independence of satisfaction from gender after income

$$\mathrm{H}_0 : \mathbb{P}(\texttt{Job Satisfaction}|\ \texttt{Gender}, \texttt{Income}) = \mathbb{P}(\texttt{Job Satisfaction}|\ \texttt{Income})$$

`Cochran-Mantel-Haenszel M^2 = 10.2001, df = 9, p-value = 0.3345`

NB

1. the popular CMH-test is not to be confused with the (asymptotically optimal) LR-test
2. it is always risky to apply asymptotic mle methods to table with 0 observations in some of its cells.

2.4.3 Fisher Test

remember: pop: $N = K + (N - K)$ sample: $n = x + (n - x)$

$$X \sim \text{HypG}(N, K, n) \implies \mathbb{P}(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

apply to 2x2 contingency table:

consider a 2×2 table of binary variables X, Y. consider the case of **fixed marginals** in both directions

Y : color blindness

X : gender

	colb	notcolb	
male	n_{11}	n_{12}	$n_{1\cdot}$
fem	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

hypothesis

$$\text{H}_0 : X, Y \text{ independent}$$

for fixed marginals $n_{1\cdot}, n_{2\cdot}, n_{\cdot 1}, n_{\cdot 2}$ the only free variable remaining is n_{11}

test statistic

$$n_{11}$$

Distribution under H_0

Compute the prob of obtaining n_{11} colb males if X, Y are independent

Idea: suppose, gender & colb independent labels on the individuals.

1. we have $n_{\cdot 1}$ color blind. These labels can be handed out among $n_{\cdot\cdot}$ persons. There are $\binom{n_{\cdot\cdot}}{n_{\cdot 1}}$ such ways.
2. among these $n_{\cdot 1}$ colb persons we observe n_{11} males and n_{21} females. Under H_0 these can be drawn equally likely from $n_{1\cdot}$ males and $n_{2\cdot}$ females. There are $\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}$ such ways.
3. hence

$$\mathbb{P}(n_{11} \text{ colb males} \mid n_{1\cdot} \text{ males}, n_{2\cdot} \text{ fem}, n_{\cdot 1} \text{ colb}, n_{\cdot 2} \text{ nocolb}) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}$$

This is a hypergeometric distribution.

Note: this test has the advantage of the Fisher test over the LR-test is that **it does not rely on the asymptotic** χ^2-**distribution**. Therefore, it is some times called the **exact** Fisher test

Normal approximation to HypG:

$$X \sim \text{HypG}\,(N, m, n)$$
$$\mathbb{P}\,(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$
$$\mathbb{E}X = n\frac{m}{N}$$
$$\mathbb{V}X = n\frac{m\,(N-n)\,(N-m)}{N^2\,(N-1)}$$

asymptotics:

$$\frac{X - \mathbb{E}X}{\sqrt{\mathbb{V}X}} \to \mathcal{N}\,(0, 1)$$

http://en.wikipedia.org/wiki/Hypergeometric_distribution:

application to 2x2 tables:

n_{11}	n_{12}	$n_{1\cdot}$
n_{21}	n_{22}	$n_{2\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

$\Longrightarrow$

$$P\,(X = n_{11}) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}$$

$$X \sim \text{HypG}\,(n_{\cdot\cdot}\,, n_{1\cdot}\,, n_{\cdot 1})$$
$$\mathbb{E}X = \frac{n_{\cdot 1} n_{1\cdot}}{n_{\cdot\cdot}}$$
$$\mathbb{V}X = \frac{n_{\cdot 1} n_{1\cdot} n_{\cdot 2} n_{2\cdot}}{n_{\cdot\cdot}^2\,(n_{\cdot\cdot}^2 - 1)}$$

asymptotics

$$\frac{n_{11} - \mathbb{E}n_{11}}{\sqrt{\mathbb{V}n_{11}}} \to \mathcal{N}\,(0, 1)$$

as $n_{\cdot\cdot} \to \infty$

generalization to binomial-multinomial leads to generalized hypergeometric distribution.

$Y:$

$X:$

	y_1	y_2	
x_1	n_{11}	n_{12}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_K	n_{K1}	n_{K2}	$n_{K\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

$$\mathbb{P}\left(n_{11},\ldots,n_{K1} \mid n_{\cdot 1}, n_{\cdot 2},\; n_{1\cdot}, n_{2\cdot}, \cdots, n_{K\cdot}\right) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}\cdots\binom{n_{K\cdot}}{n_{K1}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}$$

generalization to multinomial-multinomial possible

generalization to three variables X, Y, Z leads to Mantel-Haenszel test.

Mantel-Haenszel-Test

A widely popular alternative is the *Mantel-Haenszel-Test* which is asymptotically equivalent to the LR-test:

Consider (X, Y, Z) as a $2{\times}2 \times K$-table

$$\begin{array}{ccccc} & & Z & & \\ & Y & & & Y \\ X & \begin{array}{|c|c|}\hline n_{111} & n_{121} \\ \hline n_{211} & n_{221} \\ \hline\end{array} & \cdots & X & \begin{array}{|c|c|}\hline n_{112} & n_{122} \\ \hline n_{212} & n_{222} \\ \hline\end{array}\end{array}$$

test statistic:

$$CMH = \frac{\left(\sum_k \left(n_{11k} - \mu_{11k}\right)\right)^2}{\sum_k \mathbb{V}\left(n_{11k}\right)}$$

$$\mu_{11k} = \frac{n_{1\cdot k} n_{\cdot 1k}}{n_{\cdot\cdot k}}$$

$$\mathbb{V}\left(n_{11k}\right) = \frac{n_{1\cdot k} n_{2\cdot k} n_{\cdot 1k} n_{\cdot 2k}}{n_{\cdot\cdot k}^2 \left(n_{\cdot\cdot k} - 1\right)}$$

Agresti (2002). Asymptotically:

$$CMH \to \chi^2(1) \text{ for } n_{\cdots} \to \infty$$

Example: `mantelhaen.test {stats}`
smoking×survival-status×age (femsmoke data, Faraway)

2.5 Appendix:

2.5.1 Approx test statistic for point hypothesis

by certain approximations we cast the LR into a more intuitive form
$(N_1, ..., N_q) \sim \text{Mnom}(n, p_1, ..., p_q)$

$$\begin{aligned} &\mathrm{H}_0: \; p_i = p_{i0} \;, \quad i = 1, ..., q \\ &\mathrm{H}_1: \; \text{"no constraints"} \end{aligned}$$

computation of LR-test statistic:

$$-2\log LR = -2\sum_i n_i \left(\log p_{i0} - \log \widehat{p}_i\right)$$

hence

$$\log LR = \sum_i n_i \left(\log p_{i0} - \log \widehat{p}_i\right) = \sum_i n \frac{n_i}{n} \left(\log p_{i0} - \log \widehat{p}_i\right) = \sum_i n\widehat{p}_i \left(\log p_{i0} - \log \widehat{p}_i\right)$$

Taylor developing around 1 : $\boxed{\log x = (x-1) - \frac{1}{2}(x-1)^2 \pm \cdots}$

$$\log p_{i0} - \log \widehat{p}_i = \log\left(\frac{p_{i0}}{\widehat{p}_i}\right) \approx \left(\frac{p_{i0}}{\widehat{p}_i} - 1\right) - \frac{1}{2}\left(\frac{p_{i0}}{\widehat{p}_i} - 1\right)^2$$

get

$$\begin{aligned} \log LR &\approx \sum_i n\widehat{p}_i \left[\frac{p_{i0} - \widehat{p}_i}{\widehat{p}_i} - \frac{1}{2}\left(\frac{\widehat{p}_i - p_{i0}}{\widehat{p}_i}\right)^2\right] \\ &= n\sum_i \left[\widehat{p}_i - p_{i0} - \frac{1}{2}\frac{\left(\widehat{p}_i - p_{i0}\right)^2}{\widehat{p}_i}\right] \end{aligned}$$

noting that $\boxed{\sum_i \left(\widehat{p}_i - p_{i0}\right) = \sum_i \widehat{p}_i - \sum_i p_{i0} = 1 - 1 = 0}$

$$\begin{aligned} -2\log LR &= -2\sum_i n_i \left(\log p_{i0} - \log \widehat{p}_i\right) \approx n\sum_i^q \frac{\left(\widehat{p}_i - p_{i0}\right)^2}{\widehat{p}_i} \\ &= \sum_i^q \frac{n^2\left(\widehat{p}_i - p_{i0}\right)^2}{n\widehat{p}_i} = \sum_i^q \frac{\left(n_i - np_{i0}\right)^2}{n\widehat{p}_i} \approx \chi^2(q-1) \end{aligned}$$

2.5.2 Approx test statistic for independence

by certain approximations we cast the LR into a more intuitive form

$$-2\log LR = \sum\nolimits_{i,j} n_{ij} \left(\log \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \log \widehat{p}_{ij}\right)$$

Taylor approx

$$\log \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \log \widehat{p}_{ij} = \log\left(\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}/\,\widehat{p}_{ij}\right) \approx \left(\frac{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}}{\widehat{p}_{ij}} - 1\right) - \frac{1}{2}\left(\frac{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}}{\widehat{p}_{ij}} - 1\right)^2$$
$$= \frac{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \widehat{p}_{ij}}{\widehat{p}_{ij}} - \frac{1}{2}\left(\frac{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \widehat{p}_{ij}}{\widehat{p}_{ij}}\right)^2$$

hence

$$\sum\nolimits_{i,j} n_{ij}\left(\log \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \log \widehat{p}_{ij}\right) = \sum\nolimits_{i,j} n\widehat{p}_{ij}\left(\frac{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \widehat{p}_{ij}}{\widehat{p}_{ij}} - \frac{1}{2}\left(\frac{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \widehat{p}_{ij}}{\widehat{p}_{ij}}\right)^2\right)$$

but

$$\sum\nolimits_{i,j}\left(\widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \widehat{p}_{ij}\right) = \sum\nolimits_{i,j} \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \sum\nolimits_{i,j} \widehat{p}_{ij} = 1 - 1 = 0$$

hence

$$-2\log LR = \sum\nolimits_{i,j} n\widehat{p}_{ij}\left(\frac{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \widehat{p}_{ij}}{\widehat{p}_{ij}}\right)^2 = \sum\nolimits_{i,j} n\frac{\left(\widehat{p}_{i\cdot}\widehat{p}_{\cdot j} - \widehat{p}_{ij}\right)^2}{\widehat{p}_{ij}} = \sum\nolimits_{i,j} \frac{\left(n_{ij} - n\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}\right)^2}{n_{ij}}$$

2.5.3 Marginals and conditionals of multinomials

we show that the marginals of multinomials are **multinomial:**

$$\mathbf{N} \sim \mathrm{Mnom}_q(\mathbf{p}, n) \Longleftrightarrow \begin{cases} P(n_1) = \binom{n}{n_1}\binom{n-n_1}{n-n_1} p_1^{n_1} p_2^{n_2} & \text{for } q = 2 \\ P(n_1, n_2) = \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_1-n_2-n_3} p_1^{n_1} p_2^{n_2} p_3^{n_3} & \text{for } q = 3 \end{cases}$$

marginals [3] the marginals of multinomials are multinomial
first: $P(n_1) = \sum_{n_2} P(n_1, n_2)$

set $n_2^* = n_2{+}n_3,\ p_2^* = p_2{+}p_3$ then $P(n_1) = \binom{n}{n_1}\binom{n-n_1}{n-n_1} p_1^{n_1} p_2^{*n_2^*}$ $\sim\mathrm{Mnom}(p_1, p_2 + p_3, n)$

similar: $P(n_1) = \sum_{n_2}\sum_{n_3} P(n_1, n_2, n_3)$

$P(n_1, n_2{+}n_3{+}n_4) = \binom{n}{n_1}\binom{n-n_1}{n_2+n_3+n_4} p_1^{n_1}\,(p_2 + p_3 + p_4)^{n_2+n_3+n_4}$ $\sim\mathrm{Mnom}(p_1, p_2 + p_3 + p_4, n)$

$P(n_1, n_3) = \sum_{n_2} P(n_1, n_2, n_3)$

$P(n_1, n_3, n_2{+}n_4) = \binom{n}{n_1}\binom{n-n_1}{n_3}\binom{n-n_1-n_3}{n_2+n_4} p_1^{n_1} p_3^{n_3}\,(p_2 + p_4)^{n_2+n_4}$ $\sim\mathrm{Mnom}(p_1, p_3, p_2 + p_4, n)$

[3] marginal:
$\sum_{n_2} P\left(n_1 \in box_1, n_2 \in box_2, remainder \notin box_1 \cup box_2 |\ total = n\right) =$
$P\left(n_1 \in box_1, any\ n_2 \in box_2, remainder \notin box_1 \cup box_2 |\ total = n\right) =$
$P\left(n_1 \in box_1, n - n_1 \notin box_1 |\ total = n\right)$

conditionals: the conditionals of multinomials are multinomial

$$
\begin{aligned}
&(N_1, N_2, N_3) \sim \text{Mnom}(p_1, p_2, p_3, n) \Longrightarrow \\
P(n_2, n_3 |\ n_1) &= \frac{P(n_1, n_2, n_3)}{P(n_1, n_2 + n_3)} = \frac{\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3} p_1^{n_1} p_2^{n_2} p_3^{n_3}}{\binom{n}{n_1}\binom{n-n_1}{n_2+n_3} p_1^{n_1}\ (p_2+p_3)^{n_2+n_3}} \\
&= \frac{\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3} p_2^{n_2} p_3^{n_3}}{\binom{n-n_1}{n_2+n_3}\ (p_2+p_3)^{n_2+n_3}} \\
&= \binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\left(\frac{p_2}{p_2+p_3}\right)^{n_2}\left(\frac{p_3}{p_2+p_3}\right)^{n_3} \\
&\Longrightarrow \text{Mnom}\left(\frac{p_2}{p_2+p_3}, \frac{p_3}{p_2+p_3}, n - n_1\right)
\end{aligned}
$$

2.5.4 Relation between Multinomial and Poisson models

we show that a multivariate poisson conditional on the total number of counts is multinomial

multinomial model: $\mathbf{N} = (N_1, ..., N_q)' \sim$Mnom$(\mathbf{p}, n)$

$$\mathbb{P}(n_1, ..., n_q \mid \mathbf{p}) = \frac{n!}{n_1! \cdots n_q!}\ p_1^{n_1} \cdots p_q^{n_q}$$

$N_1, ..., N_q$ independent, $N_i \sim$Pois(λ_i)

$$\mathbb{P}(n_1, ..., n_q \mid \lambda) = \prod_i^q \frac{\lambda_i^{n_i}}{n_i!} \exp(-\lambda_i)$$

We show that independent Poissons factor into the product of an aggregated Poisson and a Multinomial

$$
\begin{aligned}
\text{Pois}(n_1, ..., n_q |\ \lambda) &= \text{Mnom}(n_1, ..., n_q |\ \mathbf{p})\ \text{Pois}(n. |\ \lambda.) \\
n. &= \sum\nolimits_i^q n_i \\
\lambda. &= \sum\nolimits_i^q \lambda_i \\
p_j &= \frac{\lambda_j}{\sum_i \lambda_i}
\end{aligned}
$$

Consider Poisson probabilities conditional on total outcome: $n.$

$$
\begin{aligned}
\prod_i^q \left\{ \frac{\lambda_i^{n_i}}{n_i!} \exp(-\lambda_i) \right\} &= \prod_i^q \frac{\lambda_i^{n_i}}{n_i!} \prod_i^q \exp(-\lambda_i) = \exp\left(-\sum\nolimits_i \lambda_i\right) \prod_i^q \frac{\lambda_i^{n_i}}{n_i!} \\
&= \exp\left(-\lambda.\right) \frac{1}{n.!} \frac{n.!}{n_1! \cdots n_q!} \prod_i^q \lambda_i^{n_i} \quad \text{bring in } n. \\
&= \frac{\lambda.^{n.}}{n.!} \exp\left(-\lambda.\right) \frac{n.!}{n_1! \cdots n_q!} \frac{1}{\lambda^{n.}} \prod_i^q \lambda_i^{n_i} \quad \text{bring in} \quad \lambda.^{n.} \\
&= \frac{\lambda.^{n.}}{n.!} \exp\left(-\lambda.\right) \frac{n.!}{n_1! \cdots n_q!} \left(\frac{\lambda_1}{\lambda.}\right)^{n_1} \cdots \left(\frac{\lambda_q}{\lambda.}\right)^{n_q} \\
&= \frac{\lambda.}{n.!} \exp\left(-\lambda.\right) \frac{n.!}{n_1! \cdots n_q!} p_1^{n_1} \cdots p_q^{n_q}
\end{aligned}
$$

This implies that the product of independent Poissons counts contitional on the total number of counts is Multinomial

$$
\mathbb{P}(n_1, ..., n_q | \ \lambda, n.) = \frac{\mathbb{P}(n_1, ..., n_q, n. | \ \lambda)}{\mathbb{P}(n. | \ \lambda)}
$$

but

$$
\mathbb{P}(n_1, ..., n_q, n. | \ \lambda) = \mathbb{P}(n_1, ..., n_q | \ \lambda)
$$

since $n.$ is redundant (contained in $n_1, ..., n_q$)

$$
\mathbb{P}(n. | \ \lambda) = \mathbb{P}(n. | \ \lambda.)
$$

since the sum of indep Poisson is Poisson

$$
\begin{aligned}
&\mathrm{Pois}(n_1, ..., n_q | \ \lambda) = \mathrm{Mnom}(n_1, ..., n_q | \ \mathbf{p}) \ \mathrm{Pois}(n. | \ \lambda.) \implies \\
&\frac{\mathrm{Pois}(n_1, ..., n_q | \ \lambda)}{\mathrm{Pois}(n. | \ \lambda.)} = \mathrm{Mnom}(n_1, ..., n_q | \ \mathbf{p}) = \mathbb{P}(n_1, ..., n_q | \ \lambda, n.)
\end{aligned}
$$

This result has a consequence for testing the **Independence hypothesis:** $X, Y, \ \mathbb{P}(X = x_i, Y = y_j) = p_{ij}$

$$
\mathrm{H}_0 : \ p_{ij} = p_{i\cdot} p_{\cdot j} \ , \ \ i = 1, ..., I, \ \ j = 1, ..., J
$$

to compute the LR-test we observe

$$
\begin{aligned}
l(\lambda | \ n_1, ..., n_q) &= \log \mathrm{Pois}(n_1, ..., n_q | \ \lambda) = \\
&= \log \mathrm{Mnom}(n_1, ..., n_q | \ \mathbf{p}) + \log \mathrm{Pois}(n | \ \lambda)
\end{aligned}
$$

the LR-test is the same for Poisson and for Multinomial observations because *the second factor of the likelihood function does not contribute information to testing the hypothesis of independence.*

This implies that the hypothesis of conditional independence can be tested no matter if the data are generated by the multinomial or under the Poisson model.

This argument holds for higher dimensional tables and many hypotheses.

2.5.5 Simpson's paradox in linear models: sign reversals (missing variable bias)

True model (suppose $\mathbb{E}x = \mathbb{E}z = 0$)

$$y = \beta x + \gamma z + \varepsilon$$

Mispecified model

$$y = \beta x + \varepsilon^*$$

note: In this case the exogenuity condition is violated:

$$\left.\begin{array}{r} y = \beta x + \varepsilon^* \\ \varepsilon^* = \gamma z + \varepsilon \end{array}\right\} \Longrightarrow cov(x, \varepsilon^*) = cov(x, \gamma z + \varepsilon) = \gamma cov(x, z) + cov(x, \varepsilon) = \gamma cov(x, z)$$

OLS estimator:

$$\begin{aligned} \widehat{\beta} &= \frac{\widehat{cov}(x, y)}{\widehat{var}(x)} \to \frac{cov(x, y)}{var(x)} \\ &= \frac{cov(x, \beta x + \gamma z + \varepsilon)}{var(x)} \\ &= \beta \frac{cov(x, x)}{var(x)} + \frac{cov(x, \gamma z + \varepsilon)}{var(x)} \\ &= \beta + \gamma \frac{cov(x, z)}{var(x)} + \frac{cov(x, \varepsilon)}{var(x)} \\ &= \beta + \gamma \frac{cov(x, z)}{var(x)} \neq \beta \end{aligned}$$

a sign change can be implied if

$$|\beta| < \left| \gamma \frac{cov(x, z)}{var(x)} \right|$$

and of opposite sign $\widehat{\beta}$ will be of wrong sign (at least for large samples).

The hidden variable bias occurrs if Z is corelated with X and Z is corelated with Y

Note: no biased by an omitted variable Z if $cov(x, z) = 0$ (marginal independence) or $\gamma = 0$ (conditional independence $(Z \perp Y) \mid X$)

2.5.6 Additional literature:

1. *Categorical Data Analysis* by Alan Agresti Wiley-Interscience; 2 edition (July 22, 2002)

2. *Statistical Methods for Categorical Data Analysis* by Daniel A. Powers and Yu Xie Academic Press; 1 edition (November 26, 1999)

3. *Discrete Multivariate Analysis:* Theory and Practice von Stephen Fienberg (Autor), Paul W. Holland (Autor), Yvonne M. Bishop (Autor) Springer, Berlin; Auflage: 1 (25. Juli 2007)

4. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* Julian Faraway (Autor) Publisher: Chapman and Hall/CRC; 1 edition (December 20, 2005)