目录

——分类问题的检验——

1. **分类数据的 Pearson χ^2 检验（分类问题）**

H₀：类 A_i 占比 p_i

H₁：反之

在 H₀ 成立条件下，检验统计量：

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\overset{\text{L}}{\sim}\chi^2(k-1)$$

$$\text{差平方的加权}=\frac{\text{差的平方}}{\text{期望频数}}$$

其中 k 为状态（类）的数目，n_i 为观察到类 A_i 的频数，n 为总频数.

拒绝域：$\chi^2>\chi_\alpha^2(k-1)$ 或者 p-value $\leqslant\alpha$，p-value $=P(\chi^2(k-1)\geqslant\chi^2)$

2. **分类数据的似然比检验（分类问题）**

H₀：类 A_i 占比 p_i

H₁：反之

在 H₀ 成立条件下，检验统计量：

$$\text{似然比 }\Lambda=\frac{\frac{n!}{n_1!\dots n_k!}p_1^{n_1}\dots p_k^{n_k}}{\sup\limits_{p_1,\dots,p_k}\frac{n!}{n_1!\dots n_k!}p_1'^{n_1}\dots p_k'^{n_k}}=\prod_{i=1}^{k}\left(\frac{np_i}{n_i}\right)^{n_i}$$

$$-2\log\Lambda=-2\sum_{i=1}^{k}n_i\log\left(\frac{p_i}{\frac{n_i}{n}}\right)\overset{\text{L}}{\sim}\chi^2(k-1)$$

其中 p_i' 为似然函数中类 A_i 占比的变量，是未知的，在 sup 意义下等于其 MLE，即 $\left(\frac{n_i}{n}\right)^{n_i}$.

拒绝域：$-2\log\Lambda>\chi_\alpha^2(k-1)$

* 其实这两种方法的不同，实质是选用不同的方式衡量实际频数与期望频数的偏差.

3. **带参数的分类数据的 χ^2 检验**

H₀：类 A_i 占比 p_i

H₁：反之

在 H₀ 成立条件下，检验统计量：

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\overset{\text{L}}{\sim}\chi^2(k-m-1)$$

$$\text{差平方的加权}=\frac{\text{差的平方}}{\text{期望频数}}$$

其中 m 为参数个数.

拒绝域：$\chi^2>\chi_\alpha^2(k-1)$

4. **带参数的分类数据的似然比检验**

H₀：类 A_i 占比 p_i

H_1：反之

在 H_0 成立条件下，检验统计量：

$$-2\log\Lambda = -2\sum_{i=1}^{k} n_i \log\left(\frac{p_i}{\frac{n_i}{n}}\right) \overset{L}{\sim} \chi^2(k-m-1)$$

拒绝域：$-2\log\Lambda > \chi_\alpha^2(k-1)$

——四格表——

1. **单侧给定的四格表独立性的 U 检验与 χ^2 检验** （两个相互独立的随机变量）

下证明四格表中独立与不相关等价：

proof: 对属性 A、B 赋值以便计算相关系数

$$X=\begin{cases}a_1 & \text{若观察值}\in A\\ a_2 & \text{若观察值}\notin A\end{cases}\qquad Y=\begin{cases}b_1 & \text{若观察值}\in B\\ b_2 & \text{若观察值}\notin B\end{cases}$$

$\mathbb{E}(X)=a_1p_{1+}+a_2p_{2+}$，$\mathbb{E}(Y)=b_1p_{+1}+b_2p_{+2}$

$\mathbb{E}(XY)=a_1b_1p_{11}+a_1b_2p_{12}+a_2b_1p_{21}+a_2b_2p_{22}$

$\Rightarrow Cov(X,Y)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)=(a_1-a_2)(b_1-b_2)(p_{11}-p_{1+}p_{+1})$

证毕.

此外易计算 $Var(X)=(a_1-a_2)^2p_{1+}p_{2+}$、$Var(Y)=(b_1-b_2)^2p_{+1}p_{+2}$，进而相关系数 $r=\frac{p_{11}-p_{1+}p_{+1}}{\sqrt{p_{1+}p_{2+}p_{+1}p_{+2}}}$.

记 $p_1=P(B\mid A)=\frac{p_{11}}{p_{1+}}$，$p_2=P(B\mid A^c)=\frac{p_{21}}{p_{2+}}$,

H_0：$p_1=p_2$（属性 A、B 相互独立）$\Leftrightarrow$ $p_{11}=p_{1+}p_{+1}$

H_1：① $p_1\neq p_2$（属性 A、B 不相互独立）

② $p_1>p_2$（属性 A 中有属性 B 的比例高）

③ $p_1<p_2$（属性 A 中有属性 B 的比例低）

①属无方向检验，②、③属有方向检验.

在 H_0 成立条件下，检验统计量：

$$U=\frac{\sqrt{n}(n_{11}n_{22}-n_{12}n_{21})}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}\overset{L}{\sim}N(0,1)$$

$$\chi^2=U^2\overset{L}{\sim}\chi^2(1)$$

	原假设 H_0	备择假设 H_1	水平α拒绝域	p-value
有方向检验	$p_1=p_2$ 属性 A、B 相互独立	$p_1>p_2$ 属性 A 中有属性 B 的比例高	$U\geqslant U_\alpha$	$P(N(0,1)\geqslant U)=\Phi(-U)\leqslant\alpha$
		$p_1<p_2$ 属性 A 中有属性 B 的比例低	$U\leqslant U_{1-\alpha}$	$P(N(0,1)\leqslant U)=\Phi(U)\leqslant\alpha$
无方向检验 （独立性检验）		$p_1\neq p_2$ 属性 A、B 不相互独立	$\chi^2\geqslant\chi^2_\alpha(1)$	$P(\chi^2(1)\geqslant\chi^2)\leqslant\alpha$

2. **连续性修正**

Yates 修正：

原假设 H_0	备择假设 H_1	修正检验统计量
$p_1=p_2$	$p_1>p_2$	$U=\frac{\sqrt{n}\left(n_{11}n_{22}-n_{12}n_{21}-\frac{n}{2}\right)}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}$
	$p_1<p_2$	$U=\frac{\sqrt{n}\left(n_{11}n_{22}-n_{12}n_{21}+\frac{n}{2}\right)}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}$
	$p_1\neq p_2$	$U=\frac{n\left(\lvert n_{11}n_{22}-n_{12}n_{21}\rvert-\frac{n}{2}\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$

3. **单侧给定的四格表独立性的似然比检验** （两个相互独立的随机变量）

有方向的似然比检验较繁琐，仅讨论无方向的似然比检验.

H_0：$p_1 = p_2$（属性 A、B 相互独立）$\Leftrightarrow$ $p_{11} = p_{1+}p_{+1}$

H_1：$p_1 \neq p_2$（属性 A、B 不相互独立）$\Leftrightarrow$ $p_{11} \neq p_{1+}p_{+1}$

在 H_0 成立条件下，检验统计量：

$$\text{似然比 } \Lambda = \frac{\sup\limits_{p}(p^{n_{+1}}(1-p)^{n_{+2}})}{\sup\limits_{p_1,p_2}({p_1}^{n_{11}}(1-p_1)^{n_{12}}{p_2}^{n_{21}}(1-p_2)^{n_{22}})} = \frac{\left(\frac{n_{+1}}{n}\right)^{n_{+1}}\left(\frac{n_{+2}}{n}\right)^{n_{+2}}}{\left(\frac{n_{11}}{n_{1+}}\right)^{n_{11}}\left(\frac{n_{12}}{n_{1+}}\right)^{n_{12}}\left(\frac{n_{21}}{n_{2+}}\right)^{n_{21}}\left(\frac{n_{22}}{n_{2+}}\right)^{n_{22}}} = \prod_{i=1}^{2}\prod_{j=1}^{2}\left(\frac{n_{i+}n_{+j}}{nn_{ij}}\right)^{n_{ij}}$$

$$-2\log\Lambda = -2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log\left(\frac{n_{i+}n_{+j}}{nn_{ij}}\right) \overset{L}{\sim} \chi^2(1)$$

拒绝域：$-2\log\Lambda > \chi_\alpha^2(1)$

4. **双侧给定时四格表检验问题** （一个随机变量）

同上，方法是一致的. +++ A

5. **总体样本容量给定时四格表检验问题** （三个随机变量）

同上，方法是一致的.

6. **完全随机的四格表独立性的似然比检验** （四个相互独立的随机变量）

同上，方法是一致的.

7. **Fisher 精确检验**

略

8. **优比检验法**

称 $\frac{p_{11}}{p_{12}} = \frac{P(B \mid A)}{P(B^c \mid A)} = \frac{\frac{p_{11}}{p_{1+}}}{\frac{p_{12}}{p_{1+}}}$ 为当个体拥有属性 A 时，有属性 B 相较于无属性 B 的优势；

称 $\frac{p_{21}}{p_{22}} = \frac{P(B \mid A^c)}{P(B^c \mid A^c)} = \frac{\frac{p_{21}}{p_{2+}}}{\frac{p_{22}}{p_{2+}}}$ 为当个体没有属性 A 时，有属性 B 相较于无属性 B 的优势.

称这两个优势的比为优比(OR，即 odds ratio)，记作 $\theta = \frac{p_{11}p_{22}}{p_{12}p_{21}}$，四格表的优比是相合性的体现.

优比有一些应用的意义：

① $OR = 1$，说明属性 A 对属性 B 的发生不起作用，四格表独立；

② $OR > 1$，说明属性 A 是危险因素，四格表正相合；

③ $OR < 1$，说明属性 A 是保护因素，四格表负相合.

有下述结论：

⑴ 若在属性 A 的个体有属性 B 的比例比没有属性 A 的个体中有属性 B 的比例高，则 $\theta > 1$；

⑵ 若在属性 A 的个体有属性 B 的比例比没有属性 A 的个体中有属性 B 的比例低，则 $\theta < 1$；

⑶ 若属性 A 与属性 B 相互独立，则 $\theta = 1$. $\Rightarrow$ $p_{11}p_{22} = p_{12}p_{21}$

记 θ 的估计 $\frac{n_{11}n_{22}}{n_{12}n_{21}}$ 为 $\hat{\theta}$，有

$$\sqrt{n}(\log\hat{\theta} - \log\theta) \overset{L}{\sim} N\left(0, \frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}\right)$$

用 p_{ij} 的 MLE 代替 p_{ij}，则

$$\frac{\log\hat{\theta}-\log\theta}{\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}} \overset{L}{\sim} N(0,1)$$

H_0：$\theta=1$（属性 A、B 相互独立）

H_1：① $\theta\neq1$（属性 A 与 B 有关）

② $\theta>1$（有属性 A 的个体中有属性 B 的比例高）

③ $\theta<1$（有属性 A 的个体中有属性 B 的比例低）

在 H_0 成立条件下，检验统计量：

$$U=\frac{\log\hat{\theta}}{\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}} \overset{L}{\sim} N(0,1)$$

$$\chi^2=U^2 \overset{L}{\sim} \chi^2(1)$$

	原假设 H_0	备择假设 H_1	水平α拒绝域	p-value
有方向检验	$\theta=1$ 属性 A、B 相互独立	$\theta>1$ 属性 A 中有属性 B 的比例高	$U\geqslant U_\alpha$	$P(N(0,1)\geqslant U)=\Phi(-U)$ $\leqslant\alpha$
		$\theta<1$ 属性 A 中有属性 B 的比例低	$U\leqslant -U_\alpha$	$P(N(0,1)\leqslant U)=\Phi(U)$ $\leqslant\alpha$
无方向检验 （独立性检验）		$p_1\neq p_2$ 属性 A、B 有关	$\chi^2\geqslant\chi^2_\alpha(1)$	$P(\chi^2(1)\geqslant\chi^2)$ $\leqslant\alpha$

9. **边缘齐性检验（McNemar χ^2 检验与似然比检验）**

边缘齐性检验针对属性 A、B 不相互独立的特殊情况，用以检验两属性在所有个体中的比例是否相同；除了边缘齐性，属性 A、B 不相互独立时一致性也是一个值得考虑的问题.

边缘齐性指 $p_{1+}=p_{+1}$，即所有个体中拥有属性 A 的比例与拥有属性 B 的比例相一致，对于四格表而言边缘齐性等价于对称性，这是由于 $p_{12}=p_{1+}-p_{11}=p_{+1}-p_{11}=p_{12}$.

最典型的例子有：① 同一批疑似确诊的患者，分别用方法 A 和方法 B 进行检验，问两种不同的方法诊断病人为阳性患者的效果是否一致；② 现有某专业一批学生的期中考试与期末考试不及格人数与及格人数，检验期中考试挂科率与期末考试挂科率是否一致.

H_0：$p_{1+}=p_{+1}$（所有个体中拥有属性 A 的比例与拥有属性 B 的比例相一致）

H_1：$p_{1+}\neq p_{+1}$

检验统计量：

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{ij}-n\hat{p}_{ij}\right)^2}{n\hat{p}_{ij}}=\frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}} \overset{L}{\sim} \chi^2(1)$$

$$\text{似然比 }\Lambda=\prod_{i=1}^{2}\prod_{j=1}^{2}\left(\frac{n\hat{p}_{ij}}{n_{ij}}\right)^{n_{ij}}$$

$$-2\log\Lambda=-2\left(n_{12}\log\frac{n_{12}+n_{21}}{2n_{12}}+n_{21}\log\frac{n_{12}+n_{21}}{2n_{21}}\right) \overset{L}{\sim} \chi^2(1)$$

拒绝域：$\chi^2>\chi^2_\alpha(1)$ 或者 $-2\log\Lambda>\chi^2_\alpha(1)$，p-value $\leqslant\alpha$，p-value $=P(\chi^2(1)\geqslant\chi^2\ or\ -2\log\Lambda)$

——二维列联表——

1. **二维列联表的 Pearson χ^2 检验（无方向检验、独立性检验）**

如果属性 A、B 相互独立，则应有 $\frac{p_{1j}}{p_{1+}}=\ldots=\frac{p_{rj}}{p_{r+}}=\frac{p_{1j}+\ldots+p_{rj}}{p_{1+}+\ldots+p_{r+}}=p_{+j}$ ⇒ 二维列联表齐性与独立性等价.

H_0：两个属性相互独立 $(\frac{p_{1j}}{p_{1+}}=\ldots=\frac{p_{rj}}{p_{r+}}=p_{+j})$

H_1：两个属性不相互独立

在 H_0 成立条件下，检验统计量：

$$\hat{p}_{ij}=\frac{n_{i+}n_{+j}}{n^2}$$

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-n\hat{p}_{ij}\right)^2}{n\hat{p}_{ij}}=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{\frac{n_{i+}n_{+j}}{n}}-n\overset{L}{\sim}\chi^2\left((r-1)(k-1)\right)$$

其中 r、c 分别是二维列联表的行数与列数.

拒绝域：$\chi^2\geqslant\chi_\alpha^2\left((r-1)(k-1)\right)$ 或者 p-value $\leqslant\alpha$，p-value $=P(\chi^2\left((r-1)(k-1)\right)\geqslant\chi^2)$

2. **二维列联表的似然比检验**

H_0：两个属性相互独立 $(\frac{p_{1j}}{p_{1+}}=\ldots=\frac{p_{rj}}{p_{r+}}=p_{+j})$

H_1：两个属性不相互独立

检验统计量：

$$\text{似然比 }\Lambda=\prod_{i=1}^{r}\prod_{j=1}^{c}\left(\frac{\hat{p}_{ij}}{\frac{n_{ij}}{n}}\right)^{n_{ij}}$$

$$-2\log\Lambda=-2\sum_{i=1}^{r}\sum_{i=1}^{c}n_{ij}\log\left(\frac{n_{i+}n_{+j}}{nn_{ij}}\right)\overset{L}{\sim}\chi^2\left((r-1)(k-1)\right)$$

拒绝域：$-2\log\Lambda\geqslant\chi_\alpha^2\left((r-1)(k-1)\right)$ 或者 p-value $\leqslant\alpha$，p-value $=P(\chi^2\left((r-1)(k-1)\right)\geqslant-2\log\Lambda)$

3. **二维 $r\times c$ 列联表的相关系数**

对于来自某二元连续性随机向量的成对数据 $\binom{x_1}{y_1},\ldots,\binom{x_n}{y_n}$，定义：

① Pearson 矩相关系数

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sqrt{(y_i-\bar{y})^2}}}$$

矩相关系数通常要求正态分布的假设，是参数统计方法，而秩相关系数与 τ 相关系数是非参数统计方法；其大小与属性赋的值有关，因此不适合度量属性数据的相合关系.

② Kendall τ 相关系数

$$\tau = \frac{2}{n(n-1)}z$$

$$z = \sum_{1\le i<j\le n} \operatorname{sign}\Big((x_i - x_j)(y_i - y_j)\Big)$$

τ 相关系数与 $x_1,\dots,x_n$ 和 $y_1,\dots,y_n$ 数值大小无关，仅与其顺序有关.

τ 相关系数值介于 -1 与 1，通常认为值越接近 1，$x_1,\dots,x_n$ 和 $y_1,\dots,y_n$ 越趋向正相关；值越接近 -1，$x_1,\dots,x_n$ 和 $y_1,\dots,y_n$ 越趋向负相关.

$$z = G - H$$

其中

$$G = \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} n_{ij}\left(\sum_{k=i+1}^{r}\sum_{t=j+1}^{c} n_{kt}\right) = \sum_{i<k}\sum_{j<t} n_{ij}n_{kt}$$

$$H = \sum_{i=1}^{r-1}\sum_{j=2}^{c} n_{ij}\left(\sum_{k=i+1}^{r}\sum_{t=1}^{j-1} n_{kt}\right) = \sum_{i<k}\sum_{j>t} n_{ij}n_{kt}$$

G 与 H 有简便的计算方式，例：

a_1	b_1	c_1
a_2	b_2	c_2
a_3	b_3	c_3

$$\begin{cases} G = a_1(b_2 + c_2 + b_3 + c_3) + b_1(c_2 + c_3) + a_2(b_3 + c_3) + b_2(c_3) \\ H = c_1(a_2 + b_2 + a_3 + b_3) + b_1(a_2 + a_3) + c_2(a_3 + b_3) + b_2(a_3) \end{cases}$$

再根据 $z = G - H$，这便容易计算了.

③ Spearman 秩相关系数

略

4. 二维列联表相合性的度量

相合性：正相合——属性 A 值比较大的个体，属性 B 往往也比较大；

负相合——属性 A 值比较大的个体，属性 B 往往却比较小.

四格表的相合性可以用 $n_{11}n_{22} - n_{12}n_{21}$ 衡量：当 $n_{11}n_{22} - n_{12}n_{21} > 0$，倾向于认为四格表正相合；

当 $n_{11}n_{22} - n_{12}n_{21} < 0$，倾向于认为四格表负相合.

① Kendall τ 系数

$$T_A = \sum_{i=1}^{r}\binom{n_{i+}}{2} = \sum_{i=1}^{r}\frac{n_{i+}(n_{i+}-1)}{2}$$

$$T_B = \sum_{j=1}^{c}\binom{n_{+j}}{2} = \sum_{j=1}^{c}\frac{n_{+j}(n_{+j}-1)}{2}$$

$$\tau = \frac{z}{\sqrt{\left(\frac{n(n-1)}{2} - T_A\right)\left(\frac{n(n-1)}{2} - T_B\right)}}$$

τ 系数的值介于 -1 与 1，值越接近 1 越倾向认为正相合；值越接近 -1 越倾向认为负相合.

当且仅当 $H = 0$、$T_A = T_B$ 时 $\tau = 1$，即每行只有一个非零值，所以只在 $r = c$，即列联表是方表的时候，除主对角线元素全部等于 0，此时方表完全正相合，$\tau = 1$；除副对角线元素全部等于 0，此时方表完全负相合，$\tau = -1$.

但非方表也可能是完全正或负相合的，但此时 τ 系数取不到 -1 和 1，这是 τ 系数的一个缺陷，如

下表的 τ 便是小于 1 的.

*	*	0	0	0
0	0	*	*	0
0	0	0	0	0

② Gamma 系数

$$\gamma = \frac{G-H}{G+H}$$

Gamma 系数的值介于 -1 与 1，值越接近 1 越倾向认为正相合；值越接近 -1 越倾向认为负相合.

当且仅当 $H=0$ 时 $\tau=1$；当且仅当 $G=0$ 时 $\tau=-1$.

Kendall τ 系数等于 1 时 Gamma 系数也等于 1，Kendall τ 系数等于 -1 时 Gamma 系数也等于 -1，但反之不一定.

对于上表 $\gamma=1$，但对于并不完全正相合的下表 γ 仍等于 1，这是 Gamma 系数的一个缺陷.

*	*	0	0	0
0	0	*	*	0
0	0	0	0	0

③ Somers d 系数

d 系数通常被应用于 $2\times c$ 列联表问题，当然于 $r\times 2$ 列联表也是可行的，其中只有两种状态的属性可以是无序的，例如“性别”、“阳性与阴性”等，通常认为他们是自变量，而另一个属性是因变量. 以 $2\times c$ 列联表为例，我们认为列属性 B 依赖于行属性 A，或者说列属性 B 是行属性 A 的响应.

仍以 $2\times c$ 列联表为例，假定列联表是单侧 n_{1+}、n_{2+} 给定的，行属性 A 可以无序，但列属性 B 一定是有序的；无论行属性 A 是否有序（如若无序赋值即可），我们认为 A_1 到 A_2 有一个由小到大的顺序关系，于是可以定义 d 系数.

d 系数有两种：

$$d \text{ 系数} \begin{cases} d_{B|A} = \dfrac{G-H}{\dfrac{n(n-1)}{2} - T_A} \\ d_{A|B} = \dfrac{G-H}{\dfrac{n(n-1)}{2} - T_B} \end{cases}$$

d 系数的值介于 -1 与 1，值越接近 1 越倾向认为正相合；值越接近 -1 越倾向认为负相合.

当列属性 B 是行属性 A 的响应用 $d_{B|A}$ 度量相合性；当行属性 A 是列属性 B 的响应用 $d_{A|B}$ 度量相合性.

度量系数的 $1-\alpha$ 置信区间：

用 $se(\,)$ 表示标准误，分别为 $\tau \pm U_{\frac{\alpha}{2}}\, se(\tau)$，$\gamma \pm U_{\frac{\alpha}{2}}\, se(\gamma)$，$d \pm U_{\frac{\alpha}{2}}\, se(d)$.

5. **二维列联表相合性的检验**

H_0：属性 A、B 相互独立

H_1：① 属性 A、B 相合

② 属性 A、B 正相合

③ 属性 A、B 负相合

在 H_0 成立条件下，检验统计量：

$$U=\frac{z}{\sigma(z)} \overset{L}{\sim} N(0,1)$$

$$\chi^2=U^2=\frac{z^2}{\sigma^2(z)} \overset{L}{\sim} \chi^2(1)$$

其中 $\sigma(z)$ 是 $z=G-H$ 的标准误，即标准差除样本容量 n 的平方根.

$$\sigma^2(z)=\frac{n(n-1)(2n+5)-\sum n_{i+}(n_{i+}-1)(2n_{i+}+5)-\sum n_{+j}(n_{+j}-1)(2n_{+j}+5)}{18}$$

$$+\frac{[\sum n_{i+}(n_{i+}-1)(n_{i+}-2)][\sum n_{+j}(n_{+j}-1)(n_{+j}-2)]}{9n(n-1)(n-2)}+\frac{[\sum n_{i+}(n_{i+}-1)][\sum n_{+j}(n_{+j}-1)]}{2n(n-1)}$$

计算繁琐可以用近似计算公式替代

$$\sigma^2(z)\approx\frac{(n^3-\sum n_{i+}^3)(n^3-\sum n_{+j}^3)}{9n^3}$$

原假设 H_0	备择假设 H_1	水平α拒绝域	p-value
属性 A、B 相互独立	属性 A、B 相合	$U\geqslant U_\alpha$	$P(N(0,1)\geqslant U)=\Phi(-U)\leqslant\alpha$
	属性 A、B 正相合	$U\leqslant -U_\alpha$	$P(N(0,1)\leqslant U)=\Phi(U)\leqslant\alpha$
	属性 A、B 负相合	$\chi^2\geqslant\chi_\alpha^2(1)$	$P(\chi^2(1)\geqslant\chi^2)\leqslant\alpha$

6. **方表一致性的度量**

方表的一致性指对角线上元素值一致的性质，如果说我们可以用边缘齐性检验是否一致，那么我们还需要一致性度量与检验这个“一致”是否是偶然的、偶然一致的.

最典型的例子有：① 同一批产品由两个质检员分别质检，每人按要求将每个产品划分为 n 个品质等级，试问他们检验结果的一致性是否是偶然的；② 两个医生分别对一批病人作某项身体健康检查，将每个病人患病情况诊断为“阴性”、“阳性”与“强阳性”，问他们的诊断结果是否偶然一致.

Keppa κ 系数：

定义 q_1：如果 q_1 足够大，说明一致性很高，但 q_1 只衡量了“一致”的部分，没有参考权衡“非一致”的部分；

$$q_1=\frac{n_{11}+n_{22}+\ldots+n_{rr}}{n}$$

定义 q_2：q_1 的期望；

偶然一致且两侧给定的情况下：

$$\mathbb{E}(n_{ii})=\frac{n_{i+}n_{+i}}{n},\ Var(n_{ii})=\frac{n_{i+}(n-n_{i+})n_{+i}(n-n_{+i})}{n^2(n-1)}$$

$$q_2=\frac{1}{n}\sum_{i=1}^{r}\mathbb{E}(n_{ii})=\sum_{i=1}^{r}\left(\frac{n_{i+}}{n}\frac{n_{+i}}{n}\right)$$

二者作差再乘规范系数，可以得到 Keppa κ 系数：

$$\kappa=\frac{q_1-q_2}{1-q_1}=\frac{\sum\frac{n_{ii}}{n}-\sum\frac{n_{i+}n_{+i}}{n^2}}{1-\sum\frac{n_{i+}n_{+i}}{n^2}}$$

κ 系数最大值为 1，这时表示方表完全一致，可以在方表除对角线元素都等于 0 时取到；当属性 A、B 的状态完全随机时$\mathbb{E}(\kappa)=0$.

如果 $\kappa < 0$ 则可以立刻认为结果是偶然一致的，但当 $\kappa \geqslant 0$ 时需讨论 κ 是否充分得大于 0，此时不能轻易判断是否偶然一致，需要进行进一步的一致性检验.

7. **方表一致性的检验**

H_0：偶然达到一致

H_1：并非偶然达到一致

在方表两侧给定且 H_0 成立条件下，检验统计量：

$$Cov(n_{ii}, n_{jj}) = \frac{n_{i+}n_{j+}n_{+i}n_{+j}}{n^2(n-1)}$$

$$Var(q_1) = \frac{1}{n-1}\left(q_2 + q_2^2 - \sum_{i=1}^{r}\frac{n_{i+}}{n}\frac{n_{+i}}{n}\left(\frac{n_{i+}}{n} + \frac{n_{+i}}{n}\right)\right)$$

$$Var(\kappa) = \frac{Var(q_1)}{(1-q_2)^2} = \frac{1}{(n-1)(1-q_2)^2}\left(q_2 + q_2^2 - \sum_{i=1}^{r}\frac{n_{i+}}{n}\frac{n_{+i}}{n}\left(\frac{n_{i+}}{n} + \frac{n_{+i}}{n}\right)\right)$$

$$U = \frac{\kappa}{\sqrt{Var(\kappa)}} \overset{L}{\sim} N(0,1)$$

拒绝域：$U > N_\alpha(0,1)$ 或者 p-value $\leqslant \alpha$，p-value $= P(N(0,1) \geqslant U)$

——高维列联表——

1. 高维列联表的条件独立性 Pearson χ^2 检验与条件独立性似然比检验

如果属性 A、B 相互独立，则应有 $\frac{p_{1j}}{p_{1+}}=\ldots=\frac{p_{rj}}{p_{r+}}=\frac{p_{1j}+\ldots+p_{rj}}{p_{1+}+\ldots+p_{r+}}=p_{+j}$ ⇒ 二维列联表齐性与独立性等价.

以下以三维列联表中属性 C 给定后属性 A 与属性 B 的条件独立性为例：

H_0：属性 C 给定后，属性 A 与属性 B 条件独立

H_1：属性 C 给定后，属性 A 与属性 B 不条件独立

按属性 C 分层，分为 t 个二维 $r\times c$ 列联表检验，在 H_0 成立条件下每个列联表都是相互独立的：

第 k 个二维 $r\times c$ 列联表的 χ^2 检验统计量：

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{njk}-\frac{n_{i+k}n_{+ik}}{n_{++k}}\right)^2}{\frac{n_{i+k}n_{+jk}}{n_{++k}}}=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ijk}^2}{\frac{n_{i+k}n_{+jk}}{n_{++k}}}-n_{++k}\overset{L}{\sim}\chi^2\big((r-1)(c-1)\big)$$

第 k 个二维 $r\times c$ 列联表的似然比检验统计量：

$$-2\log\Lambda=-2\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ijk}\log\left(\frac{n_{i+k}n_{+jk}}{n_{++k}n_{ijk}}\right)\overset{L}{\sim}\chi^2\big((r-1)(c-1)\big)$$

在 H_0 成立条件下，检验统计量：

χ^2 检验统计量的渐进分布：$\sum_{k=1}^{t}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ijk}^2}{\frac{n_{i+k}n_{+jk}}{n_{++k}}}-n\overset{L}{\sim}\chi^2\big(t(r-1)(c-1)\big)$

似然比统计量的渐进分布：$-2\sum_{k=1}^{t}\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ijk}\log\left(\frac{n_{i+k}n_{+jk}}{n_{++k}n_{ijk}}\right)\overset{L}{\sim}\chi^2\big(t(r-1)(c-1)\big)$

其中 r、c、t 分别是三维列联表属性 A、B 和 C 状态的个数.

拒绝域：$\chi^2\geqslant\chi_\alpha^2\big(t(r-1)(k-1)\big)$ 或者 p-value $\leqslant\alpha$，p-value $=P(\chi^2\big(t(r-1)(k-1)\big)\geqslant\chi^2)$

2. 高维列联表的独立性检验

• 由于似然比统计量具有可分解性，所以三维或三维以上的高维列联表的独立性检验问题常常考虑似然比方法；可分解性是指：

$$-2\log\Lambda_{(A,B,C)}\geq\begin{cases}-2\log\Lambda_{(A,BC)}\geq\begin{cases}-2\log\Lambda_{(BA,BC)}\\-2\log\Lambda_{(CA,CB)}\end{cases}\\-2\log\Lambda_{(B,AC)}\geq\begin{cases}-2\log\Lambda_{(AB,AC)}\\-2\log\Lambda_{(CA,CB)}\end{cases}\\-2\log\Lambda_{(C,AB)}\geq\begin{cases}-2\log\Lambda_{(AB,AC)}\\-2\log\Lambda_{(BA,BC)}\end{cases}\end{cases}$$

若某两个似然比检验统计量满足如上的顺序关系，则其较大值减去较小值的差值为一个似然比检验统计量，且其原假设 H_0 与较大值所检验问题的 H_0 相同，备择假设 H_1 为“较小值所检验问题的 H_0 成立但较小值所检验问题的 H_0 不成立”，自由度为较大值自由度减去较小值自由度.

比如 $-2\left(\log\Lambda_{(A,B,C)}-\log\Lambda_{(AB,AC)}\right)$ 是原假设 H_0 为“A、B、C 相互独立”、备择假设 H_1 为“A 给定后 B 与 C 条件独立，但 A、B、C 不相互独立”假设检验问题的似然比检验统计量，这个差值的自由度为 $(rct-r-t-c+2)-\big(r(c-1)(t-1)\big)=(r-1)(c+t-2)$.

三维列联表独立性检验问题的解

情况	编号	原假设 H_0	备择假设 H_1	检验统计量 $-2\log\Lambda$	渐进分布 χ^2 自由度
Ⅰ相互独立	①	A、B、C 相互独立	A、B、C 不相互独立	$-2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t}n_{ijk}\log\left(\frac{n_{i++}n_{+j+}n_{++k}}{n^2n_{ijk}}\right)$	$rct-r-t-c+2$
Ⅱ整体独立	②	A 和 (B,C) 相互独立	A 和 (B,C) 不相互独立	$-2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t}n_{ijk}\log\left(\frac{n_{i++}n_{+jk}}{nn_{ijk}}\right)$	$(r-1)(ct-1)$
	③	B 和 (A,C) 相互独立	B 和 (A,C) 不相互独立	$-2\sum_{j=1}^{c}\sum_{i=1}^{r}\sum_{k=1}^{t}n_{ijk}\log\left(\frac{n_{+j+}n_{i+k}}{nn_{ijk}}\right)$	$(c-1)(rt-1)$
	④	C 和 (A,B) 相互独立	C 和 (A,B) 不相互独立	$-2\sum_{k=1}^{t}\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ijk}\log\left(\frac{n_{++k}n_{ij+}}{nn_{ijk}}\right)$	$(t-1)(rc-1)$
Ⅲ条件独立性	⑤	A 给定后，B 与 C 条件独立	A 给定后，B 与 C 不条件独立	$-2\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t}n_{ijk}\log\left(\frac{n_{i+k}n_{ij+}}{n_{i++}n_{ijk}}\right)$	$r(c-1)(t-1)$
	⑥	B 给定后，A 与 C 条件独立	B 给定后，A 与 C 不条件独立	$-2\sum_{j=1}^{c}\sum_{i=1}^{r}\sum_{k=1}^{t}n_{ijk}\log\left(\frac{n_{ij+}n_{+jk}}{n_{+j+}n_{ijk}}\right)$	$c(r-1)(t-1)$
	⑦	C 给定后，A 与 B 条件独立	C 给定后，A 与 B 不条件独立	$-2\sum_{k=1}^{t}\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ijk}\log\left(\frac{n_{i+k}n_{+jk}}{n_{++k}n_{ijk}}\right)$	$t(r-1)(c-1)$

• Ⅰ、Ⅱ和Ⅲ的难度是依次递增的，通常倾向于选取更简单的模型，因此可以先考虑检验 Ⅰ，若 Ⅰ 被拒绝则考虑检验Ⅱ，Ⅱ都被拒绝再考虑检验Ⅲ；也可以对七种情况都进行讨论，再选取最优的模型.

其中Ⅲ便是 1.中讨论的问题，即条件独立性.

七种情况原假设 H_0 成立情况下期望频数 p_{ijk} 的估计分别为：

① $\frac{n_{i++}n_{+j+}n_{++k}}{n^2}$，② $\frac{n_{i++}n_{+jk}}{n}$，③ $\frac{n_{+j+}n_{i+k}}{n}$，④ $\frac{n_{++k}n_{ij+}}{n}$，⑤ $\frac{n_{ij+}n_{i+k}}{n_{i++}}$，⑥ $\frac{n_{ij+}n_{i+k}}{n_{+j+}}$，⑦ $\frac{n_{i+k}n_{+jk}}{n_{++k}}$

——Logistic 回归——

Logistic 回归总体而言是一种广义线性模型，用以解决单个因变量的分类问题，对应分类变量的问题；区别于对数线性模型，对应关于列联表的变量的问题.

1. **Logistic 变换及 Logistic 线性回归模型**

Logistic 回归是一种对数几率模型 (Logit model)，为了因变量为离散变量时进行回归而提出，即对分类问题进行回归，例如利用 Logistic 回归，可以通过样本的回归判断任意指定的一个拥有某 BMI 指数的人是否大概率患有心血管病（二分类、二值回归）、仅由某人的一些面部特征数据判断他是成年女士、成年男士还是未成年儿童等等，这些都可以通过 Logistic 回归实现.

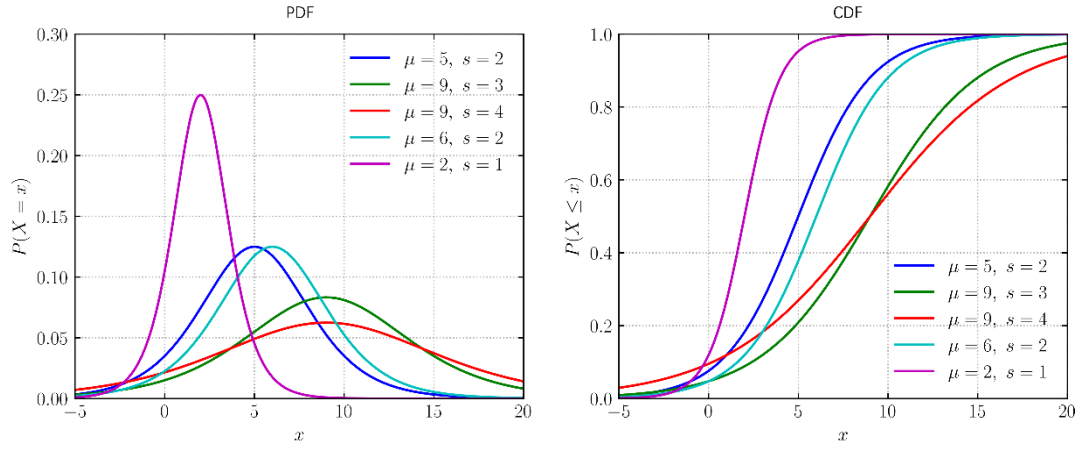
Logistic 分布：一种指数族分布，设 X 为连续型随机变量 CDF 与 PDF 分别为：

$$F(x)=\frac{1}{1+e^{-\frac{x-\mu}{\gamma}}}=\frac{1}{2}+\frac{1}{2}\tanh\left(\frac{x-\mu}{2\gamma}\right),\ x\in\mathbb{R}$$

$$f(x)=\frac{e^{-\frac{x-\mu}{\gamma}}}{\gamma(1+e^{-\frac{x-\mu}{\gamma}})^2}=\frac{1}{4\gamma}\operatorname{sech}^2\left(\frac{x-\mu}{2\gamma}\right),\ x\in\mathbb{R}$$

其中称 μ 为位置参数，$\gamma>0$ 为形状参数；$F(x)$ 是一条 S 形曲线，$f(x)$ 关于 $x=\mu$ 对称；γ 越大越厚尾，而在 $x=\mu$ 附近增长越慢.

特别地，对于二值 Logistic 回归，常常令$(Y=1\mid x)=\frac{e^{\boldsymbol{\beta X}}}{1+e^{\boldsymbol{\beta X}}}$，$P(Y=0\mid x)=\frac{1}{1+e^{\boldsymbol{\beta X}}}$即服从 Logistic 分布.



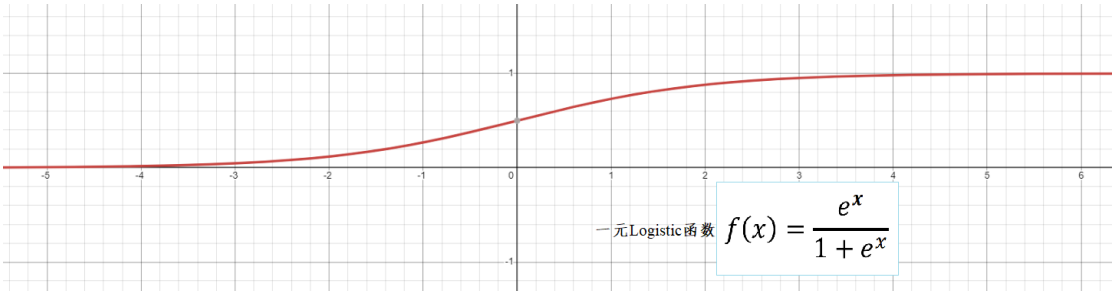
由于理想的分类函数是符号函数 sign( ) 并不可微，(对二值 Logistic 回归)提出对 Y 进行 Logistic 变换，等价于“优势比的对数等于 $\boldsymbol{\beta X}$”. Logistic 变换是针对因变量 Y 为离散变量（属性数据）且回归函数为 Logistic 函数的变换，通过这种方式使原函数线性化、成为关于 Z 的线性函数；假设 X 为自变量而 Y 为因变量，记 $\boldsymbol{\beta}$ 为系数向量、$\boldsymbol{X}$ 为设计矩阵，其中 $\boldsymbol{\beta}$ 常取极大似然估计 MLE，定义 Logistic 函数：

$$P(Y\mid X)=\frac{e^{\boldsymbol{\beta X}}}{1+e^{\boldsymbol{\beta X}}}\ (\text{Logistic 函数})$$

Logistic 函数是一种定义域为 $\mathbb{R}$、值域为 $(-1,1)$ 的非线性函数，但通过 Logistic 变换可以线性化：

$$Z=\log\left(\frac{Y}{1-Y}\right)$$

这就是 Logistic 变换.



由于概率 $y=P(Y=1\mid X)$ 的值取值在 0 到 1 之间，如若把 y 假设为多项式函数等取值在 0 到 1 之间的函数是不合适的，需要做一个映射处理，这个目的常用一些变换得到，例如：

① Logistic 变换：$f(y)=\log\frac{y}{1-y}$，特别地在 Logistic 线性模型中 $f(y)=\boldsymbol{\beta X}$

② Probit 变换：$f(y)=\Phi^{-1}(y)$

③ 双对数变换：$f(y)=\log(-\log(1-y))$

其中就包括 Logistic 变换，实质是 Logistic 函数的逆. 这样变换以后再假设 $f(y)=\boldsymbol{\beta X}$ 服从某回归模型，譬如在线性回归模型中 $f(y)=\boldsymbol{\beta X}$，从而拟合 p. 容易看出，相合性检验比独立性检验更深入，进一步地 Logistic 回归比相合性检验更深入，他直接给出了一个可能的关系式.

二值 Logistic 回归模型：假设响应变量即因变量 Y 仅有两个状态，我们分别用 0 和 1 表示，现研究 $y=P(Y=1)$，若一共有 k 个因素 $x_1,x_2,\dots,x_k$ 影响Y 的取值，则称

$$\log\frac{y}{1-y}=g(x_1,x_2,\dots,x_k)$$

为二值 Logistic 回归模型，简称 Logistic 回归模型.

当 $g(x_1,x_2,\dots,x_k)$ 为一个线性函数时称为 Logistic 线性回归模型，即：

$$\log\frac{y}{1-y}=\boldsymbol{\beta X}=\beta_0+\sum_{i=1}^{k}\beta_i x_i$$

容易知道 y 服从 Logistic 分布，即 $y=P(Y=1|X)=\frac{\exp(\boldsymbol{\beta X})}{1+\exp(\boldsymbol{\beta X})}$，$1-y=P(Y=0|X)=\frac{1}{1+\exp(\boldsymbol{\beta X})}$，这样设置可以让 $P(Y=1|X)$ 的回归系数使得 $\boldsymbol{\beta X}=0$；$\boldsymbol{\beta}$ 取其 MLE.

当协变量向量一共有 t 种而第 i 种有 n_i 个，其中有 r_i 个响应变量值为 1 而有 n_i-r_i 个取值为 0，参数 $\boldsymbol{\beta}$ 的似然函数为

$$\prod_{i=1}^{t}\left(P(Y=1|X)\right)^{r_i}\left(P(Y=0|X)\right)^{n_i-r_i}=\prod_{i=1}^{t}\left(\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\dots+\beta_nx_n}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\dots+\beta_nx_n}}\right)^{r_i}\left(\frac{1}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\dots+\beta_nx_n}}\right)^{n_i-r_i}$$

若将上式记为 $\sup L$，则有

$$-2\log\Lambda=-2\log\frac{\left(\frac{n_{Y=1}}{n}\right)^{n_{Y=1}}\left(\frac{n_{Y=0}}{n}\right)^{n_{Y=0}}}{\sup L}\overset{L}{\sim}\chi^2(1)$$

MLE 准则下 Logistic 回归损失函数：交叉熵 记 $P(Y=1\mid X)=p(x)$

$$\textbf{交叉熵}=\textbf{KL散度}+\textbf{信息熵}:\ -\sum_{i=1}^{n}p(x_i)\log\big(1-p(x_i)\big)=D_{KL}(p||1-p)-\sum_{i=1}^{n}p(x_i)\log p(x_i)$$

∗ 为什么 Logistic 回归中 $\boldsymbol{\beta}$ 取其 MLE 而非于线性回归中更常用的、一定条件下具有优良性质的普通最小二乘估计 OLS？换句话说，在均方误差 MSE 意义下得到的最佳估计便是 OLS，那为什么损失函数不再选用 MSE（而考虑交叉熵），进而用 MLE 代替 OLS？

答：

① 本质原因是分类问题中属性的分布是**多项分布**（二值 Logistic 回归中是二项分布），并没有做残差正态分布的假设；在普通线性回归中我们有正态残差假设.

注意：此处值得指出的是 OLS 是非参数方法，MLE 是参数统计方法，这正好适用于 Logistic 回归问题；最小二乘方法是一个凸优化的问题，MSE 综合了无偏性与 SSE 的值，MSE $=n\cdot$ SSE，然而这个问题下 MSE 准则得到的损失函数不是凸优化问题，极大似然的方法不一定是凸优化问题，不过在此处成立——凸性拥有极好的优化性质；OLS 可以视作使得残差在 L^2 范数意义下最小的最优解，即使得 SSE 最小，MLE 是使得似然函数最大的最优解.

正态残差假设是普通线性回归中 OLS 具有很好的一些性质所必要的前提（参考 Gauss–Markov theorem 的条件），这时 OLS 是一致最小方差无偏估计，是最佳线无偏估计；甚至在一些 MLE 如残差是有偏的，因此 Gauss–Markov 假设、正态残差假设成立情况下，在线性回归中一般考虑 OLS 而非 MLE.

但对于分类问题，在属性服从多项分布的假定下 OLS 并没有 Gauss–Markov 假设下普通线性回归的那些优良性质，而参数的 MLE 具有此时具有无偏性、渐进正态性等性质，所以考虑 MLE，而 MLE 准则下的损失函数就是交叉熵，当然这只是考虑 MLE 一方面的原因. 也可以说，他们的不同可以讲是源自提出的假设不同，这是问题实际情况决定的.

② 最小二乘损失函数，或者讲 MSE 在这种情况下是非凸的，Hessian 矩阵非正定，难以数值迭代最优化，算法很有可能收敛到某局部最优解；但似然函数是可导凸函数，拥有唯一的最优解，局部最优解等价于全局最优解，总能收敛到最优点.

③ MSE 损失函数在靠近 0 和 1 时存在梯度消失的现象，同时似然损失函数的梯度只与参数有关，与 Logistic 函数的梯度无关：一元 Logistic 函数导数最大值点在四分之一处，可能会导致其他损失函数参数更近变慢.

④ MSE 损失函数的学习比交叉熵慢很多，对错误分类惩罚不够重，当损失函数为交叉熵（负对数似然）时样本分类错误越严重惩罚会越严重，呈，也会在导数最大点附近急剧减小，而最小二乘损失函数相对而言变化平缓，并不“陡峭”，响应不够敏感.

$proof$：

① $w=\boldsymbol{\beta X}$，二项分布假设，记 $P(Y=1\mid X)=p(x)$，$P(Y=0\mid X)=1-p(x)$

$$l(w)=\log\big(L(w)\big)=\log\left(\prod_{i=1}^{n}\big(p(x_i)\big)^{y_i}\big(1-p(x_i)\big)^{1-y_i}\right)$$

$$=\sum_{i=1}^{n}\left(y_i\log\left(\frac{p(x_i)}{1-p(x_i)}\right)+\log\big(1-p(x_i)\big)\right)$$

$$= \sum_{i=1}^{n} [y_i w x_i - \log(1 + e^{w x_i})] = -n\,J(w) \quad \Leftrightarrow \quad \text{交叉熵损失函数梯度为 0}$$

② 利用下文③的结果，

$$\frac{\partial^2 MSE}{\partial w^2} = \sum_{i=1}^{n} \hat{y}_i(1-\hat{y}_i)x_i^2\left(-y_i + 2(1+y_i)\hat{y}_i - 3\hat{y}_i^2\right)$$

由于 y_i 只取 0 或 1，可以导出 MSE 二阶导数不一定总大于 0.

③ 假设损失函数为 MSE

$$\frac{\partial MSE}{\partial w} = \frac{\partial \sum(\hat{y}_i - y_i)^2}{\partial w} = \sum_{i=1}^{n} 2(\hat{y}_i - y_i)\frac{\partial(\hat{y}_i - y_i)}{\partial w} = \sum_{i=1}^{n} 2(\hat{y}_i - y_i)\frac{\partial\left(\frac{1}{1+e^{-wx_i}}\right)}{\partial w}$$

$$= \sum_{i=1}^{n} 2(\hat{y}_i - y_i)\left(\frac{1}{1+e^{-wx_i}}\right)^2 e^{-wx_i}x_i = \sum_{i=1}^{n} 2(\hat{y}_i - y_i)[\hat{y}_i(1-\hat{y}_i)x_i]$$

可以看出在 $\hat{y}_i$ 靠近 0 或 1 时梯度趋近于 0，不能有效迭代.

Logistic 回归模型的解释

$$\log\frac{y}{1-y} = \boldsymbol{\beta X} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

我们认为，若 x_i 增加 1，则优势比 $\frac{y}{1-y} = \frac{P(Y=1\mid X)}{P(Y=0\mid X)}$ 增长至原来的 e^{β_i} 倍.

对于自变量都是定量数据的问题， $\log\frac{y}{1-y}$ 一般用 $\log\frac{\hat{y}}{1-\hat{y}} = \log\frac{r_k}{n_i - r_k}$ 进行估计，其中 r_k 是 k 组观测数据的组合中取 1 的自变量的个数、n_k 为观测值个数，但当 $r_k = 0\ or\ 1$ 时情况有些麻烦，通常在 $r_k = 0$ 时用 $\log\frac{0.5}{n_k+0.5}$ 代替 $\log\frac{r_k}{n-r_k}$，在 $r_k = n_i$ 时用 $\log\frac{n_k+0.5}{0.5}$ 代替 $\log\frac{r_k}{n-r_k}$，在 $0 < r_k < n_k$ 时用 $\log\frac{r_k+0.5}{n_i-r_k+0.5}$ 代替 $\log\frac{r_k}{n-r_k}$.

$y = P(Y=1\mid X) \approx 0$ 或 $1\left(y \approx 0\right.$ 即 $\left.1 - y = P(Y=0\mid X) \approx 1\right)$时我们认为自变量对因变量 Y 影响很小而且非常有可能实际的因变量取值就是 0 或 1，而当 $y = P(Y=1\mid X) = 0.5$ 时认为自变量对因变量取值的影响很大，毕竟直观理解起来是因变量取两个可能值的概率都是 0.5.

得到模型后，做预测时再将 $y = P(Y=1\mid X)$ 作映射，根据实际问题情况规定大于某值时取因变量为 1，反之取为 0，比方可以取这个分界点为 0.5，这样便得到了一个分类.

2. **含有名义数据的二分类 Logistic 线性回归模型**

例如要研究年龄、血型与死亡率之间的关系，年龄为定量数据而血型为名义数据. 通常血型大致可以分为四种：A 型、B 型、AB 型和 O 型，不能在 Logistic 回归中简单令离散随机变量 Q，让 A 型$\Leftrightarrow Q = 1$，B 型$\Leftrightarrow Q = 2$，… 因为这样使四种血型间有了顺序关系，事实上他们只作为名义数据，相互之间并没有大小之分，更没有“3 倍的 A 型血型等于 AB 型血型”的荒谬说法；应设三个随机变量 Q_1、Q_2、Q_3，选取一个名义数据作为基线，例如O型血：A型$\Leftrightarrow Q_1 = 1 \Leftrightarrow Q = (1,0,0)$，B型$\Leftrightarrow Q_2 = 2 \Leftrightarrow Q = (0,1,0)$，AB 型$\Leftrightarrow Q_3 = 1 \Leftrightarrow Q = (0,0,1)$，O 型$\Leftrightarrow Q = (-1,-1,-1)$，Logistic 回归模型为：

$$\log\frac{y}{1-y} = \mu + \beta_1 \cdot age + \gamma_1 Q_1 + \gamma_2 Q_2 + \gamma_3 Q_3$$

此外，由上式还可以分别得到四种血型情况下的年龄与死亡率的关系的回归方程.

因此对于含有名义数据的二分类 Logistic 线性回归模型，设有 n 维定量数据、j 维名义数据，$\boldsymbol{\beta}$ 表示定量数据的系数向量、$\boldsymbol{\gamma}$ 表示名义数据的系数向量，$\boldsymbol{\beta}$、$\boldsymbol{\gamma}$ 的估计取其 MLE，**每种定性数据只需要引进一个变量，而 j 维名义数据需要引进 $j-1$ 个变量**，这种办法叫基线法，第 j 个变量既可以取 $j-1$ 个变量都取零也可以都取负一，根据问题要求而定，有

Logistic 线性回归模型：

$$\log\frac{y}{1-y}=\mu+\beta_1x_1+\cdots+\beta_nx_n+\gamma_1\lambda_1+\ldots+\gamma_j\lambda_{j-1}$$

Logistic 线性回归方程：

$$\log\frac{\hat{y}}{1-\hat{y}}=\hat{\mu}+\hat{\beta}_1x_1+\cdots+\hat{\beta}_nx_n+\hat{\gamma}_1\lambda_1+\ldots+\hat{\gamma}_j\lambda_{j-1}$$

配合四格表独立性检验，当不独立性显著时进行 Logistic 回归更有说服力.

3. **含有有序数据的二分类 Logistic 线性回归模型**

例如文化程度，小学以下、小学、初中、高中、大学及以上分别可以用 0、1、2、3、4 表示，这是一组有序数据，可以认为他们之间是有大小的，**每种有序数据也仅引进一个变量**.

方法一致.

Logistic 线性回归模型：

$$\log\frac{y}{1-y}=\mu+\boldsymbol{\beta}A+\boldsymbol{\gamma}B+\boldsymbol{\nu}C$$

Logistic 线性回归方程：

$$\log\frac{\hat{y}}{1-\hat{y}}=\hat{\mu}+\hat{\boldsymbol{\beta}}A+\hat{\boldsymbol{\gamma}}B+\hat{\boldsymbol{\nu}}C$$

C 即有序数据.

4. **Logistic 判别分析**

对于 Logistic 回归方程，假设 $\log\frac{\hat{y}}{1-\hat{y}}=\hat{\mu}+\hat{\boldsymbol{\beta}}A+\hat{\boldsymbol{\gamma}}B+\hat{\boldsymbol{\nu}}C$，令 $u(A,B,C)=\hat{\mu}+\hat{\boldsymbol{\beta}}A+\hat{\boldsymbol{\gamma}}B+\hat{\boldsymbol{\nu}}C$，当 u 较大认为 $Y=1$ 概率较大，若取分界点为 0.5，则当 $u>0$ 时可以简单认为 $Y=1$ 会发生；具体的判别方式与分界点的选取视实际问题的情况与需求而定.

5. **多项 Logistic 回归**

多项 Logistic 回归中，假设因变量有多个可能的取值 $(0,1,\ldots,N)$，基线法要求选取一个（常常是最后一个）使得其 $w=\boldsymbol{\beta X}$ 值为 0，其中 $\boldsymbol{X}$ 为设计矩阵. 一般模型：

$$\log\frac{P(Y=0\mid X)}{P(Y=N\mid X)}=\mu^{(0)}+\beta_1^{(0)}x_1+\cdots+\beta_n^{(0)}x_n$$

$$\log\frac{P(Y=1\mid X)}{P(Y=N\mid X)}=\mu^{(1)}+\beta_1^{(1)}x_1+\cdots+\beta_n^{(1)}x_n$$

…

$$\log\frac{P(Y=k\mid X)}{P(Y=N\mid X)}=\mu^{(k)}+\beta_1^{(k)}x_1+\cdots+\beta_n^{(k)}x_n$$

$$0=\log\frac{P(Y=N\mid X)}{P(Y=N\mid X)}=\mu^{(N)}+\beta_1^{(N)}x_1+\cdots+\beta_n^{(N)}x_n$$

有约束$\sum_{i=1}^{N}P(Y=i\mid X)=1$

$\boldsymbol{\beta}^{(k)}$ 用极大似然估计 MLE 代替，

$$P(Y=k\mid X)=\frac{\exp\left(\mu^{(k)}+\beta_1^{(k)}x_1+\cdots+\beta_n^{(k)}x_n\right)}{\sum_{i=1}^{N}\exp\left(\mu^{(i)}+\beta_1^{(i)}x_1+\cdots+\beta_n^{(i)}x_n\right)}=\frac{\exp\left(\mu^{(k)}+\beta_1^{(k)}x_1+\cdots+\beta_n^{(k)}x_n\right)}{1+\sum_{i=1}^{N-1}\exp\left(\mu^{(i)}+\beta_1^{(i)}x_1+\cdots+\beta_n^{(i)}x_n\right)}$$

特别地$P(Y=N\mid X)=\frac{1}{\sum_{i=1}^{N}\exp\left(\mu^{(i)}+\beta_1^{(i)}x_1+\cdots+\beta_n^{(i)}x_n\right)}=\frac{1}{1+\sum_{i=1}^{N-1}\exp\left(\mu^{(i)}+\beta_1^{(i)}x_1+\cdots+\beta_n^{(i)}x_n\right)}$

6. **如何利用计算机做 Logistic 回归?**

R 语言：线性回归

有必要先介绍一下线性回归如何操作，这就要先了解 lm() 函数.

lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)

重要参数的选择：

· formula：模型的关系式，如 formula = Z~X+Y 表示拟合 $Z=X+Y+intercept$ 的线性模型，也可以写作 formula = Z~X+Y+1，注意此时模型中有截距项；而 formula = Z~X+Y-1 表示拟合 $Z=X+Y$ 的线性模型，此时不含截距项，事实上 formula 遵循 R 表达式的语法：

“~”为变量类型的分隔，左边是响应变量，右边是解释变量；

“+”用以隔开两个解释变量；

“:”即连接两个变量表示他们的交互项，比如 formula = Z~X+Y+X:Y；

“*”表示两个(或多个)变量自己与他们之间所有可能的交互项，例如 formula = Z~A*B*C，等价于 formula = Z~A+B+C+A:B+ A:C+ B:C+ A:B:C；

“^”表示交互项次数，例如 formula = Z~(A+B+C)^2，等价于 formula = Z~ Z~A+B+C+A:B+ A:C+ B:C；

“.”表示除因变量外所有变量，可以极大简化写法，比如 formula = Z~.；

“-”移除某变量，例如 formula = Z~(A+B+C)^2- A:C，特别地“-1”表示删除截距项；

“I()”将某变量进行算术平方，比如有 formula = Z~A+I((B+C)^2)，此时等价于 formula = Z~A+D，D 是一个新变量，值为 B 与 C 的和的平方；

“*function*”可以在模型中使用一些函数，例如 formula = log(Z)~X+Y.

· data：数据集，要求是数据框类型.

· weights：应该被赋值“NULL”或一个数值向量，当赋值为“NULL”使用普通最小二乘法 OLS，当赋值为数值向量则把他作为权重使用加权最小二乘估计 WLS（即最小化 sum(weights * e^2)）.

· na.action：当数据包含“NA”时应该如何处理.

a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. The 'factory-fresh' default is na.omit. Another possible value is NULL, no action. Value na.exclude can be useful.

・method：目前只有 method = qr 可用，也可以选择 method = "model.frame"，会返回模型数据库且不进行拟合，和 model = TRUE 效果相同.

・data：数据集，要求是数据框类型.

・data：数据集，要求是数据框类型.

・剩下的一些参数不是很重要，比如 subset 可以选择一个用来观察的子集.

lm() 除了进行线性回归还可以做单因素方差分析、协方差分析，不过后者有封装好的函数 aov()，这个函数实质也是在调用 lm().

lm() 参考文档 https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm

aov() 参考文档 https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aov

---

关于模型的函数：

・summary(fit)：提供一个详细的结果，包括多种参数与指标，含部分以下函数列出的参数.

```
Call:
glm(formula = CAD ~ college * TC * SBP, family = binomial(link = "logit"),
    data = dataSet)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.73574  -0.48116  -0.31985   0.00008   2.81342

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.938e+00  5.012e-01  -7.858  3.9e-15 ***
college         2.350e+01  1.484e+03   0.016   0.9874
TC              4.956e-01  2.049e-01   2.419   0.0156 *
SBP             4.187e-01  2.028e-01   2.064   0.0390 *
college:TC     -4.956e-01  9.577e+02  -0.001   0.9996
college:SBP    -4.187e-01  9.160e+02   0.000   0.9996
TC:SBP          3.012e-03  8.425e-02   0.036   0.9715
college:TC:SBP -3.012e-03  7.834e+02   0.000   1.0000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3034.5  on 2396  degrees of freedom
Residual deviance: 1112.0  on 2389  degrees of freedom
AIC: 1128

Number of Fisher Scoring iterations: 18
```

・coefficients(fit)：列出拟合的参数，包括截距(如果有).

・cofint(fit)：列出拟合参数的置信区间.

・fitted(fit)：列出预测值.

・residuals(fit)：列出残差.

・vcov(fit)：列出拟合参数的协方差阵.

・AIC(fit)、BIC(fit)：列出 AIC、BIC 统计量的值.

・plot(fit)：绘制一系列用以评价模型的回归诊断图，包含 Q-Q 图.

・predict(fit, data)：预测.

R 语言线性回归的例子：

```r
dataSet <- read.csv("data.csv")

fit <- lm(Effect ~ Gender + Age * Method, data = dataSet)
summary(fit)
```

R 语言：Logistic 回归

前文提到了线性回归函数 lm()，这里我们需要广义线性回归函数 lm() 函数.

glm(formula, family = gaussian, data, weights, subset, na.action, start = NULL, etastart, mustart, offset, control = list(…), model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, …)

glm.fit(x, y, weights = rep(1, nobs), start = NULL, etastart = NULL, mustart = NULL, offset = rep(0, nobs), family = gaussian(), control = list(), intercept = TRUE, singular.ok = TRUE)

重要参数的选择：

・family：选择相应变量分布的假设与连接函数，比如 family = binomal(link = 'logit') 表示响应变量分布假设为二项分布，用 Logistic 函数作为连接函数，这时做的便是逻辑回归；family = binomal(link = 'probit') 表示响应变量分布假设为二项分布，用 Probit 函数作为连接函数；family = poisson(link = 'identity') 表示响应变量分布假设为泊松分布，用 $f(x) = x$ 作为连接函数.

・control：控制误差与最大迭代次数，例如 control = list(epsilon=1e-8, maxit=25) 中 epsilon 为终止准则的误差，maxit 为最大迭代次数.

glm() 参考文档 https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm

---

R 语言 Logistic 回归的例子：

```
dataSet <- read.csv("data.csv")

fit <- glm(Effect ~ Gender + Age + Method, family = binomial(link = "logit"), data = dataSet,
control = list(maxit = 200))
summary(fit)
```

Python：Logistic 回归 (sklearn 库)

```
1.  import pandas as pd
2.  import sklearn as sl
3.
4.
5.  data = pd.read_csv('data.csv')
6.
7.  X = data.loc[:, ['Gender']]
8.  Y = data.loc[:, ['Effect']]
9.
10.
11.  model = sl.linear_model.LogisticRegression()  # Logistic Regression (逻辑回归)
12.  model.fit(X, Y)
13.  pre = model.predict(X)
14.  print("Line regression predict result: ", pre)
15.
16.  epsilon = pd.sqrt(sl.metrics.mean_squared_error(Y, pre))
17.  print("Line regression mean squared error: ", epsilon)
```

7. 参数的选取：信息准则

AIC

AIC 建立在信息论上，又称赤池信息量准则，是 KL 散度的估计量；一般假设模型误差服从相互独立的正态分布，记 k 为参数的数量、RSS 为残差平方和、C 是依赖于数据的常量，

$$AIC = 2k - 2\log\hat{L} = 2k + n\log RSS - (n\log n + 2C)$$

由于 $(n\log n + 2C)$ 不影响相同一批数据不同模型 AIC 值的差别，所以可以只考虑 AIC 的一部分用以对比

$$AIC \doteq 2k + n\log\left(\frac{RSS}{n}\right) \text{或 } 2k + n\log RSS$$

倾向于选择 AIC 较小的模型，当 k 增大通常能“更多地拟合”，这会让似然函数最大值 $\hat{L}$ 的值也增大而使得 AIC 的值减小，但 k 太大的时候会大大影响 AIC 的值使之也变得庞大，与其同时出现的问题是过拟合，这是应该避免的情况，因此倾向于选择 AIC 小的模型.

当样本量较小的时候一般更正 AIC 为 AICc，即更正后的赤池信息量准则，而 AICc 在样本量增加的时候又会收敛到 AIC，可以证明在任何大小的样本量下都可以使用 AICc，同时还有另一种指标 AICu.

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \doteq \log\left(\frac{RSS}{n}\right) + \frac{n+k}{n-k-2}$$

$$AICu = \log\left(\frac{RSS}{n-k}\right) + \frac{n+k}{n-k-2}$$

BIC

BIC 又称贝叶斯信息量准则，BIC 对过量参数的惩罚比 AIC 更重，AIC 的惩罚是 $2k$ 而 BIC 的惩罚是 $k\log n$，因此应用 BIC 更容易选出一个参数更少的模型.

$$BIC = k\log n + n\log\left(\frac{RSS}{n}\right)$$

其他信息准则

我们还有 FIC、HQC 等准则与多种散度用来解决模型选择问题.

AIC: https://en.wikipedia.org/wiki/Akaike_information_criterion

BIC: https://en.wikipedia.org/wiki/Bayesian_information_criterion

FIC: https://en.wikipedia.org/wiki/Focused_information_criterion

HQC: https://en.wikipedia.org/wiki/Hannan%E2%80%93Quinn_information_criterion

KL 散度: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

JS 散度: https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence

WAIC: https://en.wikipedia.org/wiki/Watanabe%E2%80%93Akaike_information_criterion

8. 统计检验

一般用 F 检验、Wald 检验、似然比检验和拉格朗日乘子检验来假设检验回归系数与 0 是否存在显著差异（实质是对若干约束假设进行的检验），其中 Wald 检验在某些情况下会给出错误的结论，如标准误 SE 较大时（比方数据比较极端，$P(Y=1\mid X)$ 在某点激增）.

这些内容过于复杂了，暂不打算深入讨论，仅简单说明.

F 检验：原假设为 k 个回归系数都为 0，备择假设为回归系数不全为 0，则

$$\frac{\frac{ESS}{k}}{\frac{RSS}{n-k-1}} \overset{\text{L}}{\sim} F(k, n-k-1)$$

F 检验针对线性约束且需要满足随机扰动项满足误差正态分布，对误差分布没有假设时 F 检验失效，

这时 Wald 检验会是个可能的不错选择. Wald 统计量是一个标准化后的二次型，Wald 检验是一致参数统计方法，直接检验了参数的 MLE 与原假设的差异，这与似然比检验、得分检验底层逻辑上不同；更一般的 Wald 检验不仅能用来检验若干回归系数是否显著为 0，设原假设为 $\left(\beta_{i_1},\beta_{i_2},\ldots,\beta_{i_k}\right)=\mathbf{0}$，即约束 $R\beta-q=0$ 成立，$W=\left(R\hat{\beta}-q\right)^T(R\hat{\sigma}^2(X^TX)R^T)^{-1}\left(R\hat{\beta}-q\right)\overset{L}{\sim}\chi^2(k)$，$k$ 为约束个数，$\hat{\sigma}^2$ 为残差方差的估计；特别地，对于检验某一个回归系数是否显著为 0，Wald 检验的渐进分布为

$$\frac{\hat{\beta}_j}{SE_{\hat{\beta}_j}}\overset{L}{\sim}N(0,1)$$

剩下的许多检验，不再赘述.

——对数线性模型——

Logistic 回归一般用来解决仅有一种因变量但其有 2 个或 2 个以上取值的分类问题，如果分类问题的因变量有多个，可以考虑对数线性模型.

Logistic 回归模型考虑了 $P(Y=1\mid X)$ 与协变量间的关系，对数线性模型描述期望频数与协变量间的联系；同 Logistic 回归模型一样地，Logistic 回归模型中为了将介于 0 和 1 的概率取值映射到 $\mathbb{R}$ 做了 Logistic 变换，对数线性模型中为了将大于等于 0 的频数取值映射到 $\mathbb{R}$ 取了其对数.

$$\log m=\beta_0+\beta_1x_1+\ldots+\beta_kx_k$$

Logistic 模型中认为 Y 服从两点分布，对数线性模型中认为 Y 服从泊松分布.

二者都属于**广义线性模型**.

关于广义线性模型的 canonical link function：

Link function [edit]

The link function provides the relationship between the linear predictor and the mean of the distribution function. There are many commonly used link functions, and their choice is informed by several considerations. There is always a well-defined *canonical* link function which is derived from the exponential of the response's density function. However, in some cases it makes sense to try to match the domain of the link function to the range of the distribution function's mean, or use a non-canonical link function for algorithmic purposes, for example Bayesian probit regression.

When using a distribution function with a canonical parameter θ, the canonical link function is the function that expresses θ in terms of μ, i.e. $\theta=b(\mu)$. For the most common distributions, the mean μ is one of the parameters in the standard form of the distribution's density function, and then $b(\mu)$ is the function as defined above that maps the density function into its canonical form. When using the canonical link function, $b(\mu)=\theta=\mathbf{X}\boldsymbol{\beta}$, which allows $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ to be a sufficient statistic for $\boldsymbol{\beta}$.

Following is a table of several exponential-family distributions in common use and the data they are typically used for, along with the canonical link functions and their inverses (sometimes referred to as the mean function, as done here).

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta}=g(\mu)$	Mean function
Normal	real: $(-\infty,+\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta}=\mu$	$\mu=\mathbf{X}\boldsymbol{\beta}$
Exponential Gamma	real: $(0,+\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta}=-\mu^{-1}$	$\mu=-(\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	real: $(0,+\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta}=\mu^{-2}$	$\mu=(\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0,1,2,\ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta}=\ln(\mu)$	$\mu=\exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0,1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{1-\mu}\right)$	$\mu=\frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}=\frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0,1,\ldots,N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0,K)$ K-vector of integer: $[0,1]$, where exactly one element in the vector has the value 1	outcome of single K-way occurrence		$\mathbf{X}\boldsymbol{\beta}=\ln\left(\frac{\mu}{1-\mu}\right)$	
Multinomial	K-vector of integer: $[0,N]$	count of occurrences of different types (1 .. K) out of N total K-way occurrences			

In the cases of the exponential and gamma distributions, the domain of the canonical link function is not the same as the permitted range of the mean. In particular, the linear predictor may be positive, which would give an impossible negative mean. When maximizing the likelihood, precautions must be taken to avoid this. An alternative is to use a noncanonical link function.

In the case of the Bernoulli, binomial, categorical and multinomial distributions, the support of the distributions is not the same type of data as the parameter being predicted. In all of these cases, the predicted parameter is one or more probabilities, i.e. real numbers in the range $[0,1]$. The resulting model is known as *logistic regression* (or *multinomial logistic regression* in the case that K-way rather than binary values are being predicted).

For the Bernoulli and binomial distributions, the parameter is a single probability, indicating the likelihood of occurrence of a single event. The Bernoulli still satisfies the basic condition of the generalized linear model in that, even though a single outcome will always be either 0 or 1, the *expected value* will nonetheless be a real-valued probability, i.e. the probability of occurrence of a "yes" (or 1) outcome. Similarly, in a binomial distribution, the expected value is Np, i.e. the expected proportion of "yes" outcomes will be the probability to be predicted.

For categorical and multinomial distributions, the parameter to be predicted is a K-vector of probabilities, with the further restriction that all probabilities must add up to 1. Each probability indicates the likelihood of occurrence of one of the K possible values. For the multinomial distribution, and for the vector form of the categorical distribution, the expected values of the elements of the vector can be related to the predicted probabilities similarly to the binomial and Bernoulli distributions.

来源：https://en.wikipedia.org/wiki/Generalized_linear_model#Link_function

1. 二维列联表的对数线性模型

记二维列联表期望频数为 $m_{ij}=\mathbb{E}(n_{ij}),\mathrm{i}=1,\ldots,\mathrm{r},\ \mathrm{j}=1,\ldots,\mathrm{c}$，$\mu_{a(i)}$ 代表属性 A 在 A_i 时的效应，$\mu_{b(j)}$ 代表属性 B 在 B_j 时的效应，$\mu_{ab(ij)}$ 代表属性 A 在 A_i 时、属性 B 在 B_j 时的交互作用效应.

① **饱和模型**：$\log m_{ij}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{ab(ij)}$

其中，$\sum_{i=1}^{r}\mu_{a(i)}=\sum_{\mathrm{j}=1}^{\mathrm{c}}\mu_{\mathrm{b(j)}}=0$，$\begin{cases}\sum_{j=1}^{c}\mu_{ab(ij)},\ i=1,\ldots,r\\ \sum_{j=1}^{r}\mu_{ab(ij)},\ j=1,\ldots,c\end{cases}$

② **非饱和模型**：$\log m_{ij}=\mu+\mu_{a(i)}+\mu_{b(j)}$，即不考虑交互效应.

$\log m_{ij} = \mu + \mu_{a(i)}$，即只考虑属性 A 的效应，$A$、$B$ 相互独立且 B 边际分布为均匀分布.

$\log m_{ij} = \mu + \mu_{b(j)}$，即只考虑属性 B 的效应，A、B 相互独立且 B 边际分布为均匀分布.

模型分类	被检验模型	期望频数的估计 $\hat{m}_{ij}$	似然比检验统计量	Pearson χ^2 检验统计量	渐进 χ^2 分布自由度
饱和模型	$\log m_{ij} = \mu + \mu_{a(i)} + \mu_{b(j)} + \mu_{ab(ij)}$	n_{ij}			
非饱和模型	$\log m_{ij} = \mu + \mu_{a(i)} + \mu_{b(j)}$	$\frac{n_{i+}n_{+j}}{n}$	$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\hat{m}_{ij})^2}{\hat{m}_{ij}}$	$-2\log\Lambda = -2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log\left(\frac{\hat{m}_{ij}}{n_{ij}}\right)$	$(r-1)(c-1)$
	$\log m_{ij} = \mu + \mu_{a(i)}$	$\frac{n_{i+}}{c}$			$r(c-1)$
	$\log m_{ij} = \mu + \mu_{b(j)}$	$\frac{n_{+j}}{r}$			$c(r-1)$

2. 高维列联表的对数线性模型

*§7.4 高维列联表的对数线性模型

我们仍以三维列联表为例，介绍高维列联表的对数线性模型. 设三维列联表(见第五章的表5.1)的期望频数为 $m_{ijk}=E(n_{ijk}), i=1,\cdots,r; j=1,\cdots,c; k=1,\cdots,t$. 它的饱和对数线性模型为

$$\ln m_{ijk}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{c(k)}+\mu_{ab(ij)}+\mu_{bc(jk)}+\mu_{ac(ik)}+\mu_{abc(ijk)} \quad (7.4.1)$$

而它的非饱和对数线性模型有：

$$\ln m_{ijk}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{c(k)}, \quad (7.4.2)$$

$$\ln m_{ijk}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{c(k)}+\mu_{bc(jk)}, \quad (7.4.3)$$

$$\ln m_{ijk}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{c(k)}+\mu_{ac(ik)}, \quad (7.4.4)$$

$$\ln m_{ijk}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{c(k)}+\mu_{ab(ij)}, \quad (7.4.5)$$

$$\ln m_{ijk}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{c(k)}+\mu_{ab(ij)}+\mu_{ac(ik)}, \quad (7.4.6)$$

$$\ln m_{ijk}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{c(k)}+\mu_{ab(ij)}+\mu_{bc(jk)}, \quad (7.4.7)$$

$$\ln m_{ijk}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{c(k)}+\mu_{bc(jk)}+\mu_{ac(ik)}, \quad (7.4.8)$$

$$\ln m_{ijk}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{c(k)}+\mu_{ab(ij)}+\mu_{bc(jk)}+\mu_{ac(ik)} \quad (7.4.9)$$

其中 μ 是总的平均，$\mu_{a(i)}$、$\mu_{b(j)}$ 与 $\mu_{c(k)}$ 分别是属性 A 在 A_i 时、属性 B 在 B_j 时和属性 C 在 C_k 时的效应，$\mu_{ab(ij)}$、$\mu_{bc(jk)}$ 与 $\mu_{ac(ik)}$ 分别是属性 A 在 A_i 和属性 B 在 B_j 时、属性 B 在 B_j 和属性 C 在 C_k 时与属性 A 在 A_i 和属性 C 在 C_k 时的二次交互作用效应，而 $\mu_{abc(ijk)}$ 是属性 A 在 A_i、属性 B 在 B_j 和属性 C 在 C_k 时的三次交互作用效应. 类似于(7.3.2)、(7.3.3)和(7.3.4)等式，它们满足条件：

$$\sum_{i=1}^{r}\mu_{a(i)}=\sum_{j=1}^{c}\mu_{b(j)}=\sum_{k=1}^{t}\mu_{c(k)}=0,$$

$$\sum_{j=1}^{c}\mu_{ab(ij)}=\sum_{k=1}^{t}\mu_{ac(ik)}=0,\quad i=1,\cdots,r,$$

$$\sum_{i=1}^{r}\mu_{ab(ij)}=\sum_{k=1}^{t}\mu_{bc(jk)}=0,\quad j=1,\cdots,c,$$

$$\sum_{i=1}^{r}\mu_{ac(ik)}=\sum_{j=1}^{c}\mu_{bc(jk)}=0,\quad k=1,\cdots,t,$$

$$\sum_{i=1}^{r}\mu_{abc(ijk)}=0,j=1,\cdots,c,\quad k=1,\cdots,t,$$

$$\sum_{j=1}^{c}\mu_{abc(ijk)}=0,i=1,\cdots,r,\quad k=1,\cdots,t,$$

$$\sum_{k=1}^{t}\mu_{abc(ijk)}=0,i=1,\cdots,r,\quad j=1,\cdots,c$$

首先考察(7.4.2)式.它是可加模型,由第五章的(5.6.1)式知,可加模型等价于属性A、B和C相互独立.第五章的(5.6.2)式给出了期望频数m_{ijk}的估计.

然后考察(7.4.3)、(7.4.4)与(7.4.5)式.它们都只有一个二次交互作用效应,由第五章的(5.6.3)式和表5.38知,只有一个二次交互作用效应的(7.4.3)、(7.4.4)与(7.4.5)式模型分别等价于属性A和(B,C)相互独立、属性B和(A,C)相互独立与属性C和(A,B)相互独立.第五章的(5.6.4)式和表5.38给出了它们的期望频数m_{ijk}的估计.

接下来考察(7.4.6)、(7.4.7)与(7.4.8)式.它们都只有二个二次交互作用效应,由第五章的(5.6.6)式和表5.38知,只有二个二次交互作用效应的(7.4.6)、(7.4.7)与(7.4.8)式模型分别等价于属性A给定后B和C条件独立、属性B给定后A和C条件独立与属性C给定后A和B条件独立.第五章的(5.6.7)式和表5.38给出了它们的期望频数m_{ijk}的估计.

上面考察的这些模型给出了三维列联表的各种独立性,而最后考察的(7.4.9)式有三个二次交互作用效应,它给出了由第五章的(5.6.10)式所表示的三维列联表的相关模型.通常用第五章§5.6.2小节所给出的迭代算法计算这个相关模型的期望频数的估计.我们还可以用统计软件包,例如SAS,计算此相关模型的期望频数的估计.事实上,软件也是用迭代方法求解方程组的.显然,手工进行迭代计算,非常琐碎冗长,而用统计软件求解既快又精确.

第五章的(5.6.8)式和(5.6.9)式统一地给出了这些模型是否成立的似然比检验统计量和Pearson χ^2检验统计量.它们的渐近χ^2分布的自由度,以及期望频数m_{ijk}的估计$\hat{m}_{ijk}$的计算公式,或计算方法见表7.3.

表 7.3　三维列联表对数线性模型的检验问题

被检验的模型		期望频数 m_{ijk} 的估计	渐近 χ^2 分布的自由度
可加模型(7.4.2)		$n_{i++}n_{+j+}n_{++k}/n^2$	$rct-r-c-t+2$
只有一个二次交互作用效应的模型	(7.4.3)	$n_{i+k}n_{+jk}/n$	$(r-1)(ct-1)$
	(7.4.4)	$n_{+j+}n_{i+k}/n$	$(c-1)(rt-1)$
	(7.4.5)	$n_{++k}n_{ij+}/n$	$(t-1)(rc-1)$
有二个二次交互作用效应的模型	(7.4.6)	$n_{ij+}n_{i+k}/n_{i++}$	$r(c-1)(t-1)$
	(7.4.7)	$n_{ij+}n_{+jk}/n_{+j+}$	$c(r-1)(t-1)$
	(7.4.8)	$n_{i+k}n_{+jk}/n_{++k}$	$t(r-1)(c-1)$
相关模型(7.4.9)		迭代计算	$(r-1)(c-1)(t-1)$

二维列联表 $m_{ij}=E(n_{ij})$

令 $\mu_{ij}=\ln m_{ij}$ $i=1,\cdots,r$, $j=1,\cdots,c$

$\mu=\frac{1}{rc}\sum_{i}\sum_{j}\mu_{ij}$ 为对数 $\mu_{i+}=\frac{1}{c}\sum_{j=1}^{c}\mu_{ij}$ 行均值 $\mu_{+j}=\frac{1}{r}\sum_{i=1}^{r}\mu_{ij}$ 列均值.

令 $\mu_{a(i)}=\mu_{i+}-\mu$ $\mu_{b(j)}=\mu_{+j}-\mu$ $\mu_{ab(ij)}=\mu_{ij}-\mu_{i+}-\mu_{+j}+\mu$

$$\begin{cases}\mu_{ij}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{ab(ij)} \\ \ln m_{ij}=\mu+\mu_{a(i)}+\mu_{b(j)}+\mu_{ab(ij)}\end{cases}$$ 饱和模型形式

$\sum_{i}\mu_{a(i)}=0$, $\sum_{j}\mu_{b(j)}$ $\sum_{i}\mu_{ab(ij)}=\sum_{j}\mu_{ab(ij)}=0$

由 $\sum_{i=1}^{r}\mu_{a(i)}=0 \Rightarrow$ $r-1$ 个独立行效应参数.

$\sum_{j=1}^{c}\mu_{b(j)}=0 \Rightarrow$ $c-1$ 列

$\sum_{i}\mu_{ab(ij)}=\sum_{j}\mu_{ab(ij)}=0 \Rightarrow (r-1)(c-1)$ 个交互效应独立参数

模型②：$1+r-1+c-1+(r-1)(c-1)=rc$ 列联表单元格总数

饱和模型

① $1+r-1+c-1=r+c-1$ 个参数 饱和

$\mu_{a(i)}>0 \Rightarrow$ 第 i 行 > 整个列联表的总体

$\mu_{i+}-\mu$ $\mu_{b(j)}>0$

$\mu_{ab}(ij)$：变量的交互效应，反映了变量偏离变量独立的程度

$\mu_a(i)$，$\mu_b(j)$

§3 饱和模型　　行分布

$p_{+1}=\cdots=p_{+c}=\frac{1}{c}$

③ $\ln m_{ij}=\mu+\mu_a(i) \Rightarrow$ A与B独立，B的边际分布是均匀分布

④ $\ln m_{ij}=\mu+\mu_b(j) \Rightarrow$ A与B独立，A的边际……

$p_{1+}=\cdots=p_{r+}=\frac{1}{r}$

① $\hat{m}_{ij}=\frac{n_{i+}\cdot n_{+j}}{n}$　　$\chi^2=\sum\sum\frac{(n_{ij}-\hat{m}_{ij})^2}{\hat{m}_{ij}} \xrightarrow{L} \chi^2((r-1)(c-1))$

$-2\ln\Lambda=-2\sum\sum n_{ij}\ln\left(\frac{\hat{m}_{ij}}{n_{ij}}\right) \to$

③ $\hat{m}_{ij}=\frac{n_{i+}}{c}$　　$\chi^2(r(c-1))$

④ $\hat{m}_{ij}=\frac{n_{+j}}{r}$　　$\chi^2(c(r-1))$

三维列联表　　$A\times B\times C$ $(r\times c\times t)$

$E(n_{ijk})=m_{ijk}$.　　$\ln m_{ijk}=\mu_{ijk}$

$\mu=\frac{1}{rct}\sum\sum\sum\mu_{ijk}$　　$\mu_{ij+}=\frac{1}{t}\sum_k\mu_{ijk}$　　$\mu_a(i)=\mu_{i++}-\mu$

$\mu_{i++}=\frac{1}{ct}\sum_j\sum_k\mu_{ijk}$　　$\mu_{i+k}=\frac{1}{c}\sum\mu_{ijk}$　　$\mu_b(j)=\mu_{+j+}-\mu$

$\mu_{+j+}=\frac{1}{rt}\sum_i\sum_k\mu_{ijk}$　　$\mu_{+jk}=\frac{1}{r}\sum\mu_{ijk}$　　$\mu_c(k)=\mu_{++k}-\mu$

$\mu_{++k}=\frac{1}{rc}\sum_i\sum_j\mu_{ijk}$　　$\mu_{ab}(ij)=\mu_{ij+}-\mu_{i++}-\mu_{+j+}+\mu$

$$\mu_{abc}(ijk) = \mu_{ijk} - \mu_{ij+} - \mu_{i+k} - \mu_{+jk} + \mu_{i++} + \mu_{+j+} + \mu_{++k} - \mu$$

$$\Rightarrow \mu_{ijk} = \mu + \mu_a(i) + \mu_b(j) + \mu_c(k) + \mu_{ab}(ij) + \mu_{ac}(ik) + \mu_{bc}(jk) + \mu_{abc}(ijk) \quad (1)$$

$\sum_i \mu_a(i) = 0$ $\sum_j \mu_b(j) = 0$ $\sum_k \mu_c(k) = 0$

$\sum \mu_{ab}(ij) = 0$ $\sum \mu_{ac}(ik) = 0$ $\sum \mu_{bc}(jk) = 0$ $\sum \mu_{abc}(ijk) = 0$

$(r-1)(c-1)$ $(r-1)(t-1)$ $(c-1)(t-1)$ $(r-1)(c-1)(t-1)$

$r-1$ $c-1$ $t-1$

$$1 + (r-1) + (c-1) + (t-1) + (r-1)(c-1) + (r-1)(t-1) + (c-1)(t-1) + (r-1)(c-1)(t-1)$$

$$= rct$$

Solution: (A, B, C)

② $\ln m_{ijk} = \mu + \mu_a(i) + \mu_b(j) + \mu_c(k)$

① $\mu_{ijk} = \mu + \mu_a(i) + \mu_b(j) + \mu_c(k) + \mu_{ab}(ij) \Rightarrow$ C 与 (A, B) 独立.

④ $\mu_{ijk} = \mu + \mu_a(i) + \mu_b(j) + \mu_c(k) + \mu_{ac}(ik) \Rightarrow$ B 与 (A, C) 独立

⑤ $\mu_{ijk} = \mu + \mu_a(i) + \mu_b(j) + \mu_c(k) + \mu_{bc}(jk) \Rightarrow$ A 与 (B, C) 独立.

⑥ $\mu_{ijk} = \mu + \mu_a(i) + \mu_b(j) + \mu_c(k) + \mu_{ab}(ij) + \mu_{ac}(ik) \Rightarrow$ A 给定后, B 与 C 条件独立

⑦ $\mu_{ijk} = \cdots + \mu_{ab}(ij) + \mu_{bc}(jk) \Rightarrow$ B 给定后, A 与 C 条件独立

⑧ $\mu_{ijk} = \cdots + \mu_{ac}(ik) + \mu_{bc}(jk) \Rightarrow$ C 给定后, A 与 B 条件独立

⑨ $\mu_{ijk} = \mu + \mu_a(i) + \mu_b(j) + \mu_c(k) + \mu_{ab}(ij) + \mu_{ac}(ik) + \mu_{bc}(jk)$