第七节 · 参数估计

司继春

上海对外经贸大学统计与信息学院

在这一节中我们将学习统计推断中参数估计的相关内容，包括点估计和区间估计两部分。其中，我们将介绍两种点估计的方法：矩估计和极大似然估计。

1 参数估计

1.1 参数估计的基本概念

在参数模型中，我们假设总体 P 属于某一个参数族 $\{P_\theta, \theta \in \Theta\}$，从而推断总体等价于找到一个参数 θ_0，使得 $P_{\theta_0} = P$。我们一般把 θ_0 成为真值（true value）。而由于总体是不可观测的，我们只能通过样本对总体进行推断，因而我们不可能得到 θ_0 的精确值，只能对其进行估计，即**参数估计**（**Estimation**）。

参数估计包含两部分，即**点估计**（**Point estimation**）和**区间估计**（**interval estimation**）。其中点估计即找到一个统计量 $\hat{\theta}(x)$，对总体参数 θ_0 进行推断，而统计量 $\hat{\theta}$ 我们一般称为**估计量**（**estimator**）。而区间估计即找到一组统计量 $L(x), U(x)$，使得由其组成的区间包含总体参数 θ_0 的概率为已知的，即 $P(L(x) \leq \theta_0 \leq U(x)) = p$。其中统计量 $L(x), U(x)$ 被成为**区间估计量**（**interval estimator**）。

此外，类似于统计量及其实现的差别，我们还需要区分估计量和估计。如上所述，估计量即样本的一个函数，即用于估计参数的统计量，而估计（estimate）是对于某一个样本，估计量的实现。

1.2 评价估计量的标准

对于同一个参数，经常我们有不同的估计量，比如对于正态总体 $x_i \sim N\left(\mu, \sigma^2\right)\ i.i.d$，自然地，$\sigma^2$ 的一个估计量为：

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

而类似的，我们也可以使用：

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{1}$$

作为 σ^2 的一个估计。以上两个估计量的差别在于分母的不同，很显然，以上两个估计量具有不同的抽样分布。那么，在有很多统计量可供选择时，该如何评价这些统计量呢？

一个常用的标准是**均方误差**（**Mean squared error**, MSE），即对于一个参数 θ 和它的估计量 $\hat{\theta}$，其误差平方的期望 $\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2$ 为估计量 $\hat{\theta}$ 的均方误差。注意由于：

$$\begin{aligned}
\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2 &= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}+\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\
&= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)^2 + \mathbb{E}\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2 + 2\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)\left(\mathbb{E}\hat{\theta}-\theta_0\right) \\
&= \mathrm{Var}\left(\hat{\theta}\right) + \left(\mathbb{E}\hat{\theta}-\theta_0\right)^2 + 2\left(\mathbb{E}\hat{\theta}-\theta_0\right)\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right) \\
&= \mathrm{Var}\left(\hat{\theta}\right) + \left(\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\
&= \mathrm{Var}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2
\end{aligned}$$

其中定义**偏差**（**bias**）$\mathrm{Bias}\left(\hat{\theta}\right) = \mathbb{E}\left(\hat{\theta}-\theta_0\right)$，从而 $\mathrm{MSE}\left(\hat{\theta}\right) = \mathrm{Var}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2$，即均方误差等于估计量的方差与偏差平方的和。

因而，降低均方误差有两种途径：降低估计的方差以及降低偏差。此外，根据均方收敛的定义，只要 $\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2 = \mathrm{Var}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2 \to 0$，那么 $\hat{\theta} \xrightarrow{L^2} \theta_0$，从而 $\hat{\theta} \xrightarrow{p} \theta_0$。尽管小样本情况下估计量的偏差不为 0，但是我们希望当样本量趋向于无穷时，估计量收敛到真值，也是可以接受的。这就引出了评价估计量的三条标准：**无偏性**（**unbiasedness**）、**有效性**（**efficiency**）、**一致性**（**consistency**）。

1.2.1 无偏性

无偏性要求估计量的偏差为 0。当估计量的偏差为 0，即 $\mathbb{E}\left(\hat{\theta}\right) = \theta_0$ 时，我们称估计量 $\hat{\theta}$ 为无偏的（unbiased）。无偏性意味着，尽管对于每个样本，对 θ_0 的估计不可能完全准确，但是平均而言，估计量 $\hat{\theta}$ 总是围绕在真值 θ_0 的周围，不会有系统性的偏差。

例 1. 如果 $\mathbb{E}x_i = \mu$，那么样本均值的期望：

$$\mathbb{E}\left(\bar{x}\right) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right) = \mu$$

因而 $\bar{x}$ 是总体均值 μ 的无偏估计。

例 2. 根据之前的计算，我们知道对于 $x_i \sim \left(\mu, \sigma^2\right)$ $i.i.d$:

$$\mathbb{E}\left(s^2\right) = \sigma^2$$

因而式 (1) 中定义的 $\hat{\sigma}^2$ 的期望为:

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \mathbb{E}\left(\frac{N-1}{N}s^2\right) = \frac{N-1}{N}\sigma^2$$

因而 $\text{Bias}\left(s^2\right) = 0, \text{Bias}\left(\hat{\sigma}^2\right) = \frac{1}{N}\sigma^2$，只有 s^2 是 σ^2 的无偏估计量。

1.2.2 有效性

为了降低 MSE，除了降低偏差以外，降低估计量的方差 $\text{Var}\left(\hat{\theta}\right)$ 也是非常重要的手段。一般而言，如果两个估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$，如果 $\text{Var}\left(\hat{\theta}_1\right) < \text{Var}\left(\hat{\theta}_2\right)$，那么我们称 $\hat{\theta}_1$ 相对于 $\hat{\theta}_2$ 是有效的。

例 3. 在例 (2) 中，因为 $\hat{\sigma}^2 = \frac{N-1}{N}s^2$，从而

$$\text{Var}\left(\hat{\sigma}^2\right) = \left(\frac{N-1}{N}\right)^2 \text{Var}\left(s^2\right) < \text{Var}\left(s^2\right)$$

因而 $\hat{\sigma}^2$ 是相对于 s^2 更有效的估计量。

我们注意到，尽管 s^2 比 $\hat{\sigma}^2$ 的偏差更小，但是 s^2 比 $\hat{\sigma}^2$ 的方差更大。在很多应用问题，比如非参数回归或者监督学习（supervised learning）中，我们都会碰到类似的偏差-方差权衡（bias-variance tradeoff），即很多时候，同时降低偏差和方差是不可能的。

1.2.3 一致性

很多时候，尽管在有限样本下，一个估计量的偏差不为零，但是如果样本量足够大时，估计量与真值之间的误差充分的小，我们也可以接受。如果一个估计量 $\hat{\theta}$ 依概率收敛到真值 θ_0，即 $\hat{\theta} \xrightarrow{p} \theta_0$，那么我们称估计量 $\hat{\theta}$ 为一致估计量。如果一个估计量是不一致的，也就是说即便我们拥有无限多的样本，我们也不能获得真值 θ_0 的估计，因而一致性是对一个估计量的最低要求。

例 4. 在例 (1) 中，如果对样本 $\{x_i\}$ 做额外的假设（比如 x_i 独立同分布且可积），那么根据大数定律，有

$$\bar{x} \xrightarrow{p} \mathbb{E}\left(x_i\right) = \mu$$

因而样本均值是总体均值的一致估计量。

例 5. 在例 (2) 中，由于：

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2$$

其中 $\bar{x} \xrightarrow{p} \mu$，而：

$$\frac{1}{N}\sum_{i=1}^{N}x_i^2 \xrightarrow{p} \mathbb{E}\left(x^2\right) = \mu^2 + \sigma^2$$

从而 $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$，即 $\hat{\sigma}^2$ 是 σ^2 的一致估计量。而：

$$s^2 = \frac{N}{N-1}\hat{\sigma}^2 \xrightarrow{p} \sigma^2$$

因而 s^2 也是 σ^2 的一致估计量。

需要注意的是，无偏性关注的是估计量的期望，而一致性则是当样本足够大时估计量的性质，两者并没有任何必然联系，无偏性和一致性并不是彼此的充分或者必要条件。

2 区间估计

在上一节中，无偏性、一致性都是使用单个估计量（如样本均值、样本方差）对未知参数进行估计，这种估计被称为点估计（point estimation）。尽管我们可以使用点估计方法对参数值进行推断，然而我们知道，参数的点估计值 $\hat{\theta}$ 与真值 θ_0 相等的概率一般为 0，即 $P\left(\hat{\theta} = \theta_0\right) = 0$。因而更进一步的，我们很多时候希望得到一个区间，使得这个区间能够以正的概率包含真值 θ_0。这就诞生了区间估计（interval estimation）的概念。

区间估计，即对于样本 $x = (x_1, ..., x_N)$，通过一对统计量 $L(x)$ 和 $U(x)$，满足 $L(x) \leq U(x)$，我们可以使用区间 $[L(x), U(x)]$ 对未知参数 θ_0 进行推断。

例 6. 如果样本 $x_i \sim N(\mu_0, 1)$ $i.i.d, i = 1, ..., N$，那么区间 $[\bar{x} - 0.5, \bar{x} + 0.5]$ 包含真值 μ_0 的概率为：

$$\begin{aligned} P(\mu_0 \in [\bar{x} - 0.5, \bar{x} + 0.5]) &= P(\mu_0 - 0.5 \leq \bar{x} \leq \mu_0 + 0.5) \\ &= P\left(-\frac{0.5}{\sqrt{\frac{1}{N}}} \leq \frac{\bar{x} - \mu_0}{\sqrt{\frac{1}{N}}} \leq \frac{0.5}{\sqrt{\frac{1}{N}}}\right) \end{aligned}$$

由于 $\bar{x} \sim N\left(\mu_0, \frac{1}{N}\right)$，因而

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{1}{N}}} \sim N(0, 1)$$

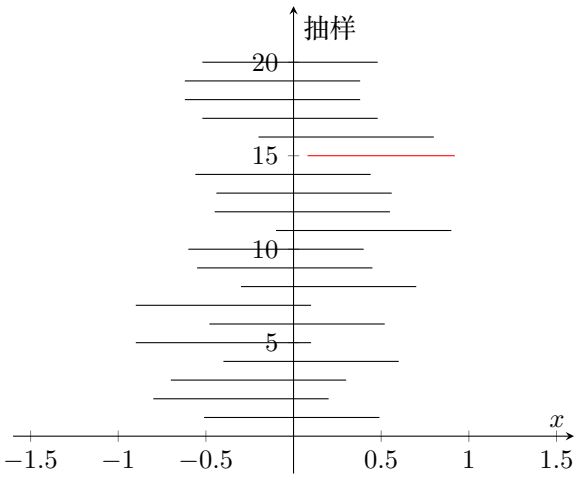

图 1: 不同抽样下长度为 1 的置信区间估计

因而

$$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = \Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - \Phi\left(-\frac{0.5}{\sqrt{\frac{1}{N}}}\right) = 2\Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - 1$$

例如，当 $N = 16$ 时，查表可得，$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = 2\Phi(2) - 1 \approx 2\times 0.9772 - 1 = 0.9544$，即区间 $[\bar{x}-0.5, \bar{x}+0.5]$ 包含真值 μ_0 的概率为 95.44%。如图 (1) 画出了当 $\mu_0 = 0$ 时，20 次不同抽样的区间。平均而言，每抽 100 次大概有 5 次区间 $[\bar{x}-0.5, \bar{x}+0.5]$ 不能包含真实的参数 μ_0（图中红色区间）。

注意由于未知参数 θ_0 是一个未知的常数，而统计量 $L(x)$ 和 $U(x)$ 是随着抽样的变化而变化的，因此我们不能说「θ_0 落入区间 $[\bar{x}-0.5, \bar{x}+0.5]$ 的概率是多少」，而只能说「区间 $[\bar{x}-0.5, \bar{x}+0.5]$ 包含 θ_0 的概率是多少」。我们把概率 $P\left(\theta_0 \in [L(x), U(x)]\right)$ 称为覆盖概率（coverage probability）。注意由于总体参数 θ_0 未知，因而概率 $P\left(\mu_0 \in [L(x), U(x)]\right)$ 可能依赖于未知的参数 θ_0，因而我们通常将而覆盖概率的下界，即 $\inf_\theta P_\theta\left(\theta_0 \in [L(x), U(x)]\right)$ 称为**置信度**（**confidence coefficient**）或者**置信水平**，通常用 $1-\alpha$ 表示。在某一置信度下，区间 $[L(x), U(x)]$ 又被称为**置信区间**（**confidence interval**）。因而在例 (6) 中，我们可以说在 95.44% 的置信水平下，置信区间为 $[\bar{x}-0.5, \bar{x}+0.5]$。

此外还需要注意的是，在例 (6) 中，为了求得置信区间和覆盖概率，我们首先将统计量 $\bar{x}$ 做了标准化处理，即使用 $\frac{\bar{x}-\mu_0}{\sqrt{\frac{1}{N}}}$ 推算概率，而不是直接使用 $\bar{x}$。使用 $\frac{\bar{x}-\mu_0}{\sqrt{\frac{1}{N}}}$ 的好处是，此统计量不依赖于任何未知参数，因而其分布不会随着未知参数的变化而变化，即服从一个标准的分布，这样一来，我们得到的覆盖概率不依赖于任何未知参数，因而就等于置信度。一般的，我们把分布不依赖于未知参数的统计量成为**基准统计量**（**pivotal statistic**）。

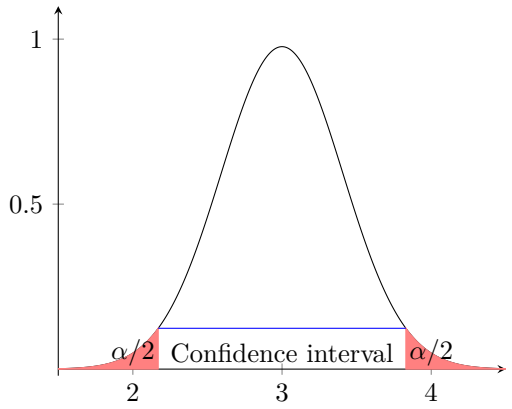

图 2: 正态分布或 t 分布置信区间

例 7. 如果样本 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i = 1, ..., N$，那么：

1. 统计量 $\bar{x} \sim N\left(\mu_0, \frac{\sigma_0^2}{N}\right)$，其分布依赖于两个未知参数；
2. 统计量 $\bar{x} - \mu_0 \sim N\left(0, \frac{\sigma_0^2}{N}\right)$，其分布仍然依赖于未知参数 μ_0；
3. 统计量
$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma_0^2}{N}}} \sim N(0, 1)$$
分布不依赖于任何未知参数，因而是基准统计量；
4. 统计量
$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \sim t(N-1)$$
分布不依赖于任何未知参数，因而是基准统计量；

例 8. 如果样本 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i = 1, ..., N$，那么

$$(N-1)\frac{s^2}{\sigma_0^2} \sim \chi^2(N-1)$$

，其分布不依赖于任何未知参数，因而是基准统计量。

在例 (6) 中，我们首先给出了区间，进而计算了该区间的置信度。然而现实中，我们经常希望得到在一定置信水平下的置信区间，即一般的区间估计过程。

例 9. 如果样本 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i = 1, ..., N$，为了得到 μ_0 的 95% 的置信区间，我们首先找到基准统计量，要求在基准统计量中，只有 μ_0 是未知的，其他都是已知的（包括已知常数以及已知统计量）。在例 (7) 中，只有统计量

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \sim t(N-1)$$

满足以上条件。如果令 $t_{\alpha/2}=F_t^{-1}\left(\frac{\alpha}{2}\right)$，我们有：

$$\begin{aligned}P\left(-t_{\alpha/2}\leq\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}}\leq t_{\alpha/2}\right)&=F_t\left(t_{\alpha/2}\right)-F_t\left(-t_{\alpha/2}\right)\\&=1-2F_t\left(t_{\alpha/2}\right)\\&=1-2F_t\left(F_t^{-1}\left(\frac{\alpha}{2}\right)\right)\\&=1-\alpha\end{aligned}$$

因而我们可以得到：

$$P\left(\bar{x}-t_{\alpha/2}\sqrt{\frac{s^2}{N}}\leq\mu_0\leq\bar{x}+t_{\alpha/2}\sqrt{\frac{s^2}{N}}\right)=1-\alpha$$

从而 $\left[\bar{x}-t_{\alpha/2}\sqrt{\frac{s^2}{N}},\bar{x}+t_{\alpha/2}\sqrt{\frac{s^2}{N}}\right]$ 就是我们想要的置信区间。例如，对于一个 $N=30$ 的正态样本，$\bar{x}=3$，$s^2=5$，如果我们想要得到 95% 置信水平下的置信区间，查表$(d.f.=29)$得到 $t_{\alpha/2}=2.0452$，因而置信下界为 $3-2.0452\times\sqrt{5/30}\approx 2.17$，置信上界为 $3+2.0452\times\sqrt{5/30}\approx 3.83$。如图 (2) 所示，其中红色区域为左右两个概率为 $\alpha/2$ 的区域，中间的一块即为所要求的置信区间。

例 10. 如果样本 $x_i\sim N\left(\mu_0,\sigma_0^2\right)\ i.i.d, i=1,...,N$，为了得到 σ_0^2 的 95% 的置信区间，我们首先找到基准统计量，要求在基准统计量中，只有 σ_0^2 是未知的，其他都是已知的。在例 (8) 中，统计量

$$\left(N-1\right)\frac{s^2}{\sigma_0^2}\sim\chi^2\left(N-1\right)$$

满足以上条件。如果令 $\chi^2_{\alpha/2}=F^{-1}_{\chi^2}\left(\frac{\alpha}{2}\right)$，$\chi^2_{1-\alpha/2}=F^{-1}_{\chi^2}\left(1-\frac{\alpha}{2}\right)$，我们有：

$$P\left(\chi^2_{\alpha/2}\leq\left(N-1\right)\frac{s^2}{\sigma_0^2}\leq\chi^2_{1-\alpha/2}\right)=1-\alpha$$

因而：

$$P\left(\frac{\left(N-1\right)s^2}{\chi^2_{1-\alpha/2}}\leq\sigma_0^2\leq\frac{\left(N-1\right)s^2}{\chi^2_{\alpha/2}}\right)=1-\alpha$$

σ_0^2 的 95% 的置信区间为：$\left[\frac{(N-1)s^2}{\chi^2_{1-\alpha/2}},\frac{(N-1)s^2}{\chi^2_{\alpha/2}}\right]$。

总结上述两个置信区间的计算，一般而言我们得到置信区间的步骤如下：

1. 给定置信度 $1-\alpha$；
2. 找到一个基准统计量，其中只有所要求的参数是未知的，其他都是已知的；

3. 找到这个基准统计量的分布函数 $F(\cdot)$;

4. 查表或使用计算机计算 $F^{-1}\left(\frac{\alpha}{2}\right)$ 以及 $F^{-1}\left(1-\frac{\alpha}{2}\right)$;

5. 通过不等式变换得到置信区间。

因而，计算置信区间最关键的步骤即找到基准统计量，并得到此统计量的分布。

尽管上两例给出了正态总体的均值和方差的置信区间的计算方法，然而很多时候我们的总体并不是一定来自于正态总体，很多时候我们很难计算在非正态总体下样本均值的精确分布。然而根据中心极限定理，在一定条件下，有:

$$\sqrt{N}\left(\bar{x}-\mu_0\right) \overset{a}{\sim} N\left(0, \operatorname{Var}(x)\right)$$

因而大样本条件下，我们可以使用中心极限定理近似样本均值的分布，从而得到区间估计。

例 11. 根据 2009 年中国城镇住户调查，在 37480 户家庭中，已知家庭年收入均值为 54157.63 元，标准差为 38533.96 元，那么全国家庭家庭平均收入的 95% 置信区间是多少? 首先一般而言，收入一般不服从正态分布，但是在大样本条件下，我们知道样本均值近似服从正态分布，因而:

$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \overset{a}{\sim} N(0,1)$$

为基准统计量。如果记 $z_{\alpha/2}=\Phi_t^{-1}\left(\frac{\alpha}{2}\right)$，查表知 $z_{2.5\%}=1.96$，因而置信下界为: $54157.63-1.96\times\sqrt{\frac{38533.96^2}{2963}}\approx 53766.88$，同理置信上界约为 54547.12，因而全国家庭家庭平均收入的 95% 置信区间为 $[53766.88, 54547.12]$。

例 12. 根据 2013 年中国家庭金融调查，样本 7711 户家庭中，有 6% 的家庭有信用卡，请问全国持有信用卡的家庭比例的 95% 置信区间是多少? 同样的，比例一般不服从正态分布，但是如果把每个家庭是否持有信用卡假设为独立同分布的伯努利分布，即 $x_i\sim Ber(p_0)$，那么 $x_i^2=x_i$，因而 $\overline{x^2}=\bar{x}$，从而:

$$\frac{s^2}{N}=\frac{N-1}{N}\frac{\overline{x^2}-\bar{x}^2}{N}\approx\frac{\overline{x^2}-\bar{x}^2}{N}=\frac{\bar{x}-\bar{x}^2}{N}=\frac{\bar{x}(1-\bar{x})}{N}$$

其中比例 $\hat{p}=\bar{x}$，从而

$$\frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} \overset{a}{\sim} N(0,1)$$

从而置信下界为 $\hat{p}-z_{2.5\%}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}=0.06-1.96\times\sqrt{\frac{0.06\times(1-0.06)}{7711}}\approx 5.47\%$，同理置信上界约为 6.53%，因而全国家庭持有信用卡比例的 95% 置信区间为 $[5.47\%, 6.53\%]$。

此外，很多时候我们还对两个样本的差值感兴趣。如果假设两个独立的样本 x_1 和 x_2，其均值分别为 $\bar{x}_1$ 和 $\bar{x}_2$，且 $x_{1i}\sim N\left(\mu_1,\sigma_1^2\right)$，$x_{2i}\sim N\left(\mu_2,\sigma_2^2\right)$，

那么 $\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{N_1}\right)$，$\bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{N_2}\right)$，其差值：

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)$$

因而可以使用以上分布对 $\mu_1 - \mu_2$ 进行区间估计。即使两个样本不来自于正态总体，仍然可以使用上面的中心极限定理，通过渐进正态性得到相似的结论。

例 13. 在 2009 年中国城镇住户调查中，共有 23440 位 20-50 岁的男性，以及 21184 位 20-50 岁的女性。已知男性年平均收入为 28367.96 元，标准差为 21811.88 元；女性年平均收入为 20145.77 元，标准差为 16541.08 元。如果假设男女收入独立，请问男女收入差异的 95% 置信区间是多少？同上，尽管收入不服从正态分布，但是大样本情况下可以使用正态分布近似。从而：

$$\bar{x}_1 - \bar{x}_2 \overset{a}{\sim} N\left(28367.96 - 20145.77, \frac{21811.88^2}{23440} + \frac{16541.08^2}{21184}\right)$$

因而其差值的置信区间为 $[7864.99, 8579.38]$。

最后，根据区间估计，我们还能确定为了达到某一精度所需要样本量的大小。根据中心极限定理，在一定条件下，有：

$$\sqrt{N}\left(\bar{x} - \mu_0\right) \overset{a}{\sim} N\left(0, \mathrm{Var}\left(x\right)\right)$$

因而大样本条件下，均值在 $1-\alpha$ 置信水平下的置信区间为：

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right]$$

区间长度应为：$\frac{2z_{\alpha/2}\sigma}{\sqrt{N}}$。可以看到，区间大小随着样本量的增加而减小。如果我们要求在 $1-\alpha$ 置信水平下的置信区间的长度为 l，那么样本量应为 $N = \left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2$，即样本量与精度成二次方关系。

例 14. 在例 (12) 中，为了使 95% 置信区间长度不超过 1%，需要的样本量为：

$$N = \left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2 = \left[\frac{2 \times 1.96 \times \sqrt{0.06 \times (1 - 0.06)}}{0.01}\right]^2 \approx 8667$$

即需要 8667 户样本。然而实际情况中，我们不太可能知道 σ，所以有时预调研是非常重要的。不过在这个例子中，我们可以得到一个样本量的上界：

$$N = \left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2 = \left[\frac{2z_{\alpha/2}\sqrt{p\left(1-p\right)}}{l}\right]^2 \leq \left[\frac{2z_{\alpha/2}\sqrt{0.5\left(1-0.5\right)}}{l}\right]^2 = \left[\frac{z_{\alpha/2}}{l}\right]^2$$

即在这个例子中，如果我们不知道持有信用卡的家庭约为 6%，那么需要的样本

数量上界为 38416 户家庭。

3 矩估计

矩估计（**method of moments**）是使用历史最长的参数估计方法，其思路是使用样本矩代替总体矩对参数进行估计。

接下来我们将介绍经典的矩估计方法，并对此方法做进一步推广。

3.1 经典矩估计

如果样本 $x=(x_1,...,x_N)'$ 是来自于总体 P_{θ_0} 的独立同分布的样本，那么其一阶样本矩和一阶总体矩可以分别定义为：

$$\begin{cases} m_1(x)=\frac{1}{N}\sum_{i=1}^{N}x_i \\ \mu_1(\theta)=\mathbb{E}_\theta x_i \end{cases}$$

其中 $\mathbb{E}_\theta$ 表示给定一个参数 θ，使用总体 P_θ 计算得到的理论的总体期望。由于真值为 θ_0，因而真实的期望 $\mathbb{E}x_i=\mathbb{E}_{\theta_0}x_i$。

我们知道，在一定比较宽松的条件下，根据大数定律有：

$$m_1(x)=\frac{1}{N}\sum_{i=1}^{N}x_i \xrightarrow{p} \mathbb{E}_{\theta_0}x_i=\mu_1(\theta_0)$$

如果 $\mu_1(\cdot)$ 是一个连续且可逆的函数，那么真实参数 θ_0 可以写为：

$$\theta_0=\mu_1^{-1}(\mu_1(\theta_0))$$

那么我们可以使用样本矩 $m_1(x)$ 代替上式中的总体矩 $\mu_1(\theta_0)$，由于 $m_1(x)\xrightarrow{p}\mu_1(\theta_0)$，而 $\mu_1^{-1}(\cdot)$ 为连续函数，从而估计量：

$$\hat{\theta}\triangleq\mu_1^{-1}(m_1(x))\xrightarrow{p}\mu_1^{-1}(\mu_1(\theta_0))=\theta_0$$

从而 $\hat{\theta}$ 是 θ_0 的一致估计。

更加形象的理解是，给定任何一个 θ，总体 P_θ 是一个确定的概率函数，因而可以计算在 θ 情况下的样本矩 $\mathbb{E}_\theta x_i$。理论上，样本矩 $m_1(x)$ 和总体矩 $\mu_1(\theta)$ 在样本量足够大的情况下应该是充分接近的，那么我们可以找到一个 $\hat{\theta}$ 使得 $\mu_1\left(\hat{\theta}\right)$ 与 $m_1(x)$ 的差距最小，从而得到对真值 θ_0 的估计。以上就是矩估计的思想。

例 15. 如果样本 $x_i\sim P(\lambda_0)$ $i.i.d$，我们知道样本矩 $m_1(x)=\bar{x}$，比如，如果我

们的样本观测值为 $x=(3,5,7,2,3)$，那么样本矩为

$$m_1(x)=\bar{x}=\frac{3+5+7+2+3}{5}=4$$

加入任意给定一个 λ，比如令 $\lambda=2$，总体的期望为 $\mathbb{E}_\lambda x_i=\lambda=2\neq 4$，因而如果认为 $\lambda_0=2$，那么总体 $P(2)$ 所产生的总体矩与样本矩仍然有差异。只有当 $\lambda=4$ 时，总体矩 $\mathbb{E}_\lambda x_i=4=m_1(x)$，总体矩与我们观察到的样本矩相等，因而我们可以推断 $\hat{\lambda}=4$。一般的，对于泊松分布总体，我们可以直接令总体矩等于样本矩得到估计，即：

$$\hat{\lambda}=m_1(x)=\bar{x}$$

下面我们分别讨论该估计量的无偏性和一致性。首先，对于无偏性，由于：

$$\mathbb{E}\hat{\lambda}=\mathbb{E}\bar{x}=\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}x_i=\lambda_0$$

因而 $\hat{\lambda}$ 是 λ_0 的无偏估计。而对于一致性，根据大数定理：

$$\hat{\lambda}=\bar{x}\xrightarrow{p}\mathbb{E}x_i=\lambda_0$$

因而 $\hat{\lambda}$ 是 λ_0 的一致估计。当然，一致性还可以通过分析 $\hat{\lambda}$ 的偏差与方差来证明。根据以上讨论，该估计量的偏差为 $\text{Bias}\left(\hat{\lambda}\right)=\mathbb{E}\left(\hat{\lambda}\right)-\lambda_0=0$，而其方差为：

$$\text{Var}\left(\hat{\lambda}\right)=\text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)=\frac{1}{N^2}\sum_{i=1}^{N}\text{Var}(x_i)=\frac{\lambda_0}{N}$$

从而 $\mathbb{E}\left(\hat{\lambda}-\lambda_0\right)^2=\text{Var}\left(\hat{\lambda}\right)+\left[\text{Bias}\left(\hat{\lambda}\right)\right]^2\to 0$，从而 $\hat{\lambda}\xrightarrow{L^2}\lambda_0$，从而 $\hat{\lambda}\xrightarrow{p}\lambda_0$。进一步，根据中心极限定理，有：

$$\sqrt{N}\left(\hat{\lambda}-\lambda_0\right)=\sqrt{N}\left(\bar{x}-\lambda_0\right)\xrightarrow{D}N(0,\lambda_0)$$

因而 $\hat{\lambda}\xrightarrow{D}N\left(\lambda_0,\frac{\lambda_0}{N}\right)$。

例 16. 如果样本 $x_i\sim LN(\mu_0,2)\ \ i.i.d$，即总体为对数正态分布，且一个参数 $\sigma^2=2$ 已知。类似的，样本矩 $m_1(x)=\bar{x}$，而总体矩 $\mathbb{E}_\mu x_i=e^{\mu_0+1}$。根据矩估计的思想，令总体矩等于样本矩，即：

$$e^{\hat{\mu}+1}=m_1(x)=\bar{x}$$

可以得到 μ_0 的矩估计值：

$$\hat{\mu}=\ln\bar{x}-1$$

现在讨论该估计量的无偏性和一致性。首先根据 Jensen 不等式：

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(\ln \bar{x}) - 1 \leq \ln(\mathbb{E}\bar{x}) - 1 = \ln e^{\mu_0+1} - 1 = \mu_0$$

因而 $\hat{\mu}$ 并不是 μ_0 的无偏估计。而根据大数定律，$\bar{x} \xrightarrow{p} \mathbb{E}x_i = e^{\mu_0+1}$，由于 ln 为连续函数，从而：

$$\hat{\mu} = \ln \bar{x} - 1 \xrightarrow{p} \ln e^{\mu_0+1} - 1 = \mu_0$$

因而 $\hat{\mu}$ 是 μ_0 的一致估计。此外，我们还可以使用 delta 方法计算 $\hat{\mu}$ 的极限分布。根据中心极限定理，$\sqrt{N}\left(\bar{x} - e^{\mu_0+1}\right) \xrightarrow{D} N(0, \mathrm{Var}(x_i))$，其中 $\mathrm{Var}(x_i) = e^{2(\mu_0+2)} - e^{2\mu_0+2}$。因而，对估计量在 e^{μ_0+1} 处进行泰勒展开：

$$\begin{aligned}
\sqrt{N}(\hat{\mu} - \mu_0) &= \sqrt{N}\left(\ln \bar{x} - 1 - \ln e^{\mu_0+1} + 1\right) \\
&= \sqrt{N}\left(\ln \bar{x} - \ln e^{\mu_0+1}\right) \\
&= \sqrt{N}\left(\ln e^{\mu_0+1} + \frac{1}{e^{\mu_0+1}}\left(\bar{x} - e^{\mu_0+1}\right) + O\left(\left(\bar{x} - e^{\mu_0+1}\right)^2\right) - \ln e^{\mu_0+1}\right) \\
&= \sqrt{N}\frac{1}{e^{\mu_0+1}}\left(\bar{x} - e^{\mu_0+1}\right) + o_p(1) \\
&\xrightarrow{D} \frac{1}{e^{\mu_0+1}}\sqrt{N}\left(\bar{x} - e^{\mu_0+1}\right) \overset{a}{\sim} N\left(0, e^{-2(\mu_0+1)}\mathrm{Var}(x_i)\right)
\end{aligned}$$

其中第三个等号使用了泰勒展开；而第四个等号是由于 $\bar{x} - e^{\mu_0-1} = o_p(1)$，同时 $\sqrt{N}\left(\bar{x} - e^{\mu_0-1}\right) = O_p(1)$，所以 $\sqrt{N}O\left(\left(\bar{x} - e^{\mu_0+1}\right)^2\right) = o_p(1)$；最后使用了中心极限定力。最终，$\hat{\mu} \xrightarrow{D} N\left(\mu_0, \frac{e^2-1}{N}\right)$。

进一步，如果我们有 k 个未知参数，即 θ 为 k 维向量，那么我们可以联立前 k 个样本矩和总体矩对 θ 进行估计，其中前 k 个样本矩和总体矩定义为：

$$\begin{cases}
m_1(x) = \frac{1}{N}\sum_{i=1}^N x_i^1 & \mu_1(\theta) = \mathbb{E}_\theta x_i^1 \\
m_2(x) = \frac{1}{N}\sum_{i=1}^N x_i^2 & \mu_2(\theta) = \mathbb{E}_\theta x_i^2 \\
\vdots & \vdots \\
m_k(x) = \frac{1}{N}\sum_{i=1}^N x_i^k & \mu_k(\theta) = \mathbb{E}_\theta x_i^k
\end{cases}$$

一般而言，如果我们有 k 个参数，$\theta = (\theta_1, ..., \theta_k)'$，那么我们使用前 k 个矩，解方程：

$$\begin{cases}
m_1(x) = \mu_1\left(\hat{\theta}\right) \\
m_2(x) = \mu_2\left(\hat{\theta}\right) \\
\vdots \\
m_k(x) = \mu_k\left(\hat{\theta}\right)
\end{cases}$$

如果该联立方程有解，即可得到参数 θ_0 的估计。

例 17. 对于正态总体 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d$，其中未知总体参数 $\theta = \left(\mu, \sigma^2\right)$，其一阶样本矩为 $m_1(x) = \bar{x}$，二阶样本矩为 $m_2(x) = \overline{x^2}$。我们知道对于正态分布，$\mu_1(\theta) = \mu, \mu_2(\theta) = \mu^2 + \sigma^2$，从而矩估计为：

$$\begin{cases} m_1(x) = \bar{x} = \hat{\mu} \\ m_2(x) = \overline{x^2} = \hat{\mu}^2 + \hat{\sigma}^2 \end{cases}$$

解得：

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \overline{x^2} - \bar{x}^2 \end{cases}$$

下面分析其无偏性和一致性。根据之前的结论，$\mathbb{E}\hat{\mu} = \mu_0, \mathbb{E}\left(\hat{\sigma}^2\right) = \frac{N-1}{N}\sigma_0^2$，因而 $\hat{\mu}$ 是无偏估计量而 $\hat{\sigma}^2$ 并非无偏估计量。而由于 $\bar{x} \xrightarrow{p} \mu_0$，$\overline{x^2} \xrightarrow{p} \mu_0^2 + \sigma_0^2$，从而 $\hat{\sigma}^2 \xrightarrow{p} \mu_0^2 + \sigma_0^2 - \mu_0^2 = \sigma_0^2$，因而 $\hat{\mu}$ 和 $\hat{\sigma}^2$ 都是一致估计量。

3.2 矩估计

接下来，我们将讨论矩估计的一般形式。对于一个统计模型，如果我们关心真实参数 $\theta_0 \in \Theta \subset \mathbb{R}^K$，只要我们可以找到 k 个矩条件，使得 θ_0 为矩条件方程：

$$\mathbb{E}\left[g\left(w_i, \theta\right)\right] = \mathbb{E}\left(\begin{bmatrix} g_1\left(w_i, \theta\right) \\ \vdots \\ g_K\left(w_i, \theta\right) \end{bmatrix}\right) = 0$$

的唯一解，那么我们就可以解以上总体矩方程组的样本方程等价形式：

$$\frac{1}{N}\sum_{i=1}^{N} g\left(w_i, \hat{\theta}\right) = 0$$

解得 $\hat{\theta}$，那么可以证明，在一些额外比较宽松的条件下，$\hat{\theta} \xrightarrow{p} \theta_0$。

注意我们要求 θ_0 为矩条件方程组的唯一解，即模型是**可识别**的（**identifiable**），这要求矩条件方程不仅有解，而且解唯一。如果矩条件方程有不止一组解，那么我们无法区分真实参数究竟是哪一组解，导致该统计问题无法准确回答。**识别**（**identification**）问题是计量经济学的核心问题。

例 18. （一元线性回归）给定数据 $w_i = (y_i, x_i)'$，我们希望估计条件期望：$\mathbb{E}(y_i|x_i)$。如果我们假定条件期望为线性函数形式，即 $\mathbb{E}(y_i|x_i) = \alpha_0 + \beta_0 x_i$，即得到 y_i 对 x_i 的线性投影：$L(y_i|x_i) = \alpha_0 + \beta_0 x_i$。使用线性函数的假设，我们把条件期望划归为一个线性函数，因而只要估计得到 α 和 β 就得到了条件期望的估计。如果我们令

$$u_i = y_i - \mathbb{E}(y_i|x_i) = y_i - \alpha_0 - \beta_0 x_i$$

那么根据条件期望的性质，有：

$$\mathbb{E}\left(u_i|x_i\right)=0$$

从而真值 α_0,β_0 满足：

$$\begin{cases}\mathbb{E}\left(y_i-\alpha_0-\beta_0x_i\right)=\mathbb{E}\left(u_i\right)=\mathbb{E}\left[\mathbb{E}\left(u_i|x_i\right)\right]=0\\ \mathbb{E}\left[x_i\left(y_i-\alpha_0-\beta_0x_i\right)\right]=\mathbb{E}\left(x_iu_i\right)=\mathbb{E}\left[\mathbb{E}\left(x_iu_i|x_i\right)\right]=\mathbb{E}\left[x_i\mathbb{E}\left(u_i|x_i\right)\right]=0\end{cases}$$

因而可以使用矩条件方程：

$$\begin{cases}\mathbb{E}\left(y_i-\alpha-\beta x_i\right)=0\\ \mathbb{E}\left[x_i\left(y_i-\alpha-\beta x_i\right)\right]=0\end{cases}\tag{2}$$

在使用以上矩条件之前需要讨论识别问题。如果 $x_i=c$ 为一个常数，那么对于任意的实数 $b\in\mathbb{R}$，$\alpha=\alpha_0+bc,\beta=\beta_0+b$ 也一定是以上矩条件方程的解，因而该问题是不可识别的。但是如果 x_i 不是常数，那么以上方程有唯一解。

假设 x_i 不是常数，那么我们可以用样本的等价形式：

$$\begin{cases}\frac{1}{N}\sum_{i=1}^N\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)=0\\ \frac{1}{N}\sum_{i=1}^N\left[x_i\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)\right]=0\end{cases}$$

解得：

$$\begin{cases}\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}\\ \hat{\beta}=\frac{\sum_{i=1}^N(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^N(x_i-\bar{x})^2}\end{cases}\tag{3}$$

可以证明以上得到的 $\hat{\alpha}\xrightarrow{p}\alpha_0,\hat{\beta}\xrightarrow{p}\beta_0$。使用式 (2) 解得：

$$\begin{cases}\alpha_0=\mathbb{E}\left(y_i\right)-\beta_0\mathbb{E}\left(x_i\right)\\ \beta_0=\frac{\mathbb{E}(x_iy_i)-\mathbb{E}(x_i)\mathbb{E}(y_i)}{\mathbb{E}\left(x_i^2\right)-[\mathbb{E}(x_i)]^2}=\frac{\mathrm{Cov}(x_i,y_i)}{\mathrm{Var}(x_i)}\end{cases}$$

使用大数定律可以得到，式 (3) 为一致估计量，即 $\hat{\alpha}\xrightarrow{p}\alpha_0,\hat{\beta}\xrightarrow{p}\beta_0$。

例 19. （多元线性回归）给定数据 $w_i=(y_i,x_i')'$，其中 $x_i\in\mathbb{R}^K$，即：

$$x_i=\begin{bmatrix}x_{i1}\\ x_{i2}\\ \vdots\\ x_{iK}\end{bmatrix}_{K\times1}$$

如果令 $x_{i1}=1$，假设 $\mathbb{E}\left(y_i|x_i\right)=x_i'\beta_0=\beta_{10}+\beta_{20}x_{i2}+\cdots+\beta_{K0}x_{iK}$，那么令

$u_i = y_i - x_i'\beta_0$，有：

$$\mathbb{E}\left(u_i|x_i\right) = 0$$

从而：

$$\mathbb{E}\left(u_i x_i\right) = \mathbb{E}\left(\begin{bmatrix} u_i \\ u_i x_{i2} \\ \vdots \\ u_i x_{iK} \end{bmatrix}\right) = \mathbb{E}\left[\mathbb{E}\left(u_i x_i|x_i\right)\right] = \mathbb{E}\left[x_i \mathbb{E}\left(u_i|x_i\right)\right] = 0$$

将 $u_i = y_i - x_i'\beta_0$ 带入，有：

$$\mathbb{E}\left[x_i\left(y_i - x_i'\beta_0\right)\right] = \mathbb{E}\left[x_i y_i\right] - \mathbb{E}\left[x_i x_i'\beta_0\right] = 0$$

如果假设 $\mathbb{E}\left(x_i x_i'\right)$ 可逆，那么：

$$\beta_0 = \left[\mathbb{E}\left(x_i x_i'\right)\right]^{-1}\left[\mathbb{E}\left(x_i y_i\right)\right]$$

在这里，我们的矩条件方程为：$\mathbb{E}\left[x_i y_i\right] - \mathbb{E}\left[x_i x_i'\beta_0\right] = 0$，使用样本矩代替总体矩，即

$$\frac{1}{N}\sum_{i=1}^{N}\left[x_i y_i\right] - \frac{1}{N}\sum_{i=1}^{N}\left[x_i x_i'\hat{\beta}\right] = 0$$

解以上方程可得：

$$\hat{\beta} = \left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i x_i'\right)\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i y_i\right)\right] = \left[\sum_{i=1}^{N}\left(x_i x_i'\right)\right]^{-1}\left[\sum_{i=1}^{N}\left(x_i y_i\right)\right]$$

如果令：

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix}$$

以及：

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

那么:

$$\begin{cases} X'X = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{bmatrix} = x_1x_1' + \cdots + x_Nx_N' = \sum_{i=1}^N (x_ix_i') \\ X'Y = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = x_1y_1 + \cdots + x_Ny_N = \sum_{i=1}^N (x_iy_i) \end{cases}$$

从而估计量可以写为 $\hat{\beta} = (X'X)^{-1} X'Y$。以上就是所谓的最小二乘估计（Ordinary least squares, OLS）。

3.3 * 大样本性质

得到统计量以后，我们经常还需要讨论统计量的无偏、一致、有效等的统计性质。一般而言，矩估计并不能够完全保证估计量是无偏的，但是一致性基本上是满足的。

针对矩估计的一致性，我们有如下结论:

定理 1. （矩估计的一致性）如果 $w_i \in \mathbb{R}^p$ 为一系列独立同分布的随机向量，假设:

1. $\theta_0 \in \Theta \subset \mathbb{R}^K$，其中 Θ 为紧集;
2. （连续性条件）函数 $g(w, \theta) \in \mathbb{R}^K$ 为 *Borel* 可测函数，且对于任意的 w，$g(w, \theta)$ 在 Θ 上对 θ 为连续函数;
3. （收敛性条件）存在一个函数 $K(w)$，使得对于任意的 $\theta \in \Theta$，$|g_j(w, \theta)| \leq K(w)$，其中 $\mathbb{E}[K(w)] < \infty$;
4. （识别条件）θ_0 为方程: $\mathbb{E}[g(w_i, \theta)] = 0$ 的唯一解

那么方程:

$$\frac{1}{N} \sum_{i=1}^N g\left(w_i, \hat{\theta}\right) = 0$$

的解 $\hat{\theta} \xrightarrow{p} \theta_0$。

在以上条件中，最为重要的条件是 $g(w, \theta)$ 函数对 θ 的连续性条件以及识别条件。比如，在例 (16) 中，矩条件为: $\mathbb{E}\left(x_i - e^{\mu+1}\right) = 0$，因而 $g(x, \mu) = x - e^{\mu+1}$，

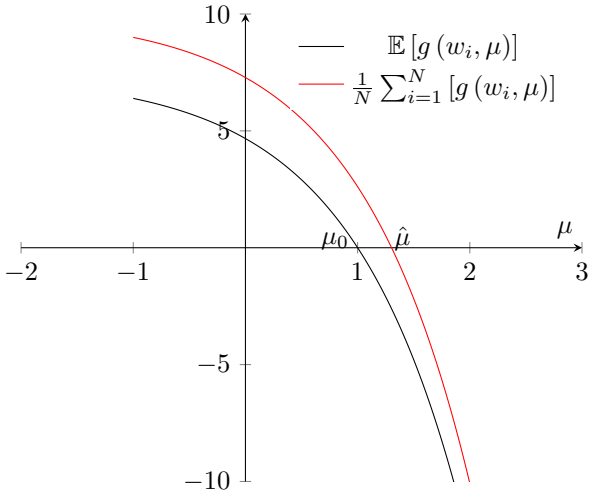

图 3: 矩估计的一致性

对 μ 为连续（且单调）的函数，且当 $\mathbb{E}\left[g\left(x_i,\mu\right)\right]=0$ 时，具有唯一解 $\mu=\mu_0$，因而满足以上条件，根据以上定理可以得到矩估计量 $\hat{\mu}\xrightarrow{p}\mu_0$。

图 (3) 画出了例 (16) 中（令 $\mu_0=1$）的总体矩条件方程

$$\mathbb{E}\left(x_i-e^{\mu+1}\right)=e^{\mu_0+1}-e^{\mu+1}$$

和样本矩条件方程：

$$\frac{1}{N}\sum_{i=1}^{N}\left(x_i-e^{\mu+1}\right)$$

识别条件意味着总体矩条件方程等于 0 的解必须为 μ_0，而样本矩条件方程等于 0 的点即我们的估计量 $\hat{\theta}$。一般而言，这两者并不相等。根据之前介绍的一致收敛定理，定理 (1) 的条件 (1)-(3) 保证了 θ 的函数 $\frac{1}{N}\sum_{i=1}^{N}g\left(w_i,\mu\right)$ 一致收敛到 θ 的函数 $\mathbb{E}\left[g\left(w_i,\mu\right)\right]$，即当样本量 $N\to\infty$ 时，样本方程曲线（红色）一致收敛到总体矩条件方程（黑线），因而自然有估计量 $\hat{\mu}$ 收敛到 μ_0。

在得到未知参数 θ 的估计量 $\hat{\theta}$ 之后，我们经常还需要知道估计量 $\hat{\theta}$ 的抽样分布，比如其大样本分布，才能够在此基础之上完成区间估计、假设检验等任务。估计量的精确分布一般是非常难以计算的，因而我们经常诉诸于估计量的大样本分布进行近似。我们接下来就讨论一下矩估计量的大样本分布。

首先考虑一维情形。我们假设 $\theta\in\mathbb{R}$，那么在矩条件：

$$\mathbb{E}\left[g\left(w_i,\theta_0\right)\right]=0$$

成立的条件下，矩估计量即解如下方程：

$$\sum_{i=1}^{N}\left[g\left(w_i,\hat{\theta}\right)\right]=0$$

如果我们假设函数 $g(\cdot,\cdot)$ 对 $\hat{\theta}$ 是可微的，那么我们可以对其在 θ_0 处进行泰勒展开：

$$0=\frac{1}{N}\sum_{i=1}^{N}\left[g\left(w_i,\theta_0\right)\right]+\frac{1}{N}\sum_{i=1}^{N}\frac{dg}{d\theta}\left(w_i,\theta_0\right)\left(\hat{\theta}-\theta_0\right)+O\left(\left(\hat{\theta}-\theta_0\right)^2\right) \tag{4}$$

两边乘以 $\sqrt{N}$，得到：

$$\begin{aligned}-\sqrt{N}\frac{\sum_{i=1}^{N}\left[g\left(w_i,\theta_0\right)\right]}{N}=&\left[\frac{\sum_{i=1}^{N}\frac{dg}{d\theta}\left(w_i,\theta_0\right)}{N}\right]\left[\sqrt{N}\left(\hat{\theta}-\theta_0\right)\right]\\&+O\left(\sqrt{N}\left(\hat{\theta}-\theta_0\right)^2\right)\end{aligned} \tag{5}$$

其中，根据一致性，有 $\hat{\theta}-\theta_0=o_p(1)$，而 $\sqrt{N}\left(\hat{\theta}-\theta_0\right)=O_p(1)$，从而 $\sqrt{N}\left(\hat{\theta}-\theta_0\right)^2=O_p(1)\cdot o_p(1)=o_p(1)$，因而最后一项可以忽略；而根据中心极限定理，由于 $\mathbb{E}\left[g\left(w_i,\theta_0\right)\right]=0$，因而有：

$$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}\left[g\left(w_i,\theta_0\right)\right]\overset{a}{\sim}N\left(0,\mathrm{Var}\left(g\left(w_i,\theta_0\right)\right)\right)$$

其中

$$\mathrm{Var}\left(g\left(w_i,\theta_0\right)\right)=\mathbb{E}\left(g\left(w_i,\theta_0\right)^2\right)-\left[\mathbb{E}\left(g\left(w_i,\theta_0\right)\right)\right]^2=\mathbb{E}\left(g\left(w_i,\theta_0\right)^2\right)\triangleq B$$

此外，根据大数定律：

$$\frac{1}{N}\sum_{i=1}^{N}\frac{dg}{d\theta}\left(w_i,\theta_0\right)\xrightarrow{p}\mathbb{E}\left(\frac{dg}{d\theta}\left(w_i,\theta_0\right)\right)\triangleq A$$

根据以上结论，式 (5) 可以写为：

$$N(0,B)=A\cdot\sqrt{N}\left(\hat{\theta}-\theta_0\right)+o_p(1)$$

从而：

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)\overset{a}{\sim}N\left(0,\frac{B}{A^2}\right)$$

注意到 A 和 B 都依赖于未知参数 θ_0，因而如果需要计算 $\hat{\theta}$ 的渐进方差，可以

将 $\hat{\theta}$ 带入到 A 和 B 的表达式中进行计算，由于 $\hat{\theta}$ 是 θ_0 的一致估计量，所以大样本条件下对渐进方差的估计仍然是准确的。

例 20. 在例 (16) 中，矩条件为：

$$\mathbb{E}\left[g\left(x_i,\mu\right)\right]=\mathbb{E}\left(x_i-e^{\mu+1}\right)=0$$

从而其中：

$$\begin{cases}A=\mathbb{E}\left(\frac{dg}{d\mu}\left(x_i,\mu_0\right)\right)=-e^{\mu_0+1}\\ B=\mathbb{E}\left(g\left(x_i,\mu_0\right)^2\right)=\mathbb{E}\left[\left(x_i-e^{\mu_0+1}\right)^2\right]=\operatorname{Var}\left(x_i\right)=e^{2(\mu_0+2)}-e^{2\mu_0+2}\end{cases}$$

因而

$$\frac{B}{A^2}=\frac{\operatorname{Var}\left(x_i\right)}{e^{2\mu_0+2}}$$

根据以上结论，有：

$$\sqrt{N}\left(\hat{\mu}-\mu_0\right)\overset{a}{\sim}N\left(0,\frac{\operatorname{Var}\left(x_i\right)}{e^{2\mu_0+2}}\right)$$

与例 (16) 的结论一致。在实际计算中，注意到 $\sqrt{N}\left(\hat{\theta}-\theta_0\right)$ 的渐进方差 $\frac{\operatorname{Var}(x_i)}{e^{2\mu_0+2}}$ 为未知参数 μ_0 的函数，因而为了计算渐进方差，我们可以将 μ_0 的估计值 $\hat{\mu}$ 带入到渐进方差公式中进行计算。

对于多元的情形，同样可以使用 delta 方法。如果我们的总体矩条件方程为：

$$\mathbb{E}\left[g\left(w_i,\theta\right)\right]=\mathbb{E}\left(\begin{bmatrix}g_1\left(w_i,\theta\right)\\ \vdots\\ g_K\left(w_i,\theta\right)\end{bmatrix}\right)=0$$

那么估计值 $\hat{\theta}$ 使得其样本矩条件方程

$$\frac{1}{N}\sum_{i=1}^{N}\left[g\left(w_i,\hat{\theta}\right)\right]=0$$

对以上方程在 θ_0 处进行泰勒展开：

$$\begin{aligned}0&=\frac{1}{N}\sum_{i=1}^{N}\left[g\left(w_i,\hat{\theta}\right)\right]\\ &=\frac{1}{N}\sum_{i=1}^{N}\left[g\left(w_i,\theta_0\right)\right]+\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial}{\partial\theta}g\left(w_i,\theta_0\right)\right]\left(\hat{\theta}-\theta_0\right)+O\left(\left\|\hat{\theta}-\theta_0\right\|^2\right)\end{aligned}\quad(6)$$

其中：

$$\frac{\partial}{\partial\theta}g\left(w_i,\theta_0\right)=\left[\begin{array}{ccc}\frac{\partial}{\partial\theta_1}g_1\left(w_i,\theta_0\right) & \cdots & \frac{\partial}{\partial\theta_K}g_1\left(w_i,\theta_0\right)\\ \vdots & \ddots & \vdots\\ \frac{\partial}{\partial\theta_1}g_K\left(w_i,\theta_0\right) & \cdots & \frac{\partial}{\partial\theta_K}g_K\left(w_i,\theta_0\right)\end{array}\right]_{K\times K}$$

根据大数定律，令 $\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial}{\partial\theta}g\left(w_i,\theta_0\right)\right]\xrightarrow{p}G_0$；由于 $\mathbb{E}\left[g\left(w_i,\theta_0\right)\right]=0$，而协方差矩阵：

$$\begin{aligned}\operatorname{Var}\left[g\left(w_i,\theta_0\right)\right]&=\mathbb{E}\left[g\left(w_i,\theta_0\right)-\mathbb{E}g\left(w_i,\theta_0\right)\right]\left[g\left(w_i,\theta_0\right)-\mathbb{E}g\left(w_i,\theta_0\right)\right]'\\&=\mathbb{E}\left[g\left(w_i,\theta_0\right)g\left(w_i,\theta_0\right)'\right]\end{aligned}$$

因而根据中心极限定理：

$$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}\left[g\left(w_i,\theta_0\right)\right]\xrightarrow{D}N\left(0,\mathbb{E}\left[g\left(w_i,\theta_0\right)g\left(w_i,\theta_0\right)'\right]\right)$$

因而方程 (6) 可以写为：

$$G_0\sqrt{N}\left(\hat{\theta}-\theta_0\right)+o_p\left(1\right)\overset{a}{\sim}N\left(0,\mathbb{E}\left[g\left(w_i,\theta_0\right)g\left(w_i,\theta_0\right)'\right]\right)$$

最后得到：

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)\overset{a}{\sim}N\left(0,G_0^{-1}\mathbb{E}\left[g\left(w_i,\theta_0\right)g\left(w_i,\theta_0\right)'\right]\left(G_0^{-1}\right)'\right)$$

例 21. 在例 (19) 中，$g\left(w_i,\beta\right)=x_iu_i=x_i\left(y_i-x_i'\beta\right)$，因而：

$$G_0=\mathbb{E}\left[\frac{\partial}{\partial\beta}g\left(w_i,\beta\right)\right]=\mathbb{E}\left(x_ix_i'\right)$$

以及：

$$\begin{aligned}\mathbb{E}\left[g\left(w_i,\theta_0\right)g\left(w_i,\theta_0\right)'\right]&=\mathbb{E}\left[x_iu_iu_i'x_i'\right]\\&=\mathbb{E}\left[u_i^2x_ix_i'\right]\end{aligned}$$

如果假设同方差，即 $\mathbb{E}\left(u_i^2|x_i\right)=\sigma^2$，那么

$$\mathbb{E}\left[g\left(w_i,\theta_0\right)g\left(w_i,\theta_0\right)'\right]=\mathbb{E}\left[u_i^2x_ix_i'\right]=\mathbb{E}\left(\mathbb{E}\left[u_i^2x_ix_i'|x_i\right]\right)=\sigma^2\mathbb{E}\left(x_ix_i'\right)$$

因而根据以上结论，有：

$$\sqrt{N}\left(\hat{\beta}-\beta_0\right)\overset{a}{\sim}N\left(0,\sigma^2\mathbb{E}\left(x_ix_i'\right)^{-1}\right)$$

4 极大似然估计

4.1 极大似然估计量

极大似然估计量（**maximum likelihood estimator**）是目前为止最常见的得到估计量的方法，其思想是，如果我们要对未知参数总体 P_θ 做推断，估计 θ_0，那么我们就寻找一个 $\hat{\theta}$，使得这组数据出现的概率最高，则 $\hat{\theta}$ 理应是 θ_0 的一个合理估计。

如果假设一组独立同分布的样本 $x = (x_1, ..., x_N)$ 来自于参数总体 P_θ，且密度函数为 $f(x_i|\theta)$，那么样本的联合分布函数为：

$$f(x|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

现在，将未知参数 θ 视为变量，x 为给定的样本，由于对数函数为单调函数，因而可以将联合分布函数取对数，得到对数似然函数（log-likelihood function）：

$$L(\theta|x) = \ln f(x|\theta) = \sum_{i=1}^{N} \ln f(x_i|\theta)$$

极大似然估计即找到一个 $\hat{\theta}$ 使得对数似然函数最大化：

$$\hat{\theta} = \arg\max_{\theta} L(\theta|x)$$

从而我们得到了极大似然估计量 $\hat{\theta}$。

例 22. 如果 $x_i \sim Ber(p_0)$ $i.i.d$，那么其联合密度函数为：

$$f(x|p) = \prod_{i=1}^{N} p^{x_i} (1-p)^{1-x_i}$$

对数似然函数为：

$$\begin{aligned} L(p|x) &= \sum_{i=1}^{N} [x_i \ln p + (1-x_i) \ln(1-p)] \\ &= \left(\sum_{i=1}^{N} x_i\right) \ln p + \left(N - \sum_{i=1}^{N} x_i\right) \ln(1-p) \end{aligned}$$

现在欲得到 p 的极大似然估计值，只要对上述对数似然函数求最大值，即：

$$\frac{\partial L(p|x)}{\partial p} = \frac{\sum_{i=1}^{N} x_i}{p} - \frac{\left(N - \sum_{i=1}^{N} x_i\right)}{1-p} = 0$$

从而得到：

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

现在讨论该估计量的无偏性和一致性。对于无偏性，我们有：

$$\mathbb{E}\left(\hat{p}\right) = \mathbb{E}\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}x_i = p_0$$

而对于一致性，根据大数定律，我们有：

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} x_i \xrightarrow{p} \mathbb{E}x_i = p_0$$

因而极大似然估计量 $\hat{p}$ 是真值 p_0 的无偏、一致估计量。

例 23. 如果 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d$，其中 $\theta = \left(\mu, \sigma^2\right)$，那么其联合密度函数为：

$$f\left(x|\theta\right) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right\}$$

对数似然函数为：

$$\begin{aligned}
L\left(\theta|x\right) &= \sum_{i=1}^{N}\left[-\frac{1}{2}\ln\left(2\pi\right) - \ln\sigma - \frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right] \\
&= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i-\mu\right)^2 \\
&= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i^2+\mu^2-2\mu x_i\right) \\
&= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{N\mu^2}{2\sigma^2} - \frac{N}{2\sigma^2}\overline{x^2} + \frac{N\mu}{\sigma^2}\bar{x}
\end{aligned}$$

对其求极大值，得到：

$$\frac{\partial L\left(\theta|x\right)}{\partial\theta} = \begin{pmatrix} -\frac{\mu}{\sigma^2} + \frac{\bar{x}}{\sigma^2} \\ -\frac{1}{\sigma} + \frac{\mu^2}{\sigma^3} + \frac{\overline{x^2}}{\sigma^3} - \frac{2\mu\bar{x}}{\sigma^3} \end{pmatrix} = 0$$

解得：

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \overline{x^2} - \bar{x}^2 \end{cases}$$

这与矩估计的估计量是一样的。我们之前已经证明了，$\hat{\mu}$ 是 μ_0 的无偏、一致估计量，而 $\hat{\sigma}^2$ 是 σ_0^2 的一致估计量，$\frac{N}{N-1}\hat{\sigma}^2$ 是 σ_0^2 的无偏估计量。

例 24. （截尾数据）现在正在进行一项调查，其中一项调查为收入（y_i）调查，其中关于收入的问题为：

- 请问您的收入是多少?
  - 小于 1000
  - 大于 10000
  - 其他 _____（请填写具体数值）

如果假设收入的对数（$x_i^* = \log_{10} y_i$）服从正态分布，即 $x_i^* \sim N\left(\mu, \sigma^2\right)$ $i.i.d$，那么我们观察到的数据为：

$$x_i = \begin{cases} 3 & y_i \leq 1000 \\ 4 & y_i \geq 10000 \\ x_i^* & otherwise \end{cases}$$

我们称数据存在截尾（censoring）现象。为了估计以上问题，我们可以计算：

$$P\left(x_i = 3\right) = P\left(x_i^* \leq 3\right) = P\left(\frac{x_i^* - \mu}{\sigma} \leq \frac{3-\mu}{\sigma}\right) = \Phi\left(\frac{3-\mu}{\sigma}\right)$$

同理 $P\left(x_i = 4\right) = 1 - \Phi\left(\frac{4-\mu}{\sigma}\right)$。因而 x_i 的密度函数为：

$$f\left(x_i|\theta\right) = \left[\Phi\left(\frac{3-\mu}{\sigma}\right)\right]^{1\{x_i=3\}} \left[1 - \Phi\left(\frac{4-\mu}{\sigma}\right)\right]^{1\{x_i=4\}} \left[\phi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{1\{3<x_i<4\}}$$

因而其对数似然函数为：

$$\begin{aligned} L\left(\theta|x\right) = & N_3 \ln \Phi\left(\frac{3-\mu}{\sigma}\right) + N_4 \ln\left[1 - \Phi\left(\frac{3-\mu}{\sigma}\right)\right] \\ & + \sum_{i=1}^{N} 1\left\{3 < x_i < 4\right\} \ln \phi\left(\frac{x_i-\mu}{\sigma}\right) \end{aligned}$$

最大化以上对数似然函数，我们就得到了正态分布总体的极大似然估计。

4.2 一致性与 Kullback-Leiber 信息

在以上两个例子中，我们发现，尽管极大似然估计不能保证无偏性，但是所有的估计量都是一致的。那么是不是极大似然估计一定能保证一致性呢?

为了回答这个问题，我们可以从两个角度来看。我们知道，对数似然函数：$\frac{1}{N} L\left(\theta|x\right) = \frac{1}{N} \sum_{i=1}^{N} \ln f\left(x_i|\theta\right)$，对于任意一个给定的 θ（不一定是真值 θ_0），在一定的条件下，根据大数定律，有：

$$\frac{1}{N} L\left(\theta|x\right) \xrightarrow{p} \mathbb{E} \ln f\left(x_i|\theta\right) \triangleq \mathscr{L}\left(\theta\right)$$

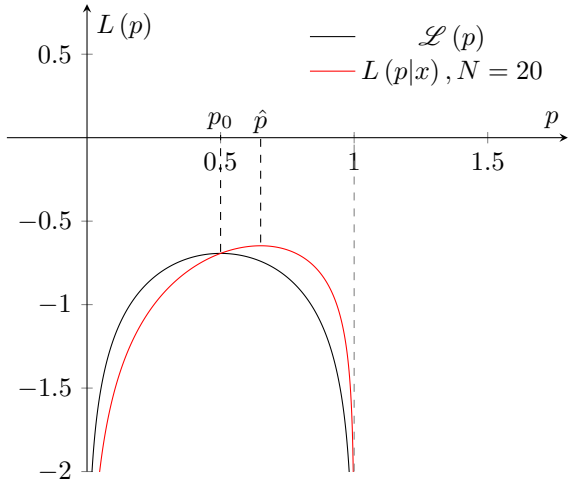

图 4: 伯努利分布的总体和样本似然函数

即样本似然函数收敛到总体的似然函数。而且根据上一章中的结论，在一定条件下，这个收敛是一致收敛的。

由于 $\frac{1}{N}L(\theta|x)$ 一致收敛到 $\mathscr{L}(\theta)$，即如图 (4) 所示，样本似然函数 $L(\theta|x)$（红线）随着样本量增大不断逼近总体似然函数 $\mathscr{L}(\theta)$（黑线），而极大似然估计的方法是最大化 $L(\theta|x)$ 获得估计，即

$$\hat{\theta} = \arg\max_{\theta} L(\theta|x)$$

那么如果真值

$$\theta_0 = \arg\max_{\theta} \mathscr{L}(\theta)$$

样本似然函数最大值 $\hat{\theta}$ 应该也会收敛到总体似然函数最大值 θ_0。

或者，如果 $\theta_0 = \arg\max_\theta \mathscr{L}(\theta)$ 成立，那么一阶条件为：

$$\frac{\partial}{\partial\theta}\mathscr{L}(\theta_0) = \frac{\partial}{\partial\theta}\mathbb{E}\left[\ln f(x_i|\theta_0)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f(x_i|\theta_0)\right] = 0$$

以上可以作为总体矩条件方程，而样本矩方程为：

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial}{\partial\theta}\ln f(x_i|\theta)\right] = 0$$

求解以上样本矩方程实际上就是得到了 $\arg\max_\theta L(\theta|x)$。因而如果 $\frac{\partial}{\partial\theta}\ln f(x_i|\theta_0)$ 满足定理 (1) 的要求，那么极大似然估计一定是一致的。

那么接下来的问题就是，我们的真值 θ_0 是不是的确能够最大化总体似然函数 $\mathscr{L}(\theta) = \mathbb{E}\ln f(x_i|\theta)$ 呢？我们先看一个例子。

例 25. 若 $x_i \sim Ber(p_0)$ $i.i.d$，那么其联合密度函数为：

$$f(x|p) = \prod_{i=1}^{N} p^{x_i} (1-p)^{1-x_i}$$

对数似然函数为：

$$L(p|x) = \sum_{i=1}^{N} [x_i \ln p + (1-x_i) \ln (1-p)]$$

根据大数定律，对于任意的 p，上述似然函数

$$\begin{aligned} \frac{1}{N} L(p|x) \xrightarrow{p} \mathscr{L}(p) &\triangleq \mathbb{E}[x_i \ln p + (1-x_i) \ln (1-p)] \\ &= \mathbb{E}(x_i) \ln p + \mathbb{E}(1-x_i) \ln (1-p) \\ &= p_0 \ln p + (1-p_0) \ln (1-p) \end{aligned}$$

其中 p_0 为真值。那么接下来的问题是，是不是只有当 $p = p_0$ 时，$\mathscr{L}(p)$ 达到了最大值呢？为求最大值，我们对 $\mathscr{L}(p)$ 求导数并令其等于 0 得到：

$$\frac{\partial \mathscr{L}(p)}{\partial p} = \frac{p_0}{p} - \frac{1-p_0}{1-p} = 0$$

从而只有当 $p = p_0$ 时，以上导数等于 0。因而真值 p_0 最大化了总体似然函数 $\mathscr{L}(p)$。

以上伯努利分布的例子并不是个例，实际上，我们可以证明，真值 θ_0 总是可以最大化总体似然函数。为了证明这一点，我们可以从总体似然函数出发：

$$\begin{aligned} \mathscr{L}(\theta) &= \mathbb{E} \ln f(x_i|\theta) \\ &= \mathbb{E}_{\theta_0} \ln f(x_i|\theta) \\ &= \int_{\mathbb{R}} \ln f(x|\theta) \cdot f(x|\theta_0) \, dx \end{aligned}$$

求其最大值，得到：

$$\begin{aligned} \frac{\partial \mathscr{L}(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} \ln f(x|\theta) \cdot f(x|\theta_0) \, dx \\ &= \int_{\mathbb{R}} \frac{\partial}{\partial \theta} \ln f(x|\theta) \cdot f(x|\theta_0) \, dx \\ &= \int_{\mathbb{R}} \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} \cdot f(x|\theta_0) \, dx \end{aligned}$$

当 $\theta = \theta_0$ 时，有：

$$\begin{aligned}\frac{\partial \mathscr{L}(\theta_0)}{\partial \theta} &= \int_{\mathbb{R}} \frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial \theta} \cdot f(x|\theta_0)\, dx \\ &= \int_{\mathbb{R}} \frac{\partial f(x|\theta_0)}{\partial \theta} dx \\ &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} f(x|\theta_0)\, dx \\ &= 0\end{aligned}$$

其中最后一步由于 $\int_{\mathbb{R}} f(x|\theta_0)\, dx = 1$。我们通常称对数似然函数的一阶导数 $\frac{\partial}{\partial \theta} \ln f(x_i|\theta)$ 为**得分函数**（**score function**），记为 $s_i(\theta) = \frac{\partial}{\partial \theta} \ln f(x_i|\theta)$。以上结论意味着得分函数的期望等于 0，即 $\mathbb{E} s_i(\theta_0) = 0$。至此，如果得分函数 $s_i(\theta)$ 满足定理 (1) 的要求，那么必然有极大似然估计量 $\hat{\theta} \xrightarrow{p} \theta_0$。

然而如果使用矩估计的思路证明极大似然的一致性，仍然需要更多的假设。为了建立极大似然方法的一致性，我们仍然需要讨论真值 θ_0 是否的确最大化了总体似然函数 $\mathscr{L}(\theta)$ 呢。

我们知道，一阶导数等于 0 是最大化的必要条件而非充分条件，因而以上的推论并不一定能够得到真值 θ_0 最大化了总体似然函数 $\mathscr{L}(\theta)$ 这一结论。为了更进一步得到真值 θ_0 是否最大化了总体似然函数，我们引入 Kullback-Leiber 信息的概念。

定义 1. 令 P 和 Q 为同一概率空间中的两个概率函数，p 和 q 分别为其概率密度函数。Kullback-Leiber 信息被定义为：

$$\mathscr{K}(P, Q) = \int_{\mathbb{R}} \ln \frac{p(x)}{q(x)} p(x)\, dx$$

当 P 和 Q 属于参数族 P_θ 和 Q_η 时，Kullback-Leiber 信息即：

$$\mathscr{K}(\theta, \eta) = \int_{\mathbb{R}} \ln \frac{f(x|\theta)}{g(x|\eta)} f(x|\theta)\, dx = \mathbb{E}_\theta \left[\ln \frac{f(x|\theta)}{g(x|\eta)} \right]$$

实际上，Kullback-Leiber 信息度量的是两个概率函数的「距离」，可以证明，Kullback-Leiber 信息 $\mathscr{K}(P, Q) \geq 0$，当且仅当 $P = Q$ 时等号成立。

对于极大似然函数，给定任意一个 θ，其代表的概率函数与真值代表的概率函数之间的距离，即 Kullback-Leiber 信息为：

$$\mathscr{K}(\theta_0, \theta) = \mathbb{E}_{\theta_0} \left[\ln \frac{f(x|\theta_0)}{f(x|\theta)} \right] \geq 0$$

当且仅当 $\theta = \theta_0$ 时等式成立，因而 θ_0 最小化了：

$$\mathbb{E}_{\theta_0} \left[\ln \frac{f(x|\theta_0)}{f(x|\theta)} \right] = \mathbb{E}_{\theta_0} \left[\ln f(x|\theta_0) - \ln f(x|\theta) \right]$$

或者等价的，最大化了 $\mathbb{E}_{\theta_0}\left[\ln f\left(x|\theta\right)\right]=\mathscr{L}\left(\theta\right)$。

最终，我们有如下结论：

定理 2. （极大似然的一致性）如果 $w_i\in\mathbb{R}^p$ 为一系列独立同分布的随机向量，假设：

1. $\theta_0\in\Theta\subset\mathbb{R}^K$，其中 Θ 为紧集；
2. （连续性条件）函数 $\ln f\left(w_i|\theta\right)\in\mathbb{R}$ 为 *Borel* 可测函数，且对于任意的 w，$\ln f\left(w|\theta\right)$ 在 Θ 上对 θ 为连续函数；
3. （收敛性条件）存在一个函数 $K\left(w\right)$，使得对于任意的 $\theta\in\Theta$，$\left|\ln f\left(w|\theta\right)\right|\leq K\left(w\right)$，其中 $\mathbb{E}\left[K\left(w\right)\right]<\infty$；
4. （识别条件）$f\left(w_i|\theta_0\right)$ 为真实的密度函数，且 θ_0 为 $\mathbb{E}\left[\ln f\left(w_i|\theta\right)\right]$ 的唯一最大值解。

那么估计量：

$$\hat{\theta}=\arg\max_{\theta}\sum_{i=1}^{N}\ln f\left(w_i|\theta\right)$$

满足：$\hat{\theta}\xrightarrow{p}\theta_0$。

注意以上定理中，我们要求 $\ln f\left(w_i|\theta\right)$ 对 θ 是连续的，而并没有要求得分函数 $s_i\left(\theta\right)$ 是连续的。此外，条件 (4) 不仅仅要求 θ_0 是唯一最大值解，而且要求我们的模型是正确设定的，也就是 $f\left(w_i|\theta_0\right)$ 是 w_i 的真实的密度函数。在这些假设的基础之上我们可以得到结论，极大似然估计是一致估计。

4.3 极限分布与 Fisher 信息

上一节中，我们知道在一定条件下，极大似然估计量 $\hat{\theta}$ 是真值 θ_0 的一致估计量，进一步的，我们希望知道估计量 $\hat{\theta}$ 的抽样分布。我们下面将从极大似然函数的一阶条件开始，使用 delta 方法得到估计量 $\hat{\theta}$ 的极限分布。

由于我们计算极大似然估计量时最大化了极大似然函数，其一阶条件为：

$$\frac{\partial L\left(\hat{\theta}|x\right)}{\partial\theta}=\frac{\partial}{\partial\theta}\sum_{i=1}^{N}\ln f\left(x_i|\hat{\theta}\right)=\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\ln f\left(x_i|\hat{\theta}\right)=\sum_{i=1}^{N}s_i\left(\hat{\theta}\right)=0$$

即样本得分函数的均值等于 0。我们对上式在 $\theta=\theta_0$ 处进行泰勒展开，得到：

$$0=\sum_{i=1}^{N}s_i\left(\hat{\theta}\right)=\sum_{i=1}^{N}s_i\left(\theta_0\right)+\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}s_i\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right)+O\left(\left(\hat{\theta}-\theta_0\right)^2\right)$$

我们记 $H_i\left(\theta\right)=\frac{\partial}{\partial\theta'}s_i\left(\theta\right)=\frac{\partial}{\partial\theta\partial\theta'}\ln f\left(x_i|\theta\right)$ 为对数似然函数的海塞矩阵，我们

有：

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}s_i\left(\theta_0\right)\xrightarrow{p}\mathbb{E}H_i\left(\theta_0\right)\triangleq -H_0$$

而由于 $\mathbb{E}s_i\left(\theta_0\right)=0$，因而 $\mathrm{Var}\left(s_i\left(\theta_0\right)\right)=\mathbb{E}\left[s_i\left(\theta_0\right)s_i'\left(\theta_0\right)\right]\triangleq\mathcal{I}_0$，因而根据中心极限定理：

$$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}s_i\left(\theta_0\right)\xrightarrow{D}N\left(0,\mathcal{I}_0\right)$$

因而：

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)=H_0^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}s_i\left(\theta_0\right)+o_p\left(1\right)\xrightarrow{D}N\left(0,H_0^{-1}\mathcal{I}_0H_0^{-1}\right)\tag{7}$$

注意其中：

$$\begin{aligned}
-H_0&=\mathbb{E}H_i\left(\theta_0\right)\\
&=\int_{\mathbb{R}}\frac{\partial}{\partial\theta\partial\theta'}\ln f\left(x|\theta_0\right)\cdot f\left(x|\theta_0\right)dx\\
&=\int_{\mathbb{R}}\frac{\partial}{\partial\theta'}\left[\frac{1}{f\left(x|\theta_0\right)}\frac{\partial f\left(x|\theta_0\right)}{\partial\theta}\right]\cdot f\left(x|\theta_0\right)dx\\
&=\int_{\mathbb{R}}\left[\frac{1}{f\left(x|\theta_0\right)}\frac{\partial^2 f\left(x|\theta_0\right)}{\partial\theta\partial\theta'}-\frac{1}{f^2\left(x|\theta_0\right)}\frac{\partial f\left(x|\theta_0\right)}{\partial\theta}\frac{\partial f\left(x|\theta_0\right)}{\partial\theta'}\right]\cdot f\left(x|\theta_0\right)dx\\
&=\int_{\mathbb{R}}\frac{\partial^2 f\left(x|\theta_0\right)}{\partial\theta\partial\theta'}-\frac{1}{f\left(x|\theta_0\right)}\frac{\partial f\left(x|\theta_0\right)}{\partial\theta}\frac{\partial f\left(x|\theta_0\right)}{\partial\theta'}dx\\
&=\int_{\mathbb{R}}\frac{\partial^2 f\left(x|\theta_0\right)}{\partial\theta\partial\theta'}dx-\int_{\mathbb{R}}\left(\frac{1}{f\left(x|\theta_0\right)}\frac{\partial f\left(x|\theta_0\right)}{\partial\theta}\frac{\partial f\left(x|\theta_0\right)}{\partial\theta'}\right)^2 f\left(x|\theta_0\right)dx\\
&=\frac{\partial^2}{\partial\theta\partial\theta'}\int_{\mathbb{R}}f\left(x|\theta_0\right)dx-\mathbb{E}_{\theta_0}\frac{\partial\ln f\left(x|\theta_0\right)}{\partial\theta}\frac{\partial\ln f\left(x|\theta_0\right)}{\partial\theta'}\\
&=-\mathbb{E}_{\theta_0}\left[s_i\left(\theta_0\right)s_i\left(\theta_0\right)'\right]\\
&=-\mathrm{Var}\left[s_i\left(\theta_0\right)\right]\\
&=-\mathcal{I}_0
\end{aligned}$$

因而我们有：$H_0=\mathcal{I}_0$，即对数似然函数海塞矩阵的期望等于得分函数的方差。将以上等式带入式 (7)，可以得到：

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)\xrightarrow{D}N\left(0,\mathcal{I}_0^{-1}\right)$$

即大样本条件下，极大似然估计量的极限分布为正态分布，且其渐进方差为对

数似然函数海塞矩阵倒数的逆矩阵。特别的，当 θ 为一维标量时，

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right) \xrightarrow{D} N\left(0,-\frac{1}{\mathbb{E}\left(\frac{d^2}{d\theta^2}\ln f\left(x|\theta\right)\right)}\right)$$

例 26. 在例 (22) 中，我们已经得到伯努利分布的极大似然估计为：

$$\hat{p}=\bar{x}$$

根据中心极限定理，$\hat{p}=\bar{x} \xrightarrow{D} N\left(p_0,\frac{p_0(1-p_0)}{N}\right)$。另一方面，直接使用以上结论也可以得到 $\hat{p}$ 的极限分布。注意对数似然函数为：

$$L\left(p|x\right)=\sum_{i=1}^{N}\left[x_i\ln p+\left(1-x_i\right)\ln\left(1-p\right)\right]$$

因而得分函数为：

$$s_i\left(p\right)=\frac{x_i}{p}-\frac{1-x_i}{1-p}$$

进而海塞矩阵（二阶导）：

$$-H_i\left(p\right)=-\frac{x_i}{p^2}-\frac{1-x_i}{\left(1-p\right)^2}$$

因而带入真值之后其期望为：

$$\mathcal{I}_0=H_0=\mathbb{E}H_i\left(p_0\right)=\frac{1}{p_0}+\frac{1}{1-p_0}=\frac{1}{p_0\left(1-p_0\right)}$$

从而 $H_0^{-1}=p_0\left(1-p_0\right)$，从而 $\sqrt{N}\left(\hat{p}-p\right)\sim N\left(0,p_0\left(1-p_0\right)\right)$。

实际上，可以证明，以上的 $\mathcal{I}_0^{-1}$ 是渐进无偏估计量所能达到的最小方差，我们称方差为 $\mathcal{I}_0^{-1}$ 的估计量为渐进有效（asymptotic efficient）估计量。而 $\mathcal{I}_0$ 实际上度量了数据中所包含的「信息量」的大小，因而我们称 $\mathcal{I}_0$ 为**费雪信息矩阵**（**Fisher information matrix**）。$\mathcal{I}_0$ 越大，意味着所包含的信息越多，极大似然估计所得到的方差也越小。如图 (5) 所示，当 $p_0=0.5$ 时，信息矩阵 $\mathcal{I}_0$ 达到了最小值，而 $\hat{p}$ 的方差达到了最大值，表现在图中即对数似然函数在真值 p_0 处非常平缓。当真值 p_0 逐渐接近 0 或者 1 时，信息矩阵 $\mathcal{I}_0$ 逐渐变大，$\hat{p}$ 的方差也逐渐变小，图中对数似然函数在真值 p_0 处也更加尖锐。因而同样是伯努利分布，真值越接近于 0 或者 1 的伯努利分布实际上携带了更多的信息。

4.4 条件极大似然估计

以上介绍了极大似然估计法，需要设定数据 X 的完整的分布情况才能得到估计。然而很多时候，我们观察到一系列数据 $w_i\in\mathbb{R}^k, i=1,...,N$，其中 $w_i=\left(y_i',x_i'\right)', y_i\in\mathbb{R}^{k_1}, x_i\in\mathbb{R}^{k_2}$，很多时候我们仅仅希望研究 x 和 y 之间的关

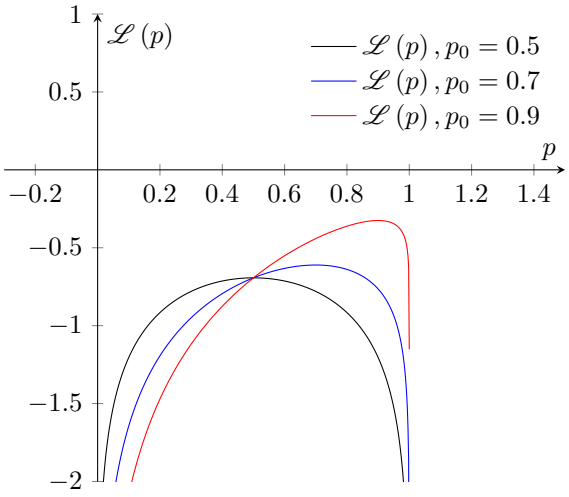

图 5: 不同 p_0 下伯努利分布的总体似然函数

系，而不关心随机向量 x 之间的关系，如果使用极大似然估计，我们就必须设定 x 的联合分布。然而，设定 x 的联合分布很多时候是多于的，实际上，如果我们能够找到 y 给定 x 的条件分布，即 $f(y|x,\theta)$，那么基于条件分布的极大似然估计仍然能够得到参数 θ_0 的一致估计。

例 27. （线性回归）如果 $(y_i, x_i')', i = 1, ..., N, x_i \in \mathbb{R}^K$ 为一系列独立同分布的随机向量。为了使用 x_i 预测或者拟合 y_i，我们可以假设 $y_i|x_i \sim N\left(x_i'\beta_0, \sigma^2\right)$，即给定 x，y 服从正态分布[1]。或者等价的，以上模型也可以写成：

$$y_i = x_i'\beta_0 + u_i$$

其中 $u_i|x_i \sim N\left(0, \sigma^2\right)$。在上述条件下，条件密度函数为：

$$f(y_i|x_i, \beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - x_i'\beta)^2}{\sigma^2}\right\}$$

因而似然函数为：

$$L(\beta|y, x) = -\frac{N}{2}\ln(2\pi) - N\ln\sigma - \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - x_i'\beta)^2$$

如果对 β 求导，可以得到：

$$-\frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y_i - x_i'\hat{\beta}\right)x_i = 0$$

[1]注意尽管 $y|x$ 服从正态分布，y 有可能不服从正态分布。

解得：

$$\hat{\beta} = \left[\sum_{i=1}^{N}(x_i x_i')\right]^{-1} \sum_{i=1}^{N}(x_i y_i) = (X'X)^{-1} X'Y$$

我们再次得到了最小二乘估计（Ordinary least squares, OLS）。特别的，如果 $K=2$，$x_{i1}=1$，即解释变量中存在截距项（常数项），以及一个额外的解释变量，我们称这种情况为一元线性回归，可以得到：

$$\begin{cases} \hat{\beta}_2 & = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \\ \hat{\beta}_1 & = \bar{y} - \hat{\beta}_2 \bar{x} \end{cases}$$

例 28. （Logistic 回归）如果 $(y_i, x_i')', i=1,...,N, x_i \in \mathbb{R}^K$ 为一系列独立同分布的随机向量，而其中 y_i 为二元变量，即 $y_i \in \{0,1\}$，那么此时使用线性回归模型就不再适合。如果我们继续使用线性回归，那么其条件期望：

$$\mathbb{E}(y_i|x_i) = x_i'\beta_0 = 1 \cdot P(y_i|x_i) + 0 \cdot (1 - P(y_i|x_i)) = P(y_i|x_i)$$

即条件期望为 $y_i=1$ 的条件概率，但是线性函数 $x_i'\beta$ 不能够保证一定在 $(0,1)$ 区间范围以内。为了避免以上问题，一个常用的假设是：

$$P(y_i=1|x_i,\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$$

由于函数 $F(x) = \frac{e^x}{1+e^x}$ 为一个分布函数，因而其函数值一定是单调的且在 $(0,1)$ 之间的。我们将 $y_i=1$ 的概率与 $y_i=0$ 的概率的比值成为几率（odds），那么根据以上设定，几率为：

$$odds = \frac{P(y_i=1|x_i,\beta)}{P(y_i=0|x_i,\beta)} = \frac{\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}}{1-\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}} = e^{x_i'\beta}$$

而对数几率（log odds，也成为 logit）为：

$$logit = log(odds) = x_i'\beta$$

因而以上模型被称为对数几率回归（Logistic 回归或者 Logit 回归）。以上模型可以等价地写成：

$$\begin{aligned} y_i^* &= x_i'\beta - u_i \\ u_i|x_i &\sim LG(0,1) \\ y_i &= 1\{y_i^* \geq 0\} \end{aligned}$$

其中 y_i^* 为不可见的**潜变量**（**latent variable**），只有 $(y_i, x_i')'$ 可见。实际上，从

潜变量模型出发，有：

$$\begin{aligned} P(y_i = 1) &= P(y_i^* \geq 0) \\ &= P(x_i'\beta - u_i \geq 0) \\ &= P(u_i \leq x_i'\beta) \\ &= F_u(x_i'\beta) \end{aligned}$$

由于 $u_i|x_i \sim LG(0,1)$，因而

$$F_u(x_i'\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$$

以上模型的条件密度函数为：

$$f(y_i|x_i,\beta) = \left[\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right]^{1\{y_i=1\}} \left[\frac{1}{1+e^{x_i'\beta}}\right]^{1\{y_i=0\}}$$

因而极大似然函数为：

$$L(\beta|y,x) = \sum_{i=1}^{N}\left[1\{y_i=1\}\ln\left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right) + 1\{y_i=0\}\ln\left(\frac{1}{1+e^{x_i'\beta}}\right)\right]$$

最大化以上似然函数，就可以得到 β 的一致估计，进而得到 $p(x_i) = P(y_i = 1|x_i)$，即给定 x_i，$y_i = 1$ 的概率的估计。

例 29. （Probit 回归）在上例中，如果将其中的逻辑斯蒂分布换成标准正态分布，即：

$$\begin{aligned} y_i^* &= x_i'\beta - u_i \\ u_i|x_i &\sim N(0,1) \\ y_i &= 1\{y_i^* \geq 0\} \end{aligned}$$

那么我们称该模型为 Probit 回归。其极大似然函数为：

$$L(\beta|y,x) = \sum_{i=1}^{N}\left[1\{y_i=1\}\ln\left(\Phi(x_i'\beta)\right) + 1\{y_i=0\}\ln\left(1-\Phi(x_i'\beta)\right)\right]$$

例 30. （Type I Tobit 回归）如果 $(y_i, x_i')', i = 1, ..., N, x_i \in \mathbb{R}^K$ 为一系列独立

同分布的随机向量，其中：

$$\begin{aligned} y_i^* &= x_i'\beta + u_i \\ u_i|x_i &\sim N\left(0, \sigma^2\right) \\ y_i &= \max\left(y_i^*, 0\right) \end{aligned}$$

其中 y_i^* 为潜变量，我们只能观察到 $y_i^* > 0$ 时的 y_i^*，对于 $y_i^* \leq 0$，则只能观察到 0。我们称这种情况为截尾数据（censored data）。可以计算，当 $y_i = 0$ 时，其概率：

$$P\left(y_i = 0|x_i\right) = P\left(x_i'\beta + u_i \leq 0\right) = P\left(\frac{u_i}{\sigma} \leq -\frac{x_i'\beta}{\sigma}\right) = 1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)$$

而当 $y_i > 0$ 时，其密度为：

$$f\left(y_i|x_i, y_i^* > 0\right) = \frac{1}{\sigma}\phi\left(\frac{x_i'\beta}{\sigma}\right)$$

其极大似然函数为：

$$L\left(\beta|y, x\right) = \sum_{i=1}^{N}\left[1\left\{y_i > 0\right\}\ln\left(\frac{1}{\sigma}\phi\left(\frac{x_i'\beta}{\sigma}\right)\right) + 1\left\{y_i = 0\right\}\ln\left(1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right)\right]$$

习题

1. 计算例 (2) 中两个估计量的 MSE。
2. 求以下分布总体的矩估计，并验证其无偏性和一致性。
   (a) $x_i \sim Ber\left(p\right)$
   (b) $x_i \sim N\left(\mu, \sigma^2\right)$
   (c) $x_i \sim P\left(\lambda\right)$
3. 若 $x_i \sim U\left(a, b\right)$ $i.i.d$，求其矩估计，并验证其一致性。
4. 若 $x_i \sim P\left(\lambda_0\right), i = 1, ..., N$，请完成以下步骤：
   (a) 写出极大似然函数 $L\left(\lambda|x\right)$
   (b) 写出极大似然函数的概率极限 $\mathscr{L}\left(\lambda\right) = \mathbb{E}\left(\frac{1}{N}L\left(\lambda|x\right)\right)$
   (c) 证明 $\lambda_0 = \arg\max_\lambda \mathscr{L}\left(\lambda\right)$
5. 求以下分布总体的极大似然估计，证明其一致性并计算估计量的极限分布。
   (a) $x_i \sim P\left(\lambda\right)$

(b) $x_i \sim N(\mu, 1)$

(c) $x_i \sim N\left(0, \sigma^2\right)$

6. 若 $x_i^* \sim N\left(\mu, \sigma^2\right)$，但是当 $x_i^* \leq 100$ 时，我们观察不到 x_i^*，即我们观察到 x_i 满足：

$$x_i = \begin{cases} 100 & x_i^* \leq 100 \\ x_i^* & otherwise \end{cases}$$

请写出以上问题的对数似然函数。

7.（泊松回归）如果 $(y_i, x_i')', i = 1, ..., N, x_i \in \mathbb{R}^K$ 为一系列独立同分布的随机向量，且 $y_i \in \mathbb{Z}$，经常使用的模型为泊松回归（Possion regression），即假设：

$$y_i | x_i \sim P\left(e^{x_i'\beta}\right)$$

请写出其条件对数似然函数。

8. 编程题：若 $x_i \sim Beta(\alpha, \beta)$ $i.i.d$，请写出其矩估计和极大似然估计的实现。（Beta 分布随机数可以使用 random.betavariate(alpha, beta) 来生成）。

参考文献

[1] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[2] Schervish, M.J., 1995. Theory of Statistics. Springer-Verlag, New York.

[3] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.