1) 若与变量x与y满足线性关系$y = kx + b$，现有一组测量值$(x_1, y_1), (x_2, y_2)$,(x_N, y_N)，若在每一个测量点，对x的测量没有误差，对y的测量误差相同，试推导出用最小二乘法得到的k和b的估计值，用x与y 测量值表示。
（选做）若在每个测量点，对x的测量没有误差，对y的测量误差各不相同（第i个测量点y的误差为δ_i），试推导最小二乘法给出的k和b的估计值，用x与y 的测量值及y的测量误差表示

答: 若每个测量点y值测量误差都相等，$\chi^2(k, b|x, y) = \frac{1}{\delta^2}\sum_{i=1}^{i=N}(y_i - kx_i - b)^2$。最小二乘法要求：

$$\begin{aligned}\frac{\partial\chi^2}{\partial k} &= 2\sum_{i=1}^{i=N}(y_i - kx_i - b)x_i = 0\\ \frac{\partial\chi^2}{\partial b} &= 2\sum_{i=1}^{i=N}(y_i - kx_i - b) = 0\end{aligned}$$

上式两边除以N，可以表示为:

$$\begin{aligned}\bar{xy} - k\bar{x^2} - \bar{x}b = 0\\ \bar{y} - k\bar{x} - b = 0\end{aligned}$$

解得:

$$\begin{aligned}k &= (\bar{xy} - \bar{x}\bar{y})/(\bar{x^2} - \bar{x}^2)\\ b &= \bar{y} - k\bar{x}\end{aligned}$$

其中k还有多种用协方差和关联函数表示方法，如$k = \mathrm{Cov}(x, y)/\sigma_x^2$，或是$k = \mathrm{Corr}(x, y)\frac{\sigma_y}{\sigma_x}$。

如每个点的误差不相等，那么根据最小二乘法:

$$\begin{aligned}\frac{\partial\chi^2}{\partial k} &= 2\sum_{i=1}^{i=N}\frac{(y_i - kx_i - b)x_i}{\delta_i^2} = 0\\ \frac{\partial\chi^2}{\partial b} &= 2\sum_{i=1}^{i=N}\frac{(y_i - kx_i - b)}{\delta_i^2} = 0\end{aligned}$$

解得：

$$\begin{aligned}k &= (\bar{xy}_\delta C - \bar{x}_\delta\bar{y}_\delta)/(\bar{x}_\delta^2 C - \bar{x}_\delta{}^2)\\ b &= (\bar{y}_\delta - k\bar{x}_\delta)/C\end{aligned}$$

，其中：

$$\begin{aligned}
\bar{x_\delta} &= \frac{1}{N}\sum_{i=1}^{i=N}\frac{x_i}{\delta_i^2} \\
\bar{y_\delta} &= \frac{1}{N}\sum_{i=1}^{i=N}\frac{y_i}{\delta_i^2} \\
\bar{xy_\delta} &= \frac{1}{N}\sum_{i=1}^{i=N}\frac{x_i y_i}{\delta_i^2} \\
\bar{x_\delta^2} &= \frac{1}{N}\sum_{i=1}^{i=N}\frac{x_i^2}{\delta_i^2} \\
C &= \frac{1}{N}\sum_{i=1}^{i=N}\frac{1}{\delta_i^2}
\end{aligned}$$

2) 泊松分布的PDF为$P(n|\mu)=\frac{\mu^n}{n!}e^{-\mu}$，其中参数$\mu$为非负实数，变量$n$为非负整数。若对某个满足泊松分布的变量进行了$m$次独立测量，测量结果分别为$n_1, n_2, , , n_m$，试根据这组数据，是用最大似然法估计该分布的参数$\mu$。

答：

$$\begin{aligned}
&-\ln L = m\mu - \sum_{i=1}^{i=m} n_i \ln\mu + \sum_{i=1}^{i=m}\ln n_i! \\
&\frac{\partial(-\ln L)}{\partial\mu}|_{\mu=\hat{\mu}} = m - \frac{1}{\hat{\mu}}\sum_{i=1}^{i=m} n_i = 0 \\
&\hat{\mu} = \frac{1}{m}\sum_{i=1}^{i=m} n_i
\end{aligned}$$

3) a) 估算满足泊松分布变量的期望值$E(n)$和协方差: $E[(n-\mu)^2]$。

答：

$$\begin{aligned}
E(n) &= \sum_{k=0}^{k=+\infty} k\frac{\mu^k}{k!}e^{-\mu} \\
&= \sum_{k=1}^{k=+\infty}\frac{\mu^k}{(k-1)!}e^{-\mu} \\
&= \sum_{k=0}^{k=+\infty}\frac{\mu^{k+1}}{k!}e^{-\mu} \\
&= \mu\sum_{k=0}^{k=+\infty}\frac{\mu^k}{k!}e^{-\mu} = \mu
\end{aligned}$$

其中最后一步用了泊松分布的归一化性质：$\sum_{k=0}^{k=+\infty}\frac{\mu^k}{k!}e^{-\mu}=e^{\mu}e^{-\mu}=1$。

$$\begin{aligned}
E[(n-\mu)^2] &= E(n^2-2n\mu+\mu^2)\\
&= \sum_{k=0}^{k=+\infty}k^2\frac{\mu^k}{k!}e^{-\mu}-2\mu E(n)+\mu^2\\
&= \sum_{k=0}^{k=+\infty}k(k-1)\frac{\mu^k}{k!}e^{-\mu}+\sum_{k=0}^{k=+\infty}k\frac{\mu^k}{k!}e^{-\mu}-\mu^2\\
&= \sum_{k=2}^{k=+\infty}\frac{\mu^k}{(k-2)!}e^{-\mu}+\mu-\mu^2\\
&= \mu^2\sum_{k=0}^{k=+\infty}\frac{\mu^k}{k!}e^{-\mu}+\mu-\mu^2\\
&= \mu^2+\mu-\mu^2=\mu
\end{aligned}$$

b) 若k个独立的随机变量$(n_1,n_2,n_3,...,n_k)$分别满足参数为$(\mu_1,\mu_2,,\mu_k)$的泊松分布，证明这k个随机变量的和$N=n_1+n_2+,+n_k$满足参数为$\mu=\mu_1+\mu_2+...+\mu_k$的泊松分布。

答： 考虑两个独立的随机变量：n_1和n_2，分别满足参数为μ_1和μ_2的泊松分布。则他们的联合PDF函数为$f(n_1,n_2)=\frac{\mu_1^{n_1}\mu_2^{n_2}}{n_1!n_2!}e^{-(\mu_1+\mu_2)}$。观察到$n_1$和$n_2$之和为$N$的机率为:

$$\begin{aligned}
&\sum_{n_1+n_2=N}f(n_1,n_2)\\
&= e^{-(\mu_1+\mu_2)}\sum_{n_1+n_2=N}\frac{\mu_1^{n_1}\mu_2^{n_2}}{n_1!n_2!}\\
&= \frac{e^{-(\mu_1+\mu_2)}}{N!}\sum_{n_1+n_2=N}\frac{N!}{n_1!n_2!}\mu_1^{n_1}\mu_2^{n_2}\\
&= \frac{e^{-(\mu_1+\mu_2)}}{N!}(\mu_1+\mu_2)^N
\end{aligned}$$

所以n_1+n_2满足$\mu_1+\mu_2$的分布。因此通过数学归纳法，不难证明对k个独立泊松分布随机变量，其和满足$\mu=\mu_1+\mu_2+...+\mu_k$的泊松分布。

注： 这个证明使用了二项式展开公式$(a+b)^N=\sum_{k=0}^{k=N}\frac{N!}{k!(N-k)!}a^kb^{N-k}=\sum_{k_1+k_2=N}\frac{N!}{k_1!k_2!}a_1^kb_2^k$。

若用多项式展开公式$(\sum_{i=1}^{i=n}a_i)^N=\sum_{k_1+k_2+...+k_n=N}\frac{N!}{k_1!k_2!...k_n!}a_1^{k_1}a_2^{k_2}...a_n^{k_n}$可直接得证。

4) 若一所学校有m个班级，分别有n_i个学生$(1\le i\le m)$，现在在学校里随机抽选学生（每一个学生被抽到的几率相同），调查他们所在班级的人数，最后将抽查到的每个学生所在的班级人数求平均来估计每个班级的平均人数。这样的估计偏差是多少？

答： 显然，平均每个班人数为$\bar{n}=\frac{N}{m}$，其中$N=\sum_{i=1}^{i=m}n_i$为学校总人数。而通过在学校随机抽选学生的方法，抽到第i个班的学生几率为n_i/N。因此这样调研出来的班级平均人数期

望值为$\sum_{i=1}^{i=m}\frac{n_i^2}{N}$。所以偏差为$\sum_{i=1}^{i=m}\frac{n_i^2}{N}-\bar{n}$。这个表达式可以用各个班级人数$n_i$的均方差表示。$n_i$均方差为$\sigma_n^2=\frac{\sum_{i=1}^{i=m}n_i^2}{m}-N^2/m^2=\bar{n}(\frac{\sum_{i=1}^{i=m}n_i^2}{N}-\bar{n})$。所以要算的偏差为$\sigma_n^2/\bar{n}$。

注： 这个问题是检查悖论(Inspection Paradox)的一个典型例子。这是一种由于样本空间的改变引起的偏差。在统计每个班级人数时，样本空间为这m个班级，每个班级是平权的；但是在通过随机选定的学生调查时，调查样本对每个学生是平权的，但是对每个班级并不平权（人数多的班级几率更大，所以权重更高）。类似的问题还有不少，如公交车悖论（乘客平均等待时间往往比公交平均发车间隔长）等。

5) 若一个随机变换的电压值满足1.7V-1.8V之间的均匀分布，现在用一个量程为0-2 V，显示精度为0.01V的数显电压表去测量这个电压，那么由于电压表显示精度造成的测量误差为多少？

答： 在电压真值为U （$1.7\text{V}\le U\le 1.8\text{V}$）的情况下，测量值$U_M=[100U]/100$。显然，$U_M-U$是一个周期为0.01 V的周期函数。我们只要计算它在一共周期内的误差。电压表显示精度可用平均平方(MSE)估算:

$$\begin{aligned}
E[(U_M-U)^2] &= \frac{1}{0.1\text{ V}}\int_{1.7\text{ V}}^{1.8\text{ V}}(U_M-U)^2dU\\
&= \frac{1}{0.01\text{ V}}\int_{1.70\text{ V}}^{1.71\text{ V}}(U_M\text{ V}-U)^2dU\\
&= \frac{1}{0.01\text{ V}}\int_{1.70\text{ V}}^{1.71\text{ V}}(1.70\text{ V}-U)^2dU\\
&= \frac{1}{0.01\text{ V}}\int_{0\text{ V}}^{0.01\text{ V}}u^2du\\
&= \frac{10^{-4}}{3}\text{V}^2
\end{aligned}$$

注： 因为这个电压表读数总是小于或等于电压真值，所以测量结果是有偏差的。该偏差为$E(U_M-U)=-0.005$ V。如将每次测量结果加上0.005 V，就可消除这个偏差，那么$E[(U_M-U)^2]=\frac{10^{-4}}{12}\text{V}^2$。一般的，在含有偏差的情况下，MSE为$b^2+\sigma^2$，其中$b$为偏差，$\sigma^2$为方差。在这个问题中，$b^2$和$\sigma^2$分别为$\frac{10^{-4}}{4}\text{V}^2$ 和$\frac{10^{-4}}{12}\text{V}^2$。

6) 考虑通过测量一个金属小球的直径来估算其体积，若每次测量结果满足正态分布$P(d|d_T,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(d-d_T)^2}{2\sigma^2}}$ ，其中d为小球直径的测量值，d_T 为小球直径真值，且$\sigma\ll d_T$。现在考虑两种估算方法：

   a) 对小球直径进行多次测量，得到直径测量值$(d_1,d_2...,d_N)$，将这些测量值的平均值$\bar{d}=(d_1+d_2+...+d_N)/N$当作小球直径，估算出小球体积$V=\frac{1}{6}\pi\bar{d}^3$。

   b) 对小球直径进行多次测量，得到直径测量值$(d_1,d_2,...,d_N)$，根据这些直径估算出体积$(V_1,V_2,...,V_N)$，其中$V_i=\frac{1}{6}\pi d_i^3$，最终估算小球体积为其平均值$V=\bar{V}=(V_1+V_2+...+V_N)/N$。

请判断a)和b)两种方法分别是否是对小球体积的无偏估计（如果不是无偏估计试对偏差进行估计）？它们又是否是小球体积的一致估计？

答： 两者都不是无偏估计，a)为有一致估计，b)不是一致估计。首先看下b)中估算出来的体积偏差。按照b)中的测量方法，体积的期望值为：

$$\begin{aligned}
& E(\frac{1}{6}\pi d^3) \\
= & \frac{1}{6}\pi \int_{-\infty}^{+\infty} d^3 \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(d-d_T)^2}{2\sigma^2}} \mathrm{d}d \\
= & \frac{1}{6}\pi \int_{-\infty}^{+\infty} (t+d_T)^3 \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-t^2}{2\sigma^2}} \mathrm{d}t \\
= & \frac{1}{6}\pi \int_{-\infty}^{+\infty} (t^3+3t^2d_T+3td_T^2+d_T^3) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-t^2}{2\sigma^2}} \mathrm{d}t \\
= & \frac{1}{6}\pi \int_{-\infty}^{+\infty} (3t^2d_T+d_T^3) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-t^2}{2\sigma^2}} \mathrm{d}t \\
= & \frac{3d_T}{6}\pi \int_{-\infty}^{+\infty} (t^2) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-t^2}{2\sigma^2}} \mathrm{d}t + \frac{1}{6}\pi d_T^3 \\
= & \frac{3d_T\sigma^2}{6}\pi + \frac{1}{6}\pi d_T^3 \\
= & V_T(1+3\sigma^2/d_T^2)
\end{aligned}$$

式中$V_T = \frac{1}{6}\pi d_T^3$为小球体积真值。可见测量偏差为$3V_T\sigma^2/d_T^2$。可见方法b)不是无偏的。而随着测量次数$N \to \infty$，平均体积$\bar{V}$会趋于中心值为$V_T(1+3\sigma^2/d_T^2)$，的正态分布，故这个偏差无法随着$N \to \infty$而趋于0。可见方法b)也不是小球体积的一致估计。而在a)的方法中，根据中心极限定理，$N \to \infty$时，$\bar{d}$会趋于中心值为d_T,σ^2为σ^2/N的正态分布。根据前面的讨论，用$\bar{d}$计算体积会导致$V_T(1+\frac{3\sigma^2}{Nd_T^2})$的偏差。但是这个偏差显然随着$N \to \infty$而趋于0。所以a)为一致估计，但不是无偏估计。

7) 若随机变量x与y二维正态分布：$P(x,y) = Ke^{-(x^2+3xy+3y^2)}$，其中K为归一化常数：a)计算$x$与$y$的期望值b)计算$x$与$y$的协方差矩阵。

答 x的期望值为：

$$\begin{aligned}
E(x) = & \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xKe^{-(x^2+3xy+3y^2)}\mathrm{d}x\mathrm{d}y \\
= & K\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xe^{-((x+\frac{3y}{2})^2+\frac{3}{4}y^2)}\mathrm{d}x\mathrm{d}y \\
= & K\int_{-\infty}^{+\infty} e^{-\frac{3}{4}y^2}\int_{-\infty}^{+\infty} (x+\frac{3}{2}y-\frac{3}{2}y)e^{-(x+\frac{3y}{2}y)^2}\mathrm{d}(x+\frac{3y}{2}y)\mathrm{d}y \\
= & K\int_{-\infty}^{+\infty} e^{-\frac{3}{4}y^2}\int_{-\infty}^{+\infty} -\frac{3}{2}ye^{-(x+\frac{3y}{2}y)^2}\mathrm{d}(x+\frac{3y}{2}y)\mathrm{d}y \\
= & -K\int_{-\infty}^{+\infty} e^{-\frac{3}{4}y^2}\frac{3}{2}Cy\mathrm{d}y \\
= & 0
\end{aligned}$$

同理可以计算出$E(y)=0$。x和y的协方差矩阵为：

$$S=\begin{pmatrix}\mathrm{Cov}(x,x) & \mathrm{Cov}(x,y)\\ \mathrm{Cov}(x,y) & \mathrm{Cov}(y,y)\end{pmatrix}$$

其中:

$$\begin{aligned}&\mathrm{Cov}(x,x)=\sigma_x^2=E[(x-E(x))^2]=E(x^2)\\&\mathrm{Cov}(y,y)=\sigma_y^2=E[(y-E(y))^2]=E(y^2)\\&\mathrm{Cov}(x,y)=E[(x-E(x))(y-E(y))]=E(xy)\end{aligned}$$

在计算中，我们会用到正态分布的一些性质：

$$\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}\mathrm{d}x=1 \tag{1}$$

$$\int_{-\infty}^{+\infty}\frac{x^2}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}\mathrm{d}x=\sigma^2 \tag{2}$$

其中第一个等式为正态分布归一化条件，第二个等式为正态分布方差性质。由此可得：

$$\int_{-\infty}^{+\infty}e^{-Ax^2}\mathrm{d}x=\sqrt{\frac{\pi}{A}} \tag{3}$$

$$\int_{-\infty}^{+\infty}x^2e^{-Ax^2}\mathrm{d}x=\frac{1}{2A}\sqrt{\frac{\pi}{A}} \tag{4}$$

下面计算协方差矩阵元：

$$\begin{aligned}E(x^2) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x^2Ke^{-(x^2+3xy+3y^2)}\mathrm{d}x\mathrm{d}y\\&= \int_{-\infty}^{+\infty}Ke^{-\frac{3y^2}{4}}\int_{-\infty}^{+\infty}[(x+\frac{3}{2}y)-\frac{3}{2}y]^2e^{-(x+\frac{3}{2}y)^2}\mathrm{d}(x+\frac{3}{2}y)\mathrm{d}y\\&= \int_{-\infty}^{+\infty}Ke^{-\frac{3y^2}{4}}\int_{-\infty}^{+\infty}(t^2-3ty+\frac{9}{4}y^2)e^{-t^2}\mathrm{d}t\mathrm{d}y\\&= \int_{-\infty}^{+\infty}Ke^{-\frac{3y^2}{4}}(\frac{\sqrt{\pi}}{2}+\frac{9\sqrt{\pi}}{4}y^2)\mathrm{d}y\\&= \frac{\sqrt{\pi}K}{2}\sqrt{\frac{4}{3}\pi}+\frac{9K}{4}\sqrt{\pi}\frac{1}{2\times\frac{3}{4}}\sqrt{\frac{4}{3}\pi}=\sqrt{\frac{\pi}{3}}+\sqrt{3\pi}=\frac{4K\pi}{\sqrt{3}}\end{aligned}$$

$$\begin{aligned}E(y^2) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}y^2Ke^{-(x^2+3xy+3y^2)}\mathrm{d}x\mathrm{d}y\\&= \int_{-\infty}^{+\infty}e^{-(x+\frac{3}{2}y)^2}\mathrm{d}(x+\frac{3}{2}y)\mathrm{d}y\\&= \int_{-\infty}^{+\infty}Ky^2e^{-\frac{3y^2}{4}}\sqrt{\pi}\mathrm{d}y=\frac{4K\pi}{3\sqrt{3}}\end{aligned}$$

$$
\begin{aligned}
E(xy) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyKe^{-(x^2+3xy+3y^2)}\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{+\infty} Ke^{-\frac{3y^2}{4}}\int_{-\infty}^{+\infty}(ty-\frac{3}{2}y^2)e^{-t^2}\mathrm{d}t\mathrm{d}y \\
&= -\frac{3}{2}K\sqrt{\pi}\int_{-\infty}^{+\infty} y^2e^{-\frac{3y^2}{4}}\mathrm{d}y \\
&= -\frac{3}{2}K\sqrt{\pi}\frac{1}{2\times\frac{3}{4}}\sqrt{\frac{4}{3}\pi} = -\frac{2K\pi}{\sqrt{3}}
\end{aligned}
$$

上面计算中，$t = x + \frac{3}{2}y$。所以,x与y协方差矩阵为:

$$
S = \begin{pmatrix} \frac{4K\pi}{\sqrt{3}} & -\frac{2K\pi}{\sqrt{3}} \\ -\frac{2K\pi}{\sqrt{3}} & \frac{4K\pi}{3\sqrt{3}} \end{pmatrix}
$$

注： 此处可以通过归一化条件将K值求出来，最后的协方差矩阵还可以进一步化简。归一化条件要求:

$$
\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Ke^{-(x^2+3xy+3y^2)}\mathrm{d}x\mathrm{d}y = 1
$$

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} e^{-\frac{3y^2}{4}}\int_{-\infty}^{+\infty} e^{-t^2}\mathrm{d}t\mathrm{d}y = \int_{-\infty}^{+\infty} e^{-\frac{3y^2}{4}}\mathrm{d}y\int_{-\infty}^{+\infty} e^{-t^2}\mathrm{d}t \\
&= \sqrt{\pi}\sqrt{\frac{4}{3}\pi} = \frac{2\pi}{\sqrt{3}}
\end{aligned}
$$

因此$K = \frac{\sqrt{3}}{2\pi}$。$x$与$y$的协方差矩阵可以进一步简化为:

$$
S = \begin{pmatrix} 2 & -1 \\ -1 & \frac{2}{3} \end{pmatrix}
$$