# 《概率论与数理统计》习题

第二十一讲　贝叶斯估计

1. 设一页书上的错别字个数服从泊松分布 $P(\lambda)$，$\lambda$ 有两个可能取值：1.5 和 1.8，且先验分布为

$$P(\lambda = 1.5) = 0.45, \quad P(\lambda = 1.8) = 0.55$$

现检查了一页，发现有 3 个错别字，试求 $\lambda$ 的后验分布。

解：

由题意可得，

$$P(X = 3|\lambda = 1.5) = \frac{1.5^3}{3!}e^{-1.5}.$$

$$P(X = 3|\lambda = 1.8) = \frac{1.8^3}{3!}e^{-1.8}.$$

$$\begin{aligned} P(X = 3) &= P(X = 3|\lambda = 1.8)P(\lambda = 1.8) \\ &\quad + P(X = 3|\lambda = 1.5)P(\lambda = 1.5) \\ &= \frac{1.51875e^{-1.5} + 3.207e^{-1.8}}{6}. \end{aligned}$$

接下来计算 $\lambda$ 的后验分布，

$$\begin{aligned} P(\lambda = 1.5|X = 3) &= \frac{P(X = 3|\lambda = 1.5)P(\lambda = 1.5)}{P(X = 3)} \\ &= 0.3899. \\ P(\lambda = 1.8|X = 3) &= 1 - P(\lambda = 1.5|X = 3) \\ &= 0.6101. \end{aligned}$$

2. 验证：正态分布方差（均值已知）的共轭先验分布是倒伽玛分布。(提示：若 $X$ 服从伽玛分布，那么称随机变量 $1/X$ 的分布为倒伽玛分布。)
证明：

设总体 $X|\sigma^2 \sim N(\mu_0, \sigma^2)$，其中 $\mu_0$ 已知，$x_1, x_2, \dots, x_n$ 为样本，令 $\sigma^2$ 的先验分布为倒伽马分布 $IGa(\alpha, \lambda)$，其密度函数为

$$\pi(\sigma^2) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\Big(\frac{1}{\sigma^2}\Big)^{\alpha+1} e^{-\frac{\lambda}{\sigma^2}}, \sigma^2 > 0.$$

则 $\sigma^2$ 的后验分布为，

$$\begin{aligned}
&\pi(\sigma^2|x_1, x_2, \dots, x_n)\\
&= \frac{p(x_1, x_2, \dots, x_n|\sigma^2)\pi(\sigma^2)}{\int p(x_1, x_2, \dots, x_n|\sigma^2)\pi(\sigma^2)d\sigma^2}\\
&= \frac{(2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2\right\} \frac{\lambda^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} e^{-\frac{\lambda}{\sigma^2}}}{\int_0^{+\infty} (2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2\right\} \frac{\lambda^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} e^{-\frac{\lambda}{\sigma^2}} d\sigma^2}\\
&= \frac{\left(\frac{1}{\sigma^2}\right)^{\alpha+\frac{n}{2}+1} exp\left\{-\frac{1}{\sigma^2}\left[\lambda + \frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2\right]\right\}}{\int_0^{+\infty} \left(\frac{1}{\sigma^2}\right)^{\alpha+\frac{n}{2}+1} exp\left\{-\frac{1}{\sigma^2}\left[\lambda + \frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2\right]\right\} d\sigma^2}\\
&= \frac{\left[\lambda + \frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2\right]^{\alpha+\frac{n}{2}}}{\Gamma(\alpha+\frac{n}{2})}\left(\frac{1}{\sigma^2}\right)^{\alpha+\frac{n}{2}+1} exp\left\{-\frac{1}{\sigma^2}\left[\lambda + \frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2\right]\right\}.
\end{aligned}$$

可以看出，$\sigma^2|x_1, x_2, \dots, x_n \sim IGa(\alpha + \frac{n}{2}, \lambda + \frac{1}{2}\sum_{i=1}^n (x_i-\mu_0)^2)$，即证明成立。 ■

3. 设 $x_1, x_2, \cdots, x_n$ 为来自如下幂级数分布的样本，总体分布密度为

$$p(x; c, \theta) = cx^{c-1}\theta^{-c} I_{\{0 \le x \le \theta\}}, c > 0, \theta > 0.$$

证明:

(a) 若 $c$ 已知，则 $\theta$ 的共轭先验分布为帕雷托分布；

(b) 若 $\theta$ 已知，则 $c$ 的共轭先验分布为伽玛分布。

(a) 设 $\pi(\theta) = \alpha\mu^\alpha\theta^{-(\alpha+1)} I\{\theta \ge \mu\} \quad \alpha \ge 1, \mu > 0$. 则 $\theta$ 的后验分布密度函数为

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_0^{+\infty} p(x|\theta)\pi(\theta)\mathrm{d}\theta}$$

$$= \frac{c^n \left(\prod_{i=1}^n x_i\right)^{c-1} \theta^{-nc} I\left\{\theta \geqslant x_{(n)}\right\} \cdot \alpha\mu^\alpha\theta^{-(1+\alpha)} I\left\{\theta \geqslant \mu\right\}}{\int_0^{+\infty} c^n \left(\prod_{i=1}^n x_i\right)^{c-1} \theta^{-nc} I\left\{\theta \geqslant x_{(n)}\right\} \cdot \alpha\mu^\alpha\theta^{-(1+\alpha)} I\{\theta \geqslant \mu\}\mathrm{d}\theta}$$

$$=\frac{\theta^{-nc}\cdot\theta^{-(1+\alpha)}I\left\{\theta\geqslant\theta_0\right\}}{\int_{\theta_0}^{+\infty}\theta^{-nc}\cdot\theta^{-(1+\alpha)}\mathrm{d}\theta}=(nc+\alpha)\theta_0^{nc+\alpha}\theta^{-(nc+\alpha+1)}I\left\{\theta\geqslant\theta_0\right\}$$

其中 $\theta_0=max\{x_{(n)},\mu\}$, 因此

$$\pi(\theta|x)\sim PA(nc+\alpha,\theta_0)$$

所以当 $c$ 已知时帕雷托分布为 $\theta$ 的共轭先验分布.

(b) 设 $\pi(c)=\frac{\lambda^\alpha}{\Gamma(\alpha)}e^{-\lambda c}\cdot c^{\alpha-1}I\{c>0\}\quad \alpha>0,\mu>0$. 则 $c$ 的后验密度函数为

$$\pi(c\mid x)=\frac{p(x\mid c)\pi(c)}{\int_0^\infty p(x\mid c)\pi(c)dc}$$

$$=\frac{c^n\left(\prod_{i=1}^n x_i\right)^{c-1}\theta^{-nc}\cdot e^{-\lambda c}c^{\alpha-1}}{\int_0^\infty c^n\left(\prod_{i=1}^n x_i\right)^{c-1}\theta^{-nc}\cdot e^{-\lambda c}c^{\alpha-1}\ \mathrm{d}c}$$

$$=\frac{\left(\lambda-\sum_{i=1}^n(\ln x_i-\ln\theta)\right)^{n+\alpha}}{\Gamma(n+\alpha)}c^{n+\alpha-1}\exp\left\{-c\left[\lambda-\sum_{i=1}^n(\ln x_i-\ln\theta)\right]\right\}$$

这说明 $c\mid x\sim Ga\left(n+\alpha,\lambda-\sum_{i=1}^n(\ln x_i-\ln\theta)\right)$.

4. 设 $x_1,x_2,\cdots,x_n$ 是来自参数为 $\lambda$ 的泊松分布 $P(\lambda)$ 的样本. 假定 $\lambda$ 的先验分布为伽玛分布 $Ga(\alpha,\beta)$.

(a) 计算 $\lambda$ 的后验分布。

(b) 求 $\lambda$ 的贝叶斯估计 $\hat{\lambda}_1$。

(c) 求 $\lambda$ 的极大似然估计 $\hat{\lambda}_2$。

解:

(1)

$$
\begin{aligned}
P(x|\lambda) &= \frac{\lambda^x}{x!}e^{-\lambda} \\
\pi(\lambda) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \\
h(x_1, x_2, \ldots, x_n, \lambda) &= \prod_{i=1}^{n}\frac{\lambda^{x_i}}{(x_i)!}e^{-\lambda}\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \\
m(x_1, x_2, \ldots, x_n) &= \int_0^{+\infty} h(x_1, x_2, \ldots, x_n, \lambda)d\lambda \\
\pi(\lambda|x_1, x_2, \ldots, x_n) &= \frac{h(x_1, x_2, \ldots, x_n, \lambda)}{m(x_1, x_2, \ldots, x_n)} \\
&= \frac{\prod_{i=1}^{n}\frac{\lambda^{x_i}}{(x_i)!}e^{-\lambda}\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}}{\int_0^{+\infty}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{(x_i)!}e^{-\lambda}\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}d\lambda} \\
&= \frac{(n+\beta)^{(\sum_{i=1}^{n}x_i+\alpha)}}{\Gamma(\sum_{i=1}^{n}x_i+\alpha)}\lambda^{\sum_{i=1}^{n}x_i+\alpha-1}e^{-(n+\beta)\lambda}
\end{aligned}
$$

可得，$\lambda|x_1, x_2, \ldots, x_n \sim Ga(\sum_{i=1}^{n}x_i+\alpha, n+\beta)$.

(2) 使用后验分布的均值作为参数的贝叶斯估计：

$$\lambda|X \sim Ga(\sum_{i=1}^{n}x_i+\alpha, n+\beta).$$

$$\hat{\lambda}_1 = \frac{\sum_{i=1}^{n}x_i+\alpha}{n+\beta}.$$

(3) 参数的 MLE：

$$
\begin{aligned}
L(\lambda) &= P(x_1, x_2, \ldots, x_n|\lambda) \\
&= \prod_{i=1}^{n}\frac{\lambda^{x_i}}{(x_i)!}e^{-\lambda}. \\
lnL(\lambda) &= \sum_{i=1}^{n}x_i ln\lambda - \lambda n - \sum_{i=1}^{n}ln(x_i!).
\end{aligned}
$$

$lnL(\lambda)$ 关于 $\lambda$ 求导，并令其等于 0，解得 $\hat{\lambda}_2 = \bar{x}$。

5. 设随机变量 $X$ 服从负二项分布，其概率分布为

$$f(x|p) = \binom{x-1}{k-1}p^k(1-p)^{x-k}, x = k, k+1, \cdots$$

证明其成功概率 $p$ 的共轭先验分布族为贝塔分布族。

证明:

令成功概率 p 的先验分布为 Be(a,b)，a>0，b>0，则 $x_1, x_2, \ldots, x_n$ 与 $\theta$ 的联合分布为:

$$
\begin{aligned}
h(x_1, x_2, \ldots, x_n, p) &= \prod_{i=1}^{n} \binom{x_i - 1}{k - 1} p^{nk} (1-p)^{\sum_{i=1}^{n} x_i - nk} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{\alpha-1} (1-p)^{b-1}. \\
m(x_1, x_2, \ldots, x_n) &= \int_0^1 h(x_1, x_2, \ldots, x_n, p) dp \\
&= \prod_{i=1}^{n} \binom{x_i - 1}{k - 1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(nk+a)\Gamma(\sum_{i=1}^{n} x_i - nk + b)}{\Gamma(\sum_{i=1}^{n} x_i + a + b)}. \\
\pi(p|x_1, x_2, \ldots, x_n) &= \frac{h(x_1, x_2, \ldots, x_n, p)}{m(x_1, x_2, \ldots, x_n)} \\
&= \frac{\Gamma(\sum_{i=1}^{n} x_i + a + b)}{\Gamma(nk+a)\Gamma(\sum_{i=1}^{n} x_i - nk + b)} p^{nk+a-1} (1-p)^{\sum_{i=1}^{n} x_i - nk + b - 1}.
\end{aligned}
$$

即成功概率 p 的后验分布为 $Be(nk+a, \sum_{i=1}^{n} x_i - nk + b)$，证明成立。■

## 第二十二讲　区间估计

1. 阅读书第 300 页中例 6.6.1。采用随机模拟的方法，按以下设定复现图 6.6.1。$x_1, x_2, \cdots, x_n$ 来自正态分布 $N(\mu, \sigma^2)$，我们打算构造 $\mu$ 的区间估计。假定均值 $\mu$ 的真值为 0，方差 $\sigma^2$ 的真值为 $3^2 = 9$。考虑四种情况：

   情况一：方差已知，即 $\sigma^2 = 9$ 时，样本量 $n = 10$；

   情况二：方差已知，即 $\sigma^2 = 9$ 时，样本量 $n = 30$；

   情况三：方差未知时，样本量 $n = 10$；

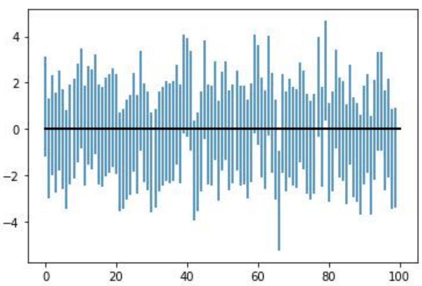
   情况四：方差未知时，样本量 $n = 30$；

   在上述的每一种情况下，对 $\mu$ 构造置信水平为 95% 的置信区间，并重复 100 次。由此，得到 100 个区间，并类似地绘制图 6.6.1。比较所绘制的四张图，你可以得到怎样的结论？(答案写清结论及理由，图像可打印上交)

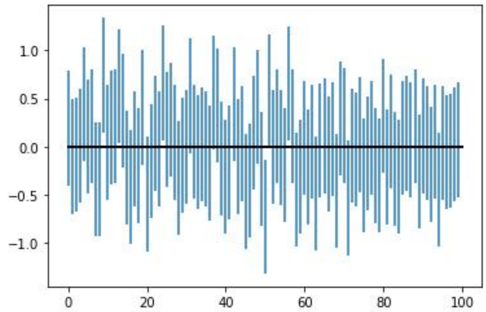
解：(仅供参考)

```
In [3]: import scipy
        from scipy import stats
        import numpy as np
        import matplotlib.pyplot as plt
```

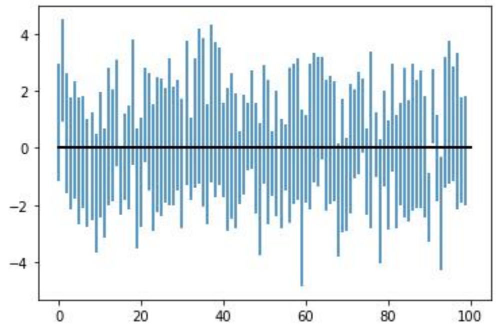
```
In [4]: alpha = 0.05
        # 方差已知为3 样本大小为10
        for i in range(100):
            n = 10
            samples=stats.norm.rvs(0,3,size=n)
            average = np.average(samples)
            std = 3
            left = average-scipy.stats.t.ppf(1-alpha/2,n-1)*std/(n**0.5)
            right = average+scipy.stats.t.ppf(1-alpha/2,n-1)*std/(n**0.5)
            region = (left,right)
            plt.plot([0,100],[0,0],color='black')
            plt.vlines(i,left,right)
```



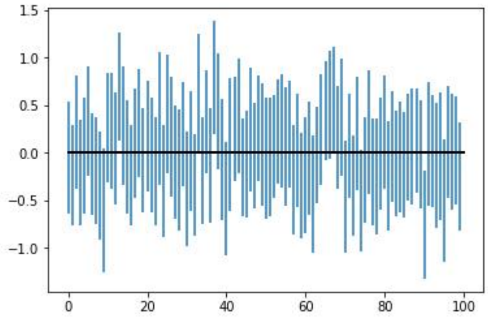
In [8]:
```python
alpha = 0.05
# 方差已知为3 样本大小为100
for i in range(100):
    n = 100
    samples=stats.norm.rvs(0,3,size=n)
    average = np.average(samples)
    std = 3
    left = average-scipy.stats.t.ppf(1-alpha/2,n-1)*std/(n**0.5)
    right = average+scipy.stats.t.ppf(1-alpha/2,n-1)*std/(n**0.5)
    region = (left,right)
    plt.plot([0,100],[0,0],color='black')
    plt.vlines(i,left,right)
```




In [6]:
```python
alpha = 0.05
# 方差未知 样本大小为10
for i in range(100):
    n = 10
    samples=stats.norm.rvs(0,3,size=n)
    average = np.average(samples)
    std = np.std(samples)
    left = average-scipy.stats.t.ppf(1-alpha/2,n-1)*std/(n**0.5)
    right = average+scipy.stats.t.ppf(1-alpha/2,n-1)*std/(n**0.5)
    region = (left,right)
    plt.plot([0,100],[0,0],color='black')
    plt.vlines(i,left,right)
```



In [7]:
```
alpha = 0.05
# 方差未知 样本大小为100
for i in range(100):
    n = 100
    samples=stats.norm.rvs(0,3,size=n)
    average = np.average(samples)
    std = np.std(samples)
    left = average-scipy.stats.t.ppf(1-alpha/2,n-1)*std/(n**0.5)
    right = average+scipy.stats.t.ppf(1-alpha/2,n-1)*std/(n**0.5)
    region = (left,right)
    plt.plot([0,100],[0,0],color='black')
    plt.vlines(i,left,right)
```




(1) 当方差和样本量确定时，区间长度一致；

(2) 当方差已知，随着样本量增大，区间长度变小，估计的精确度越高；当样本量已知、方差未知时，区间长度会受到样本方差影响；

(3) 观察得到，区间所包含的参数真值个数与置信水平大体上一致，是置信水平的一个合理解释。

2. 总体 $X \sim N(\mu, \sigma^2)$，$\sigma^2$ 已知，问样本量 $n$ 取多大时才能保证 $\mu$ 的置信水平为 95% 的置信区间的长度不大于 $k$。

解:

可得 $\mu$ 的 95% 置信区间为

$$\left[\bar{x} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

其区间长度为 $2u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. 若使得 $2u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq k$，可得:

$$n \geq \frac{4}{k^2}\sigma^2 u_{1-\frac{\alpha}{2}}^2 = (\frac{3.92\sigma}{k})^2.$$

即样本量 n 至少取 $(\frac{3.92\sigma}{k})^2$ 时，才能保证 $\mu$ 的置信水平为 95% 的置信区间的长度不小于 k。

3. 假设人体身高服从正态分布，今抽测甲、乙两地区 18 岁 ~ 25 岁女青年身高数据如下：甲地抽取 10 名，样本均值 1.64 米，样本标准差 0.2 米；乙地区抽取 10 名，样本均值 1.62 米，样本标准差 0.1 米。求
   (a) 两样本总体方差比的置信水平为 95% 的置信区间；
   (b) 两样本总体均值差的置信水平为 95% 的置信区间；

解：

设 $x_1, x_2, \ldots, x_{10}$ 为甲地区抽取的女青年身高，$y_1, y_2, \ldots, y_{10}$ 为乙地区抽取的女青年身高，且 $\bar{x} = 1.64$，$\bar{y} = 1.62$，$s_x = 0.2$，$s_y = 0.1$.

(1) $\frac{\sigma_x^2}{\sigma_y^2}$ 的 $1-\alpha$ 置信区间为：

$$\left[\frac{s_x^2}{s_y^2} \cdot \frac{1}{F_{1-\frac{\alpha}{2}}(m-1.n-1)}, \frac{s_x^2}{s_y^2} \cdot \frac{1}{F_{\frac{\alpha}{2}}(m-1.n-1)}\right].$$

其中，$\alpha = 0.05$，m=n=10，$F_{0.975}(9,9) = 4.03$，$F_{0.025}(9,9) = \frac{1}{F_{0.975}(9,9)}$.
所以 $\frac{\sigma_x^2}{\sigma_y^2}$ 的 95% 置信区间为：

$$[0.9926, 16.1200].$$

(2) 由于 (1) 的 95% 置信区间包含 1，因此有理由假定两个正态总体的方差相等，可得：

$$s_w^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} = 0.025$$

所以，两样本总体均值差的置信水平为 95% 的置信区间：

$$\bar{x} - \bar{y} \pm \sqrt{\frac{m+n}{mn} s_w^2} t_{1-\frac{\alpha}{2}}(m+n-2) = [-0.1286, 0.1686].$$

注意：(1) 本题的 $s_y = 0.1$，请以发布的作业为准；(2) 第二问也可以不假定方差相等，采用近似方法求置信区间。

4. 设总体 $X$ 的密度函数为

$$p(x;\theta) = e^{-(x-\theta)} I_{\{x>\theta\}}, -\infty < \theta < \infty$$

$x_1, x_2, \cdots, x_n$ 为抽自此总体分布的简单随机样本。

(a) 证明：$x_{(1)} - \theta$ 的分布与 $\theta$ 无关，并求出此分布；

(b) 求 $\theta$ 的置信水平的 $1-\alpha$ 置信区间。

解：

(1) 证明：

令 $y_i = x_i - \theta$，$i = 1, 2, \ldots, n$，则 $y_i \sim Exp(1)$ 且相互独立. $y_{(1)}$ 的密度函数为

$$g(y) = ne^{-ny}, y > 0.$$

即 $x_{(1)} - \theta$ 的分布与 $\theta$ 无关，其密度函数为 $g(y) = ne^{-ny}, y > 0.$ ■

(2)

取 c 和 d 使得：

$$P(c \le x_{(1)} - \theta \le d) = \int_c^d ne^{-ny} dy = 1 - \alpha.$$

由于 $g(y) = ne^{-ny}$ 在 $y > 0$ 上单调递减，为使得区间长度最短，故 $c = 0$，$d = \frac{-ln\alpha}{n}$. 所以，$\theta$ 的置信水平的 $1-\alpha$ 置信区间为 $\left[x_{(1)} + \frac{ln\alpha}{n}, x_{(1)}\right]$.

5. $0.50, 1.25, 0.80, 2.00$ 是取自总体 $X$ 的样本，已知 $Y = \ln X$ 服从正态分布 $N(\mu, 1)$.

(a) 求 $\mu$ 的置信水平为 95% 的置信区间；

(b) 求 $X$ 的数学期望的置信水平为 95% 的置信区间；

解：

(1) 将样本数据进行 $Y = lnX$ 转化，所得 Y 的样本值为：

$$-0.6931, 0.2231, -0.2231, 0.6931.$$

并将其看作是来自正态总体 $N(\mu, 1)$ 的样本，且 $\bar{y} = 0$，$\sigma = 1$ 已知，因此，$\mu$ 的置信水平为 95% 的置信区间为

$$\left[\bar{y} - u_{1-\frac{\alpha}{2}}/\sqrt{n}, \bar{y} + u_{1-\frac{\alpha}{2}}/\sqrt{n}\right] = [-0.98, 0.98].$$

(2) 可得，$X = e^Y$，$EX = e^{\mu+\frac{1}{2}}$ 是关于 $\mu$ 的增函数，所以 X 的数学期望的置信水平为 95% 的置信区间为：

$$\left[e^{\mu_l+\frac{1}{2}}, e^{\mu_h+\frac{1}{2}}\right] = [0.6188, 4.3929].$$