STA2004 Homework 2 Student ID: 121090429

(Question1): (15 points) The Hardy-Weinberg formula for the genetic variation of a population at equilibrium states that with probability π of the A allele, the probabilities of genotypes (AA, Aa, aa) are $(\pi^2, 2\pi(1-\pi), (1-\pi)^2)$. Assume a multinomial distribution for n observations with counts (y_1, y_2, y_3) of (AA, Aa, aa).

(a) Express the log-likelihood function for the multinomial and find the MLE of π, denoted as $\hat{\pi}$.

(b) Find $I(\pi)$, and thus specify the asymptotic distribution of $\hat{\pi}$ (Hint: you may need $\text{Cov}[Y_i, Y_j] = -np_ip_j$ if $(Y_1, Y_2, \ldots, Y_K) \sim multinomial(n, p_1, p_2, \ldots, p_K)$).

(c) Explain how you could use the asymptotic distribution in (b) to construct a 95% confidence interval for π.

(a) The maximum likelihood function is:

$$L(\pi) = (\pi^2)^{y_1} \cdot [2\pi(1-\pi)]^{y_2} \cdot [(1-\pi)^2]^{y_3}$$
$$= \pi^{2y_1} \cdot (2\pi)^{y_2} (1-\pi)^{y_2} (1-\pi)^{2y_3}$$
$$= 2^{y_2} \cdot \pi^{2y_1+y_2} (1-\pi)^{y_2+2y_3}$$

$$\ell(\pi) = \log L = y_2 \log 2 + (2y_1+y_2)\log\pi + (y_2+2y_3)\log(1-\pi)$$

$$\ell'(\pi) = \frac{2y_1+y_2}{\pi} - \frac{y_2+2y_3}{1-\pi}$$
$$= \frac{2y_1+y_2-2(y_1+y_2+y_3)\pi}{\pi(1-\pi)}$$
$$= \frac{2y_1+y_2-2n\pi}{\pi(1-\pi)}$$

Let $\ell'(\hat{\pi}) = 0 \Rightarrow \hat{\pi} = \frac{2y_1+y_2}{2n}$

$$\ell''(\pi) = \frac{-2y_1-y_2}{\pi^2} - \frac{y_2+2y_3}{(1-\pi)^2} < 0, \forall \pi$$

So $\hat{\pi} = \frac{2y_1+y_2}{2n}$ is the global maximizer of $L(\pi)$.

(b) $E(y_1) = n \cdot \pi^2$

$E(y_2) = 2n\pi(1-\pi)$

y_1 follows Binomial distribution: $y_1 \sim B(n, \pi^2)$,

So $Var(y_1) = n\pi^2(1-\pi^2)$.

y_2 follows Binomial distribution: $y_2 \sim B(n, 2\pi(1-\pi))$,

So $Var(y_2) = 2n\pi(1-\pi)(1-2\pi+2\pi^2)$

$$\Rightarrow E(2y_1+y_2) = 2n\pi^2 + 2n\pi(1-\pi) = 2n\pi$$

$$E(2y_1+y_2)^2 = Var(2y_1+y_2) + [E(y_1+2y_2)]^2$$
$$= 4Var(y_1) + Var(y_2) + 4Cov(y_1, y_2) + 4n^2\pi^2$$
$$= 4Var(y_1) + Var(y_2) + 2Cov(2y_1, y_2) + 4n^2\pi^2$$
$$= 4n\pi^2(1-\pi^2) + 2n\pi(1-\pi)(1-2\pi+2\pi^2) - 8n\pi^3(1-\pi) + 4n^2\pi^2$$
$$= 2n\pi - 2n\pi^2 + 4n^2\pi^2$$

$$I(\pi) = E\left(\frac{\partial \ell(\pi)}{\partial \pi}\right)^2 = E\left(\frac{2y_1+y_2-2n\pi}{\pi(1-\pi)}\right)^2$$

$$= \frac{1}{\pi^2(1-\pi)^2}\left[E(2y_1+y_2)^2 - 4n\pi E(2y_1+y_2) + 4n^2\pi^2\right]$$

$$= \frac{1}{\pi^2(1-\pi)^2}(2n\pi - 2n\pi^2)$$

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

$$\Rightarrow \frac{1}{I(\pi)} = \frac{\pi(1-\pi)}{2n}$$

The asymptotic distribution of $\hat{\pi}$ is $N(\pi, \frac{1}{I(\pi)})$

$$\Rightarrow \hat{\pi} \sim N(\pi, \frac{\pi(1-\pi)}{2n}).$$

(c) $1-\alpha = 95\% \Rightarrow \alpha = 5\% = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025.$

Since $\hat{\pi} \sim N(\pi, \frac{\pi(1-\pi)}{2n})$,

$\frac{\hat{\pi}-\pi}{\sigma} \sim N(0,1)$, where $\sigma = \sqrt{\frac{\pi(1-\pi)}{2n}}$

$$P(-Z_{0.025} < \frac{\hat{\pi}-\pi}{\sigma} < Z_{0.025}) = 95\%$$

$\Rightarrow \hat{\pi} - \sigma Z_{0.025} < \pi < \hat{\pi} + \sigma Z_{0.025}$, where $Z_{0.025} = 1.96$

$$\pi \in \left[\frac{2y_1+y_2}{2n} - 1.96\frac{\pi(1-\pi)}{2n}, \frac{2y_1+y_2}{2n} + 1.96\frac{\pi(1-\pi)}{2n}\right].$$

The 95% confidence interval for π is:

$$\left[\frac{2y_1+y_2}{2n} - 1.96\frac{\pi(1-\pi)}{2n}, \frac{2y_1+y_2}{2n} + 1.96\frac{\pi(1-\pi)}{2n}\right]$$

(**Question2**): (15 points) Suppose you flip a coin 314 times, and heads appears 159 times.

(a) Construct an approximate 90% condence interval for the probability that the coin comes up heads.

(b) Approximately how many samples would you need to obtain an approximate 90% confidence interval with width 0.02, while keeping exactly 50.64% of flips appearing heads?

(c) You give the coin to a friend, who also flips the coin 314 times, and obtains an approximate confidence interval [0.390, 0.500] instead. What confidence level did (s)he use?

(a) $\hat{p} = \frac{x}{n} = \frac{159}{314} = 0.506$

$1-\alpha = 90\% \Rightarrow \alpha = 10\% = 0.1$

$Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$

$$\frac{\sigma^2}{n} = \frac{\hat{p}(1-\hat{p})}{n} = \frac{\frac{159}{314}(1-\frac{159}{314})}{314}$$

$$\frac{\sigma}{\sqrt{n}} = 0.028$$

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{p}-p}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

The lower confidence limit $\hat{L} = \hat{p} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 0.460$

The upper confidence limit $\hat{U} = \hat{p} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 0.552$

The confidence interval is $[0.460, 0.552]$

(b) $\hat{p}' = 50.64\% = 0.5064$

width $= \hat{U}' - \hat{L}' = 2 \cdot z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 0.02 \Rightarrow z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 0.01$

$z_{\frac{\alpha}{2}} = 1.645$

$\sigma'^2 = \hat{p}'(1-\hat{p}') = 0.250$

$\Rightarrow \sigma' = 0.5$

So $1.645 \times \frac{0.5}{\sqrt{n'}} = 0.01 \Rightarrow \sqrt{n'} = 82.25$

we get: $n' = 6765$.

We need 6765 samples.

(c) The width is $0.500 - 0.390 = 0.110$

Suppose the confindence level is $1-\alpha'$

The original width of confidence interval is $0.552 - 0.460 = 0.092$

$$\frac{z_{\frac{\alpha'}{2}}}{z_{\frac{\alpha}{2}}} = \frac{0.110}{0.092} \Rightarrow z_{\frac{\alpha'}{2}} = 1.97$$

So $\frac{\alpha'}{2} = 0.02442 \Rightarrow \alpha' = 0.04884$

$1-\alpha' = 0.95116 = 95\%$

(Question3): (15 points) Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma_X^2)$, $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma_Y^2)$ and they are independent. Suppose that both σ_X^2 and σ_Y^2 are known.

(a) Construct a two-sided $100(1-\alpha)\%$ confidence interval for the difference $\mu_X - \mu_Y$.

(b) We want to obtain a 90% condence interval for the difference between true average cable strengths made by Company X and by Company Y. Suppose cable strength is normally distributed for both types of cables with $\cancel{\mu_X}\ \sigma_X = 50$ and $\cancel{\mu_Y}\ \sigma_Y = 30$. If we can make $n + m = 6000$ observations, how many of these should be on Company X cable if we want to minimize the width of the interval?

(a) $$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1),$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Rightarrow \hat{L} = \bar{X} - \bar{Y} - \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \cdot z_{\frac{\alpha}{2}}$$

$$\hat{U} = \bar{X} - \bar{Y} + \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \cdot z_{\frac{\alpha}{2}}$$

Two-sided $100(1-\alpha)\%$ confidence interval is:

$$\left[\bar{X} - \bar{Y} - \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \cdot z_{\frac{\alpha}{2}},\ \bar{X} - \bar{Y} + \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \cdot z_{\frac{\alpha}{2}} \right]$$

(b) $1-\alpha = 90\% \Rightarrow \alpha = 10\% = 0.1$

The width of the confidence interval is $2z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$

$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

To minimize the width of the CI is to minimize $\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$ i.e., minimize $\left(\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$.

Given that $n+m = 6000$

Let $f(n) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m} = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{6000-n}$

$$f'(n) = -\frac{\sigma_X^2}{n^2} + \frac{\sigma_Y^2}{(6000-n)^2}$$

$$= \frac{n^2\sigma_Y^2 - (6000-n)^2\sigma_X^2}{n^2(6000-n)^2}$$

Let $f'(n) = 0$, $(\sigma_X^2 - \sigma_Y^2)n^2 - 12000\sigma_X^2 n + 36000000\sigma_X^2 = 0$.

Solve the equation, we get

$$n = \frac{12000(\sigma_X^2 - \sigma_X\sigma_Y)}{2(\sigma_X^2 - \sigma_Y^2)}$$

$$= \frac{6000\sigma_X}{\sigma_X + \sigma_Y} = \frac{6000 \times 50}{50+30} = 3750$$

(Question4): (15 points) Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. In this question, we will construct confidence interval for σ^2. Let $C \sim \chi^2(r)$, and as usual for any $\alpha \in [0,1]$ we denote $\chi^2_\alpha(r)$ to be

$$P(C > \chi^2_\alpha(r)) = \alpha.$$

(a) Suppose that μ is known. By considering the distribution of

$$\sum_i \left(\frac{X_i - \mu}{\sigma}\right)^2,$$

prove that

$$\left[\frac{\sum_i (X_i-\mu)^2}{\chi^2_{\alpha/2}(n)}, \frac{\sum_i (X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

Proof: Since μ is known,

the sample variance $S^2 = \frac{1}{n}\sum_i (X_i - \mu)^2$

$$\sum\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum(X_i-\mu)^2 = \frac{nS^2}{\sigma^2} \sim \chi^2(n)$$

Since $P(C>\chi^2_\alpha(r))=\alpha$,

$P(C>\chi^2_{\frac{\alpha}{2}}(r)) = \frac{\alpha}{2}$, $P(C>\chi^2_{1-\frac{\alpha}{2}}(r)) = 1-\frac{\alpha}{2}$

$P(\chi^2_{1-\frac{\alpha}{2}}(n) \leqslant \frac{nS^2}{\sigma^2} \leqslant \chi^2_{\frac{\alpha}{2}}(n))$

$= P(\frac{nS^2}{\sigma^2} > \chi^2_{1-\frac{\alpha}{2}}(n)) - P(\frac{nS^2}{\sigma^2} > \chi^2_{\frac{\alpha}{2}}(n))$

$= 1-\alpha.$

i.e., $P\left(\chi^2_{1-\frac{\alpha}{2}}(n) \leqslant \frac{\sum_i(X_i-\mu)^2}{\sigma^2} \leqslant \chi^2_{\frac{\alpha}{2}}(n)\right) = 1-\alpha$

$\Rightarrow P\left(\frac{\sum_i(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)} \leqslant \sigma^2 \leqslant \frac{\sum_i(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}\right) = 1-\alpha$

So $\left[\frac{\sum_i(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}, \frac{\sum_i(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}\right]$ is a $100(1-\alpha)\%$ confidence interval for σ^2.

(b) Suppose that μ is unknown. By considering the distribution of

$$\sum_i\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2},$$

prove that

$$\left[\frac{\sum_i(X_i-\bar{X})^2}{\chi^2_{\alpha/2}(n-1)}, \frac{\sum_i(X_i-\bar{X})^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

Proof: μ is unknown.

sample variance $S^2 = \frac{1}{n-1}\sum_i\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 = \frac{1}{n-1}\cdot\frac{1}{\sigma^2}\sum(X_i-\bar{X})^2$

$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ i.e., $\frac{\sum_i(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$

Since $P(C>\chi^2_\alpha(r))=\alpha$,

$P(C>\chi^2_{\frac{\alpha}{2}}(r)) = \frac{\alpha}{2}$, $P(C>\chi^2_{1-\frac{\alpha}{2}}(r)) = 1-\frac{\alpha}{2}$

$P(\chi^2_{1-\frac{\alpha}{2}}(n-1) \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant \chi^2_{\frac{\alpha}{2}}(n-1))$

$= P(\frac{(n-1)S^2}{\sigma^2} > \chi^2_{1-\frac{\alpha}{2}}(n-1)) - P(\frac{(n-1)S^2}{\sigma^2} > \chi^2_{\frac{\alpha}{2}}(n-1))$

$= 1-\alpha.$

i.e., $P\left(\chi^2_{1-\frac{\alpha}{2}}(r) \leqslant \frac{\sum_i(X_i-\mu)^2}{\sigma^2} \leqslant \chi^2_{\frac{\alpha}{2}}(r)\right) = 1-\alpha$

$\Rightarrow P\left(\frac{\sum_i(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \leqslant \sigma^2 \leqslant \frac{\sum_i(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right) = 1-\alpha$

So $\left[\frac{\sum_i (X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{\sum_i (X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right]$ is a $100(1-\alpha)\%$

confidence interval for σ^2.

(c) In the same setting as part (b), that is, suppose that μ is unknown. Construct a $100(1-\alpha)\%$ confidence interval for σ.

Know that $\left[\frac{\sum_i (X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{\sum_i (X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right]$ is a $100(1-\alpha)\%$

confidence interval for σ^2.

$\sigma = \sqrt{\sigma^2} > 0$

$\Rightarrow$ The $100(1-\alpha)\%$ confidence interval for σ

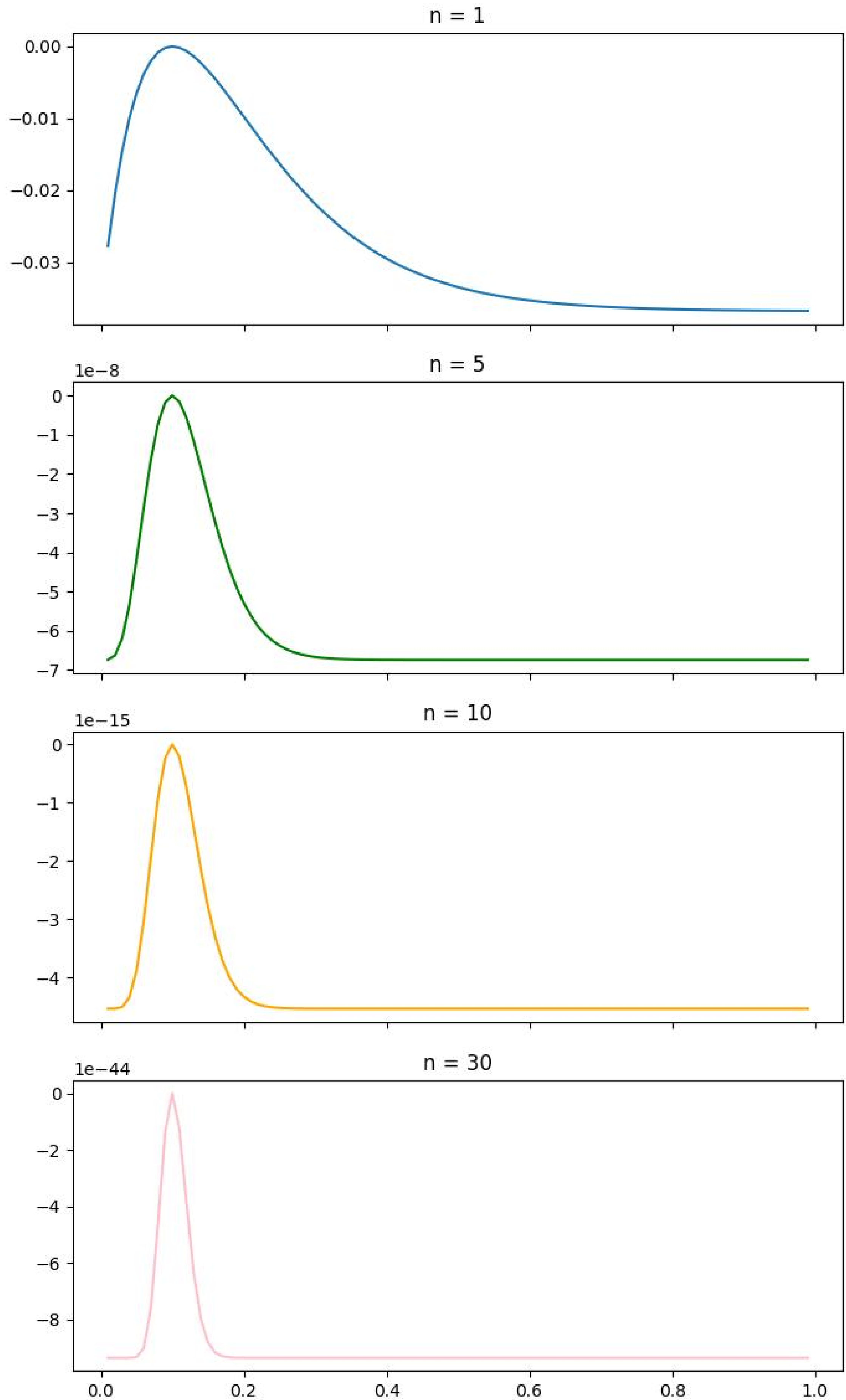
is : $\left[\sqrt{\frac{\sum_i (X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}}, \sqrt{\frac{\sum_i (X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}}\right]$

In [8]:

```
# STA2004 Programming assignment2
# Name: Ou Ziyi  Student ID:121090429

# 2.1
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from functools import reduce

## Since μ= E[Y] = 1/λ and y_bar = 10, we can know that λ_hat = 0.1
## L(λ) = λ^{2} * e^{-λny}


fig, ((ax1, ax2, ax3, ax4)) = plt.subplots(4, 1, figsize=(8, 14), sharex=True)
n_value = np.arange(0.01,1.00,0.01)
list1= []
list2 =[]
for i in n_value:
    b = pow(i, 1)*np.exp(-i*10*1) - pow(1/10, 1)*np.exp(-1/10*10*1)
    list1.append(i)
    list2.append(b)
ax1.set_title('n = {}'.format(1))
ax1.plot(list1, list2)

list1= []
list2 =[]
for i in n_value:
    b = pow(i, 5)*np.exp(-i*10*5) - pow(1/10, 5)*np.exp(-1/10*10*5)
    list1.append(i)
    list2.append(b)
ax2.set_title('n = {}'.format(5))
ax2.plot(list1, list2, color = 'green')

list1= []
list2 =[]
for i in n_value:
    b = pow(i, 10)*np.exp(-i*10*10) - pow(1/10, 10)*np.exp(-1/10*10*10)
    list1.append(i)
    list2.append(b)
ax3.set_title('n = {}'.format(10))
ax3.plot(list1, list2, color = 'orange')

list1= []
list2 =[]
for i in n_value:
    b = pow(i, 30)*np.exp(-i*10*30) - pow(1/10, 30)*np.exp(-1/10*10*30)
    list1.append(i)
    list2.append(b)
ax4.set_title('n = {}'.format(30))
ax4.plot(list1, list2, color = 'pink')
```

Out[8]:

```
[<matplotlib.lines.Line2D at 0x21fc7fb8880>]
```



n = 1
0.00
−0.01
−0.02
−0.03
n = 5
1e−8
0
−1
−2
−3
−4
−5
−6
−7
n = 10
1e−15
0
−1
−2
−3
−4
n = 30
1e−44
0
−2
−4
−6
−8
0.0
0.2
0.4
0.6
0.8
1.0

In [7]:

```
# 2.2
import numpy as np
n_value = [1,5,10,30]
for n in n_value:
    confidence_interval = np.array([1/10 - (1/10*1.96)/np.sqrt(n),1/10 + (1/10*1.96)/np.sqrt(n)
    print(confidence_interval)
```

```
[-0.096   0.296]
[0.01234614 0.18765386]
[0.03801936 0.16198064]
[0.06421546 0.13578454]
```

In [17]:

```
# 2.3
lambda_zero = 1
Z_half_alpha = 1.96
size = 10000
plt.title('Confidence Intervals')
plt.xlabel('lambda')
plt.ylabel('Samples')
plt.vlines(1, 0, 100, colors='green')
for num in range(100):
    data_set = np.random.exponential(lambda_zero, size)
    mean_of_data = np.mean(data_set)
    lambda_hat = 1/mean_of_data
    half_length = lambda_hat * Z_half_alpha/np.sqrt(size)
    if lambda_hat - half_length > 1 or lambda_hat + half_length < 1:
        plt.hlines(num, lambda_hat - half_length, lambda_hat + half_length, colors='red')
    else:
        plt.hlines(num, lambda_hat - half_length, lambda_hat + half_length, colors='black')
plt.show()
```