Question 1 (20 Points, 18 minutes)

You are interested in estimating the average compensation of all CEO’s of “e commerce” companies. Generally CEO compensation is related to “sales”. Let $Y$ be CEO compensation and $X$ be sales. Suppose you know that average sales among all “e commerce” companies is $10 million.

Part a

Use the following summary statistics to calculate the simple random sample and ratio estimate for average CEO compensation. 

\[
\bar{X} = 8,000,000 \quad S_X = 500,000 \\
\bar{Y} = 1,200,000 \quad S_Y = 200,000 \\
n = 10 \quad r_{XY} = 0.7
\]

Note $r_{XY}$ is the sample correlation coefficient.

$\bar{Y}_{SRS} = \bar{Y} = 1,200,000,$

$\bar{Y}_R = \mu_x \bar{Y} / \bar{X} = 10,000,000 \times 1,200,000 / 8,000,000 = 1,500,000$

Part b

Estimate the standard error of each estimator. Ignore finite population correction. Which estimate appears to be more accurate?

\[
SE(\bar{Y}) = S_Y / \sqrt{n} = 200,000 / \sqrt{10} = 63,245 \\
SE(\bar{Y}_R) = \sqrt{\frac{1}{10} (0.15^2 \times 500,000^2 + 200,000^2 - 2 \times 0.15 \times 0.7 \times 500,000 \times 200,000)} = 49,623
\]

Question 2 (15 Points, 13 minutes)

Suppose $X_1, X_2, \ldots, X_n \sim N(\mu_x, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_y, \sigma^2)$ with all random variables independent. The standard method to test equality of means ($\mu_x = \mu_y$) is the two sample t-test using the statistic

\[
T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \\
S_p^2 = \frac{(n - 1)S_x^2 + (m - 1)S_y^2}{n + m - 2}
\]

and $S_x^2$ and $S_y^2$ are the usual sample variance estimates. Show that $T \sim t_{n+m-2}$ provided $\mu_x = \mu_y$.

First note that

$\bar{X} \sim N(\mu_X, \sigma^2 / n), \quad \bar{Y} \sim N(\mu_Y, \sigma^2 / m)$

so

$\bar{X} - \bar{Y} \sim N(0, \sigma^2 / n + \sigma^2 / m)$ \quad (if $\mu_x = \mu_y$)

and

\[
\frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1)
\]
Next note that
\[ S_X^2 (n - 1)/\sigma^2 \sim \chi_{n-1}^2, \quad S_Y^2 (m - 1)/\sigma^2 \sim \chi_{m-1}^2 \]
so
\[ \frac{(n - 1)S_X^2 + (m - 1)S_Y^2}{\sigma^2} \sim \chi_{n+m-2}^2. \]
Therefore
\[ T = \frac{X - Y}{\sqrt{S_Y^2 (\frac{1}{n} + \frac{1}{m})}} = \frac{\frac{X - Y}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}}{\sqrt{\frac{[(n-1)S_X^2 + (m-1)S_Y^2]/\sigma^2}{n+m-2}}} = \frac{N(0, 1)}{\sqrt{\lambda_{n+m-2}/(n + m - 2)}} \sim t_{n+m-2} \]

**Question 3 (25 Points, 23 minutes)**

**Part a**

Use moment generating functions to show that the \textit{gamma}(\alpha = n, \lambda = 1/2) and \chi_{2n}^2 distributions are identical. Note you may use the gamma and chi-square mgf’s from the notes.

Let \( X \sim \text{gamma}(\alpha = n, \lambda = 1/2) \) and \( Y \sim \chi_{2n}^2 \). Then
\[ M_X(t) = \left( \frac{\lambda}{\lambda - t} \right)^\alpha = (1 - t/\lambda)^{-\alpha} = (1 - 2t)^{-n} \]
but
\[ M_Y(t) = (1 - 2t)^{-2n/2} = (1 - 2t)^{-n} \]
Therefore \( X \) and \( Y \) have the same distribution.

**Part b**

Suppose \( X_1, X_2, \ldots, X_n \sim \exp(\lambda) \) and \( Y_1, Y_2, \ldots, Y_m \sim \exp(\lambda) \) where all random variables are independent and \( n \) and \( m \) are not necessarily equal. What is the distribution of
\[ \frac{X}{Y}. \]

Hint: Use mgfs to calculate the distribution of \( \sum_{i=1}^n X_i \) where \( X_i \sim \exp(\lambda) \).

First note that
\[ M_{\sum X_i} = \left( \frac{\lambda}{\lambda - t} \right)^n \]
so $\sum X_i \sim \text{gamma}(n, \lambda)$ if $X_i \sim \text{exp}(\lambda)$. Therefore

$$\frac{X}{Y} = \frac{\sum X_i / n}{\sum Y_i / m} = \frac{2\lambda \sum X_i / n}{2\lambda \sum Y_i / m} = \frac{\text{gamma}(n, 1/2) / n}{\text{gamma}(m, 1/2) / m} \quad \text{from the hint}$$

$$= \frac{\chi^2_{2n}/2n}{\chi^2_{2m}/2m} \quad \text{from part a}$$

$$= F_{2n,2m}$$

**Question 4 (40 Points, 36 minutes)**

$(X, Y)$ are uniformly distributed within a triangle (i.e. not just on the perimeter) with vertices at $(0, 0)$, $(1, 0)$ and $(1, C)$ where $C$ is a constant.

**Part a**

Give the joint distribution and the marginal densities of $X$ and $Y$ (in terms of $C$). Be careful to state the ranges.

The area of the triangle is $C/2$ so

$$f(x, y) = 2/C, \quad 0 \leq x \leq 1, 0 \leq y \leq Cx$$

and

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{0}^{Cx} 2/C = 2x, \quad 0 \leq x \leq 1$$

$$f(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{y/C}^{1} 2/C = 2(1 - y/C)/C, \quad 0 \leq y \leq C$$

**Part b**

Calculate $E(XY)$ and $\text{Var}(XY)$ in terms of $C$.

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dy dx$$

$$= \int_{0}^{1} \int_{0}^{Cx} xy^2/C dy dx$$

$$= \int_{0}^{1} \frac{x}{C} y^2 dy dx$$

$$= \int_{0}^{1} C x^3 dx$$

$$= C/4$$
\[ E((XY)^2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 y^2 f(x,y) \, dy \, dx \]
\[ = \int_0^1 \int_0^x y^2 2/C \, dy \, dx \]
\[ = \int_0^1 2x^2 \frac{3}{3C} \, dx \]
\[ = \int_0^1 2C^2 x^5/3 \, dx \]
\[ = C^2/9 \]

Therefore \( \text{Var}(XY) = \frac{C^2}{9} - \frac{C^2}{16} = 7C^2/144 \)

**Part c**

Suppose that we sample \( n \) pairs \((X_i, Y_i)\) from the triangle. Each pair is sampled independently from the other pairs. Show that

\[ \frac{\sum_{i=1}^{n} X_i Y_i - a}{b} \Rightarrow N(0,1) \]

for some \( a \) and \( b \). Give \( a \) and \( b \)!

Since the pairs are all independent \( X_i Y_i \) is independent from \( X_j Y_j \) provided \( i \neq j \). Therefore by the CLT

\[ \frac{\sum_{i=1}^{n} X_i Y_i - nE(XY)}{\sqrt{n \text{Var}(XY)}} \Rightarrow N(0,1) \]

Therefore

\[ \frac{\sum_{i=1}^{n} X_i Y_i - a}{b} \Rightarrow N(0,1) \]

provided \( a = nC/4 \) and \( b = \sqrt{n7C^2/144} \).

**Part d**

Suppose that \( C \) is unknown. Use the results from parts b and c to give an unbiased estimate for \( C \) and construct an approximate 95% confidence interval for \( C \).

From b we know that \( \hat{C} = \frac{4XY}{n} \) is an unbiased estimate of \( C \) because \( E(4XY) = 4E(XY) = 4C/4 = C \).

From c we know

\[ \frac{\sum_{i=1}^{n} X_i Y_i - nC/4}{\sqrt{n7C^2/144}} = \frac{12 \sum_{i=1}^{n} X_i Y_i \sqrt{i/n}}{C \sqrt{i/n}} - 3 \sqrt{n/i} \Rightarrow N(0,1) \]

Therefore

\[ P \left( -1.96 \leq \frac{12 \sum_{i=1}^{n} X_i Y_i \sqrt{i/n}}{C \sqrt{i/n}} - 3 \sqrt{n/i} \leq 1.96 \right) \approx 0.95 \]

or

\[ P \left( -1.96 + 3 \sqrt{n/i} \leq \frac{12 \sum_{i=1}^{n} X_i Y_i \sqrt{i/n}}{C \sqrt{i/n}} \leq 1.96 + 3 \sqrt{n/i} \right) \approx 0.95 \]
or

\[
P \left( \frac{-1.96 + 3 \sqrt{n/\hat{t}}}{12 \sum_{i=1}^{n} X_i Y_i} \leq \frac{1}{C \sqrt{\hat{t}/n}} \leq \frac{1.96 + 3 \sqrt{n/\hat{t}}}{12 \sum_{i=1}^{n} X_i Y_i} \right) \approx 0.95
\]

or

\[
P \left( \frac{12 \sum_{i=1}^{n} X_i Y_i}{3n + 1.96 \sqrt{\hat{t}/n}} \leq C \leq \frac{12 \sum_{i=1}^{n} X_i Y_i}{3n - 1.96 \sqrt{\hat{t}/n}} \right) \approx 0.95
\]

or

\[
P \left( \frac{12 \bar{XY}}{3 + 1.96 \sqrt{\hat{t}/n}} \leq C \leq \frac{12 \bar{XY}}{3 - 1.96 \sqrt{\hat{t}/n}} \right) \approx 0.95
\]

So

\[
\begin{bmatrix}
\frac{12 \bar{XY}}{3 + 1.96 \sqrt{\hat{t}/n}} & \frac{12 \bar{XY}}{3 - 1.96 \sqrt{\hat{t}/n}} \\
\end{bmatrix}
\]

is an approximate 95% confidence interval for \( C \).