Q1 [15 points]:
Use the method of maximum likelihood to estimate θ in the pdf

$$f_Y(y \mid \theta) = \frac{\theta}{2\sqrt{y}} e^{-\theta\sqrt{y}}, \quad y \geq 0$$

Evaluate θ_e for the following random sample of size 4 : $Y_1 = 6.2, Y_2 = 7.0, Y_3 = 2.5$, and $Y_4 = 4.2$.

Solution:

$$\ell(\theta) = \prod_{i=1}^{4} \frac{\theta}{2\sqrt{y_i}} e^{-\theta\sqrt{y_i}} = \frac{\theta^4}{16\prod_{i=1}^{4}\sqrt{y_i}} e^{-\theta\sum_{i=1}^{4}\sqrt{y_i}}$$

$$\ln \ell(\theta) = 4\ln\theta - \ln\left(16\prod_{i=1}^{4}\sqrt{y_i}\right) - \theta\sum_{i=1}^{4}\sqrt{y_i}$$

$$\frac{d\ln\ell(\theta)}{d\theta} = \frac{4}{\theta} - \sum_{i=1}^{4}\sqrt{y_i}.$$

$$\frac{d\ln\ell(\theta)}{d\theta} = 0 \quad \text{implies} \quad \widehat{\theta} = \frac{4}{\sum_{i=1}^{4}\sqrt{y_i}} = \frac{4}{8.766} = 0.456$$

Q2 [15 points]:
Suppose the random samples are obtained from a two-parameter uniform pdf $Y \sim \mathcal{U}[\theta_1, \theta_2]$. Based on the random sample $Y_1 = 6.3, Y_2 = 1.8$, $Y_3 = 14.2$, and $Y_4 = 7.6$, find the maximum likelihood estimates for θ_1 and θ_2.

$$f_Y\left(y \mid \theta_1, \theta_2\right) = \frac{1}{\theta_2 - \theta_1}, \quad \theta_1 \leq y \leq \theta_2.$$

Solution:

$$\ell(\theta) = \left(\frac{1}{\theta_2 - \theta_1}\right)^n, \text{ if } \theta_1 \leq y_1, y_2, \ldots, y_n \leq \theta_2, \text{ and 0 otherwise}$$

Or, we may write

$$\ell(\theta) = \left(\frac{1}{\theta_2 - \theta_1}\right)^n, \text{ if } \min\{y_i\}_{i=1}^n > \theta_1, \text{ and } \max\{y_i\}_{i=1}^n < \theta_2, \text{ and it is zero otherwise.}$$

Using step function $u(y - y_0)$

$$u(y - y_0) = \begin{cases} 0 & y < y_0 \\ 1 & y > y_0 \end{cases}$$

we may write

$$\ell(\theta) = \left(\frac{1}{\theta_2 - \theta_1}\right)^n u\big(\min\{y_i\} - \theta_1\big)\, u\big(\theta_2 - \max\{y_i\}\big)$$

Clearly, the sufficient statistics for θ_1 and θ_2 are $\min\{y_i\}$ and $\max\{y_i\}$, respectively. So, we obtain

$$\widehat{\theta_1} = \min\{y_i\},$$
$$\widehat{\theta_2} = \max\{y_i\}.$$

From the given data, we obtain $\theta_1 = 1.8$, and $\theta_2 = 14.2$. Note that we have discussed in detail in a problem session that $\widehat{\theta_2} = \max\{y_i\}$ is a biased estimator using order statistics, and how the bias may be removed. Similarly, $\widehat{\theta_1} = \min\{y_i\}$ is also a biased estimator, and the bias may be removed. Please prepare yourself and show your work to me that how may you prove that the estimator $\widehat{\theta_1} = \min\{y_i\}$ is biased in nature.

Q3 [30 points]:
(a) Use the method of moments to estimate θ in the pdf

$$f_Y(y \mid \theta) = \left(\theta^2 + \theta\right) y^{\theta-1}(1-y), \quad 0 \leq y \leq 1$$

Assume that a random sample of size n has been collected.

(b) Using Taylor's series based method, obtain the bias and variance of the estimator obtained in part **(a)**.

Solution: First of all, we show that

$$E(Y) = \int_0^1 y\left(\theta^2+\theta\right) y^{\theta-1}(1-y)dy = \left(\theta^2+\theta\right)\int_0^1 y^{\theta}(1-y)dy = \frac{\theta}{\theta+2}$$

This is sufficient to obtain an estimator of θ from the first-order sample moment of Y. This gives

$$\begin{aligned}
\theta &= \frac{2E[Y]}{1-E[Y]} \\
\widehat{\theta} &= \frac{2\widehat{\mu_Y}}{1-\widehat{\mu_Y}} \\
\text{where } \ \widehat{\mu_Y} &= \frac{1}{n}\sum_{i=1}^{n} Y_i \\
\text{and } \ E[\widehat{\mu_Y}] &= \frac{\theta}{\theta+2}
\end{aligned}$$

For the sake of analysis, we would need to know $E[Y^2]$ for the computation of variance of Y.

$$E(Y^2) = \int_0^1 y^2\left(\theta^2+\theta\right) y^{\theta-1}(1-y)dy = \left(\theta^2+\theta\right)\int_0^1 y^{\theta+1}(1-y)dy = \frac{\theta^2+\theta}{\theta^2+5\theta+6}$$

This gives

$$\text{Var}(Y) = E[Y^2] - (E[Y])^2 = \frac{2\theta}{(\theta+2)^2(\theta+3)}$$

For the ease of analysis, we also define

$$\begin{aligned}
\widehat{\mu_Y} &=: T \\
\widehat{\theta} &= \frac{2T}{1-T} =: h(T)
\end{aligned}$$

In the sequel, we would need $h'(T)$ and $h''(T)$, so we compute it here

$$h'(T) = \frac{d}{dT} h(T) = \frac{2}{(1-T)^2}$$
$$h''(T) = \frac{d}{dT} h'(T) = \frac{4}{(1-T)^3}$$

We may easily show that the estimator $\widehat{\theta}$ is biased as follows:

$$E[\widehat{\theta}] = E\left[\frac{2\widehat{\mu_Y}}{1-\widehat{\mu_Y}}\right] \neq \frac{2E[\widehat{\mu_Y}]}{1-E[\widehat{\mu_Y}]} = \frac{2\theta/(2+\theta)}{1-\theta/(2+\theta)} = \theta$$

Next we compute the bias and variance of $\widehat{\theta}$. From the lecture 08's slides, we have

Using Taylor's series based expansion, we may obtain

$$\widehat{\theta} = h(T) = h\left(\mu_T\right) + \left(T-\mu_T\right) h'\left(\mu_T\right) + \frac{1}{2}\left(T-\mu_T\right)^2 h''\left(\mu_T\right)$$

where the statistic $T = T(Y)$ is the function of given random variables Y, $\mu_T = E[T(Y)]$; next, we take mean of both sides to obtain

$$E\left[\widehat{\theta}\right] \approx h\left(\mu_T\right) + \frac{1}{2} \operatorname{var}(T) h''\left(\mu_T\right)$$

and

$$\operatorname{var}(\widehat{\theta}) = \left[h'\left(\mu_T\right)\right]^2 \operatorname{var}(T)$$

The estimation error is defined as $\widetilde{\theta} = \widehat{\theta} - \theta$, where θ is the true value, constant in nature, to be estimated, therefore the variance of $\widetilde{\theta}$ is same as that of $\widehat{\theta}$, this gives $\operatorname{var}(\widetilde{\theta}) = \left[h'\left(\mu_T\right)\right]^2 \operatorname{var}(T)$.

where $\mu_T = E[T] = E[\widehat{\mu_Y}] = \theta/(2+\theta)$, this gives

$$h(\mu_T) = \frac{2\mu_T}{1-\mu_T} = \theta$$
$$h'(\mu_T) = \frac{2}{\left(1-\mu_T\right)^2} = \frac{1}{2}(2+\theta)^2$$
$$h''(\mu_T) = \frac{4}{\left(1-\mu_T\right)^3} = \frac{1}{2}(2+\theta)^3$$

We compute $\mathrm{var}(T) = \mathrm{var}(\widehat{\mu_Y}) = \frac{1}{n}\mathrm{var}(Y)$, where $\mathrm{var}(Y)$ has been computed above. Finally, combinung the earlier results, we obtain

$$E[\widehat{\theta}] \approx \theta + \frac{1}{2n}\left(\frac{\theta+2}{\theta+3}\right)\theta$$
$$\mathrm{var}(\widehat{\theta}) \approx \frac{1}{2n}\frac{(2+\theta)^2}{3+\theta}\theta, \qquad \mathrm{var}(\widehat{\theta}) \propto \frac{\theta^2}{n}$$

Estimator is asymptotically consistent because

$$\lim_{n\to\infty} E[\widehat{\theta}] \to \theta$$
$$\lim_{n\to\infty} \mathrm{var}(\widehat{\theta}) \approx 0.$$

Q4 [40 points]:
(a) Let $X_1, X_2, \ldots, X_n$ be a random sample from $f_X(x \mid \theta) = \frac{1}{\theta} e^{-x/\theta}, x > 0$. Find the MLE $\widehat{\theta}$.

By definition, the likelihood function $\ell(\theta)$ is

$$\ell(\theta) = \log\left[\prod_{i=1}^{n} f\big(X_i \mid \theta\big)\right] = \sum_{i=1}^{n} \log\left[f\big(X_i \mid \theta\big)\right]$$

(b) Obtain the bias of the MLE $\widehat{\theta}$.

(c) Obtain the variance of the MLE by computing it explicity as follows:

$$\operatorname{var}\left(\widehat{\theta}\right) = \mathrm{E}\left[\widehat{\theta}^2\right] - \left(\mathrm{E}\left[\widehat{\theta}\right]\right)^2$$

(d) Obtain the asymptotic variance of MLE $\widehat{\theta}$ as follows:

$$\text{asymptotic var}\left(\widehat{\theta}\right) \approx \frac{1}{nI(\theta)}$$
$$\text{where } I(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2} \ell(\theta)\right] = -\mathrm{E}\left[\ell''(\theta)\right]$$

(e) Are the variances obtained in **(c)** and **(d)** equal? Is the ML estimator a best estimator for θ?

Solution: To find the MLE of θ, we first define the likelihood function:

$$\operatorname{lik}(\theta) = f\left(x_1, \ldots, x_n \mid \theta\right) = f\left(x_1 \mid \theta\right) \cdots f\left(x_n \mid \theta\right)$$

Substituting the definition of the density function of X yields

$$\operatorname{lik}(\theta) = \left(\frac{1}{\theta} \cdot e^{-\frac{x_1}{\theta}}\right) \cdot \ldots \cdot \left(\frac{1}{\theta} \cdot e^{-\frac{x_n}{\theta}}\right) = \frac{1}{\theta^n} \cdot e^{-\frac{x_1 + \cdots + x_n}{\theta}}$$

It's easier to work with the natural logarithm of the given expression, so we define

$$l(\theta) = \ln(\operatorname{lik}(\theta)) = -n \cdot \ln(\theta) - \frac{1}{\theta} \cdot \sum_{i=1}^{n} x_i$$

and we need to find its global maximum on the interval $\langle 0, +\infty \rangle$ (where θ can take on values).

The derivative of l is

$$l'(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \cdot \sum_{i=1}^{n} x_i.$$

Stationary points are the null points of the above derivative, so

$$l'(\theta) = 0 \Longleftrightarrow -\frac{n}{\theta} + \frac{1}{\theta^2} \cdot \sum_{i=1}^{n} x_i = 0 \Longleftrightarrow \frac{n}{\theta} = \frac{1}{\theta^2} \cdot \sum_{i=1}^{n} x_i \Longleftrightarrow n \cdot \theta = \sum_{i=1}^{n} x_i$$
$$\Longleftrightarrow \theta = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i.$$

Therefore, the MLE of θ is

$$\widehat{\theta} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = \widehat{\mu_X}.$$

(b) There is no bias. Because $E[\widehat{\theta}] = E[X] = \int_0^\infty x \frac{1}{\theta} e^{-x/\theta} dx = \theta$ (Unbiased).

(c) The variance is computed below:

$$\begin{aligned}\operatorname{var}(\widehat{\theta}) = \operatorname{var}(\widehat{\mu_X}) &= \operatorname{var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{var}(X_i) \\ &= \frac{1}{n} \operatorname{var}(X) = \frac{1}{n}\left(E[X^2] - (E[X])^2\right) = \frac{1}{n}\left(2\theta^2 - \theta^2\right) = \frac{\theta^2}{n}\end{aligned}$$

(d) According to the Cramer-Rao Theorem, no unbiased estimator of θ can have variance less than $\frac{1}{n \cdot I(\theta)}$ (this is the Cramer-Rao lower bound), where

$$I(\theta) = E\left(\left[\frac{\partial}{\partial \theta} \ln f(X \mid \theta)\right]^2\right)$$

(This is called Fisher's information). Remember that $I(\theta)$ can also be calculated as

$$I(\theta) = -E\left(\frac{\partial^2}{\partial \theta^2} \ln f(X \mid \theta)\right).$$

$$\ln f(x \mid \theta) = \ln\left(\frac{1}{\theta} \cdot e^{-\frac{\theta}{\theta}}\right) = -\ln(\theta) - \frac{x}{\theta}.$$

Furthermore,

$$\frac{\partial}{\partial \theta} \ln f(X \mid \theta) = -\frac{1}{\theta} + \frac{x}{\theta^2}$$

from which it follows that

$$\frac{\partial^2}{\partial \theta^2} \ln f(X \mid \theta) = \frac{1}{\theta^2} - \frac{2x}{\theta^3}$$

Since X is an exponential random variable with the parameter $\frac{1}{\theta}$ (this is one of the X_i 's), then its expected value is θ, so

$$E\left(\frac{\partial^2}{\partial\theta^2}\ln f(X \mid \theta)\right) = E\left(\frac{1}{\theta^2} - \frac{2X}{\theta^3}\right) = \frac{1}{\theta^2} - \frac{2 \cdot E(X)}{\theta^3} = \frac{1}{\theta^2} - \frac{2}{\theta^2} = -\frac{1}{\theta^2}$$

from which we can conclude that the value of $I(\theta)$ is

$$I(\theta) = \frac{1}{\theta^2}$$

Finally, the Cramer-Rao lower bound is

$$\operatorname{var}(\widetilde{\theta}) = \frac{1}{n \cdot I(\theta)} = \frac{\theta^2}{n}$$

Notice that this is exactly the variance of $\overline{X} = \widehat{\mu_X}$ (which in this case is our MLE for θ), so we have that the variance of $\widetilde{\theta}$ reaches the Cramer-Rao lower bound, which means that no unbiased estimator for θ can have lower variance than that of $\widetilde{\theta}$ (we say that $\widehat{\theta}$ is an efficient estimator, and also remember that, here, $\operatorname{var}(\widehat{\theta}) = \operatorname{var}(\widetilde{\theta})$).

Q1 [25 points]: According to the National Association of Colleges and Employers, the average hourly wage of an undergraduate college student working as a co-op is \$17.3 and the average hourly wage of a college student working as an intern is \$16.6. Assume that such wages are normally distributed in the population and that the population variances are equal. Suppose these figures were actually obtained from the data below.

(a) Use these data and $\alpha = 0.10$ to test to determine if there is a significant difference in the mean hourly wage of a college co-op student and the mean hourly wage of a college intern.

(b) Using these same data, construct a 90% confidence interval to estimate the difference in the population mean hourly wages of college co-ops and interns.

Co-ops	Interns
16.97	16.23
16.38	15.58
17.51	17.34
18.55	16.04
18.47	14.93
19.20	17.25
15.68	17.38
17.04	17.02
18.37	15.12
16.08	17.21
16.88	16.98
16.27	17.55

Q2 [20 points]: The vice president of marketing brought to the attention of sales managers that most of the company's manufacturer representatives contacted clients and maintained client relationships in a disorganized, haphazard way. The sales managers brought the reps in for a three-day seminar and training session on how to use an organizer to schedule visits and recall pertinent information about each client more effectively. Sales reps were taught how to schedule visits most efficiently to maximize their efforts. Sales managers were given data on the number of site visits by sales reps on a randomly selected day both before and after the seminar. Use the following data to test whether significantly more site visits were made after the seminar ($\alpha = .05$). Assume the differences in the number of site visits are normally distributed.

Rep	Before	After
1	2	4
2	4	5
3	1	3
4	3	3
5	4	3
6	2	5
7	2	6
8	3	4
9	1	5

Q3 [25 points]: Using the given sample information, test the following hypotheses:

(a) $\mathcal{H}_0 : p_1 - p_2 = 0 \quad \mathcal{H}_a : p_1 - p_2 \neq 0$. Let $\alpha = 0.05$.

	Sample 1	Sample 2
c	$n_1 = 350$	$n_2 = 410$
	$x_1 = 160$	$x_2 = 190$

Note that x is the number in the sample having the characteristic of interest.

(b) $\mathcal{H}_0 : p_1 - p_2 = 0 \quad \mathcal{H}_a : p_1 - p_2 > 0$. Let $\alpha = 0.1$.

	Sample 1	Sample 2
c	$n_1 = 700$	$n_2 = 600$
	$\widehat{p_1} = 0.4$	$\widehat{p_2} = 0.25$

Q4 [20 points]: How long are resale houses on the market? One survey by the Houston Association of Realtors reported that in Houston, resale houses are on the market an average of 112 days. Of course, the length of time varies by market. Suppose random samples of 13 houses in Houston and 11 houses in Chicago that are for resale are traced. The data shown here represent the number of days each house was on the market before being sold. Use the given data and a 1% level of significance to determine whether the population variances for the number of days until resale are different in Houston than in Chicago. Assume the numbers of days resale houses are on the market are normally distributed.

$$\text{data}_{\text{Houstan}} = [132 \;\; 138 \;\; 131 \;\; 127 \;\; 99 \;\; 126 \;\; 134 \;\; 126 \;\; 94 \;\; 161 \;\; 133 \;\; 119 \;\; 88]$$
$$\text{data}_{\text{Chicago}} = [118 \;\; 85 \;\; 113 \;\; 81 \;\; 94 \;\; 93 \;\; 56 \;\; 69 \;\; 67 \;\; 54 \;\; 137];$$

Q5 [30 points]: Sketch a scatter plot from the following data, and determine the equation of the regression line.

x	12	21	28	8	20
y	17	15	22	19	24

Test the slope of the regression line. Use $\alpha = 0.05$.

Note: Data is also available at LMS.

Q1. $\mathcal{H}_0: \mu_1 - \mu_2 = 0$

$\mathcal{H}_1: \mu_1 - \mu_2 \neq 0$

Co-op	Intern
$n_1 = 12$	$n_2 = 12$
$\bar{X}_1 = 17.2833$	$\bar{X}_2 = 16.5525$
$S_1 = 1.1322$	$S_2 = 0.9350$

(a) For two-tail test $\alpha/2 = 0.05$
$df = 12 + 12 - 2 = 22$

Critical $t_{0.05, 22} = \pm 1.717$

$$\text{Observed } t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2(n_1-1) + S_2^2(n_2-1)}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = 1.5812$$

Since $t = 1.5812 < 1.717 = t_{0.05, 22}$,
the decision is to fail to reject the null hyp.

Q1 (b) $t_{0.05,22} = \pm 1.717$ ②

$$\bar{X}_1 - \bar{X}_2 \pm t \sqrt{\frac{S_1^2(n_1-1) + S_2^2(n_2-1)}{n_1+n_2-2}} \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= 0.7308 \pm 1.717 * 0.4622$$

$$= 0.7308 \pm 0.7936$$

$$-0.0628 \leq \mu_1 - \mu_2 \leq 1.5244$$

Q2. $H_0 : D = 0$
$H_1 : D < 0$

Let $d = X_1 - X_2 = X_{before} - X_{after}$.

$n = 9, \quad \bar{d} = -1.7778$
$S_d^2 = 2.9444 \Rightarrow S_d = 1.7159$

$\alpha = 0.05, \ df = n - 1 = 9 - 1 = 8$

For onetail test, $t_{0.05,8} = -1.86$

$$t_{observed} = \frac{\bar{d} - D}{S_d / \sqrt{n}} = \frac{-1.7778 - 0}{1.7159/\sqrt{9}}$$

$t_{obs} = -3.1082$

Since $t_{obs} < t_{crit}$
$-3.11 < -1.86$

The decision is to reject the null hypothesis

Q3.

Sample 1

$n_1 = 350$

$x_1 = 160$

Sample 2

$n_2 = 410$

$x_2 = 190$

$\Rightarrow \hat{p}_1 = \frac{x_1}{n_1} = 0.4571 \qquad \hat{p}_2 = \frac{x_2}{n_2} = 0.4634$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{160+190}{350+410} = 0.4605$$

$$\bar{q} = 1 - \bar{p} = 0.5395$$

$H_0: p_1 - p_2 = 0$

$H_a: p_1 - p_2 \neq 0$

For two-tail test, $\frac{\alpha}{2} = 0.025$

$$Z_{0.025} = \pm 1.96$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)}} = \frac{0.4571 - 0.4634}{\sqrt{0.4605 \times 0.5395 \left(\frac{1}{350} + \frac{1}{410}\right)}}$$

$Z_{obs} = -0.1737, \quad |Z_{obs}| < 1.96$

Fail to reject H_0.

Q3 (b)

Sample 1	Sample 2
$\hat{p}_1 = 0.4$	$\hat{p}_2 = 0.25$
$n_1 = 700$	$n_2 = 600$

$H_0 : p_1 - p_2 = 0$

$H_a : p_1 - p_2 > 0$

Let $\alpha = 0.1 \Rightarrow Z_{0.1} = 1.28$ (single-tail)

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = 0.3308, \quad \bar{q} = 0.6692$$

$$Z_{obs} = \frac{0.4 - 0.25}{\sqrt{\bar{p}\cdot\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = 5.7304 > 1.28$$

Reject the null hypothesis.

Q4. Let Houston be group 1, & Chicago be group 2.

$H_0: \sigma_1^2 = \sigma_2^2$

$H_a: \sigma_1^2 \neq \sigma_2^2$

$\alpha = 1\% = 0.01$.

$df_1 = 13-1 = 12$
$df_2 = 11-1 = 10$

This is a two-tail problem. $\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$

$F_{0.005,12,10} = 5.66$

$F_{0.995,10,12} = 0.177 = \frac{1}{5.66}$

If the observed value is greater than 5.66 or less than 0.177, we reject the null hypo.

$$\left.\begin{array}{l} s_1^2 = 393.3974 \\ s_2^2 = 702.6909 \end{array}\right\} \Rightarrow F = \frac{s_1^2}{s_2^2} = 0.5598$$

Since, $0.177 < 0.5598 < 5.66$ is true
We accept H_0.

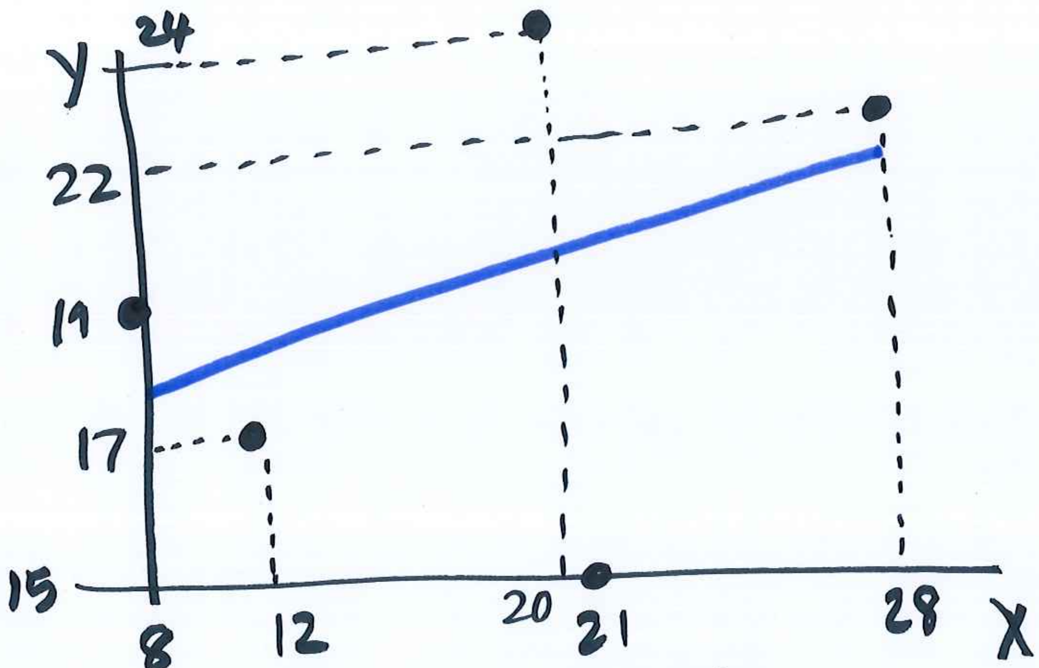
Q5. X = [12, 21, 28, 8, 20]';

Y = [17, 15, 22, 19, 24]';

XX = [ones(5,1) X];

$$Bs = (XX' * XX)^{-1} * XX' * Y$$

$$= \begin{bmatrix} 16.5096 \\ 0.1624 \end{bmatrix} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix}$$



b0 = 16.5096; b1 = 0.1624;

line([8, 28], b1 * [8, 28] + b0)

$$S_b = \frac{S_e}{\sqrt{\Sigma X^2 - (\Sigma X)^2/n}}$$

$\hat{Y} = b_0 * b_1 X = [18.46, 19.92, 21.06, 17.81 19.76];$

$$S_e^2 = \frac{SSE}{n-2}$$

$$SSE = \Sigma_{i=1}^{5} (Y_i - \hat{Y}_i)^2 = \text{sum}((Y - \text{Yhat}).\hat{}2)$$

$$= 46.6399$$

$$S_e = \sqrt{\frac{SSE}{n-2}} = 3.9429.$$

$$S_b = \frac{3.9429}{\sqrt{1833 - 89^2/5}} = 0.25$$

$b_1 = 0.1624$

$\alpha = 0.05$

$H_0: \quad \beta_1 = 0$

$H_a: \quad \beta_1 \neq 0$

This is a two-tail problem

$$\alpha/2 = 0.025$$

$$df = n - 2 = 5 - 2 = 3$$

$$t_{0.025,3} = \pm 3.182$$

$$t = \frac{b_1 - \beta_1}{S_b} = \frac{0.1624 - 0}{0.25} = 0.6496$$

Since $t = 0.6496 < t_{0.025,3} = 3.182$, the decision is to fail to reject H_0. (Slope is significant).

HABIB UNIVERSITY	**Midterm-I Exam (Part A)** **CE 362: Statistics and Inference** **Dr. Shafayat Abrar**	**Timing: 02:25 PM - 03:40 PM** **Oct 26, 2022** **Duration: 75 min**

Q1 [25 points]: A random sample of 51 items is taken. The data is shown below (and also shared in a separate file). Use this data to test the following hypotheses, assuming you want to take only a 1% risk of committing a Type I error and that the data is known to be normally distributed.

$$\begin{aligned} H_0: &\quad \mu = 60 \\ H_a: &\quad \mu < 60 \end{aligned}$$

59.37	61.56	61.66	58.57
55.45	61.02	58.02	57.52
62.15	62.82	54.04	50.74
54.74	61.99	59.74	58.39
52.86	52.40	62.43	54.60

Q2 [25 points]: Suppose you are testing $H_0: p = 0.3$ versus $H_a: p \neq 0.3$. A random sample of 740 items shows that 205 have this characteristic. With a 0.05 probability of committing a Type-I error, test the hypothesis.

(a) Using p-value method, find the probability of the observed z value for this problem. What is your decision about the hypothesis test?

(b) If you had used the critical value method, what would the two critical values be?

(c) How do the sample results compare with the critical values?

Q3 [25 points]: A savings and loan averages about $\$100,000$ in deposits per week. However, because of the way pay periods fall, seasonality, and erratic fluctuations in the local economy, deposits are subject to a wide variability.

In the past, the variance for weekly deposits has been about $\$199,996,164$. In terms that make more sense to managers, the standard deviation of weekly deposits has been $\$14,142$ (which is simply the square root of $\$199,996,164$).

Shown here are data from a random sample of 15 weekly deposits for a recent period. Assume weekly deposits are normally distributed. Use these data and $\alpha = 0.10$ to test to determine whether the variance for weekly deposits has changed from its past value $\$199,996,164$.

$95,000	135,000	115,000
70,000	45,000	105,000
130,000	140,000	130,000
110,000	95,000	70,000
85,000	100,000	120,000

Q4 [25 points]: Suppose a hypothesis states that the mean is exactly 60. If a random sample of 30 items is taken to test this hypothesis, what is the value of Type-II error probability, β, if the population standard deviation is 8 and the alternative mean is 65? Use Type-I error probability, $\alpha = 0.01$

Q1 Solution:

$n = 20$. $\alpha = 0.01$ (Type-I error)

$\bar{X} = 58$ (using Matlab)

$S^2 = 14.298$ (using Matlab).

$H_0: \mu = 60$

$H_a: \mu < 60$

One-tailed problem:

we use t-statistic because variance is not given, assuming it to be unknown.

$df = n-1 = 20-1 = 19$.

$t_{\alpha,df} = t_{0.01,19} = -2.5395$

(minus sign to be used)

$$t = \frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{58-60}{\sqrt{\frac{14.298}{20}}} = -2.3654$$

Observed $t = -2.3654 > t_{0.01,19} = -2.539$

The decision is to fail to reject the null hypothesis.

Q2. Solution:

$$H_0: \quad p = 0.3$$
$$H_a: \quad p \neq 0.3$$

A two-tailed problem. $\alpha = 0.05$.

$$n = 740$$
$$n_0 = 205$$
$$\hat{p} = \frac{205}{740} = 0.277$$

We use Z-statistic for this proportion problem.

$$\text{observed } Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.277 - 0.3}{\sqrt{\frac{0.3 \times 0.7}{740}}} = -1.365.$$

For two-tail, $\alpha/2 = 0.025$

$$Z_{critical} = Z_{0.025} = \pm 1.96$$

Since, observed $Z = -1.365 > -1.96 = Z_c$

The decision is to fail to reject the null hypothesis.

Q2. <u>p-value method.</u>

observed $Z = -1.365$

from the table of z-statistic, we obtain

$$\Phi(-1.365) = \Phi(z) = \text{Area} \cong \frac{\Phi(-1.36) + \Phi(-1.37)}{2}$$

$$\cong \frac{0.0869 + 0.0853}{2}$$

$$\cong 0.0861 \text{ (using interpolation)}$$

A true value, however, may be computed using computer

$$\Phi(-1.365) = 1 - \frac{1}{2}\operatorname{erfc}\left(\frac{-1.365}{\sqrt{2}}\right)$$

$$= 0.086126525$$

The simple interpolation (as shown above) provides a pretty good approximation.

(4)

Since the p-value = 0.0861 > $\frac{\alpha}{2}$ = 0.025, the decision is to fail to reject the null hypothesis.

Critical values method:

$$Z_c = \frac{\hat{p}_c - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$\pm 1.96 = \frac{\hat{p}_c - 0.3}{\sqrt{\frac{0.3 \times 0.7}{740}}}$$

$$\hat{p}_c = 0.3 \pm 0.03301$$

0.267 & 0.333 are the critical values.

Since $\hat{p} = 0.277$ (the observed value) is not outside critical values in tails, the decision is to fail to reject the null hypothesis.

Q3 Solution:

$$\mathcal{H}_0: \sigma^2 = \$\ 199{,}996{,}164$$

$$\mathcal{H}_a: \sigma^2 \neq \$\ 199996{,}164$$

This is a two-tailed problem.

$\alpha = 0.1$ (given) $\Rightarrow \frac{\alpha}{2} = 0.05$

$n = 15$ (values given).

$df = n-1 = 14$

$S^2 = var(\text{given-data-values})$

$= 738571428.57$

This is variance test problem, we use χ^2-statistic.

$$\chi^2_{0.05,14} = 23.6848$$

$$\chi^2_{0.95,14} = 6.5706$$

The observed statistic is

$$\chi^2 = \frac{(15-1)738571428.57}{199996164} = 51.7$$

Since. $\chi^2 = 51.7 > 23.684$ The decision is to REJECT the null hypothesis. The variance has changed.

Q4 Solution.

$$H_0 : \mu = 60$$
$$H_a : \mu \neq 60$$

This is two-tailed problem.

$n = 30,$

$\sigma^2 = 8,$

$\alpha = 0.01 \Rightarrow \alpha/2 = 0.005$

For β, assume $\mu_a = 65.$

$Z_{0.005} = \pm 2.5758$

Let us find critical values of sample mean.

$$Z_c = \frac{\bar{X}_c - \mu}{\sigma/\sqrt{n}}$$

$$\bar{X}_c = \mu + Z_c \frac{\sigma}{\sqrt{n}}$$

$$\bar{X}_c = 60 \pm 2.5758 \times \frac{8}{\sqrt{30}}$$

56.2378 and 63.7622

Finding critical Z-values for alternate hypothesis.

$$Z_1^{u} = \frac{63.7622 - 65}{8/\sqrt{30}} = -0.8474$$

$$Z_1^{L} = \frac{56.2378 - 65}{8/\sqrt{30}} = -5.999 = -6$$

$$\beta = \Phi(-0.8474) - \underbrace{\Phi(-6)}_{\approx 0}$$

$$\beta = 0.\underbrace{19838610225021 5}_{\text{using Matlab.}}$$

$$1 - \frac{1}{2}\operatorname{erfc}(-0.8474/\sqrt{2})$$

The probability of Type-II error is 19.84%.

Q1 [20 points]: Consider identically and independently distributed (iid) samples X_i for $i = 1, 2, \ldots, n$ from the Rayleigh probability density function (pdf)

$$f(x_i \mid \sigma) = \frac{x_i}{\sigma^2} \exp\left(-\frac{1}{2}\frac{x_i^2}{\sigma^2}\right).$$

Derive the Neyman-Pearson test

$$L(\boldsymbol{X}) = \frac{f(\boldsymbol{X}; \mathcal{H}_1)}{f(\boldsymbol{X}; \mathcal{H}_0)} > \gamma$$

for the hypothesis testing problem

$$\begin{aligned} \mathcal{H}_0 &: \sigma^2 = \sigma_0^2 \\ \mathcal{H}_1 &: \sigma^2 = \sigma_1^2 > \sigma_0^2. \end{aligned}$$

Q2 [30 points]: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ where $\mu \in (-\infty, \infty)$ and $\sigma^2 \in (0, \infty)$ are unknown parameters. Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

be the sample mean which aims to estimate the unknown μ.
Note that $\overline{X}$ is a random variable.

(a) Prove that the pdf of $\overline{X}$ is given by

$$f(x) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{n}{2\sigma^2}(x-\mu)^2\right) \qquad \text{for } x \in (-\infty, \infty)$$

Note: You need to evaluate mean and variance of $\overline{X}$.

(b) Evaluate the Cramer-Rao lower bound (crlb), i.e., $\operatorname{var}(\widehat{\mu} - \mu) \approx \dfrac{1}{nI(\mu)}$, where $nI(\mu) = -E[\ell''(\mu)]$, and $\ell(\mu) = \sum_{i=1}^{n} \log f(x_i \mid \mu)$ is the log likelihood function.

(c) Based on crlb, argue if $\overline{X} = \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the optimal estimator of μ.

Q3 [20 points]: Suppose that $X_1, \ldots, X_n$ are iid from the Weibull pdf

$$f(x \mid \alpha) = \alpha^{-1} \beta\, x^{\beta-1} \exp\left(-x^{\beta}/\alpha\right), \quad x > 0$$

where $\alpha(> 0)$ is the unknown parameter, but $\beta(> 0)$ is assumed known.

(a) Using Neyman factorization theorem, obtain the sufficient statistic for α.

(b) Using maximum likelihood method, obtain an estimator of α.

Q4 [20 points]: Suppose that $X_1, \ldots, X_n$ are iid from the Beta pdf

$$f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1} (1-x)^{\beta-1},$$
$$0 < x < 1, \alpha > 0, \beta > 0$$

where α is the unknown parameter, but β is assumed known.

(a) Using Neyman factorization theorem, obtain the sufficient statistic for α.

(b) Using maximum likelihood method, obtain an estimator of α.

Q5 [30 points]: This problem is concerned with the estimation of the variance of a normal distribution with unknown mean from a sample $X_1, \ldots, X_n$ of i.i.d. normal random variables. During problem sessions, we have discussed the evaluation of mean-squared errors (mse) of the following two estimators of variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2, \quad \text{and} \quad \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

Now consider another estimator of variance as given by

$$P = \rho \cdot \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

where $\rho > 0$. We can instantly notice that P can be written in terms of S^2 or $\widehat{\sigma^2}$ as

$$P = \rho \cdot (n-1) \cdot S^2 = \rho \cdot n \cdot \widehat{\sigma^2}$$

Neatly proving all steps, prove that the value of ρ which minimizes the mse of P is

$$\rho = \frac{1}{n+1}.$$

Q1

$$\mathcal{H}_0 : \sigma^2 = \sigma_0^2$$

$$\mathcal{H}_1 : \sigma^2 = \sigma_1^2$$

$$p(x_i | \mathcal{H}_0) = \frac{x_i}{\sigma_0^2} \exp\left(-\frac{1}{2}\frac{x_i^2}{\sigma_0^2}\right)$$

$$p(\bar{x} | \mathcal{H}_0) = \frac{1}{(\sigma_0^2)^n}\left[\prod_{i=1}^{n} x_i\right] \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i}^{n} x_i^2\right)$$

$$p(x_i | \mathcal{H}_1) = \frac{x_i}{\sigma_1^2} \exp\left(-\frac{1}{2}\frac{x_i^2}{\sigma_1^2}\right)$$

$$p(\bar{x} | \mathcal{H}_1) = \frac{1}{(\sigma_1^2)^n}\left[\prod_{i=1}^{n} x_i\right] \exp\left(-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n} x_i^2\right)$$

$$\mathcal{L}(x) = \frac{p(\bar{x} | \mathcal{H}_1)}{p(\bar{x} | \mathcal{H}_0)} > \gamma$$

$$= \frac{(\sigma_0^2)^n \exp\left(-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n} x_i^2\right)}{(\sigma_1^2)^n \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n} x_i^2\right)}$$

$$\left(\frac{\sigma_0}{\sigma_1}\right)^{2n} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\sum_{i=1}^{n} x_i^2\right) > \gamma$$

$$\frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)\sum_{i=1}^{n} x_i^2 > \ln\left(\frac{\sigma_1^{2n}}{\sigma_0^{2n}}\gamma\right)$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 > \frac{\ln\left(\frac{\sigma_1^{2n}}{\sigma_0^{2n}}\gamma\right)}{\frac{N}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)} = \gamma'$$

Q2. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$E\bar{X} = \frac{1}{n}\sum_{i=1}^{n} EX_i$

since $X_i \sim N(\mu, \sigma^2)$, $EX_i = \mu$.

$E\bar{X} = \frac{1}{n} \cdot n \cdot \mu = \mu$ (unbiased)

$$\mathrm{var}(\bar{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{var}\left(\sum_i X_i\right)$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n} \mathrm{var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq n}}^{n} \mathrm{cov}(X_i, X_j)\right]$$

$$\mathrm{cov}(X_i, X_j) = E(X_i - \bar{X}_i)(X_j - \bar{X}_j) = 0$$

because X_i & X_j are independent.

$$\mathrm{var}(\bar{X}) = \frac{1}{n^2} \cdot n \cdot \mathrm{var}(X_i) = \frac{\mathrm{var}(X_i)}{n}$$

$$\mathrm{var}(\bar{X}) = \sigma^2/n$$

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

which immediately gives.

$$f_{\bar{X}}(x) = \frac{1}{\sqrt{\frac{2\pi\sigma^2}{n}}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2/n}\right).$$

(b) $$\text{lik}(x) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2/n}\right)$$

$$\begin{aligned} l(x) &= \log \text{lik}(x) \\ &= \frac{1}{2}\log\left(\frac{n}{2\pi\sigma^2}\right) - \frac{n(x-\mu)^2}{2\sigma^2} \end{aligned}$$

$$l'(x) = \frac{\partial l(x)}{\partial \mu} = \frac{+2n(x-\mu)}{2\sigma^2}$$

$$l''(x) = \frac{\partial^2 l(x)}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

$$\text{var}(\mu - \hat{\mu}) = \frac{1}{-E\, l''(\mu)} = \frac{\sigma^2}{n} = \text{C.R.L.B.}$$

CRLB stands for Cramer - Rao lower bound

Since $\text{var}(\bar{X}) = \text{var}(\mu - \hat{\mu}) = \frac{\sigma^2}{n}$

Therefore $\bar{X} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the optimal estimator of μ.

Q3. $f(x|\alpha) = \frac{\beta}{\alpha} x^{\beta-1} \exp(-\frac{x^\beta}{\alpha})$

(a) Since β is known, only α is an only unknown parameter.

$$f(x|\alpha) = \underbrace{\beta x^{\beta-1}}_{=:h(x)} \underbrace{\frac{1}{\alpha} \exp(-\frac{x^\beta}{\alpha})}_{g(x;\alpha)}$$

if $X = \{X_1, X_2, \ldots, X_n\}$, then

$$f(x|\alpha) = \underbrace{\beta^n \prod_{i=1}^n x_i^{\beta-1}}_{=:h(x)} \underbrace{\frac{1}{\alpha^n} \exp\left(-\frac{1}{\alpha}\sum_{i=1}^n x_i^\beta\right)}_{=:g(x;\alpha)}$$

The sufficient statistic for α is

$$T(X) = \sum_{i=1}^n X_i^\beta$$ (assuming β is known)

Q3 (b)

$$lik(\alpha) = \frac{1}{\alpha^n} \exp\left(-\frac{1}{\alpha} \sum_{i=1}^{n} X_i^{\beta}\right)$$

$$l(\alpha) = \log lik(\alpha)$$

$$= -n \log(\alpha) - \frac{1}{\alpha} \sum_{i=1}^{n} X_i^{\beta}$$

$$l'(\alpha) = \frac{\partial l(\alpha)}{\partial \alpha} = -\frac{n}{\alpha} + \frac{1}{\alpha^2} \sum_{i=1}^{n} X_i^{\beta}$$

Substituting $l'(\alpha) = 0$, we get

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} X_i^{\beta}$$

The m.l. estimator of α is thus found in closed-form.

Q4. (a) β is known; α is unknown.

$$f(x|\alpha) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Applying Neyman factorisation, we get

$$f(x|\alpha) = h(x)\, g(x|\alpha)$$

where

$$h(x) = \frac{(1-x)^{\beta-1}}{\Gamma(\beta)}$$

and

$$g(x|\alpha) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} x^{\alpha-1}$$

for multivariate $X = \{X_1, X_2, \dots, X_n\}$

$$g(x|\alpha) = \frac{\Gamma(\alpha+\beta)^n}{\Gamma(\alpha)^n} \prod_{i=1}^{n} x_i^{\alpha-1}.$$

Q5. $P = \rho(n-1)S^2$

$$EP = E[\rho(n-1)S^2]$$
$$= \rho(n-1)\,E[S^2]$$
$$= \rho(n-1)\sigma^2$$

For the proof of $ES^2 = \sigma^2$, refer to problem session II ppt slides.

Note $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

$$\Rightarrow \quad S^2 = \frac{P}{\rho(n-1)} \Rightarrow \frac{P}{\rho\sigma^2} \sim \chi^2_{n-1}$$

$$\mathrm{var}(P) = (\rho\sigma^2)^2\,\mathrm{var}(\chi^2_{n-1})$$
$$= \rho^2\sigma^4\, 2(n-1)$$
$$= 2\rho^2\sigma^4(n-1)$$

$$MSE(P) = var(P) + Bias^2(P)$$

where $var(P) = 2\rho^2\sigma^4(n-1)$

$Bias(P) = EP - \sigma^2 = \rho(n-1)\sigma^2 - \sigma^2$

$$MSE(P) = 2\rho^2\,\sigma^4(n-1) + [\rho(n-1)-1]^2\sigma^4$$

$$\frac{\partial MSE(P)}{\partial \rho} = 0$$ to find optimal value of ρ

$$(2\rho)(2\sigma^4(n-1)) + 2\sigma^4(\rho(n-1)-1)(n-1)$$

$$\rho\left(4\sigma^4(n-1) + 2\sigma^4(n-1)^2\right) = 2\sigma^4(n-1)$$

$$\rho = \frac{2(n-1)}{4(n-1) + 2(n-1)^2}$$

$$= \frac{2(n-1)}{2n^2 - 4n + 2 + 4n - 4}$$

$$= \frac{2(n-1)}{2(n^2-1)} = \frac{n-1}{n^2-1} = \frac{1}{n+1}$$ [proved].