Introduction to mathematical Statistics **Midterm 3 Solution**

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$.
(a). Derive the confidence interval for μ when σ^2 is known.
(b). Derive the confidence interval for μ when σ^2 is unknown.

Solution:

(a) First we derive the pivotal quantity for the inference on μ based on a random sample from a normal population $N(\mu, \sigma^2)$ when the population variance σ^2 is known.

1) Point Estimator for μ : $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

$\overline{X}$ is ***NOT*** a pivotal quantity since its distribution is not entirely known.

2) Let $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$, we can show that $Z \sim N(0,1)$ using the moment generating function method:

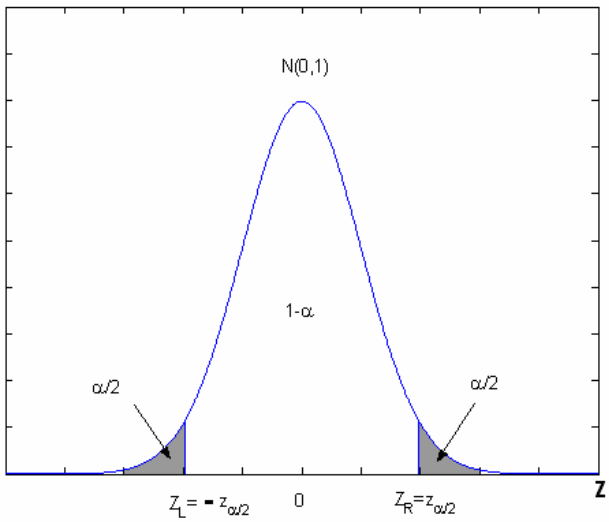
$$M_Z(t) = \exp(tZ) = \exp(t\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}) = \exp(\frac{-t\mu}{\sigma/\sqrt{n}})\exp(\frac{t}{\sigma/\sqrt{n}}\overline{X})$$

$$= \exp(\frac{-t\mu}{\sigma/\sqrt{n}})\exp(\frac{t}{\sigma/\sqrt{n}}\frac{\sum_{i=1}^{n} X_i}{n}) = \exp(\frac{-t\mu}{\sigma/\sqrt{n}})\prod_{i=1}^{n}\exp(\frac{t}{\sigma\sqrt{n}}X_i)$$

$$= \exp(\frac{-t\mu}{\sigma/\sqrt{n}})\exp\left\{n\left[\mu(\frac{t}{\sigma\sqrt{n}}) + \frac{\sigma^2}{2}(\frac{t}{\sigma\sqrt{n}})^2\right]\right\} = \exp\left[\frac{1}{2}t^2\right]$$

Thus Z is a pivotal quantity when σ is known.

3). Now we derive the confidence interval. First, we draw the pdf of our pivotal quantity Z as follows.



Let α be any small positive value less than 1 (*usually less than 0.5), in the above figure, we have:

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(\overline{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$\therefore$ The 100(1-α)% confidence interval for μ (when σ^2 is known and the population is normal) is

$$\left[\overline{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

(b) Next we derive the pivotal quantity for the inference on μ based on a random sample from a normal population N(μ, σ^2) when the population variance σ^2 is unknown.

1) Point estimator : $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

2) $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Note: the above two statistics are NOT pivotal quantities.

3) **Theorem.** Sampling from normal population

a. $Z \sim N(0,1)$

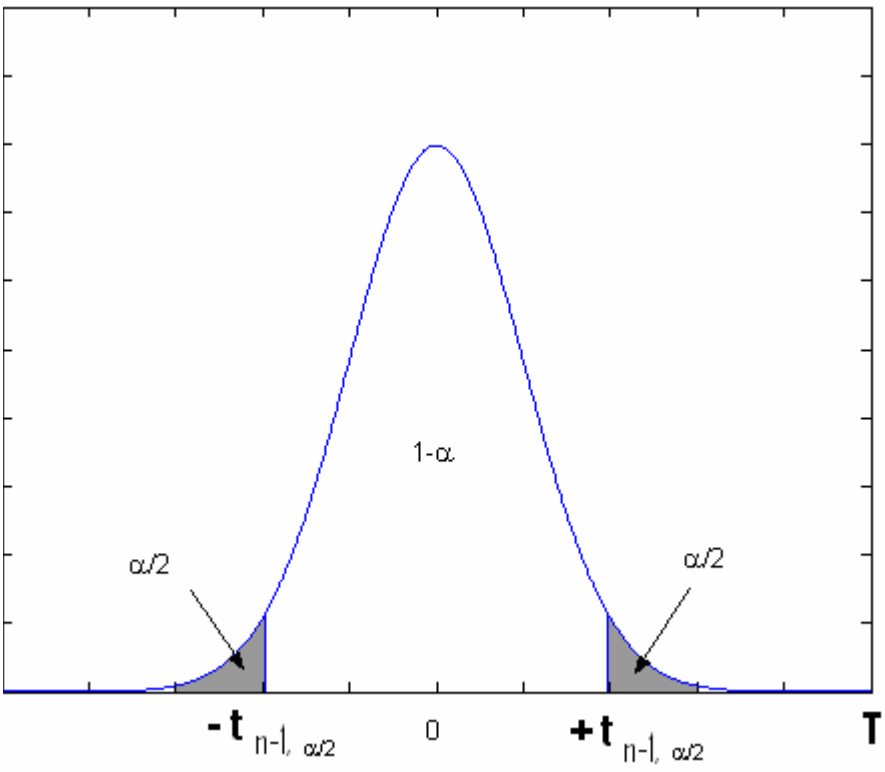
b. $W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

c. Z and W are independent

Definition. $T = \frac{Z}{\sqrt{W/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

Therefore $T = \frac{Z}{\sqrt{W/(n-1)}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ is a pivotal quantity.

4). Now we will use this pivotal quantity to derive the 100(1-α)% confidence interval for μ. We start by plotting the pdf of the t-distribution with n-1 degrees of freedom as follows:



The above pdf plot corresponds to the following probability statement:

$P(-t_{n-1,\alpha/2} \le T \le t_{n-1,\alpha/2}) = 1-\alpha$

$\Rightarrow P(-t_{n-1,\alpha/2} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{n-1,\alpha/2}) = 1-\alpha$

$\Rightarrow P(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \bar{X}-\mu \le t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$

$\Rightarrow P(-\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le -\mu \le -\bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$

$\Rightarrow P(\bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \ge \mu \ge \bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$

$\Rightarrow P(\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$

=> Thus the $100(1-\alpha)\%$ C.I. for μ when σ^2 is unknown is

$[\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}]$. (*Please note that $t_{n-1,\alpha/2} \ge Z_{\alpha/2}$)

2. Let $X_1,\ldots,X_n \overset{iid}{\sim} N(\mu_1,\sigma^2)$ and $Y_1,\ldots,Y_n \overset{iid}{\sim} N(\mu_2,\sigma^2)$ be two independent samples. Furthermore, σ^2 is known. Please derive the confidence interval for $\mu_1-\mu_2$

Solution:

1) Parameter of interest

$\mu_1-\mu_2$ (or $\frac{\mu_1}{\mu_2}$)

2) Point estimator for the parameter of interest

$\overline{X}-\overline{Y}$ (or $\frac{\overline{X}}{\overline{Y}}$)

$$\overline{X}-\overline{Y} \sim N(\mu_1-\mu_2,\frac{\sigma^2}{n}+\frac{\sigma^2}{n}) = N(\mu_1-\mu_2,\frac{2\sigma^2}{n})$$

The ratio is not used because it is much harder to derive its distribution.

3) $$Z=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{2}{n}}} \sim N(0,1)$$

$\therefore Z$ is a pivotal quantity for $(\mu_1-\mu_2)$ when σ is known.

We will prove the distribution of Z using the moment generating method as follows.

$$M_Z(t)=\exp(tZ)=\exp(t\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{2/n}})$$

$$=\exp(\frac{-t(\mu_1-\mu_2)}{\sigma\sqrt{2/n}})\exp(\frac{t}{\sigma\sqrt{2/n}}\overline{X})\exp(\frac{-t}{\sigma\sqrt{2/n}}\overline{Y})$$

$$=\exp(\frac{-t(\mu_1-\mu_2)}{\sigma\sqrt{2/n}})\exp(\frac{t}{\sigma\sqrt{2/n}}\frac{\sum_{i=1}^n X_i}{n})\exp(\frac{-t}{\sigma\sqrt{2/n}}\frac{\sum_{i=1}^n Y_i}{n})$$

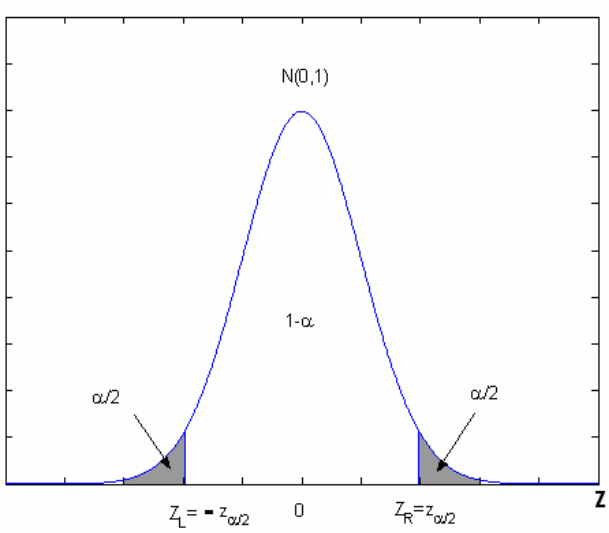
$$=\exp(\frac{-t(\mu_1-\mu_2)}{\sigma\sqrt{2/n}})\left[\prod_{i=1}^n \exp(\frac{t}{\sigma\sqrt{2n}}X_i)\right]\left[\prod_{i=1}^n \exp(\frac{-t}{\sigma\sqrt{2n}}Y_i)\right]$$

$$=\exp(\frac{-t(\mu_1-\mu_2)}{\sigma\sqrt{2/n}})\exp\left\{n\left[\mu_1(\frac{t}{\sigma\sqrt{2n}})+\frac{\sigma^2}{2}(\frac{t}{\sigma\sqrt{2n}})^2\right]\right\}\exp\left\{n\left[\mu_2(\frac{-t}{\sigma\sqrt{2n}})+\frac{\sigma^2}{2}(\frac{-t}{\sigma\sqrt{2n}})^2\right]\right\}$$

$$=\exp(\frac{-t(\mu_1-\mu_2)}{\sigma\sqrt{2/n}})\exp\left[\mu_1(\frac{t}{\sigma\sqrt{2/n}})+\frac{t^2}{4}\right]\exp\left[\mu_2(\frac{-t}{\sigma\sqrt{2/n}})+\frac{t^2}{4}\right]$$

$$=\exp\left[\frac{1}{2}t^2\right]$$

4) Now we derive the confidence interval. First, we draw the pdf of our pivotal quantity Z as follows.



Let α be any small positive value less than 1 (*usually less than 0.5), in the above figure, we have:

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2} \le \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{2}{n}}} \le Z_{\alpha/2}) = 1-\alpha$$

$$P(-Z_{\alpha/2}\sigma\sqrt{\frac{2}{n}} \le \overline{X}-\overline{Y}-(\mu_1-\mu_2) \le Z_{\alpha/2}\sigma\sqrt{\frac{2}{n}}) = 1-\alpha$$

$$P(\overline{X}-\overline{Y}-Z_{\alpha/2}\sigma\sqrt{\frac{2}{n}} \le \mu_1-\mu_2 \le \overline{X}-\overline{Y}+Z_{\alpha/2}\sigma\sqrt{\frac{2}{n}}) = 1-\alpha$$

$\therefore$ The 100(1-α)% confidence interval for μ (when σ^2 is known and the population is normal) is

$$\left[\overline{X}-\overline{Y}-Z_{\alpha/2}\sigma\sqrt{\frac{2}{n}},\ \overline{X}-\overline{Y}+Z_{\alpha/2}\sigma\sqrt{\frac{2}{n}}\right]$$

3. During one of the "beer wars" in the early 1980's, a taste test between Schlitz and Budweiser was the focus of a TV commercial. 100 people agreed to drink 2 unmarked mugs and indicate which of the two beers they liked better. The results: fifty-four chose "Bud" while the rest chose Schlitz.
(a). Please construct and interpret the corresponding 95% confidence interval for p - the proportion of beer drinkers who prefer Bud to Schlitz. (*Note: Please derive the general formula for the 100(1-α)% confidence interval for a population proportion p based on a large sample first.)
(b). How large does the sample size need to be in order for the sample proportion $\hat{p}$ to have a 95% chance of lying within 0.03 of p? Please calculate the sample size for two scenarios: (a) we have no estimate for p ; (b) we have an estimate for p in $\hat{p} = 0.54$ (*Note: Please first derive the general formula for sample size calculation based on a maximum error of E and a confidence level of 100(1-α)%.)

Solution:
(a)
Let $X_i \overset{i.i.d.}{\sim} Bernoulli(p),\ i = 1, \cdots, n$, please find the 100(1-α)% CI for p.

Point estimator : $\boxed{\hat{p} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}}$ (ex. $n = 1000$, $\hat{p} = 0.6$)

Our goal: derive a 100(1-α)% C.I. for p

By the central limit theorem (CLT), for a large sample, we have $\boxed{Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} \dot{\sim} N(0,1)}$

$$E(\bar{X}) = E(\frac{\sum X_i}{n}) = \frac{1}{n} E(\sum X_i) = \frac{1}{n} \cdot np = p\ ,\ (\because \sum X_i \sim B(n,p))$$

$$Var(\bar{X}) = Var(\frac{\sum X_i}{n}) = \frac{1}{n^2} Var(\sum X_i) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

$$Z = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

$$\boxed{Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)}$$

$$\boxed{Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \dot{\sim} N(0,1)} \text{ (By Slustky's theorem)}$$

$100(1-\alpha)\%$ **(approximate, or large sample) C.I. for p**

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) \approx 1-\alpha$$

$$\Rightarrow P(-Z_{\alpha/2} \le \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le Z_{\alpha/2}) \approx 1-\alpha$$

$$\Rightarrow P(-Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \hat{p} - p \le Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

$$\Rightarrow P(-\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le -p \le -\hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

$$\Rightarrow P(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ge p \ge \hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

$$\Rightarrow P(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

$$\Rightarrow \text{The } 100(1-\alpha)\% \text{ large sample C.I. for p is } \boxed{[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]}.$$

\# CLT => n large usually means $n \ge 30$

\# special case for the inference on p. n large means

Let $X = \sum_{i=1}^{n} X_i$, large sample means:

$n\hat{p} = X \ge 5$ (X= total # of 'S'), and $n(1-\hat{p}) = n - X \ge 5$ (n-X= total # of 'F')

For the given problem, we have $n = 100,\ X = 54,$ and we want a 95% CI for p

For a 95% confidence interval, $1-\alpha = 0.95,\ \alpha = 0.05,\ \frac{\alpha}{2} = 0.025$

$$\hat{p} = \frac{54}{100} = 0.54\ ;\ Z_{0.025} = 1.96$$

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.54)(0.46)}{100}} = 0.049$$

$$Z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times 0.049 = 0.096$$

$\therefore$ The 95% confidence interval for p is $[0.444, 0.636]$

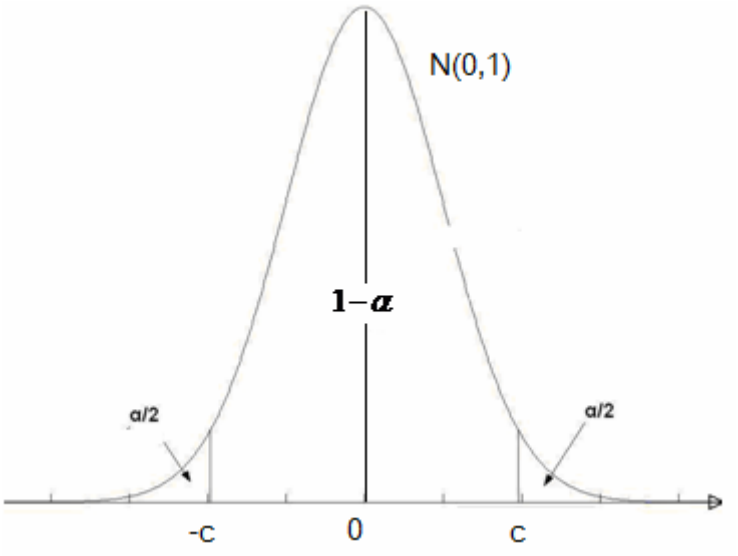
(b) **Derive the general formula for n**

$$P(|\hat{p} - p| \le E) = 1-\alpha$$

$$P(-E \le \hat{p} - p \le E) = 1-\alpha$$

$$P(-\frac{E}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le \frac{E}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}) = 1-\alpha \text{ and } Z = \frac{p-\hat{p}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

$$Thus:\ P(-\frac{E}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le Z \le \frac{E}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}) = 1-\alpha$$



$$c = Z_{\alpha/2} = \frac{E}{\sqrt{\frac{\hat{p}(1-p)}{n}}}$$

$$\therefore n = \frac{(Z_{\alpha/2})^2 \hat{p}(1-p)}{E^2} \le \frac{(Z_{\alpha/2})^2}{4 \cdot E^2}$$

Plug in
$Z_{0.025} = 1.96,\ \hat{p} = 0.5,\ E = 0.03$, we have n = 1068

Plug in
$Z_{0.025} = 1.96,\ \hat{p} = 0.54,\ E = 0.03$, we have n = 1061

4. Let $X_i \overset{i.i.d.}{\sim} Bernoulli(p)$, $i = 1, 2, \cdots, n.$ Please

(a). Derive the method of moment estimator for p
(b). Derive the maximum likelihood estimator for p
(c). Is there an efficient estimator for p? Please show the entire derivation.
Hint: Cramér-Rao Inequality: Let $\hat{\theta} = h(X_1, X_2, \cdots, X_n)$ be unbiased for θ, where $X_i,\ i = 1, \cdots, n,$ is a random sample from a population with pdf $f_X(x;\theta)$ satisfying all regularity conditions. Then

$$Var(\hat{\theta}) \ge \left\{ nE\left[\left(\frac{\partial \ln f_X(x;\theta)}{\partial \theta} \right)^2 \right] \right\}^{-1} = \left\{ -nE\left[\frac{\partial^2 \ln f_X(x;\theta)}{\partial \theta^2} \right] \right\}^{-1}$$

Solution: $P(X = x) = f(x; p) = p^x (1-p)^{1-x}\ ;\ x = 0, 1;$

(a). The population mean is p (because $E(X) = 1*p + 0*(1-p) = p$) and the sample mean is $\frac{\sum_{i=1}^{n} X_i}{n}$.

Therefore the moment estimator of p is $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$.

(b). $L = \prod_{i=1}^{n} f(x_i; p)$

$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

$= p^{\sum x_i}(1-p)^{n-\sum x_i}$

$l = \ln L = (\sum x_i)\ln p + (n - \sum x_i)\ln(1-p)$

$\frac{\partial l}{\partial p} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0$

$\therefore \hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$ is the MLE for p

(c). $E(\hat{p}) = p$

$\text{var}(\hat{p}) = \frac{p(1-p)}{n}$

Now we derive the C-R lower bound for an unbiased estimator of p:

$P(X = x) = f(x; p) = p^x(1-p)^{1-x}$; $x = 0,1;$

$\ln f(x, p) = x\ln p + (1-x)\ln(1-p)$

$\frac{\partial \ln f(x,p)}{\partial p} = \frac{x}{p} - \frac{1-x}{1-p}$

$\frac{\partial^2 \ln f(x,p)}{\partial p^2} = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$

$E\left[-\frac{X}{p^2} - \frac{1-X}{(1-p)^2}\right] = -\frac{p}{p^2} - \frac{1-p}{(1-p)^2} = -\frac{1}{p(1-p)}$

C-R lower bound

$\text{var}(\hat{p}) \geq \frac{1}{-nE\left[\frac{\partial^2 \ln f}{\partial p^2}\right]} = \frac{p(1-p)}{n}$

The MLE of p is unbiased and its variance = C-R lower bound. Thus it is an efficient estimator of p.

Definition. Efficient Estimator

If $\hat{\delta}$ is an unbiased estimator of δ and its variance = C-R lower bound, then $\hat{\delta}$ is an efficient estimator of δ.

*** That's all, folks! Good luck! ***