# Question 1.

1. If the price of unleaded petrol at UK filling stations is a random variable with mean 120.8p per litre, and standard deviation 4.9p, use the Central Limit theorem to determine the probability that the average price in a random sample of 50 filling stations is below 122p.

So we have that $\mu_F = 120.8$, $\sigma_F = 4.9$.

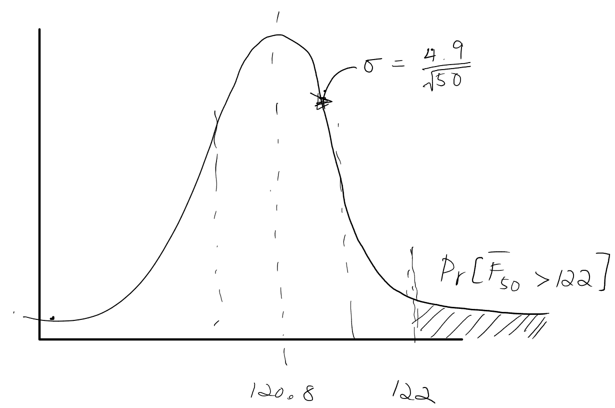
Let $\bar{F}_{50}$ be $\frac{1}{50}\sum_{i=1}^{50}(F_i)$.

$\mathbb{E}[\bar{F}_{50}] = \mu_F$, $Var(\bar{F}_{50}) = \frac{\sigma_F}{50}$

And so we are asking $Pr[\bar{F}_{50} > 122]$?

By the CLT,

$\bar{F}_{50} \dot{\sim} N(120.8, \frac{4.9^2}{50})$, and so that looks like the following:




Now we simply standardise both sides:

$$Pr[\bar{F}_{50} > 122] = Pr\left[\frac{\bar{F}_{50} - 120.8}{4.9/\sqrt{50}} > \frac{122 - 120.8}{4.9/\sqrt{50}}\right]$$

$$= Pr[Z > 1.73]$$

2. Suppose that students' marks on the economics prelims paper are normally distributed with mean 61 and standard deviation 9.5.
(Assume that the number of colleges is sufficiently large that individual observations may be considered i.i.d.)

(a) What is the distribution of the sample mean for a random sample of size n?

Denote students' marks on the paper as $M \sim N(61, 9.5^2)$.

Let the sample mean be $\bar{M}_n = \frac{1}{n}(\sum^n M_i)$

The distribution of the sample mean, has

$$E(\bar{M}_n) = \mu_M = 61 \quad \text{and}$$

$$Var(\bar{M}_n) = \frac{\sigma_M^2}{n} = \frac{9.5^2}{n}$$

Is this true?

YES, the sum of indep. normal variables is itself normal.

Given that the population is normally distributed, the sample mean is also normally distributed as $\{M_i\}$ are i.i.d from the population (what property is this?)

The sampling distribution is therefore

$$\bar{M}_n \sim N\left(61, \frac{9.5^2}{n}\right).$$

(b) In a random sample of 10 students, what is the probability that their average mark exceeds 63?

$$M_{10} \sim N\left(61, \frac{9.5^2}{10}\right).$$

We want $Pr[M_{10} > 63]$, Standardising, we obtain

$Pr\left[\frac{M_{10}-61}{9.5/\sqrt{10}} > \frac{63-61}{9.5/\sqrt{10}}\right]$, which can be simplified to give

$Pr\left[Z > \frac{2\sqrt{10}}{9.5}\right]$ Checking the standard normal table, we obtain

$Pr[Z > 0.66574] \approx 0.25$.

(c) Suppose that you have a sample of 10 students that is selected by choosing a college at random, and then choosing 10 students at random from that college.
i. What is the expected value of their average mark?
ii. Explain why you cannot determine the variance of their average mark. Is it likely to be higher or lower than the variance of the sample mean in random sample of 10 students? Explain the intuition for your answer.

(i) The expected value of their average mark is also 61.

(ii). The variance of their average mark differs from college to college; the variance of the average mark is likely to be higher in this case compared to the variance of the sample mean in a random sample of all students. This is because ...

why?

Consider drawing 10 students from a particular college ....

5. The 1165 Oxford PPE applicants in 2007 achieved an average score of 60.86 on the TSA test, with a standard deviation of 8.02. Construct a 95% confidence interval for the population mean score.

The population mean score $M$: $\mu_M = 60.86$, $\sigma_M = 8.02$.

What is the question asking us actually? Suppose we started taking samples of size $n$ and seeing the sample mean score $\bar{M}_n$. If $\bar{M}_n$ were large enough, then $\bar{M}_n \sim N\left(60.86, \frac{8.02}{\sqrt{n}}\right)$ by the CLT.

The confidence interval is a random variable: it is the low and the high values such that $\mu_M$ would fall within in 95% of samples. (Doesn't this depend on $n$, though?)

$$95\%\ CI = \{\mu_M + 1.96\,\sigma(\bar{M}_n)\}.$$

$$= \left\{60.86 + 1.96\,\frac{SE(\bar{M}_n)}{\sqrt{n}}\right\}$$ unbiased estimator of

$$= \left\{60.86 + 1.96 \cdot \frac{8.02}{\sqrt{n}}\right\}.$$

6.(a) Consider a random sample of size n from a Bernoulli distribution with parameter p. If the sample mean is X̄, show that the sample variance is given by s2 = n
X̄(1 − X̄). Compare the sample mean and variance with the
n−1
population mean and variance.

Sample mean $\bar{X} = \frac{1}{n}\sum^{n} X_i$

$\mathbb{E}(\bar{X}) = p.$

Show: $s^2 = \frac{n}{n-1}\bar{X}(1-\bar{X})$.

We have that $s^2 = \frac{1}{n-1}\sum_i^n (X_i - \bar{X})^2$

$$\Rightarrow s^2 = \frac{1}{n-1}\sum_i^n (X_i^2 - \bar{X}X_i - \bar{X}^2)$$

$$= \frac{1}{n-1}\left[\sum_i^n X_i^2 - \sum^n \bar{X}X_i - \sum^n \bar{X}^2\right]$$

$$= \frac{1}{n-1}\left[\sum^n X_i^2 - \bar{X}\sum^n X_i - nX^2\right]$$

Given that $\bar{X} = \frac{1}{n}\sum^n X_i$,

$$= \frac{1}{n-1}\left[\sum^n X_i^2 - n\bar{X}^2 - n\bar{X}^2\right]$$

$$= \frac{1}{n-1}\left[\sum^n X_i^2 - n\bar{X}^2\right]$$

Since the variable $X$ is Bern,
$X_i^2 = X_i$

$$= \frac{1}{n-1}\left[\sum^n X_i - n\bar{X}^2\right]$$

$$= \frac{1}{n-1}\left[n\bar{X} - n\bar{X}^2\right]$$

$$= \frac{n}{n-1}\bar{X}(1-\bar{X})$$

Shown. □

The population mean and variance of a Bernoulli random variable $X$:

$$\mu_X = \mathbb{E}[X] \qquad Var(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

$$= \mathbb{E}[X] - \mathbb{E}[X]^2$$

$$= \mathbb{E}[X](1 - \mathbb{E}[X])$$

Compare the pop. mean and variance with the sample mean and variance: don't know what they are asking me to compare...

(b) In an opinion poll of 300 voters, 140 say that they will vote for the incumbent, and 160 for the rival candidate. Estimate the proportion of votes that will be obtained by the incumbent in the election. Calculate the sample variance. Find 95% and 99% confidence intervals for the incumbent's proportion of votes in the election.

Let the proportion of votes obtained by the incumbent be $X$.

The sample mean $\bar{X} = \frac{1}{300} \sum_{i=1}^{300} X_i$

By the CLT, $\bar{X} \sim N\left(\frac{140}{300}, \sigma^2/n\right)$

The sample variance is an unbiased estimator of the pop. variance:

$$s^2 = \frac{1}{n-1} \sum_{i}^{n} (X_i - \bar{X})^2$$

In part a) we showed that for a Bernoulli var,

$$s_X^2 = \frac{n}{n-1} \bar{X}(1-\bar{X})$$

To find the estimate of the pop variance we plug in the results.

$$s^2_{X\,\text{actual}} = \frac{\cancel{300}}{299}\left(\frac{140}{\cancel{300}}\right)\left(\frac{160}{300}\right)$$

$$= \frac{140 \times 160}{299 \cdot 300}$$

$$= 0.250$$

The 95% and 99% confidence intervals are:

95% CI: $\left\{ \bar{X} + 1.96\, \sigma_{\bar{Y}} \right\}$

99% CI: $\left\{ \bar{X} + 2.58\, \sigma_{\bar{Y}} \right\}$

ESTIMATORS & THEIR ESTIMATES

$s_X^2 \rightarrow \sigma^2$, $s_X \rightarrow \sigma$

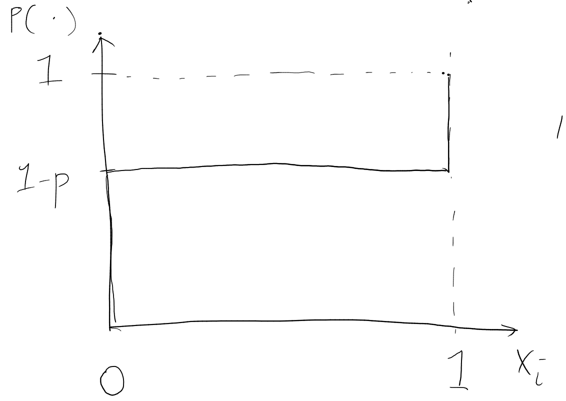
$SE(\bar{X}) \rightarrow \frac{\sigma}{\sqrt{n}}$

We know that $s_X^2$ is an unbiased estimator of $\sigma^2$, and $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$. Therefore, $s_X$ is an unbiased estimator of $\sqrt{n}(s_X)$.

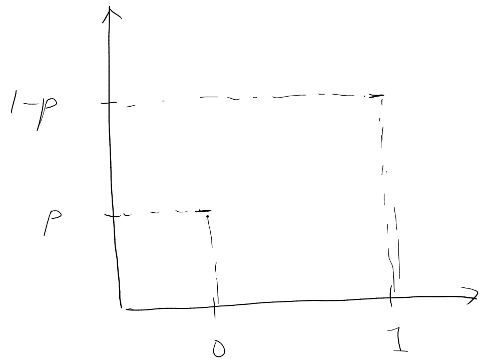
QE 2018 PYP

4. Let Xi be a Bernoulli random variable with P (Xi = 1) = p and P (Xi = 0) = 1 − p.
(a) What are the density and distribution functions of Xi ?

The CDF looks like this:




The PDF looks like this:




Not sure what else they are asking. Check tomorrow.

(b) Find the expected value, variance and skewness of Xi .

$$\mathbb{E} X_i = \sum (p_i) P(X_i = p_i)$$
$$= 0(1-p) + 1(p)$$
$$= p$$

$$Var(X_i) = \mathbb{E} X_i^2 - \mathbb{E}(X_i)^2$$
$$= \sum p_i^2 Pr(X_i = p_i) - p^2$$
$$= 0^2(1-p) + 1^2(p) - p^2$$
$$= p(1-p).$$

$$Skewness(X_i) = \frac{\mathbb{E}[(X_i - \mathbb{E} X_i)^3]}{\sigma^3_{X_i}}$$
$$=$$

how to calculate this?
what's $\sigma^3_{x_i}$ ?

$$\sigma^3_{x_i} = (\sigma_{x_i})^3 . = \sqrt{p(1-p)}^3$$

$$= \mathbb{E}\left[ X_i^3 - 3X_i^2 \mathbb{E}X_i - 3X_i(\mathbb{E}X_i)^2 - (\mathbb{E}X_i)^3 \right] \cdot \frac{1}{\sigma^3_{X_i}}$$

$$= (\mathbb{E} X_i^3 - 3p\,\mathbb{E}X_i - 3p\,\mathbb{E}X_i^2 - p^3) \cdot \frac{1}{\sigma^3_{X_i}}$$

$$= (p - \cancel{3p^2} + \cancel{3p^2} - p^3)$$

$$= p(1-p^2) \cdot \frac{1}{\sigma^3_{X_i}}$$

$$= p(1-p)(1+p) \cdot \frac{1}{\sigma^3_{X_i}}$$

$$= \left[ \frac{p^2(1-p)^2(1+p)^2}{p^3(1-p)^3} \right]^{1/2}$$

$$= \frac{(1+p)}{\sqrt{p(1-p)}}$$

Check with PPE gang or just check PYP answers on WebLearn.

For X1, ..., Xn independent, identically Bernoulli(p), let p̂ be

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

(c) What is the standard error of p̂?

$$SE(\hat{p}) = s_{\hat{p}} / \sqrt{n}$$

$$= \sqrt{\frac{n}{n-1}[(\hat{p})(1-\hat{p})]} \cdot \frac{1}{\sqrt{n}}$$

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

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

$$= \frac{1}{n-1}\left(\sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2\right)$$

$$= \frac{1}{n-1}\left(\sum X_i - 2\bar{X}\sum X_i + n\bar{X}^2\right)$$

$$= \frac{1}{n-1}\left(n\bar{X} - 2\bar{X}n\bar{X} + n\bar{X}^2\right)$$

$$= \frac{1}{n-1}\left(n\bar{X} - n\bar{X}^2\right)$$

$$= \frac{n}{n-1}\left(\bar{X}(1-\bar{X})\right)$$

(d) Explain why

$$t = \frac{\hat{p} - p}{se(\hat{p})} \to N(0,1)$$

where se(p̂) is the standard error of p̂.

By the CLT, if $n$ is large enough, $X_i$s are i.i.d, and $0 < Var(X_i) < \infty$,

$$\hat{p} \sim N(\mu_{X_i}, \sigma^2_{\hat{p}})$$

We know that $\mathbb{E}(X_i) = \mu_{X_i} = p$

and $SE(\hat{p})$ is an unbiased estimator of $\sigma_{\hat{p}}$.

standard deviation of $\hat{p}$: a number, NOT a random var.

Therefore $\frac{\hat{p} - p}{se(\hat{p})} \xrightarrow{D} N(0,1)$.

(e) Suppose that in a sample of size n = 100, we obtain $\hat{p} = 0.3$. Construct an approximate 95% confidence interval for p. State all the results being used.

Confidence interval is defined as the interval such that in 95% of samples, the population mean will fall within the CI.

$$95\% \text{ CI} = \{ \hat{p} \pm 1.96\, \sigma_{\hat{p}} \}$$

Since $\sigma_{\hat{p}}$ is unknown, we estimate it with $se(\hat{p})$.

$$= \{ \hat{p} \pm 1.96\, se(\hat{p}) \}$$

where $se(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n-1}}$

When we draw a sample of $n = 100$ and $\hat{p}_{actual} = 0.3$, the confidence interval is:

$$\Rightarrow \left\{ 0.3 \pm 1.96 \left( \sqrt{\frac{(0.3)(0.7)}{(99)}} \right) \right\}$$

$$\Rightarrow \{ 0.3 \pm 0.046 \}$$

$$\Rightarrow \{ 0.254, \; 0.346 \}$$

Check this answer

QE 2015

$Y_i = \sum_{j=1}^{i} X_j$

4. Let X j , j = 1,2, ... be a sequence of independent and identically distributed random variables with finite mean μ and variance σ 2. For i = 1,2, ... consider the random variable Y i = ∑ij=1 X j .
(a) [25%] Find E(Y i ), V(Y i ), and Cov(Y i , Y k ) for i < k stating all the properties being used.

We are given $Y_i = \sum_{j=1}^{i} X_j$ where $\mathbb{E}(X_j) = \mu$
$Var(X_j) = \sigma^2$

$$\mathbb{E}(Y_i) = \mathbb{E}\left(\sum_{j=1}^{i} X_j\right)$$

$$= \mathbb{E}X_1 + \mathbb{E}X_2 + \cdots + \mathbb{E}X_i$$

$= i\mu.$ since all $X_i$s have the same mean $\mu$.

$$Var(Y_i) = Var\left(\sum_{j=1}^{i} X_j\right)$$

$$= Var(X_1 + X_2 + \cdots + X_j)$$

$$= Var(X_1) + Var(X_2) + \cdots + Var(X_j)$$
$$+ Cov(X_1, X_2) + Cov(X_1, X_3) \cdots Cov(X_1, X_i)$$
$$+ Cov(X_{i-1}, X_i)$$

Since $X_i$s are independent of one another,
$Cov(X_a, X_b) = 0 \ \forall a, b$
$\Rightarrow i \cdot \sigma^2$

$Cov(Y_i, Y_k)$ for $i < k$:

$$Cov(Y_i, Y_k) = \mathbb{E}(Y_i Y_k) - \underbrace{\mathbb{E}(Y_i)}_{i\mu}\underbrace{\mathbb{E}(Y_k)}_{k\mu}$$

$$= \mathbb{E}(Y_i Y_k) - i \cdot k \cdot \mu^2$$

$$= \mathbb{E}\left(\sum_{j=1}^{i} X_j \cdot \sum_{j=1}^{k} X_j\right) - ik\mu^2$$

$$= \mathbb{E}(X_1 X_1 + X_1 X_2 + \cdots + X_1 X_k$$
$$\vdots$$
$$X_i X_1 + \cdots + X_i X_k) - ik\mu$$

Here $\mathbb{E}(X_a X_b)$ where $a=b$ $= \mathbb{E}(X_1^2)$

$$= Var(X_1) + \mathbb{E}(X_i)^2$$

$$= \sigma^2 + \mu^2$$

And $\mathbb{E}(X_a X_b)$ where $b \neq a$

$$= Cov(X_a, X_b) + \mathbb{E}(X_a)\mathbb{E}(X_b)$$

$$= 0 + \mu^2$$

$$= \mu^2$$

Since $i < k$, we know there are $i$ such occurrences of the first type, and $(i \times k) - i$ such occurrences of the second.

$$\begin{array}{c|l}
 & \longleftarrow \quad k \quad \longrightarrow \\
\uparrow & X_1X_1 + X_1X_2 + \cdots \cdots + X_1X_k \\
i & X_1X_2 + X_2X_2 \\
 & \quad \vdots \qquad \ddots \\
\downarrow & X_iX_1 + \cdots + X_iX_i + \cdots + X_iX_k
\end{array}$$

$$\Rightarrow (i)(\sigma^2 + \mu^2) + i(k-1)\mu^2 - ik\mu^2$$

$$\Rightarrow i(\sigma^2 + \mu^2) + i(k-2)\mu^2$$

$$\Rightarrow i\sigma^2 + i\mu^2 + i(k-2)\mu^2$$

$$\Rightarrow i\sigma^2 + i(k-1)\mu^2$$

$$\Rightarrow i(\sigma^2 + (k-1)\mu^2) \quad \square$$

[25%] State and discuss the Law of Large Numbers and the Central Limit Theorem for independent and identically distributed observations. Can these two theorems be applied to X j ? And Y i ?

The LLN states that with a large number of i.i.d. observations, you will expect the mean of those observations to approach the true population mean.

The CLT states that:

Let $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ .

If n is large enough,

$$\bar{X} \xrightarrow{d} N\left(\mu_{X_j}, \frac{\sigma^2_{X_j}}{n}\right)$$

These two theorems can be applied to $X_j$. The conditions for LLN are that the random variables must be i.i.d, and $Var(X_j) = \sigma_X^2 < \infty$, which are fulfilled.

Similarly, given that $X_j$ are i.i.d, $Y_j$ is also i.i.d.

The $Var(Y_j)$ is also finite as $i\sigma_X^2 < \infty$.

$$\sum_{j=1}^{i} a_j = i \qquad Z_i = \sum_{j=1}^{i} a_j X_j$$

(c) [25%] Consider now Zi = ∑ij=1 a j X j where a j is a sequence of real numbers satisfying ∑ij=1 a j = i and ∑ij=1 a j2 > i. Which estimator would you prefer to estimate the population mean μ: (Y n /n) or (Zn /n)? Why?

$$\sum_{j=1}^{i} a_j^2 > i$$

$$Y_i = \sum_{j=1}^{i} X_j \quad , \quad Z_i = \sum_{j=1}^{i} a_j X_j$$

Given that $\sum_{j=1}^{i} a_j = i$, this means that $\mathbb{E}(Y_i) = \mathbb{E}(Z_i)$

as $\mathbb{E}(Z_i) = \sum a_j \mathbb{E}(X_j) = \sum a_j \mu = \mu \sum a_j = i\mu$.

So the estimators $\left(\frac{Y_n}{n}\right)$ and $\frac{Z_n}{n}$ are both unbiased estimators.

However, $\frac{Y_n}{n}$ is a more <u>efficient</u> estimator.

$$Var(Y_n) = n\sigma^2 \quad , \quad Var\left(\frac{Y_n}{n}\right) = \sigma^2/n.$$

$$\begin{aligned}
Var(Z_n) &= Var(a_1 X_1 + a_2 X_2 + \dots + a_n X_n) \\
&= a_1^2 Var(X_1) + a_2^2 Var(X_2) + \dots \\
&\quad + a_1 a_2 \cancel{Cov(X_1, X_2)} + \dots = 0 \\
&\overset{i.i.d}{=} a_1^2 \sigma_x^2 + a_2^2 \sigma_x^2 + \dots + a_n^2 \sigma_x^2 \\
&= \sigma_x^2 \sum_{i=1}^{n} a_i^2 .
\end{aligned}$$

We are given $\sum_{i=1}^{n} a_i^2 > i$, so

$$> i\sigma_x^2 .$$

Hence $Var(Z_n) > Var(Y_n)$ and thus

$Var\left(\frac{Y_n}{n}\right)$ is a better estimator (more efficient) than $Var\left(\frac{X_n}{n}\right)$.

$\sigma^2 = 1$

(d) [25%] Let σ 2 = 1. A random sample is drawn and the following statistic is obtained:

$$(Y_n/n) = (Y_{300}/300) = 2.79$$

(i) Test, at the 5% level of significance, the null hypothesis that μ = 3 against the alternative that μ ≠ 3.

$H_0 : \mu = 3$

$H_1 : \mu \neq 3$.

$\frac{Y_{300}^{act}}{300} = 2.79$

$\frac{Y_n}{n} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ Simply denote this as $\bar{X}$.

$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$

By the CLT,

$\bar{X} \sim N(\mu, \sigma_{\bar{X}}^2)$ as $Var(X) = 1 < \infty$.

$\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}/\sqrt{n}}\right) \sim N(0,1)$, as $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$,

Under the null hypothesis,

$Z = \frac{\bar{X} - 3}{1/\sqrt{300}} \xrightarrow{d} N(0,1)$

Decision rule:

reject $H_0$ if $|Z| > |C_\alpha|$

where $|C_\alpha|$ satisfies $P(Z < C_\alpha) = 1 - \alpha$

Setting $\alpha = 0.05$, $C_\alpha = 1.96$.

Finally,

$Z_{act} = \frac{(2.79 - 3)\sqrt{300}}{1} = 3.64$.

Because $Z_{act} > C_\alpha$, we reject $H_0$.

(ii)
Construct and interpret a 95% confidence interval for μ. State all the relevant results being used.

The I% confidence interval is the interval whereby, in 95% of samples, the true value of the population mean $\mu$ would fall within it.

$$95\% \text{ CI} = \left\{ \bar{X} \pm 1.96\, \sigma_{\bar{X}} \right\} \qquad \text{Since } \sigma = 1, \quad \sigma^2_{\bar{X}} = \frac{1}{\sqrt{300}}$$

$$= \left\{ 2.79 \pm \frac{1.96}{\sqrt{300}} \right\}$$

$$= \left\{ 2.68,\ 2.90 \right\}$$

Because $\mu = 3$ does not fall within the 95% CI, we can reject the null hypothesis.

"State all the relevant results being used":

→ we have the CLT, because otherwise we would have no idea of the distribution of $\bar{X}$.

→ also that $\bar{X}$ is an unbiased estimator of $\mu$