Estimation

Dr. Wan Nor Arifin

Biostatistics and Research Methodology Unit
Universiti Sains Malaysia

Last update: Nov 5, 2023

Outlines

- Introduction
- Estimation
- Point Estimation
- Interval Estimation
- Calculation

Expected outcomes

- Understand the concept of estimation as one of inference methods
- Understand the concepts of point and interval estimates
- Apply the concepts to calculate point and interval estimates

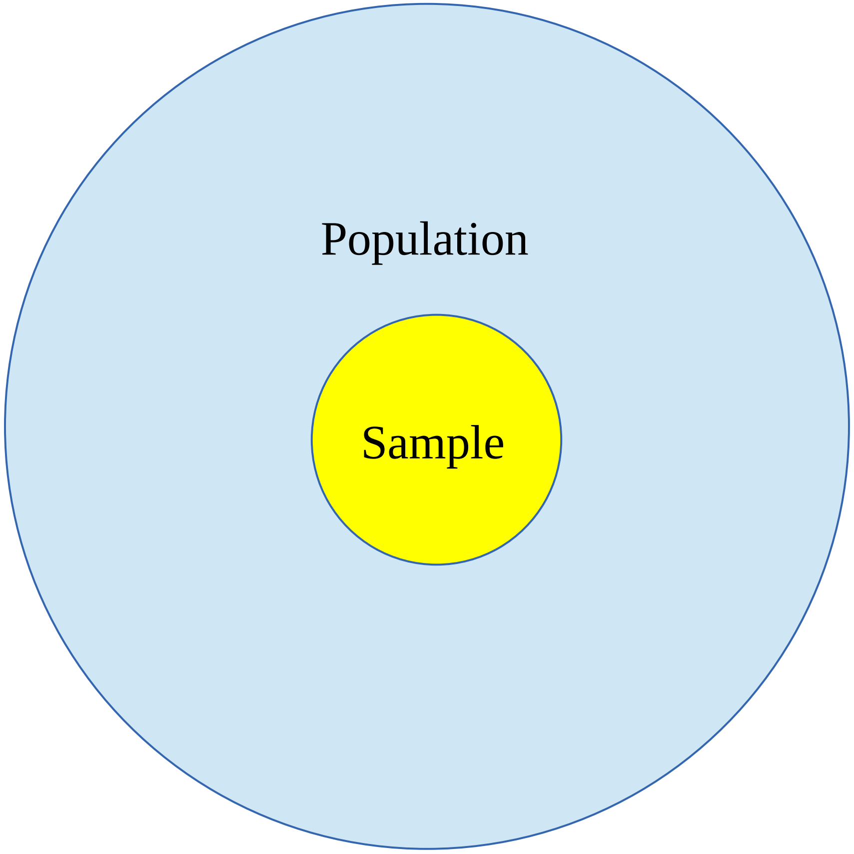
Introduction

Overview

- Statistics is a field of study dealing with (Daniel, 1995):
  1. Collection, organization, summarization and analysis of data.
  2. Making inference/conclusion about population data from sample data.

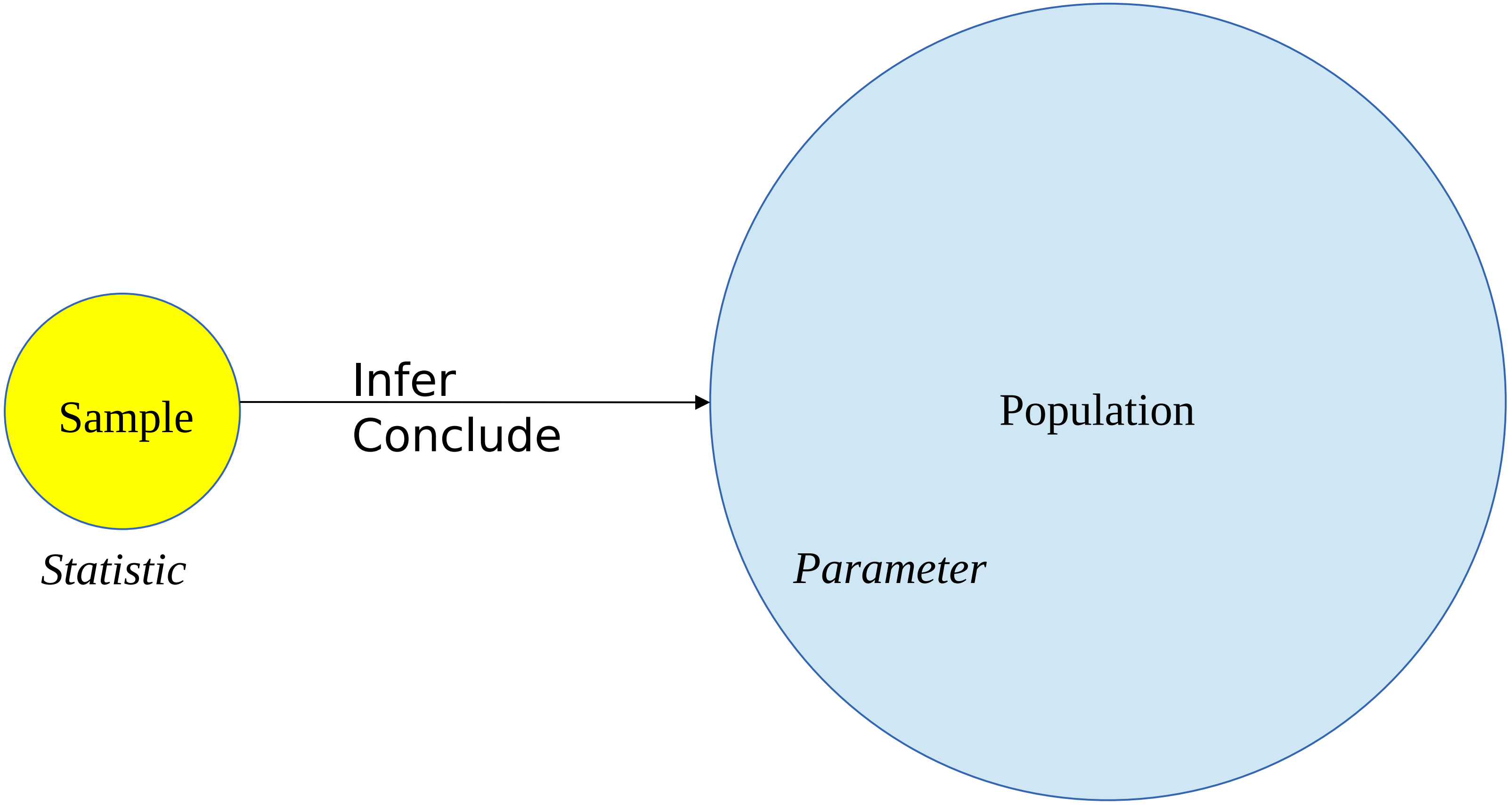
Overview

- Population *vs* sample

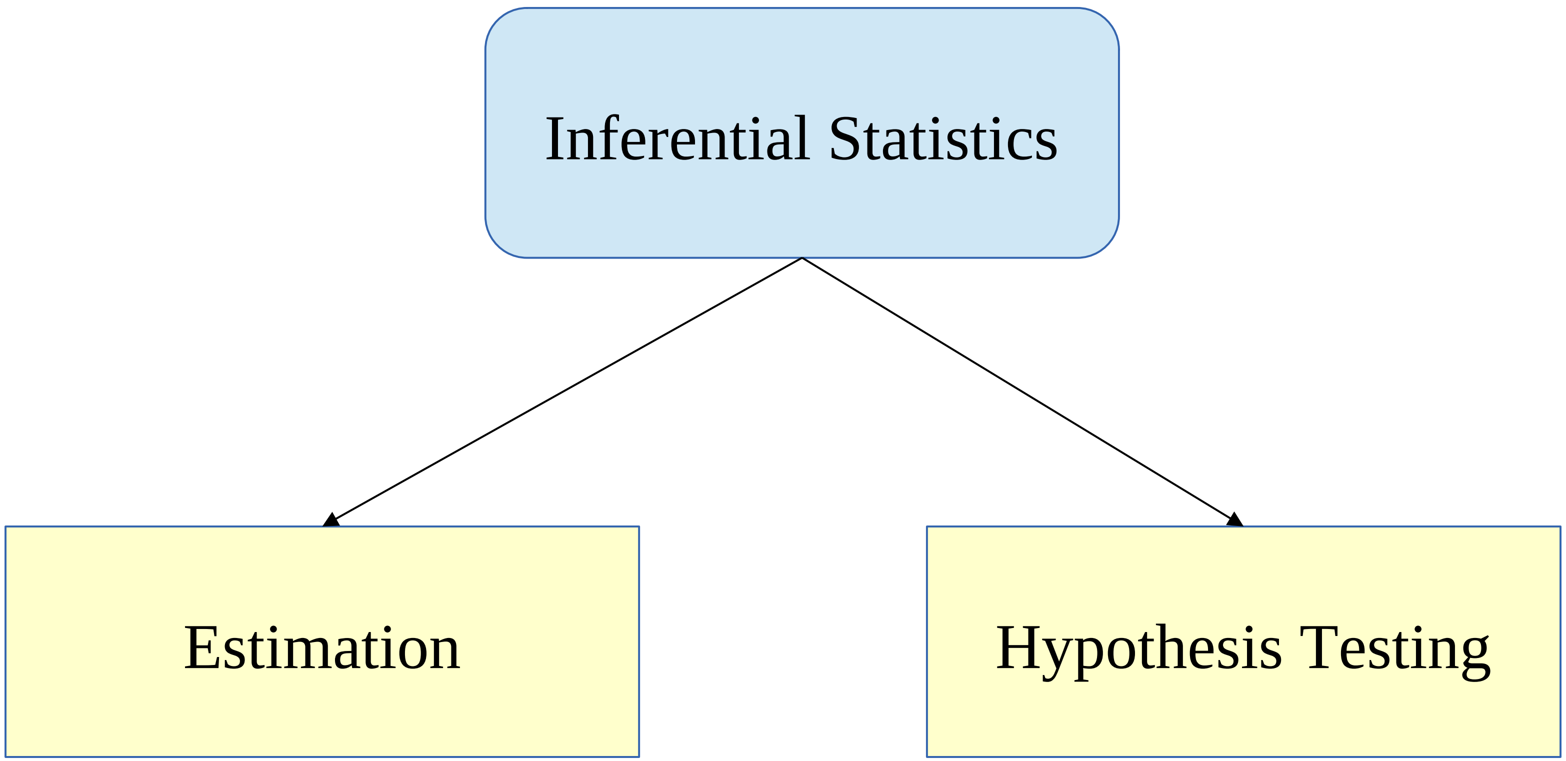


Overview

- Inference:



Overview



Estimation

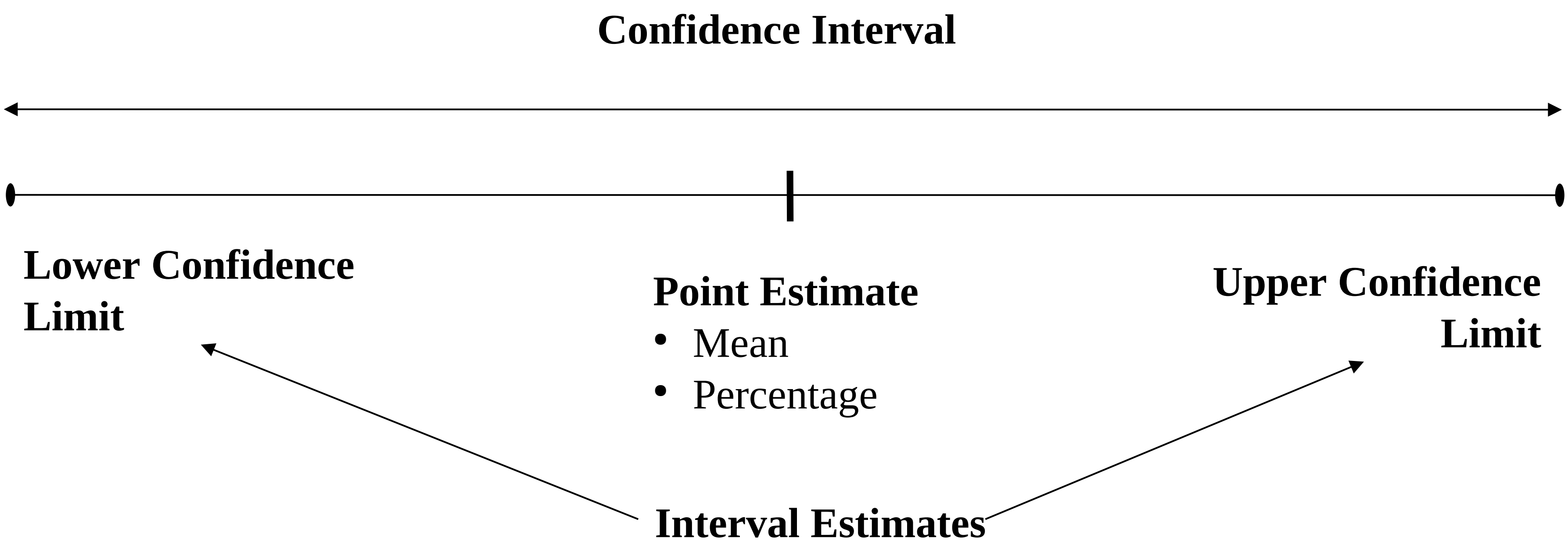
Estimation

- The process of calculating a statistic from a sample data as an approximation of a parameter of the population
- Two types of estimation:
  - Point: a single numerical value used as an estimation of a parameter value
  - Interval: two numerical values presented as range that includes the parameter value, confidence interval.

Estimation

- The process of calculating a statistic from a sample data as an approximation of a parameter of the population
- Two types of estimation:
  - Point: a single numerical value used as an estimation of a parameter value
  - Interval: two numerical values presented as range that includes the parameter value, confidence interval.

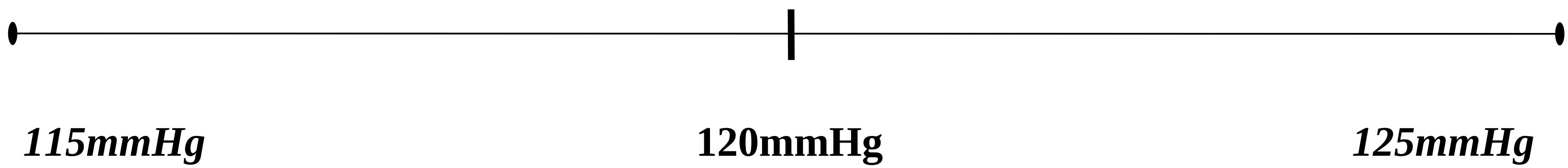
Estimation


Confidence Interval
Lower Confidence Limit
Point Estimate
• Mean
• Percentage
Upper Confidence Limit
Interval Estimates

Estimation

<u>**Mean SBP for Normal population**</u>

95% Confidence Interval

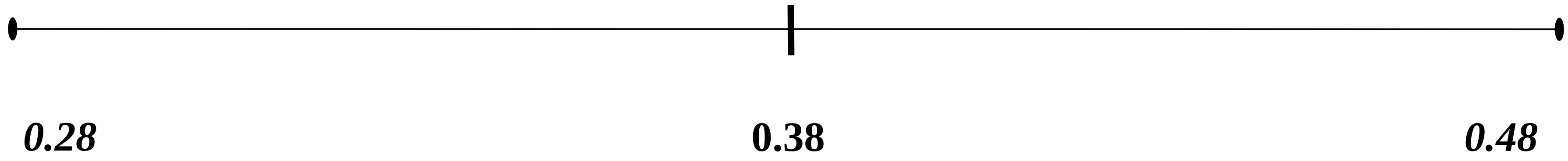


Mean SBP = 120mmHg (95% CI: 115mmHg, 125mmHg)

Estimation

<u>Proportion of Obesity among University Students' population</u>

95% Confidence Interval



Proportion of obesity = 0.38% (95% CI: 0.28, 0.48)

Point Estimation

Point Estimator

- A point estimator is a function of a sample, say

$$W(X_1, \ldots, X_n)$$

- This function is referred as a statistic
- For example, sample mean is a function

$$mean(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Point Estimate

- A point estimate is the realized value of an estimator, say

$$W(x_1, \ldots, x_n)$$

- For example, the mean of a sample is

$$mean(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Notations

$$\text{parameter} = \theta$$

$$\text{estimate} = \hat{\theta}$$

For population vs sample mean

$$\text{parameter} = \mu$$

$$\text{estimate} = \bar{x}$$

Interval Estimation

Confidence Interval

- Point estimate is not enough, should be accompanied by interval estimates – confidence interval.
- Educated guess (estimate) of the true population parameter in the form of range.
- Present as point estimate followed by its interval estimates for a given confidence level:

point estimate(% confidence level: **lower confidence limit**,upper **confidence limit**)

120mmHg (95% CI: **115mmHg**, **125mmHg**)

Confidence Interval

- Generally to obtain confidence interval:

$$\text{point estimate} \pm \text{precision}$$

$$\text{point estimate} \pm (\text{reliability coefficient}) \times (\text{standard error})$$

Confidence Interval

Confidence level	Reliability coefficient
90%	1.65
95%	1.96
99%	2.56

Interval Estimator

- An interval estimator is a pair of functions of a sample

$$[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$$

Interval Estimates

- Interval estimates are a pair of the realized values the interval estimator,

$$[L(x_1, \ldots, x_n), U(x_1, \ldots, x_n)]$$

where

$$L(x_1, \ldots, x_n) \leq \theta \leq U(x_1, \ldots, x_n)$$

Interval Estimates

- For example, the interval estimates of mean of $X \sim N(\mu, \sigma2)$

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

for $(1 - \alpha)$ confidence level

$$P_{\mu}\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Calculation

One Population Mean

- Mean of numerical data

$$\bar{x} = \frac{\sum x_i}{n}$$

- Calculate confidence interval for one sample mean using standard normal distribution, *z*

Standard normal distribution, *z*

- Data normally distributed
- Population standard deviation σ, is known OR
- σ is not known, for a large sample size (usually 30 or more), the sample standard deviation *s*, used in place of σ

Standard normal distribution, *z*

- confidence interval is given by

point estimate$\pm$(reliability coefficient)$\times$(standard error)

$$\bar{x} \pm z_{(1-\alpha/2)} \times \sigma_{\bar{x}}$$

$$\bar{x} \pm z_{(1-\alpha/2)} \times \frac{\sigma}{\sqrt{(n)}}$$

Reliability coefficient

- Commonly used reliability coefficient using *z* distribution $z_{(1-\alpha/2)}$ by $(1-\alpha) \times 100\%$

$$\alpha = 0.10,\ (1-\alpha)100\% = 90\% \rightarrow 1.65$$

$$\alpha = 0.05,\ (1-\alpha)100\% = 95\% \rightarrow 1.96$$

$$\alpha = 0.01,\ (1-\alpha)100\% = 99\% \rightarrow 2.58$$

Calculate

- From data on systolic blood pressure (SBP) collected from 30 patients, the mean SBP was 120mmHg with SD of 15mmHg. Estimate with 95% confidence of the population mean of SBP.

$$\bar{x} = 120$$
$$s = 15 \approx \sigma$$
$$n = 30$$

$$\bar{x} \pm z_{(1-\alpha/2)} \times \sigma / \sqrt{(n)}$$
$$120 \pm 1.96 \times 15 / \sqrt{(30)}$$
$$120 \pm 1.96 \times 2.739$$
$$95\%\, CI : 114.6\,, 125.4$$

mean SBP = 120mmHg (95% CI: 114.6, 125.4)

"We are 95% confident that the population mean is between 114.6 and 125.4."

Calculate

- From data on systolic blood pressure (SBP) collected from 30 patients, the mean SBP was 120mmHg with SD of 15mmHg. Estimate with 95% confidence of the population mean of SBP.

$$\bar{x} = 120$$
$$s = 15 \approx \sigma$$
$$n = 30$$

$$\bar{x} \pm z_{(1-\alpha/2)} \times \sigma / \sqrt{(n)}$$
$$120 \pm 1.96 \times 15 / \sqrt{(30)}$$
$$120 \pm 1.96 \times 2.739$$
$$95\%\, CI : 114.6\,, 125.4$$

mean SBP = 120mmHg (95% CI: 114.6, 125.4)

"We are 95% confident that the population mean is between 114.6 and 125.4."

One Population Proportion

- Common in medicine, proportion of diabetes etc.
- Mean of numerical data

$$\hat{p} = \frac{x}{n}$$

- Calculate confidence interval for one sample proportion using standard normal distribution, *z*

Standard normal distribution, *z*

- Sampling distribution of p is quite close to normal distribution when both np and $n(1-p)$ greater than 5
- E.g.

Can use z

$n = 1000$, $p = 0.05$, $1 – p = 0.95$; $np = 50$, $n(1 – p) = 950$

Cannot use z

$n = 50$, $p = 0.05$, $1 – p = 0.95$; $np = 2.5$, $n(1 – p) = 47.5$

Standard normal distribution, *z*

- confidence interval is given by

$$\text{point estimate} \pm (\text{reliability coefficient}) \times (\text{standard error})$$

$$\hat{p} \pm z_{(1-\alpha/2)} \times \sigma_{\hat{p}}$$

$$\hat{p} \pm z_{(1-\alpha/2)} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Reliability coefficient

- Same reliability coefficient using *z* distribution $z_{(1-\alpha/2)}$ by $(1-\alpha) \times 100\%$

$$90\% \rightarrow 1.65$$

$$95\% \rightarrow 1.96$$

$$99\% \rightarrow 2.58$$

Calculate

- It was found that in a study among drug addicts in Kelantan, 130 out of 200 are HIV positive. Construct 99% confidence interval for the proportion of HIV positive among the addicts.

$$\hat{p} = 130/200 = .65$$
$$n = 200$$

$$\hat{p} \pm z_{(1-\alpha/2)} \times \sqrt{\hat{p}(1-\hat{p})/n}$$
$$.65 \pm 2.58 \times \sqrt{(.65(.35)/200)}$$
$$99\%\ CI:\ 0.5623, 0.7377$$

Percentage of HIV = 65.0% (99% CI: 56.23%, 73.77%)

"We are 99% confident that the population percentage of HIV among the addicts is between 56.23% and 73.77%."