Geomatics Engineering

ENGO 361: Least Squares Estimation

Lab ONE

Multivariate Statistics, Mathematical Models, and Error Propagation

Ahmed A. Youssef

ahmed.youssef1@ucalgary.ca

UNIVERSITY OF CALGARY

SCHULICH School of Engineering

Mobile Multi-Sensors Systems Research Group

January 2018

General Introduction

- Name: Ahmed A. Youssef
- Office: Calgary Center for Innovative Technology (CCIT) – Third Floor - Room No.: 361.
- Office Hours:
  - Wednesday 12:00 AM to 02:00 PM.
  - Thursday 12:00 AM to 02:00 PM.
- E-mail: ahmed.youssef1@ucalgary.ca

Notes:

- **My office location is inaccessible. So, please send an e-mail if you are willing to drop by during office hours, or any other time, to let me know where we could meet.**
- **During the course duration, please feel free to contact me via e-mail anytime with any course-related inquiries.**

General Introduction

- **Labs Submission Date**
  - Lab due date is going to be announced whenever the lab is posted.
  - Late submissions are going to penalized. Each day delay in submission is going to result in a 10 % deduction out of the total lab grade.

- **Grading Criteria**
  - A detailed grading rubric is going to be set up for each lab, such that each step is rewarded a part of the grade.
  - So I would like to urge you to state each step you have taken to reach your solution.
  - Each step is going to be graded, so please do not state the final answer directly.

- **Submission and Feedback Method**
  - Each lab should be submitted in a Microsoft Word document of PDF file to D2L drop-box prior to the stated due date.
  - You should receive the feedback for each lab within 10 days from the submission day. The feedback is going to be in the form of comments on the submitted document.

Least Squares Estimation
Course Importance

ENGO361 – Lab ONE

Due Date: February 8th, 2018 – at 12:00 AM

LAB OVERVIEW

- **Part One**: Theory and Measures of Error
  - Problem No. 1: Basics of Univariate Statistics: Mean, Residuals, Average Error, Probable Error, Variance, Standard Deviation, Probability Density Histogram, Standard Deviation of Mean, Weighted Average, and Weighted Standard Deviation.

- **Part Two**: Concepts of Multivariate Statistics
  - Problem No. 2: Covariances and Correlation.

- **Part Three**: Mathematical Models
  - Problem No. 3: Linear vs. Non-linear Models, Direct vs. Indirect vs. Conditional Models.

- **Part Four**: Error Propagation and Pre-analysis of Survey Measurements
  - Problem No. 4: Univariate Error Propagation.
  - Problem No. 5: Multivariate Error Propagation.

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Given:**

- Distance Dataset with 50 observations → Observer A [Units: m]
- Distance Dataset with 50 observations → Observer B [Units: m]

➢ **Requirement (a):**

i. Best Estimates for Dataset A ($\overline{X_A}$), and Dataset B ($\overline{X_B}$)
ii. Residuals for Dataset A (v_A), and Dataset B (v_B)
iii. Standard Deviation for Dataset (σ_A), and Dataset (σ_B)
iv. Standard Deviation of the mean for Dataset A ($\sigma_{\bar{X}A}$), and Dataset B ($\sigma_{\bar{X}B}$)
v. Average Error for Dataset A (a_e^A), and Dataset B (a_e^B)
vi. Probable Error for Dataset A (p_e^A), and Dataset B (p_e^B)

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (a):**

i. Best Estimates for Dataset A ($\overline{X_A}$), and Dataset B ($\overline{X_B}$)

The best estimate of a one-dimensional random sample can be represented by the mean of the dataset.

The mean of the dataset can be given by the formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} l_i$$

Where:

- (l_i) is the measurement value for measurement (i).
- (n) is the total number of measurements in the dataset.
- $i = 1,2,3,\ldots\ldots,n$

For the given datasets: $(n = 50)$

- Dataset A: $\bar{x}_A = \frac{1}{50}(153.0 + 152.9 + 149.0 + \cdots + 151.6)$
- Dataset B: $\bar{x}_B = \frac{1}{50}(149.6 + 151.4 + 149.0 + \cdots + 150.0)$

Matlab Syntax:

```
xBar = mean(l); %calculates the mean of input vector [l].
```

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (a):**

ii. Residuals for Dataset A (v_A), and Dataset B (v_B)

The residuals of any given dataset is defined by the deviation of each measurement within the dataset from the best estimate (mean).

The residuals of a measurement within a dataset can be given by the formula:

$$v_i = \bar{x} - l_i$$

Where:

- (v_i) is the residual of measurement (i).
- ($\bar{x}$) is the best estimate/mean of the measurements.
- (l_i) is the measurement value for measurement (i).

For the given datasets: ($n = 50$)

- Dataset A: $v_i^A = \bar{x}_A - l_i^A$
- Dataset B: $v_i^B = \bar{x}_B - l_i^B$

Matlab Syntax:

```
v = mean(l)-l; %calculates the residuals of input vector [l].
```

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (a):**

iii. Standard Deviation for Dataset (σ_A), and Dataset (σ_B)

The standard deviation of any given dataset is a measure of data variability, or data scattering about the mean. It is considered as a measure of data precision.

The standard deviation of a given dataset is given by the formula:

$$\sigma = \pm \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} v_i^2}$$

Where:

- (v_i) is the residual of measurement (i).
- (n) is the total number of measurements in the dataset.
- $i = 1,2,3, \ldots\ldots, n$

Matlab Syntax:

```
sigma = std(l); %calculates the standard deviation of input vector [l].
```

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (a):**

iii. Standard Deviation for Dataset (σ_A), and Dataset (σ_B)

CONTINUED

For the given datasets: ($n = 50$)

- Dataset A:

$$\sigma_A = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(v_i^A\right)^2} = \sqrt{\frac{1}{50-1}((\bar{x}_A - 153.0)^2 + (\bar{x}_A - 152.9)^2 + \cdots + (\bar{x}_A - 151.6)^2)}$$

Where, ($\bar{x}_A$) is the mean value of the dataset A, as calculated from requirement (a.i)

- Dataset B:

$$\sigma_B = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(v_i^B\right)^2} = \sqrt{\frac{1}{50-1}((\bar{x}_B - 149.6)^2 + (\bar{x}_B - 151.4)^2 + \cdots + (\bar{x}_B - 150.0)^2)}$$

Where, ($\bar{x}_B$) is the mean value of the dataset B, as calculated from requirement (a.i)

Matlab Syntax:

```
sigma = std(l); %calculates the standard deviation of input vector [l].
```

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (a):**

iv. Standard Deviation of the mean for Dataset ($\sigma_{\bar{X}A}$), and Dataset ($\sigma_{\bar{X}B}$)

The standard deviation of the mean any given dataset is an estimate of the precision of the best estimate, which is derived from the standard deviation of any given dataset. Sometimes, it is referred to as "**Standard Error of the Mean**".

The standard deviation of the mean for any given dataset is given by the formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where:

- (σ) is the standard deviation of the dataset.
- (n) is the total number of measurements in the dataset.

For the given datasets: ($n = 50$)

- Dataset A: $\sigma_{\bar{X}A} = \sigma_A/\sqrt{50}$
- Dataset B: $\sigma_{\bar{X}B} = \sigma_B/\sqrt{50}$

Where, (σ_A), and (σ_B) are the standard deviations of the datasets A and B respectively, as calculated from previous requirement (a.i)

Matlab Syntax:

```
sigmaBar = std(l)/sqrt(length(l));
%calculates the standard deviation of the mean for input vector [l].
```

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (a):**

v. Average Error for Dataset A (a_e^A), and Dataset B (a_e^B)

Sometimes, it is referred to as "**Mean Absolute Deviation (MAD)**". The average error is the mean/average of the absolute values of the residuals. The average error can be considered as a measure of precision, or data scatter from the mean.

The average error for any given dataset is given by the formula:

$$a_e = \frac{1}{n-1}\sum_{i=1}^{n}|v_i|$$

Where:

- ($|v_i|$) is the absolute residual of measurement (i).
- (n) is the total number of measurements in the dataset.

Matlab Syntax:

```
ae = sum(abs(mean(l)-l)) / (length(l)-1);
%calculates the average error of input vector [l].
```

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (a):**

v. Average Error for Dataset A (a_e^A), and Dataset B (a_e^B)

CONTINUED

For the given datasets: ($n = 50$)

- Dataset A:

$$a_e^A = \frac{1}{50-1}\sum_{i=1}^{n}\left|v_i^A\right| = \frac{1}{50-1}\left(|\bar{x}_A - 153.0| + |\bar{x}_A - 152.9| + \cdots + |\bar{x}_A - 151.6|\right)$$

Where, ($\bar{x}_A$) is the mean value of the dataset A, as calculated from requirement (a.i)

- Dataset B:

$$a_e^B = \frac{1}{50-1}\sum_{i=1}^{n}\left|v_i^B\right| = \frac{1}{50-1}\left(|\bar{x}_A - 149.6| + |\bar{x}_A - 151.4| + \cdots + |\bar{x}_A - 150.0|\right)$$

Where, ($\bar{x}_B$) is the mean value of the dataset B, as calculated from requirement (a.i)

Matlab Syntax:

```
ae = sum(abs(mean(l)-l)) / (length(l)-1);
%calculates the average error of input vector [x].
```

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (a):**

vi. Probable Error for Dataset A (p_e^A), and Dataset B (p_e^B)

The probable error can be defined as the median of the absolute values of the residuals of a given dataset. In other words, the probable error is the residual whose index is midway between the maximum and minimum absolute values of residuals.

How to calculate the probable error of a given dataset:

Step 1: Calculate the residuals for the given dataset (v_i).
Step 2: Get the absolute values of the calculated residuals ($|v_i|$).
Step 3: Sort the absolute values in ascending or descending order.
Step 4: Find the residual whose index is in the middle of the sorted residuals, as follows, let n be the number of residuals:

If n is an odd number: $p_e = v_{\left(\frac{n+1}{2}\right)}$

If n is an even number: $p_e = \frac{1}{2}\left(v_{\left(\frac{n}{2}\right)} + v_{\left(\frac{n}{2}+1\right)}\right)$

Matlab Syntax:

```
ysort = sort(abs(mean(l)-l),'ascend');
y = 0.5 * (ysort(length(l)/2) + ysort((length(l)/2)+1));
%the probable error of input vector [l] of even indices.
```

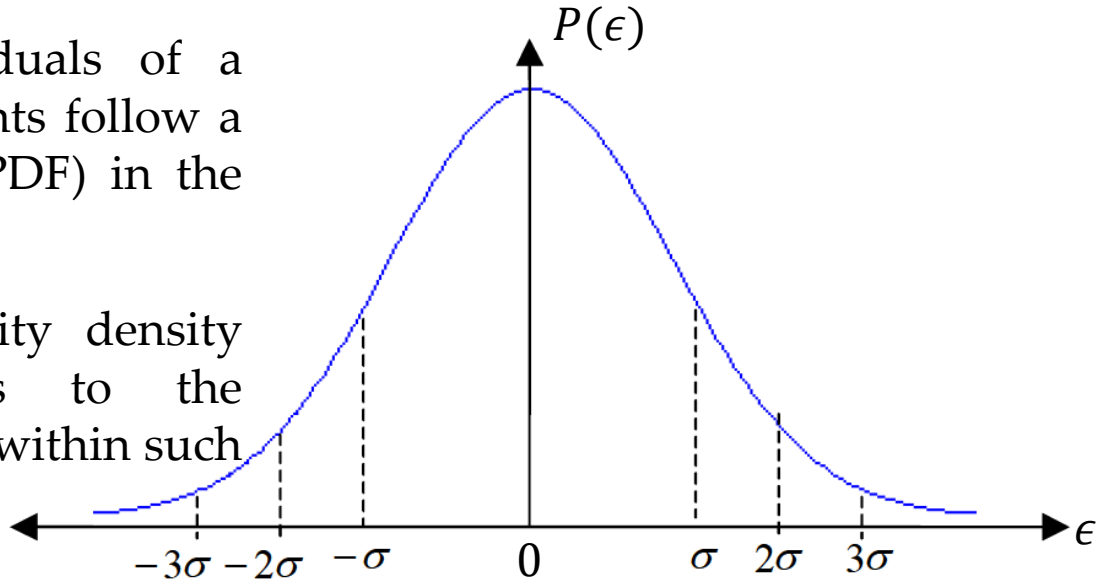
Part ONE: Theory and Measures of Error
Problem ONE

- **Requirement (b):**
  i. Best Estimate of the Dataset ($\overline{X_A}$), and Dataset B ($\overline{X_B}$) at level of confidence 95%
  ii. Standard Deviation of Dataset A (σ_A), and Dataset (σ_B) at level of confidence 95 %

- **<u>Solution for Requirement (b):</u>**

It is assumed that the residuals of a random sample of measurements follow a probability density function (PDF) in the form of a gaussian distribution.

The area under the probability density function curve corresponds to the probability of having residuals within such range values.



Part ONE: Theory and Measures of Error
Problem ONE

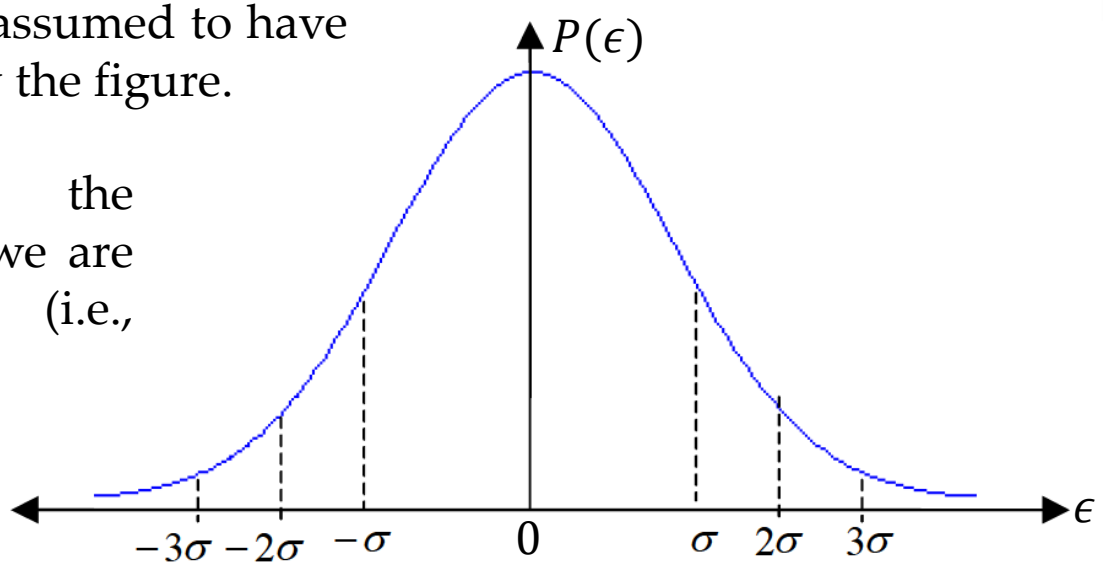
➢ **Solution for Requirement (b) – CONTINUED:**

The ratio between the area under the PDF curve, to the total area under the curve (which is equal to one) is referred to as the level of confidence, such that:

$$P(-\sigma \leq \epsilon \leq \sigma) = 0.683 \equiv 68.3\%$$
$$P(-2\sigma \leq \epsilon \leq 2\sigma) = 0.954 \equiv 95.4\%$$
$$P(-3\sigma \leq \epsilon \leq 3\sigma) = 0.997 \equiv 99.7\%$$

The residuals, by definition, is assumed to have a zero mean, which is shown by the figure.

Therefore, regardless, of the confidence interval, in which we are operating, the best estimate (i.e., mean) should remain the same.



Part ONE: Theory and Measures of Error
Problem ONE

➢ **Solution for Requirement (b) – CONTINUED:**

i. Best Estimate of the Dataset ($\overline{X_A}$), and Dataset B ($\overline{X_B}$) at level of confidence 95%

As discussed in the previous two slides, it can be concluded that:

$\overline{X_A}$ → is the same as the value calculated in requirement (a.i)

$\overline{X_B}$ → is the same as the value calculated in requirement (a.i)

ii. Standard Deviation of Dataset A (σ_A), and Dataset (σ_B) at level of confidence 95 %

As discussed in the previous two slides, it can be concluded that:

The standard deviations for datasets A, and B, as calculated in requirement (a.iii), represent a level of confidence at 68.2 %.

Therefore:

$$\sigma_A\Big|_{@95\%} = 2\sigma_A\Big|_{@68.2\%}$$

$$\sigma_B\Big|_{@95\%} = 2\sigma_B\Big|_{@68.2\%}$$

Part ONE: Theory and Measures of Error
Problem ONE

- **Requirement (c):**
  i. Plot Measurement Number vs Measurement Values for Datasets A, and B.
  ii. Plot Measurement Number vs Measurement Residuals for Datasets A, and B.
  iii. Plot Probability Density Histogram (PDH) for Residuals of Datasets A, and B.

- **<u>Solution for Requirement (c):</u>**

Using Matlab Plot Function:

i. Plot I:
   - For Dataset A: (Number of Measurements → on X-axis, Measurement Values → on Y-axis)

ii. On same Plot I:
   - For Dataset B: (Number of Measurements → on X-axis, Measurement Values → on Y-axis)

Matlab Syntax:

```
Figure; plot(A); hold on; plot(B); grid on;
xlabel('No. of Measurement'); ylabel('Distance (m)');
%Plot of Measurements Values.
```

➢ **Solution for Requirement (c) - CONTINUED:**

Using Matlab Plot Function:

i. Plot II:
   - For Dataset A: (Number of Measurements → on X-axis, Residual Values→ on Y-axis)

ii. On same Plot II:
   - For Dataset B: (Number of Measurements → on X-axis, Residual Values → on Y-axis)

Matlab Syntax:

```
Figure; plot(mean(A)-A); hold on; plot(mean(B)-B); grid on;
xlabel('No. of Measurement'); ylabel('Residuals (m)');
%Plot of Measurements Values.
```

➢ **Solution for Requirement (c) - CONTINUED:**

iii. Probability Distribution Histogram (PDH) for Residuals:

- Probability Distribution Histogram (PDH) is a bar chart that represents the relative frequency of residuals occurrence in a data set.

- As stated earlier, we assume that the residuals, ultimately represent a random variable, that follow a probability density function (PDF) in the form of a gaussian distribution.

- The PDH provides the proof, whether the residuals follow a gaussian distribution.

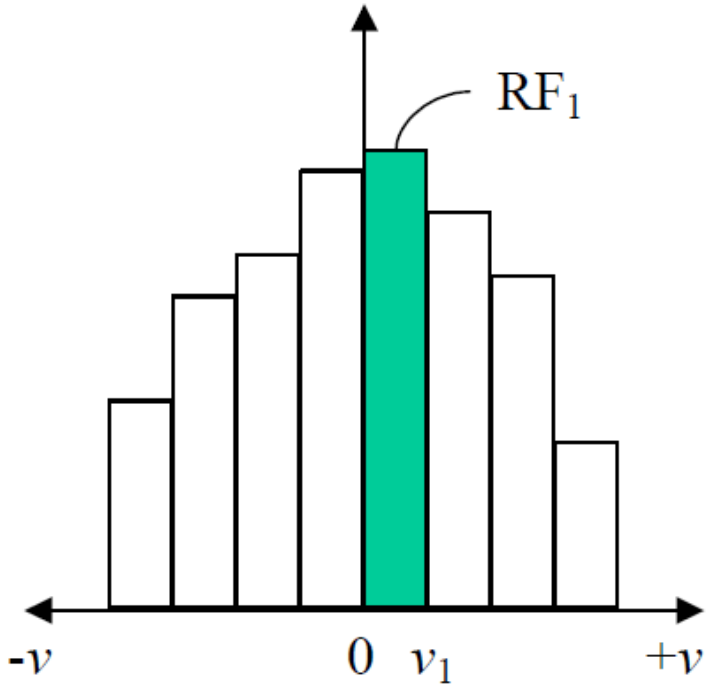
- The PDH is the discrete form of the continuous PDF for a random sample, which is the practical case for deriving a probability density representation of real measurements, which are discrete in nature.

- Remember that the area under the PDH and/or PDF is the probability of such residual value.

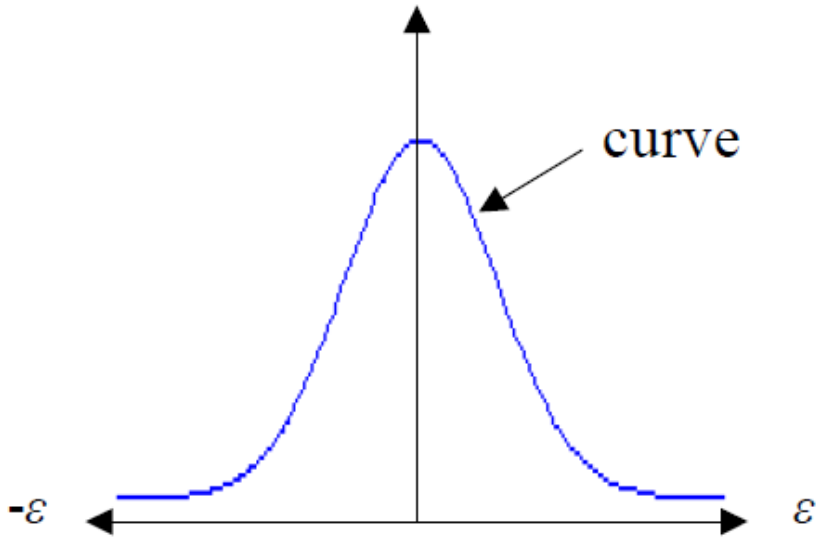
➢ **Solution for Requirement (c) - CONTINUED:**

iii. Probability Distribution Histogram (PDH) for Residuals:

- Example of PDH



- Example of PDF



➢ **Solution for Requirement (c) - CONTINUED:**

iii. Probability Distribution Histogram (PDH) for Residuals:

- You can use the Matlab function to plot the PDH for the residuals of both Datasets A, and B. However, one needs to know how the Matlab function works. The process of Matlab function to plot histograms is as follows:
  1. Calculate residuals for each dataset (v_i^A), and (v_i^B).
  2. Sort the computed residuals ascendingly.
  3. Calculate the range of sorted residuals, such that:

$$Range = argmax(v_i) - argmin(v_i)$$

  4. Divide the range of residuals by number of classes ($k = 10$), such that:

$$\Delta_j = \frac{Range}{k} = \frac{Range}{10}$$

  5. Calculate each class limits, where:

$$class_1 = argmin(v_i) + \Delta_j$$

$$class_2 = class_1 + \Delta_j$$

$$\vdots$$

$$class_n = class_{n-1} + \Delta_j$$

➢ **Solution for Requirement (c) - CONTINUED:**

iii. Probability Distribution Histogram (PDH) for Residuals:

6. Count the number of residuals (n_j) lying within each class (j).
7. Calculate the probability/relative frequency of each class.

$$f_j = \frac{n_j}{n.\Delta_j}$$

(f_j) is the relative frequency, (n_j) is the number of residuals within class, (Δ_j) is the class width

(n) is the total number of measurements/residuals

The relative frequency is calculated such that the area of each bar represents the probability of such class.

8. Plot the bar chart, where bars are placed at each class mid-values along the X-axis versus relative frequency on the Y-axis.

Matlab Syntax:

```
Figure; histogram((mean(x)-x),n,'Normalization','pdf'); grid on;
xlabel('Distances (m)'); ylabel('Relative Frequency');
%Plot of PDH for residuals of vector [x] with number of bins (n).
```

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (d):**

Compare the precision of both observers on the basis of average error, probable error, and standard deviation. Do all conclusions agree? If not, comment on the differences.

➢ **<u>Solution for Requirement (d):</u>**

In your solution, try to answer these questions:

- Based on the three measures of precision, which is more precise out of the two datasets (A) and (B).
- Does the conclusion change from one measure to the other?
- Which measure is more conservative, and reliable in case you are going to report a precision for the measurements?
- Which of them can be interpreted and used mathematically in PDFs?

Any additional insights and conclusions will be great.

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Requirement (d):**

Weighted mean and Standard Deviation of Weighted Mean from the provided datasets (A), and (B).

➢ **Solution for Requirement (d):**

Step 1: Calculate the standard deviation of the mean ($\sigma_{\bar{X}A}$) for dataset (A)

→ the same as calculated for the dataset A in requirement (a), no. (iv)

Step 2: Calculate the standard deviation of the mean ($\sigma_{\bar{X}B}$) for dataset (B)

→ the same as calculated for the dataset A in requirement (a), no. (iv)

Step 3: Calculate the weights of each datasets; or probabilities of each dataset (P_A, P_B)

$$P_A = \frac{1}{\sigma_{\bar{x}A}^2} \qquad P_B = \frac{1}{\sigma_{\bar{x}B}^2}$$

You should understand that it is intuitive that the probability/weight is inversely proportional to the precision (i.e., standard deviation). The intuition is that if the standard deviation of a measurement is high, it means that it is less precise, then the confidence in that measurement is low, or the weight of that measurement is less.

Part ONE: Theory and Measures of Error
Problem ONE

➢ **Solution for Requirement (d) - CONTINUED:**

Step 4: Calculate the weighted mean ($\bar{x}_{WM}$)

$$\bar{x}_{WM} = \frac{P_A \bar{x}_A + P_B \bar{x}_B}{P_A + P_B}$$

Step 5: Calculate the standard deviation of the weighed mean (σ_{WM})

$$\sigma_{WM}^2 = \frac{1}{(n-1)} \left[\frac{P_A \sum_{i=1}^{n} \left(v_i^A\right)^2 + P_B \sum_{i=1}^{n} \left(v_i^B\right)^2}{P_A + P_B} \right]$$

$$\sigma_{WM} = \pm \sqrt{\sigma_{WM}^2}$$

Clarification on concept of weighting: We can write the formula for weighted mean as follows:

$$\bar{x}_{WM} = \frac{P_A}{P_A + P_B} \bar{x}_A + \frac{P_B}{P_A + P_B} \bar{x}_B$$

Where $\left(\frac{P_A}{P_A+P_B}\right)$ and $\left(\frac{P_B}{P_A+P_B}\right)$ represent the weights or probabilities of means from datasets (A), and (B) respectively. Then, the dataset with higher probability will be much closer to the weighted mean than the other.

Additionally, let's assume that ($P_A = P_B$), then by substituting in the formula, we would get:

$$\bar{x}_{WM} = \frac{P_A}{P_A + P_B} \bar{x}_A + \frac{P_B}{P_A + P_B} \bar{x}_B = \frac{1}{2} \bar{x}_A + \frac{1}{2} \bar{x}_B = \frac{\bar{x}_A + \bar{x}_B}{2} = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i$$

Which is the original formula of the mean, then we can infer that the original mean formula in slide (7), is actually the same as the weighted mean formula; however, with equal weights for each measurement.

Part TWO: Concepts of Multivariate Statistics
Problem TWO

➢ **Given:**

- $n_{players} = 10$
- Weights (Units: in kg.) → $w = [73,77,71, \ldots, 85]$
- Heights (Units: in m.) → $h = [1.80,1.70,1.86, \ldots, 1.79]$
- Speed (Units: in m/sec.) → $s = [7.2,7.2,8.0, \ldots, 5.9]$
- No. of Goals (Unitless) → $g = [13,7,79, \ldots, 6]$

Let N be an array of multivariate observations, and can be expressed as:

$$N = [w, h, s, g]$$

➢ **Requirement (a):**

i. Best Estimates for each set within the multivariate set of observations: $\bar{X}_N = [\bar{x}_w, \bar{x}_h, \bar{x}_s, \bar{x}_g]$.

ii. Standard Deviation for each set within the multivariate set of observations: $\sigma_N = [\sigma_w, \sigma_h, \sigma_s, \sigma_g]$.

Part TWO: Concepts of Multivariate Statistics
Problem TWO

➢ **Solution for Requirement (a):**

i. Best Estimates for each set within the multivariate set of observations: $\bar{X}_N = [\bar{x}_w, \bar{x}_h, \bar{x}_s, \bar{x}_g]$.

The best estimate of a each sample of data within the provided multivariate dataset can be represented by the mean of the dataset. For which we can use the same formula used earlier in problem ONE.

$$\bar{x}_w = \frac{1}{n}\sum_{i=1}^{n} w_i \qquad \bar{x}_h = \frac{1}{m}\sum_{j=1}^{m} h_j \qquad \bar{x}_s = \frac{1}{l}\sum_{k=1}^{l} s_k \qquad \bar{x}_g = \frac{1}{u}\sum_{s=1}^{u} g_s$$

Where:

- (w_i, h_j, s_k, g_s) is the measurement value for each measurement for each data type.
- $(n = m = l = u = 10)$ is the total number of measurements in the dataset.

$$\bar{X}_N = [\bar{x}_w, \bar{x}_h, \bar{x}_s, \bar{x}_g]$$

Matlab Syntax:

```
xBar = mean(l); %calculates the mean of input vector [l].
```

Part TWO: Concepts of Multivariate Statistics
Problem TWO

➢ **Solution for Requirement (a):**

ii. Standard Deviation for each set within the multivariate set of observations: $\sigma_N = [\sigma_w, \sigma_h, \sigma_s, \sigma_g]$.

The standard deviation of a each sample of data within the provided multivariate dataset can be represented by the same formula used earlier in problem ONE.

$$\because n = m = l = u = 10$$

$$\sigma_w = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}_w - w_i)^2 \qquad \sigma_h = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}_h - h_i)^2$$

$$\sigma_s = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}_s - w_s)^2 \qquad \sigma_g = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}_g - w_g)^2$$

$$\sigma_N = [\sigma_w, \sigma_h, \sigma_s, \sigma_g]$$

Matlab Syntax:

```
sigma = std(l); %calculates the standard deviation of
                input vector [l].
```

Part TWO: Concepts of Multivariate Statistics
Problem TWO

- **Requirement (b):**

  The covariance matrix (C_N) for the given multivariate dataset.

- **Solution for Requirement (b):**

The covariance matrix (C_N) for the given multivariate dataset is a matrix that represents the correlation between the provided set of measurements. Even though the measurements might not related mathematically or physically, they can be correlated statistically.

Simple illustrative example:
The correlation between the number of hours spent studying, and the acquired score in an exam.

The structure of the covariance matrix (C_N) for the dataset N, can be given as follows

$$C_N = \begin{bmatrix} \sigma_w^2 & \sigma_{wh} & \sigma_{ws} & \sigma_{wg} \\ \sigma_{hw} & \sigma_h^2 & \sigma_{hs} & \sigma_{hg} \\ \sigma_{sw} & \sigma_{sh} & \sigma_s^2 & \sigma_{sg} \\ \sigma_{gw} & \sigma_{gh} & \sigma_{gs} & \sigma_g^2 \end{bmatrix}$$

Where, (σ_{ij}) is the covariance between (i) and (i) observations. For example, ($\sigma_{hw} = \sigma_{wh}$) is the covariance between the weight (w) and height (h) measurements.
The diagonal elements represent the variances (σ_i^2) of each of type of measurements in the dataset.

Part TWO: Concepts of Multivariate Statistics
Problem TWO

➢ Solution for Requirement (b) - CONTINUED:

The covariance between two elements in covariance matrix can be calculated in a similar way to the variance. For example, the covariance between the weight data, and the height data can be calculated using the following formula:

$$\sigma_{wh} = \sigma_{hw} = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}_w - w_i)(\bar{x}_h - h_i) \qquad \text{(Units: kg.m)}$$

Similarly, you can calculate the rest of the covariances to build the covariance matrix.

Notes:
1. It is important to state the units of each element in the covariance matrix.
2. Check if the covariance matrix is correct, if:
   i. Matrix is Symmetric.
   ii. Has positive diagonal elements.
   iii. Non-singular and Invertible.

Matlab Syntax:

```
x = [w h s g]; % Matrix [x] with rows as observations and columns
               % as variables.
y = cov(x); % calculates the covariance matrix of input
            % multivariate matrix [x].
```

Part TWO: Concepts of Multivariate Statistics
Problem TWO

➢ **Requirement (c):**

The correlation matrix (ρ_N) for the given multivariate dataset.

➢ **<u>Solution for Requirement (c):</u>**

- The correlation matrix (C_N) for the given multivariate dataset is a matrix that represents the normalization for the covariance matrix.
- The normalization of any quantity can be calculating the ratio its value with respect to the full scale range of that value.
- Correlation matrices usually are derived to standardize, and have the ability to compare and assess the correlation between variables within a multivariate dataset.

The structure of the correlation matrix (ρ_N) for the dataset N, can be given as follows

$$\rho_N = \begin{bmatrix} \rho_{ww} & \rho_{wh} & \rho_{ws} & \rho_{wg} \\ \rho_{hw} & \rho_{hh} & \rho_{hs} & \rho_{hg} \\ \rho_{sw} & \rho_{sh} & \rho_{ss} & \rho_{sg} \\ \rho_{gw} & \rho_{gh} & \rho_{gs} & \rho_{gg} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{wh} & \rho_{ws} & \rho_{wg} \\ \rho_{hw} & 1 & \rho_{hs} & \rho_{hg} \\ \rho_{sw} & \rho_{sh} & 1 & \rho_{sg} \\ \rho_{gw} & \rho_{gh} & \rho_{gs} & 1 \end{bmatrix}$$

Where, (ρ_{ij}) is the correlation between (i) and (i) observations. For example, ($\sigma_{hw} = \sigma_{wh}$) is the covariance between the weight (w) and height (h) measurements.

The diagonal elements represent the correlation coefficients (ρ_{ii}) of each of type of measurements in the dataset with itself, and it is always equal to **ONE**.

Part TWO: Concepts of Multivariate Statistics
Problem TWO

➢ Solution for Requirement (c) - CONTINUED:

- The correlation between two variables within a multivariate dataset can be calculates using the following formula

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

- For example, the correlation coefficient between the weight and height measurements can be calculated as follows:

$$\rho_{wh} = \frac{\sigma_{wh}}{\sigma_w \sigma_h}$$

Similarly, you can calculate the rest of the covariances to build the covariance matrix.

Notes:
1. What do you think the units of the correlation coefficients are? Please, state your answer in the lab report.
2. Similar to the covariance matrix, the matrix is correct, if:
    i. Matrix is Symmetric.
    ii. Has one values along its diagonal elements.
    iii. Since the correlation is a normalization matrix, all the values within the matrix must lie between [-1,1]

Part TWO: Concepts of Multivariate Statistics
Problem TWO

➢ Solution for Requirement (c) - CONTINUED:

- The interpretation of the correlation matrix is the most important part of this requirement, because it is the whole purpose of deriving the correlation matrix.
- It is preferable to express your interpretations in a tabular form, as the following example:

Matrix Element Indices	Correlation Coefficients Values	Discussion
$\rho_N(1,1)$, $\rho_N(2,2)$, $\rho_N(3,3)$, and $\rho_N(4,4)$	$\rho_{ww} = \rho_{hh} = \rho_{ss} = \rho_{gg} = 1$	The correlation between any variable and itself is equal to (1). Hence, there is a complete positive correlation. This is an intuitive conclusion.
⋮	⋮	⋮
⋮	⋮	⋮
$\rho_N(3,4)$, and $\rho_N(4,3)$	$\rho_{sg} = \rho_{gs} = \boldsymbol{VALUE}$	**YOUR DISCUSSION**

Part TWO: Concepts of Multivariate Statistics
Problem TWO

➢ Solution for Requirement (c) - CONTINUED:

- Your discussion should include:

i. State the sign of each correlation coefficient
ii. State your interpretation of the correlation sign
iii. State if the correlation is strong, moderate, or weak.
iv. State what is meant by a strong, moderate, or weak correlation, and state the relation between the correlation.

Recall from the Lecture Notes:

1. Strength of Correlation Definition.

$$0 < |\rho_{ab}| \leq 0.35 \rightarrow \text{Weak Correlation}$$

$$0.35 < |\rho_{ab}| \leq 0.75 \rightarrow \text{Significant Correlation}$$

$$0.75 < |\rho_{ab}| \leq 1.0 \rightarrow \text{Strong Correlation}$$

2. Geometrical Interpretation of the Covariance and Correlation.

Matlab Syntax:

```
x = [w h s g]; % Matrix [x] with rows as observations and columns
               % as variables.
y = corrcoef(x); % calculates the correlation matrix of input
                 % multivariate matrix [x].
```

Part THREE: Mathematical Models
Problem THREE

- **Given: Leveling Network**
  - Stations A, B are known [**Constants**]
  - Stations C, D are unknowns [**Unknowns**]
  - For all points are measured [**Observations**]

Note: The direction of the arrows represent the direction of terrain/land.

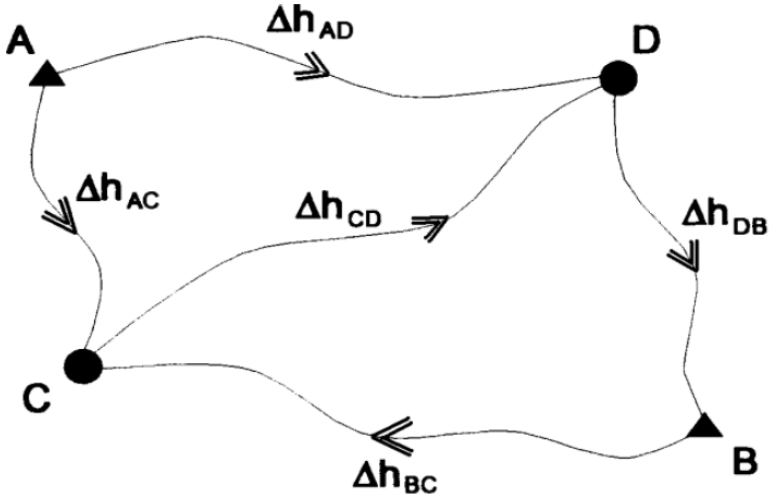


Figure 1. Levelling Network

- **Requirement (a):**

Determine whether or not it is possible to formulate the following math models to solve for the elevations of C and D.

i. Direct Model
ii. Indirect Model
iii. Conditional Model

- **Requirement (b):**

Form the math models in matrix form and state whether each model is linear or non-linear.

Part THREE: Mathematical Models
Problem THREE

➢ **Solution for Requirement (a) and (b):**

- The mathematical models, in the context of least squares estimation, are crucial, as we usually need to estimate parameters which are dependent on the measurements that we are making.

Simple illustrative example:
The calculation of the area (A) of a rectangular shape from the measurements of its length (L) and width (W). Hence, $A = L.W$. This can be considered as a mathematical model, in which we are trying to estimate the unknown parameter (i.e., area) for the observations/measurements (i.e., length and width)

- The mathematical models needs to be well-defined within the estimation process.

- Mathematical models representing the relation between the measurements and unknowns can be expressed in more than one formulation.

- Each formulation of the mathematical models has its own solution procedure, which is going to be discussed throughout the course.

➢ Solution for Requirement (a) and (b) - CONTINUED:

i. Direct Model

The direct model is a model the expresses the unknowns as a function of the measurements and constants.

$$Unknowns = f(Measurements, Constants)$$

Illustrative Example: Area of rectangle(A), of dimensions (L,W). The model is a direct model, and is written as: $A = L.W$

For the given problem:

Let the number of the unknowns be (u), number of measurements be (n), and the number of equations be (m).

$$\therefore u = 2, n = 5$$

First, can we represent the model as a direct model?

YES

What is the number of the equations, that we can write to represent the relation between the unknowns and measurements?

Since we write the unknowns as a function of the measurements, then $m = u = 2$

➢ **Solution for Requirement (a) and (b) - CONTINUED:**

i. Direct Model

One formulation of the direct model can be written as follows:

$$H_C = H_A + \Delta h_{AC}$$
$$H_D = H_B + \Delta h_{DB}$$

It is clear that the model is a linear one.

In matrix form, it can be written as follows:

$$\begin{bmatrix} H_C \\ H_D \end{bmatrix}_{2\times 1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \end{bmatrix}_{2\times 5} \begin{bmatrix} \Delta h_{AC} \\ \Delta h_{AD} \\ \Delta h_{BC} \\ \Delta h_{DB} \\ \Delta h_{CD} \end{bmatrix}_{5\times 1} + \begin{bmatrix} H_A \\ H_B \end{bmatrix}$$

➢ Solution for Requirement (a) and (b) - CONTINUED:

ii. Indirect Model

The indirect model is a model the expresses the measurements as a function of the unknowns and constants.

$$Measurements = f(Unknowns, Constants)$$

For the given problem:

Let the number of the unknowns be (u), number of measurements be (n), and the number of equations be (m).

$$\therefore u = 2, n = 5$$

First, can we represent the model as an indirect model?

YES

What is the number of the equations, that we can write to represent the relation between the unknowns and measurements?

Since we write the measurements as a function of the unkowns, then $m = n = 5$

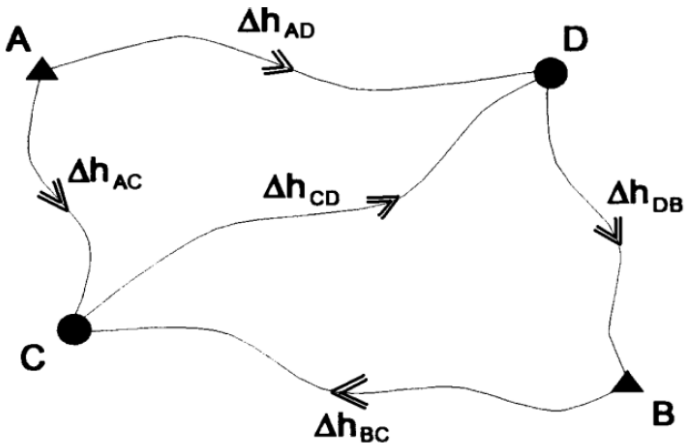
➢ Solution for Requirement (a) and (b) - CONTINUED:

ii. Indirect Model

One formulation of the indirect model can be written as follows:

$$\Delta h_{AC} = H_c - H_A$$
$$\Delta h_{AD} = H_D - H_A$$
$$\Delta h_{CD} = H_D - H_C$$
$$\Delta h_{BC} = H_c - H_B$$
$$\Delta h_{DB} = H_B - H_D$$

It is clear that the model is a linear one.



In matrix form, it can be written as follows:

$$\begin{bmatrix} \Delta h_{AC} \\ \Delta h_{AD} \\ \Delta h_{BC} \\ \Delta h_{DB} \\ \Delta h_{CD} \end{bmatrix}_{5\times1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix}_{2\times5} \begin{bmatrix} H_C \\ H_D \end{bmatrix}_{2\times1} + \begin{bmatrix} H_A \\ H_B \end{bmatrix}$$

➢ Solution for Requirement (a) and (b) - CONTINUED:

iii. Conditional Model

The conditional model is a model the expresses a function of the measurements, and constants that is equal to zero.

$$f(Unknowns, Constants) = 0$$

Illustrative Example: Sum of internal angles of a triangle (α, β, γ). The model is a conditional model, and is written as: $\alpha + \beta + \gamma - 180 = 0$.

For the given problem:

Let the number of the unknowns be (u), number of measurements be (n), and the number of equations be (m).

$$\therefore u = 2, n = 5$$

First, can we represent the model as an indirect model?

YES

What is the number of the equations, that we can write to represent the relation between the unknowns and measurements?

For conditional model, the number of equations is the maximum number of independent conditions that we can write. Then, $m = n - u = 3$.

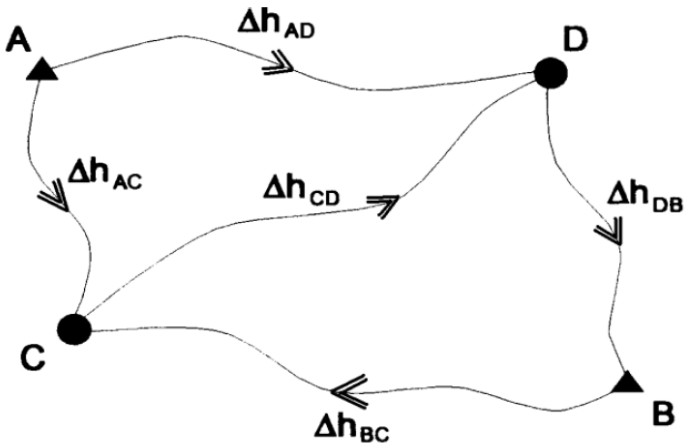
➢ Solution for Requirement (a) and (b) - CONTINUED:

ii. Conditional Model

One formulation of the conditional model can be written as follows:

$$\Delta h_{AC} + \Delta h_{CD} - \Delta h_{AD} = 0$$
$$\Delta h_{CD} + \Delta h_{DB} + \Delta h_{BC} = 0$$
$$\Delta h_{AD} + \Delta h_{DB} + \Delta h_{BC} - \Delta h_{AC} = 0$$
$$\Delta h_{AC} + \Delta h_{AD} - H_A + H_B = 0$$
$$\Delta h_{AD} + \Delta h_{DB} - H_B + H_A = 0$$

} Dependent Conditions



In matrix form, it can be written as follows:

$$\begin{bmatrix}0\\0\\0\end{bmatrix}_{3\times1} = \begin{bmatrix}1 & -1 & 0 & 0 & 1\\ 0 & 0 & 1 & 1 & 1\\ -1 & 1 & 1 & 1 & 0\end{bmatrix}_{3\times5} \begin{bmatrix}\Delta h_{AC}\\ \Delta h_{AD}\\ \Delta h_{BC}\\ \Delta h_{DB}\\ \Delta h_{CD}\end{bmatrix}_{5\times1}$$

It is clear that the model is a linear one.

Part THREE: Mathematical Models
Problem THREE

➢ **Requirement (c):**

Assuming a unique formulation, determine the number of possible solutions for the elevations of each of C and D, and list all the possible equations.

➢ **<u>Solution for Requirement (c):</u>**

- What are the maximum number of direct model equations by which H_C and H_D can be determined.

$$H_C = H_A + \Delta h_{AC}$$

$$H_C = H_A + \Delta h_{AD} - \Delta h_{CD}$$

$$H_C = H_B + \Delta h_{BC}$$

$$\vdots$$

Complete the rest similarly

Part FOUR: Error Propagation
Problem FOUR

- **Given: Race Track**
  - $L = 1607.25 \pm 6cm$
  - $W = 1227.67 \pm 7cm$

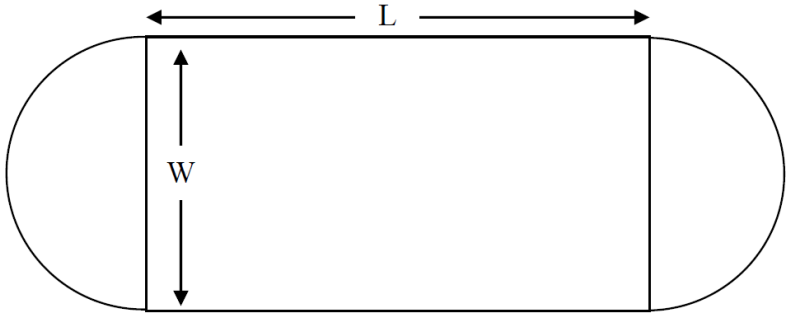


Figure: 2

- **Requirement (a):** The area enclosed by the track (A)
- **Solution for Requirement (a):**

$$A = A_{rectangle} + 2A_{semicircle} = L.W + \frac{\pi w^2}{4}$$

- **Requirement (b):** The length of the track (P)
- **Solution for Requirement (b):**

$$P = 2L + 2\pi W$$

Part FOUR: Error Propagation
Problem FOUR

- **Requirement (c):** The standard deviation of perimeter of track (A)
- **Solution for Requirement (c):**

Law of propagation of variances (or error propagation)

- Since errors are inevitable within the measurements, the computed unknown parameters are affected by such errors through the mathematical model.

- The objective of the law of propagation of variances is to calculate the errors/variances that would propagate from the measurements to the unknown parameters through the mathematical model.

$$\sigma_P^2 = \left(\frac{\partial P}{\partial L}\right)^2 \sigma_L^2 + \left(\frac{\partial P}{\partial W}\right)^2 \sigma_W^2$$

$$P = 2L + 2\pi W$$

Remember:
Partial derivative is derivative of function w.r.t one variable, and assuming the other variables constant

$$\frac{\partial P}{\partial L} = 2 \qquad \frac{\partial P}{\partial W} = 2\pi$$

Part FOUR: Error Propagation
Problem FOUR

- **Requirement (d):** The standard deviation of perimeter of track (A)
- **Solution for Requirement (d):**

From law of propagation of variances (or error propagation)

$$\sigma_A^2 = \left(\frac{\partial A}{\partial L}\right)^2 \sigma_L^2 + \left(\frac{\partial A}{\partial W}\right)^2 \sigma_W^2$$

$$A = L.W + \frac{\pi w^2}{4}$$

Remember:
Partial derivative is derivative of function w.r.t one variable, and assuming the other variables constant

$$\frac{\partial A}{\partial L} = W \qquad \frac{\partial A}{\partial W} = L + \frac{\pi W}{2}$$

Part FOUR: Error Propagation
Problem FIVE

➢ **Given:**

- **Constants (Error Free)**

$X_A = 220m$, $Y_A = 610m$, $\alpha = 55°$

- **Measurements**

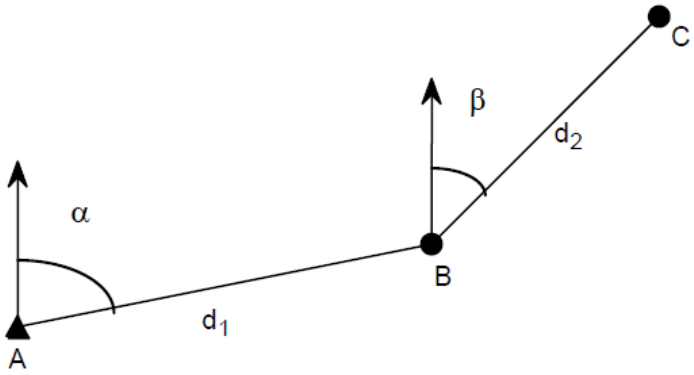
$d_1 = 760.0\ m$, $\sigma_{d1} = 4\ cm$

$d_2 = 801.0\ m$, $\sigma_{d2} = 2.5\ cm$

$\beta = 45°$, $\sigma_\beta = 5" = (\frac{5}{3600})° = \left(\frac{5}{3600} \times \frac{\pi}{180°}\right)$

- **Unknowns**

(X_B, Y_B) and (X_c, Y_c)



➢ **Requirements:** Statistical Estimates of (X_B, Y_B) and (X_c, Y_c)

i. Best Estimates.
ii. Standard Deviations.
iii. Covariance and Correlation Matrix.

Part FOUR: Error Propagation Problem FIVE

➢ **Solution for Requirements:**

i. Best Estimates of (X_B, Y_B) and (X_c, Y_c)

$$\begin{aligned} X_B &= X_A + d_1 sin\alpha \\ Y_B &= Y_A + d_1 cos\alpha \\ X_C &= X_A + d_1 sin\alpha + d_2 sin\beta \\ Y_C &= Y_A + d_1 cos\alpha + d_2 cos\beta \end{aligned} \longrightarrow \begin{bmatrix} X_B \\ Y_B \\ X_C \\ Y_C \end{bmatrix} = \begin{bmatrix} d_1 sin\alpha \\ d_1 cos\alpha \\ d_1 sin\alpha + d_2 sin\beta \\ d_1 cos\alpha + d_2 cos\beta \end{bmatrix} + \begin{bmatrix} X_A \\ Y_A \\ X_A \\ Y_A \end{bmatrix}$$

Substitute with given values into the model to get the best estimates

ii. Covariance and Correlation

Step One: Construct a covariance matrix for measurements (C_l)

Typically, we assume no statistical correlation between the measurements.

$$C_l = \begin{bmatrix} \sigma_{d_1}^2 & \sigma_{d_2 d_1} & \sigma_{d_1 \alpha} \\ \sigma_{d_1 d_2} & \sigma_{d_2}^2 & \sigma_{d_2 \alpha} \\ \sigma_{\alpha d_1} & \sigma_{\alpha d_2} & \sigma_{\alpha}^2 \end{bmatrix} = \begin{bmatrix} (0.04m)^2 & 0 & 0 \\ 0 & (0.025m)^2 & 0 \\ 0 & 0 & \left(\frac{5}{3600} \times \frac{\pi}{180°}\right)^2 \end{bmatrix}$$

It is EXTREMELY important to mind the units, when solving problems with error propagation, due to the interaction between the partial derivatives and the standard deviations.

➢ Solution for Requirements - CONTINUED:

ii. Covariance and Correlation

Step Two: Apply the law of variances propagation for multivariate statistics to calculate the covariance of unknowns ($C_{\bar{x}}$).

$$C_{\bar{x}} = J.C_l.J^T$$

Where, (J) is the Jacobian Matrix (a matrix of partial derivatives w.r.t. measurements), size of (J) is (4×3)

$$J = \begin{bmatrix} \left(\frac{\partial X_B}{\partial d_1}\right) & \left(\frac{\partial X_B}{\partial d_2}\right) & \left(\frac{\partial X_B}{\partial \alpha}\right) \\ \left(\frac{\partial Y_B}{\partial d_1}\right) & \left(\frac{\partial Y_B}{\partial d_2}\right) & \left(\frac{\partial Y_B}{\partial \alpha}\right) \\ \left(\frac{\partial X_C}{\partial d_1}\right) & \left(\frac{\partial X_C}{\partial d_2}\right) & \left(\frac{\partial X_C}{\partial \alpha}\right) \\ \left(\frac{\partial Y_C}{\partial d_1}\right) & \left(\frac{\partial Y_C}{\partial d_2}\right) & \left(\frac{\partial Y_C}{\partial \alpha}\right) \end{bmatrix} = \begin{bmatrix} sin\alpha & 0 & d_1 cos\alpha \\ cos\alpha & 0 & -d_1 sin\alpha \\ sin\alpha & sin\beta & d_1 cos\alpha \\ cos\alpha & cos\beta & -d_1 sin\alpha \end{bmatrix}$$

Substitute with given values into the above equations to get the covariance matrix

Step Three: Calculate the correlation matrix of the unknowns ($\rho_{\bar{x}}$) from the covariance matrix ($C_{\bar{x}}$).

$\rho_{\bar{x}}$ ⇒ Same as explained in Part II

Part FOUR: Error Propagation
Problem FIVE

➢ **Solution for Requirements - CONTINUED:**

iii. Standard Deviation

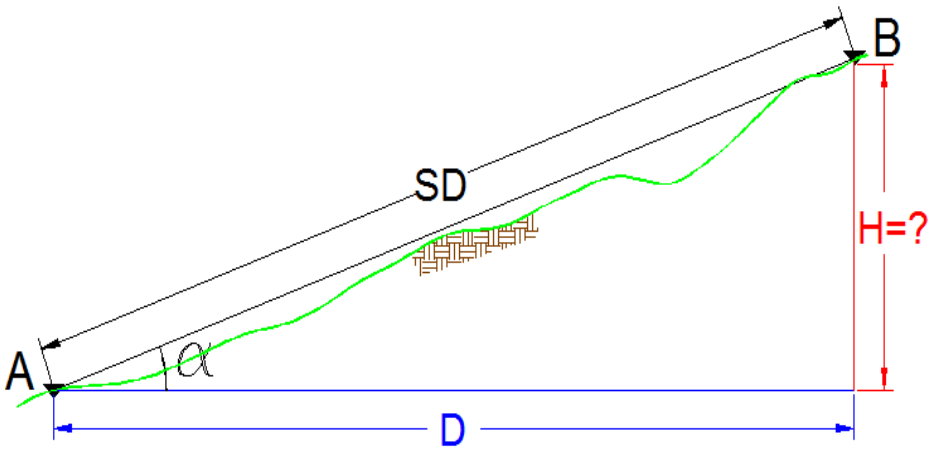
Calculate the square root of the diagonal elements of the covariance matrix of the unknowns ($C_{\bar{x}}$), such that:

$$\sigma_{X_B} = \sqrt{C_{\bar{x}}(1,1)} \qquad \sigma_{Y_B} = \sqrt{C_{\bar{x}}(1,1)} \qquad \sigma_{X_C} = \sqrt{C_{\bar{x}}(1,1)} \qquad \sigma_{Y_C} = \sqrt{C_{\bar{x}}(1,1)}$$

ADDITIONAL MATERIALS

Leveling Concept

- Say I want to determine the difference in height between two points on ground surface.
- In order to grasp the concept of leveling imagine two points A and B lying on earth's surface where there is an inclination in natural ground surface as shown.
- How to determine the height difference between the two points A and B



➢ **Alternative A:**

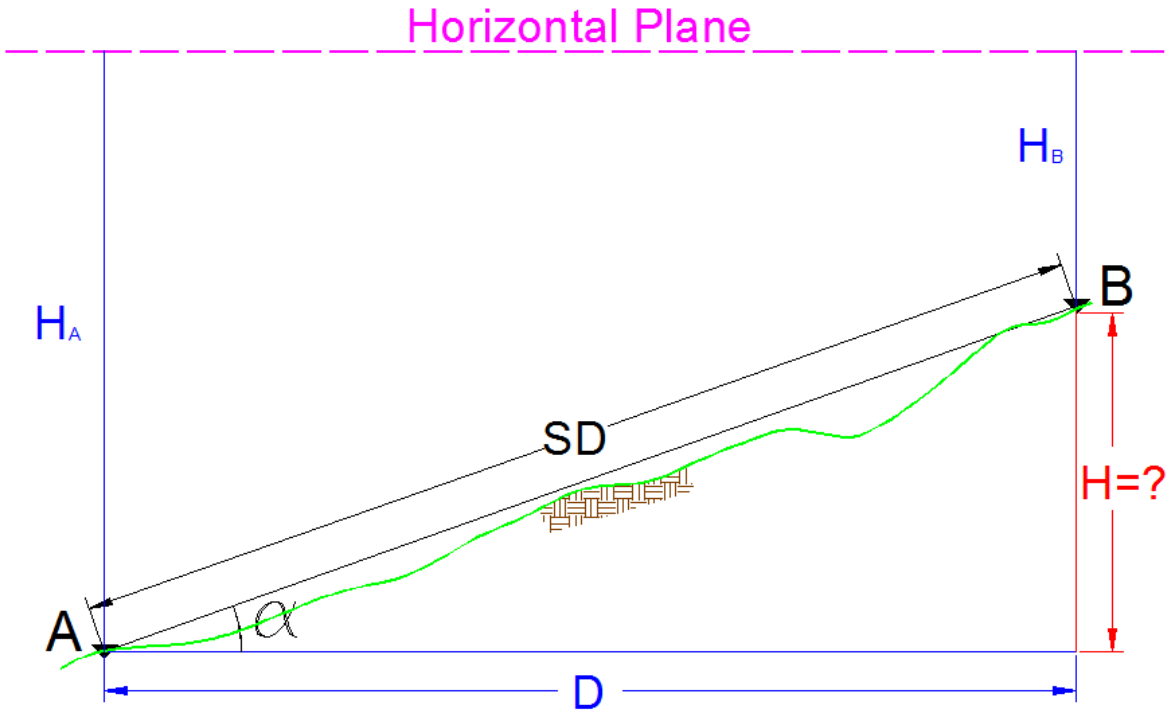
- Determine the angle of elevation α, and sloped distance SD.

∴Vertical Height Difference (H) can be given as:

$$H = SD.\tan(\alpha)$$

Leveling Concept

➢ **Alternative B:**
If we measure the height of point A relative to any horizontal surface H_A, and then measure the height of point B relative to the same surface H_B, as shown.



- Since all points in a horizontal plane have the same height.

- Therefore, the height difference between the two points can be given as:

$$H = H_A - H_B$$

Leveling Concept

➢**Alternative B is preferred more than Alternative A because as mentioned before measuring angles is not as simple as measuring distances. Also, measuring angles induces more errors in observed heights because error in angle increases with distance.**

- In order to apply **Alternative B,** we need to provide:

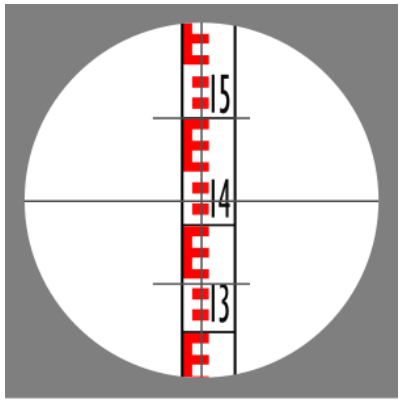
i. A Measurement Tool to measure the heights from points to the horizontal surface.
ii. A Horizontal Surface in order to measure the heights with respect to it
iii. A Magnifying Tool in order to measure those heights from distance

- **Leveling Rod**

It can be considered as the **Measurement Tool** that is used to measure the heights from points to the horizontal surface.

It is usually aluminum bar with centimeters graduation, same as a ruler but with special characteristics, that is placed on points in order to determine their height.

Leveling rod graduation is special in its nature, as shown.

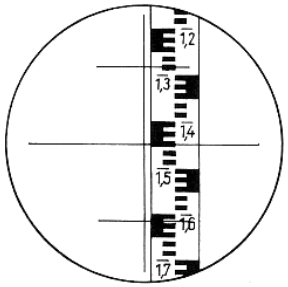


- **Level**

It can be considered as the piece of equipment that provides the **Horizontal Surface** that heights are measured to, and the **Magnifying Tool** that helps reading heights on leveling rods from distance.

It consists mainly from a **spirit level** and **foot screws** to adjust the spirit level, and these create a virtual horizontal plane to which heights are measured and is known as **Collimation Plane.**

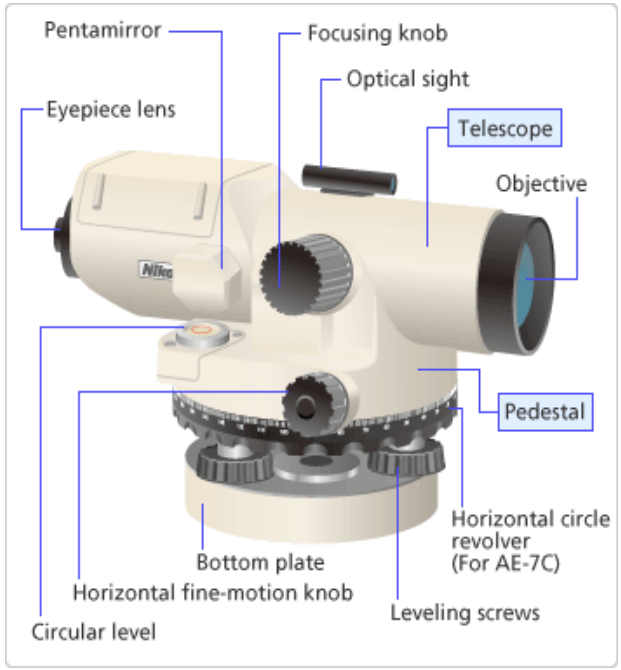
The heights are taken using crosshairs that are seen when looking through the eyepiece lens of the telescope, as shown.

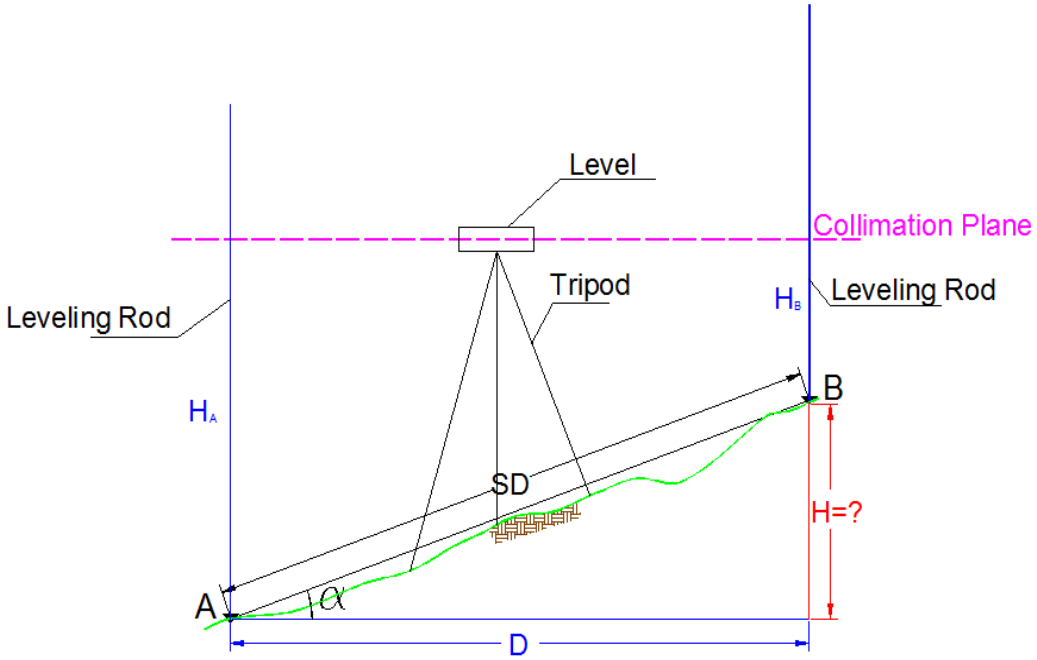


Leveling Equipment

- **Level**

Main Components





Finally, the concept is summed up with the equipment to provide a reliable method to determine height differences between points on ground surface.

Optical Axis: Line joining the centers of eyepiece and objective lens of the telescope, it is also known as line of sight and it must be always horizontal for any level.

Leveling Concept

- **Illustrative Example**