CIV3204 – Engineering Investigation

Table of Contents

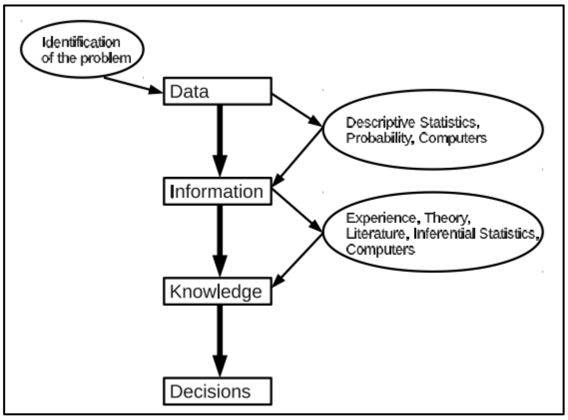
1. Basics

1.1 Definitions

- Statistics = collect/organise/analyse/interpret data
- System = no. of components logically (or physically) linked together for some purpose
- Process = set of activities operating a system that transforms inputs to outputs
- Data = Specific observations of measured numbers
- Information = processed & summarised data yielding facts/ideas
- Knowledge = Selected & organised information forming the basis for decisions

1.2 Statistics

- Decision Making Process


Identification of the problem
Data
Descriptive Statistics, Probability, Computers
Information
Experience, Theory, Literature, Inferential Statistics, Computers
Knowledge
Decisions

1 = Get Data
2 = Find information based on data
3 = Apply theory to info to get knowledge
4 = Make decisions

- Importance of Statistics
  - Statistics are seen in climate data (ave. temp trends), sport (betting odds), traffic data (road deaths
  - Statistics allows us to get data, process it & make a decision.
  - Interpreting this data could involve skewness, sampling error, conditional probability etc.

  ∴ we must take all this into consideration before arriving at a conclusion

- Risk & Uncertainty
  - Each scenario (decisions & conclusions drawn from statistics) typifies some level of risk & uncertainty
  - As engineers, we must quantify & define an acceptable level of risk, to make informed decisions & recommendations **(critical for assessing liability)**

- Statistics & Probability
  - Statistical concepts provide a means to make informed decisions in the presence of risk & uncertainty
  - Statistics are rooted in probability theory, which is the process of quantifying uncertainty, randomness, or fluctuation involved in a stochastic process

  - Why is an understanding of statistics important for Civil Engineers?

    > Mission of a Civil Engineer: ”Design, build, operate &/or improve the physical system & products.”
    > Independent assessment of what makes things ”tick”
    > Evaluate new processes
    > Forecast/predict effects of system changes
    > Knowing when to ”call the plumber”

    ∴ an engineer must use his knowledge to critically examine the work of others i.e “understand how quantitative methods are used to assist in the investigation and analysis of engineering problems”

2.3 Sampling

- Sampling Issues

  Why sample from an entire population?
  - Impracticality of recording data for the entire population i.e polls before elections.
  - Rare events
  - Futility of testing
  - Accuracy of testing: Becomes more difficult to achieve if too many objects need to be tested.
  - Randomisation is critical

- Types of Survey Errors

  - Coverage error = part of population excluded from sample
  - Non response error = part of sample may not respond
  - Sampling error = Conclusions differ from sample to sample
  - Measurement error = wrong measurements lead to wrong conclusions

- Sampling error & Sampling Bias

  Sampling error = error when sample is used to estimate population parameter
  - Function of sample size & variability
  - Can only be minimized

  Sample bias = error by bad sampling
  - Error associated w/ instrument drift, incorrect sampling frame, non-response & choice of q's
    - Can be eliminated through optimization/planning

- Statistical Methods

  Descriptive statistic = collecting & describing data
  1. Collect data (i.e survey)
  2. Present data (i.e tables/graphs)
  3. Characterise data (i.e through sample mean)

  Inferential statistics = draw conclusion/make decision concerning a population based on sample results

  Estimation: I.e estimate population mean weight using sample mean weight

  Hypothesis testing: I.e test the claim that the population mean weight is 70kg

7. Sampling Distributions & Interval Estimates

7.1 Sampling Distributions

- Why study Sampling Distributions?

  - Sample statistics are used to estimate the population parameters

  - The problem is that different samples provide different estimates

  > Large samples give better estimates but cost more

  > So we want to know how good the estimate is

  i.e how can we get a feeling of the uncertainty in the mean we've calculated

  - Want to approach the solution to the question in a theoretically consistent way

  ∴ use the sampling distribution

- Definitions

  - Sample distribution = distribution of a sample statistic, obtained using all possible different samples of the same size as opposed to population dist. = distribution of a population characteristic

  - Sample statistic = random variable

  i.e sample mean, sample standard deviation

- Developing Sampling Distributions:

  E.g.

  - Assume a population of people w/ size = 4

  > Random variable X = age

  > Values of X = 18, 20, 22 & 24

- The mean (μ):

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{18+20+22+24}{4} = 21$$

∴ distribution of ages is uniform

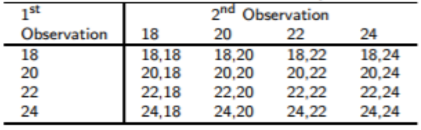
- Standard Deviation (σ):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}} = \sqrt{\frac{(-3)^2+(-1)^2+1^2+3^2}{4}} = \sqrt{\frac{20}{4}} = 2.236$$

- Now, take **WITH REPLACEMENT** all possible samples of size n=2

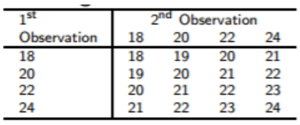
(takes all possible pairs of two, can ask same person twice)

> This means that 16 samples are taken with replacement



1st Observation	2nd Observation 18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

> Leading to 16 different means



1st Observation	2nd Observation 18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

> From the 16 means:

There is one 18, one 24, two 19's, two 23's, three 20's, three 22's & four 21's

- Mean of the sampling distribution ($\mu_{\bar{X}}$):

$$\begin{aligned}\mu_{\overline{X}} &= \frac{\sum_{i=1}^{N}\overline{X}_i}{N}\\ &= \frac{18+19+19+20+20+20+21+21+21+21+22+22+22+23+23+24}{16}\\ &= 21\end{aligned}$$

- Variance of the sampling distribution ($\sigma_{\bar{X}}$):

$$\begin{aligned}\sigma_{\overline{X}} &= \sqrt{\frac{\sum_{i=1}^{N}(\overline{X}_i-\mu_{\overline{X}})^2}{N}}\\ &= \sqrt{\frac{(18-21)^2+2(19-21)^2+3(20-21)^2+4(21-21)^2+3(22-21)^2+2(23-21)^2+(24-21)^2}{16}}\\ &= \sqrt{\frac{9+2\cdot4+3+3+2\cdot4+9}{16}}\\ &= \sqrt{\frac{40}{16}}\\ &= 1.58\end{aligned}$$

8. Hypothesis Formulations

8.1 Objectives

- To understand how to ask the right questions to apply inferential techniques
- To understand how to formulate statistical hypotheses
- To develop a structured method for testing hypotheses
- To understand how to control for and what to do about error

8.2 Hypothesis Formulation

- Statistical Inference Revisited

  - We assess population parameters by calculating sample statistics (i.e use sample to get idea of population)
  - We base our assessment on incomplete information, so this inference is subject to uncertainty & error
  - The challenge of statistical inference is ∴ to provide a means for estimation while controlling this probability of making an error
  - In this process asking the right questions is half the problem:

    i.e Company manufactures ball bearings for precise machines that should have an average diameter of 6 mm

  - **Hypothesis** = a claim (assumption) about a population parameter

    Examples of parameters are the population mean
    The parameter must be identified before analysis
    For example, I claim that the average ball bearing diameter is 6 mm

    The Null Hypothesis (H_0) states the assumption numerically to be tested

    e.g. Average no. of TV sets in U.S. homes is at least 3:

    H_0: $\mu \geq 3$

    Note: its always about a population parameter (H_0: $\mu \geq 3$) not a sample statistic (H_0: $\bar{X} \geq 3$)

- The Null Hypothesis (H_0)

  - Assumption to be tested, always begin with the assumption that the Null is true

    This refers to the status quo:

    This is similar to the notion of innocent until proven guilty
    (if you don't prove known hypothesis is wrong its assumed to be correct)

  - Must contain '=' sign (or in the form of ≤ or ≥)

- The Alternative Hypothesis (H_1)

  - What we want to prove (believed to be true)
  - Opposite of the Null Hypothesis, i.e challenges the status quo

    e.g. Prove average no. of TV sets in U.S. homes is less than three:

    H_1: $\mu < 3$

  - Never contains the "=" sign

- E.g.
- The average travel time on a 20-km stretch of road is 20 minutes. An engineer claims that by retuning the traffic lights he can reduce this. The travel authority decides to test this. After testing the stretch for a month, they conclude that the average travel time is now 19 minutes and 36 seconds.
- How would you formulate the hypotheses to show that this retuning of the traffic lights reduced the travel time? (think about the Null & alternative hypothesis)

  H1: $\mu < 20$ H0: $\mu \geq 20$

- E.g.
- A political party is taking part in the election. They have a continuous contract with a financer, who will donate funds to the party, as long as they are polled to receive at least 25% of the votes.
- Results from a specific poll predict that the party will receive 15% of the votes
- Which hypotheses does the financer need to test if he wants to stop donating funds to the party?

  H1: $\mu < 25\%$ (Financer needs to prove this) H0: $\mu \geq 25\%$

13. Multivariate Linear Regression

13.1 Introduction

- Linear Regression Model

  - General linear regression model can be written as:

  (linear regression model for single independent variable)

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i}$$

Yi = dependent variable
Xi = independent variable
β_0 = population Y intercept
β_1 = population slope coefficient
$\beta_0 + \beta_1 Xi$ = linear component of the regression model
εi = random component of the regression model

- We now extend this model to a linear regression w/ multiple variables:

(More than one independent variable)

$$\boxed{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i}$$

This is generally referred to as the multiple linear regression model with k independent variables

- Multiple Linear Regression Equation

  - Used to predict the value of a variable based on the value of two or more other variables

  - We estimate the coefficients of the multiple linear regression model using sample data (same as simple linear regressions)

  - For this, use the **multiple linear regression** equation w/ k independent variables:

$$\boxed{\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki}}$$

$\hat{Y}i$ = Estimated (predicted) value of Y (yhat)
b_0 = Estimated intercept
b_{1i} , b_{2i} , ..., b_{ki} = Estimated slope coefficients

- The idea is to again use the least squares method to estimate all these coefficients:

Calculate sum of squares (SSQ) then calc 1st derivative WRT b0 & bi
We repeat this, and calculate the 1st derivatives WRT all the b parameters

15.3 Fundamental Assumption

- Time series analysis is founded on two fundamental assumptions:

1) The first is **stationarity:**

- Stationarity = Statistical properties of the time series are constant in time
- The consequence is:

$$f[x(t_1)] = f[x(t_2)] = \ldots = f[x(t_n)] = f(x)$$

- This assumption ensures that the statistics of the time series are independent of time:

$$\boxed{\begin{array}{rclcrcl} m(t) & = & m & \quad & \text{cov}(t_1, t_2) & = & \text{cov}(\tau) \\ \sigma^2(t) & = & \sigma^2 & \quad & \rho(t_1, t_2) & = & \rho(\tau) \end{array}}$$

τ = t1 – t2

2) The second assumption is **ergodicity:**

- Calculating statistics over an ensemble is equal to calculating them over a single realization

i.e calculating variances, means, autocovariances for the observed time series is the same for the ensemble

15.4 Definitions

- White Noise (a_t):
  - Definition = A sequence of uncorrelated random variables in time

    (are Gaussianly distributed w/ mean = 0 and variance = σ_a^2)

    Correlation between a_t & a_{t-1} = 0
  - Autocovariance function (γk):

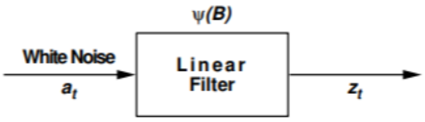
$$\gamma_k = \mathrm{E}\left[a_t a_{t+k}\right] = \begin{cases} \sigma_a^2 & k = 0 \\ 0 & k \neq 0 \end{cases}$$

  - Autocorrelation function (ρk):

    ρk = 1 for k = 0

    ρk = 0 for all other values of k

- Linear Filters:
  - Linear operator ψ(B) that transforms white noise (a_t) into a sequence of correlated variables Z_t



- Backwards Delay Operator B:
  - Operates on an element of a time series to produce the previous element
  - Definition: B*Z_t $(a_0 + a_1 B + a_2 B^2) Z_t = a_0 Z_t + a_1 Z_{t-1} + a_2 Z_{t-2}$