Need for sampling

- PDFs of RVs establishes theoretical framework. But
  - Entire sample space may not be known
  - Parameters of distribution may not be known
- From a finite sample derive conclusions about the pdf and its parameters
- Sample (or observation) set is assumed to be sufficiently representative of the entire sample space
  - Proper sampling procedures and design of experiments to be used for obtaining the sample

Basic Concepts

- Population: Set of all possible outcomes of a random experiment characterized by f(x)
- Sample set (realization) : Finite set of observations obtained through an experiment
- Inference: Conclusion derived regarding the population (pdf, parameters) from the sample set
    - Inference made from a sample set is also uncertain since it depends on the sample set which is one of many possible realizations

Statistical Analysis

- Descriptive Statistics (Analysis)
  - Graphical : Organizing and presenting the data (eg. box plots, probability plots)
  - Numerical: Summarizing the sample set (eg. mean, mode, range, variance, moments)
- Inferential
  - Estimation: Estimate parameters of the pdf along with its confidence region
  - Hypotheses testing: Making judgements about f(x) and its parameters

Measures of Central Tendency - Mean

- Represent sample set by a single value
  - Mean (or average): $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
    - Best estimate in least squares criterion
    - Unbiased estimate of population mean: $E[\bar{x}] = \mu$
    - Affected by outliers
    - Eg: Sample heights of 20 cherry trees

    [55 55 59 60 63 65 66 67 67 67 71 71 72 73 75 75 78 81 82 83]
    - Mean = 69.25 (population mean used to generate random sample was 70)
    - Mean = 71.75 (after a bias of 50 was added to first sample value)

Measures of Central Tendency – Median

- Represent sample set by a single value
  - Median: Value of x_i such that 50% of the values are less than x_i and 50% of observations are greater than x_i
    - Robust with respect to outliers in data
    - Best estimate in least absolute deviation sense
    - Eg: Sample heights of 20 cherry trees

    [55 55 59 60 63 65 66 67 67 67 71 71 72 73 75 75 78 81 82 83]
    - Median = 69 (population mean used to generate random sample was 70)
    - Median = 69 (after a bias of 50 was added to first sample value)

Measures of Central Tendency -Mode

- Represent sample set by a single value
  - Mode: Value that occurs most often (Most probable value)
  - eg. Sample heights of 20 cherry trees

[55 55 59 60 63 65 66 67 67 67 71 71 72 73 75 75 78 81 82 83]

  - Mode: 67 (three occurrences)

Measures of Spread

- Represents spread of sample set
  - Sample variance : $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
    - Unbiased estimate of population variance : $\mathrm{E}[s^2] = \sigma^2$
    - Standard deviation is sqrt of variance
  - Mean absolute deviation : $\bar{d} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$
  - Range : $R = x_{max} - x_{min}$
  - Eg. Sample heights of 20 cherry trees

s^2 = 70.5132 and 212.25 with outlier

s = 8.392 (population std used for generating numbers was 10)

MAD = 6.85 and 9.5 with outlier

Range = 83 - 55 = 28

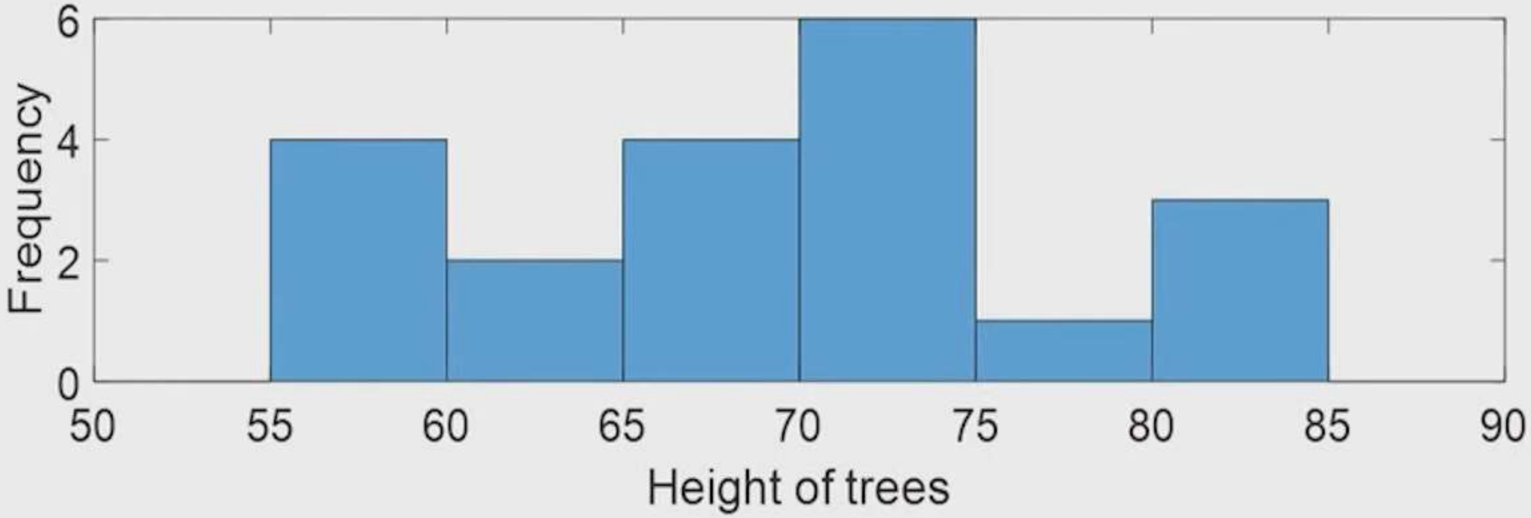
Distribution of sample mean and variance

- Sample mean
  - For any distribution sample mean is an unbiased estimate of population mean
  - If $x_i \sim \mathcal{N}(\mu, \sigma^2)$ and all observations are mutually independent, then $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$
- Sample Variance
  - For any distribution sample variance is an unbiased estimate of the population variance
  - If $x_i \sim \mathcal{N}(\mu, \sigma^2)$ and all observations are mutually independent, then $\frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{N-1}$

Graphical Analysis - Histograms

- Histograms
  - Divide the range of values in sample set into small intervals and count how many observations fall within each interval.
  - For each interval plot a rectangle with width = interval size and height equal to number of observations in interval
  - eg. Sample of 20 heights of black cherry trees

[73 75 55 60 66 71 81 67 83 75 82 71 63 55 72 78 67 65 67 59]



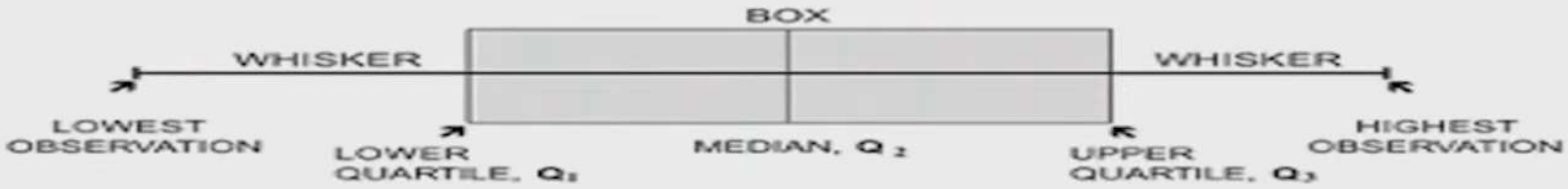
Graphical Analysis - Box Plot

- Box plot
  - Find quartiles (Q1, Q2 and Q3), minimum and maximum values in range
  - Box is between Q1 and Q3, and whiskers is between min and max values
  - eg. Sorted values of heights of 20 cherry trees

[55 55 59 60 63 65 66 67 67 67 71 71 72 73 75 75 78 81 82 83]

Q1: 64, Q2 (median): 69, Q3: 75, min: 55, max: 83

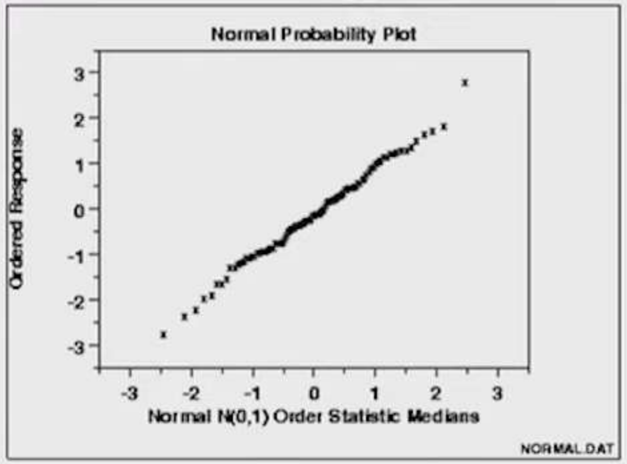
Figure 1. Box and whisker plot



Graphical Analysis – Probability Plot

- Probability plot (p-p or q-q plot)
  - Determine different quantile values from sample set. Plot computed quantiles vs theoretical quantile values from chosen distribution
  - Same example (standardized and sorted values)

[-1.697 -1.697 -1.2206 -1.1016 -0.7443 -0.5061 -0.3870 -0.2679 -0.2679 -0.2679 0.2084 0.2084 0.3275 0.4466 0.6848 0.6848 1.0420 1.3993 1.5184 1.6374]



Graphical Analysis – Scatter Plot

- Scatter plot
  - Plot of one RV (y) against another RV (x) to examine whether there is any dependence
  - Example: Marks obtained vs study time for 100 students