Lecture 15: Intro to Statistics.

**Aim Lecture** Motivate the main questions in statistics and illustrate methods for displaying data.

**Question** Are engineering

**E.g.1** Scores for

45 48 53 53 56 59 60 62 62 64 64 64 67 69 70
71 72 72 75 76 77 78 78 81 84 85 86 89 90 99

51 55 57 61 62 64 64 65 66 68
69 73 73 75 76 79 87 88 89 95

55 74 76 78 79 83 85 89 93 99
Questions 1. Has enough data been collected to answer our original question?
   2. Has the correct data been collected?
   3. How do you present the data collected in a format which helps answer the question at hand?

We’ll look at Q3 today.

Frequency Distributions

E.g.1 cont’d The range of values (in this case scores) lies between 40 and 99 so we shall break up these into class intervals. For sake of argument let’s pick class intervals

40-49, 50-59, 60-69, 70-79, 80-89, 90-99

To each class interval we record the frequency,
i.e.

<table>
<thead>
<tr>
<th>Score Range</th>
<th>40s</th>
<th>50s</th>
<th>60s</th>
<th>70s</th>
<th>80s</th>
<th>90s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq(eng)</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freq(law)</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freq(maths)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

E.g.1 cont’d
Histograms

A histogram is a

Box plots

If we were to draw the histogram for maths students on top of this the picture gets rather messy. Box plots can be used to present data in all three cases but at the price of losing some information.

Suppose we are given a sequence

\[ x_1 \leq x_2 \leq \ldots \leq x_n \]

called data points.

\textbf{Defn} (Median, Quartiles)
The sample median of the data points is the value $x$.

More precisely, if $n = 2k + 1$

If $n = 2k$

Alternatively, use linear interpolation. For $r \in (0, 1), i \in \mathbb{N}$ define

$x_{i+r} =$

& median of data points is

**e.g.** Median of $x_1, x_2, x_3, x_4$ is

The lower quartile $q_l$ is

$x_k$ where
The upper quartile $q_u$

$x_k$ where

<table>
<thead>
<tr>
<th>low</th>
<th>$q_l$</th>
<th>Med</th>
<th>$q_u$</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng</td>
<td>45</td>
<td>62</td>
<td>78</td>
<td>99</td>
</tr>
<tr>
<td>Law</td>
<td>51</td>
<td>68.5</td>
<td>77</td>
<td>95</td>
</tr>
<tr>
<td>Maths</td>
<td>55</td>
<td>76</td>
<td>81</td>
<td>88</td>
</tr>
</tbody>
</table>

E.g.1 cont’d Box Plot for Scores
The box plot shows

1. A box with median marked and

2. Whiskers showing

**N.B.** “Most” not defined here. It depends on your needs, package etc.

3. Isolated dashes showing

**Variants of Frequency Distribution**

Given data points $x_1, \ldots, x_n$ taking on values $y_1, \ldots, y_r$ (say in ascending order). Let $f_k$ be the frequency of the $k$-th value.
E.g. 2

Data Points: 1 1 3 3 3 4 5 5 7 7 7 7

<table>
<thead>
<tr>
<th>Value</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rel Freq</td>
<td>$\frac{1}{12}$</td>
<td>$\frac{2}{12}$</td>
<td>$\frac{4}{12}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cum Freq</td>
<td>6</td>
<td>8</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rel Cum Freq</td>
<td>$\frac{6}{12}$</td>
<td>$\frac{8}{12}$</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We can also record

Relative Frequency:

Cumulative Frequency:

Relative cumulative Frequency:

Measures of Location and Dispersion
Consider the following 2 frequency distribution graphs (smoothed out).

The frequency distribn often gives more info than needed. Most essential info is

a.

b.

For a. can use

**Defn (Mean,Mode)** The arithmetic mean of a set of data points \(x_1 \leq x_2 \leq \ldots \leq x_n\) is
\bar{x} := 

The sample mode

**E.g. 3** Data Points: 1,1,3,4,6,6,7

For b) use

**Defn (Variance, Standard Deviation)**

The sample variance of a set of data points \(x_1 \leq x_2 \leq \ldots \leq x_n\) is

\[s^2\]

Its square root \(s\) is called
E.g. 3 revisited

\[ s^2 = \]

\[ s = \]

**Propn** For data \( x_1, \ldots, x_n \)

\[ s^2 = \]

Proof: \( s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \)

\[ = \frac{1}{n} \sum_{i=1}^{n} (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \]