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

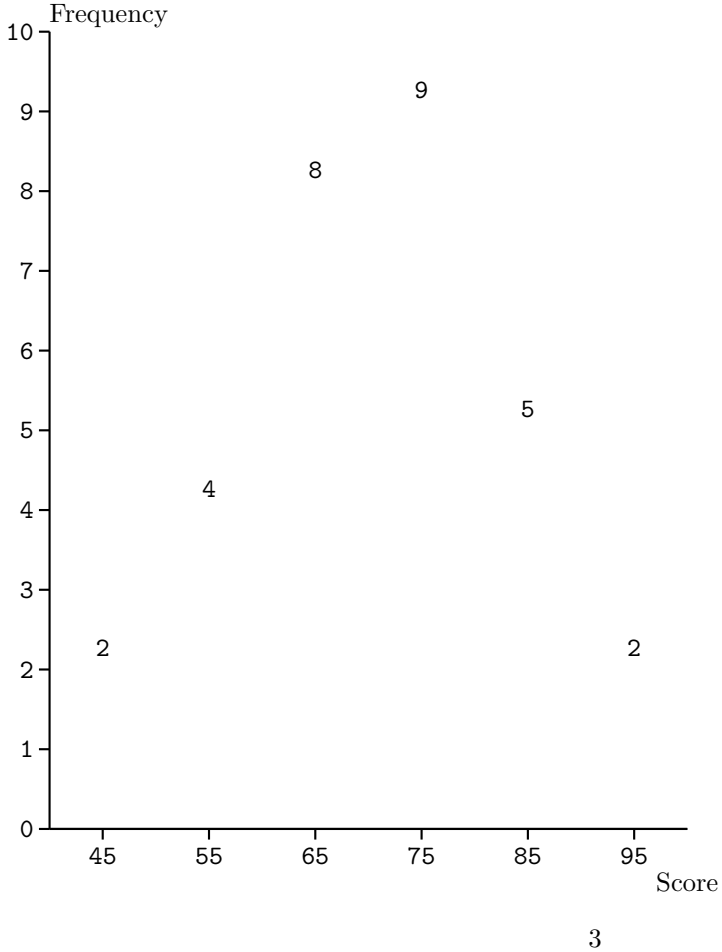
40-49,50-59,60-69,70-79,80-89,90-99

To each class interval we record the *frequency*,

i.e.

	40s	50s	60s	70s	80s	90s
$Freq(eng)$				9	5	2
$Freq(law)$				5	3	1
$Freq(maths)$	0	1	0	4	3	2

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.

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

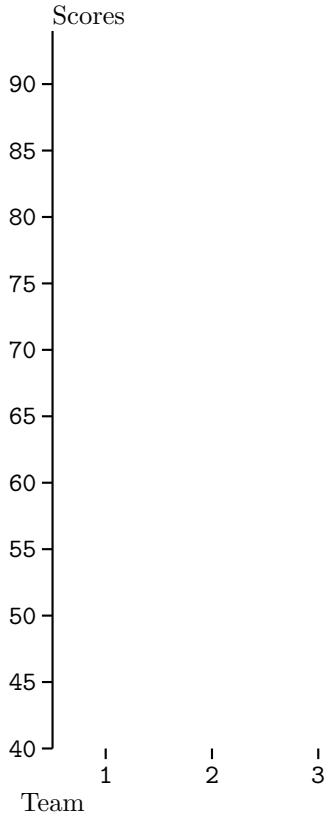
x_k where

The upper quartile q_u

x_k where

	low	q_l	Med	q_u	hi
Eng	45	62		78	99
Law	51		68.5	~ 77	95
Maths	55	~ 76	81	~ 88	99

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

Value	1	3	4	5	7
Freq			1	2	4
Rel Freq			$\frac{1}{12}$	$\frac{2}{12}$	$\frac{4}{12}$
Cum Freq			6	8	12
Rel Cum Freq			$\frac{6}{12}$	$\frac{8}{12}$	1

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

$\overline{\text{x}} :=$

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: $s^2 = \frac{1}{n}\sum(x_i - \overline{\mathrm{x}})^2$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\,\overline{\mathrm{x}} + \overline{\mathrm{x}}^2)$$