Sample surveys

Selection bias
response bias

simple random samples
quota sampling
cluster sampling

looking at data

pie chart
dot chart
bar chart
histogram
Sample Surveys.

- how to design the survey that generates the sample to get a representative sample.

- how to use that sample to make inference about the population biases - to be avoided.

Selection Bias:
- when the people you select for the sample are not representative of the whole population.
Opinion polls for 1936 election

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roosevelt</td>
<td>43%</td>
</tr>
<tr>
<td>London</td>
<td>57%</td>
</tr>
</tbody>
</table>

Mailed out 10 million surveys
- addresses from telephone directories
- club mailing lists.

1936 - Poor people were unlikely to have telephones or be club members.

1936 - Great Depression - poor people likely to vote differently.

Taking a large sample with a biased procedure does not improve the results.

Response Bias.
98% of respondents were dissatisfied with their marriage.

75% had extramarital affairs.

Only 4% of survey forms were returned.

93% satisfied
7% extramarital affairs
‘1 in 2’ admits to plagiarism

Michael Stothard
Chief News Editor

Half of Cambridge students have committed plagiarism as defined by the university, according to a Varsity survey.

49 per cent of students admitted that they have plagiarised work, although this differed radically between subjects and colleges. Ironically, students of the Law faculty plagiarised the most out of any subject, with 62 per cent of them breaking the university rules. The second highest was the Archaeology and Anthropology department with 59 per cent.

“It is a depressing set of statistics,” said Robert Foley, a Professor in Biological Anthropology at King’s College.

The college at the bottom of the Tompkins table, St Edmundsbury, had the highest proportion of plagiarising students, with 67 per cent admitting to breaking the university rules. Selwyn, at the top of the Tompkins table, had the fewest number of plagiarising students.

“It stands to reason that those students who are performing less well will resort to more underhand means to get by,” said a member of the University Council, the principal executive and policy making body of the university.

It is perhaps not surprising that 80 per cent of students said that the university is doing enough to punish plagiarism. “You can see why students, a great number of whom are frequently breaking the rules to their own benefit, would be keen to uphold the impression that the system is working,” said a member of the General Board, the body responsible for education policy at the University.

“Sometimes when I am really fed up,” said a Land Economy student at Pembroke, “I Google the essay title, copy and throw everything on to a blank word document and flog the order list. They usually end up being the best essays.”

60 per cent of Land Economy students admitted to plagiarism. “I set out to do it because less than five per cent of the student population replied to the survey.”

Cambridge application levels hit record high of 15,000

Vicky Woolley
News Reporter

The University has announced that the number of applicants for undergraduate courses has reached a record high this year. For the past four years, around 13,500 students have applied to study at Cambridge, whereas this year the figure is set to be closer to 15,000. This equates to an increase of around 12 per cent, compared with an average rise of only 6.5 per cent across all UK universities.

Some subjects have fared better than others. Computer Science has seen a 53 per cent rise in applications, with the next biggest rise being in Philosophy, which is up by 43 per cent. Theology and AS/A level applications drop by 18.5 and 10 per cent respectively.

This rise in applicant numbers comes at the end of a year in which many changes have been made to the admissions process. The requirement for a foreign language GCSE has been removed because of fears it disadvantaged applicants from state schools, where foreign language teaching has declined rapidly in recent years. The Cambridge Application Form was also abolished: the form, which students had to fill in alongside their UCAS application, was felt to be off-putting for some students.

Continued on page 3
How to choose a good sample?

Simple random sampling:
- decide how large a sample is needed.
- from the list of the entire population
  - draw an element at random
  - delete that element from the list.
  - repeat until desired number in the sample.
- often difficult to do.
- often pretend we're doing this, as it makes the math easier.

Other types of samples.

Quota Sampling / stratified sampling
- try to make the sample match the population.
  - specify # men + # women.
  - age distribution
  - race of men/women.
But - final choice is left to the interviewer - introduces biases.

eg in a shopping mall, the people may fit the specified criteria, but are unlikely to be representative of the population.
(eg too many non-working mothers)

Cluster Sampling

Arrange for the people in the sample to be physically more co-located than in a simple random sample.

Multiple stages - each subject to chance.
Gallup Poll.

Finally - the actual person to be interviewed is determined in advance. - removes selection bias.

few 1000's in sample gives errors of ±3%
Telephone surveys.

- generate phone numbers randomly.
  - under samples groups (e.g., students) who only have cellphones.

Convenience sample.

- easy to generate
- hard to analyse statistically.
Looking at Data.

Graphical methods / representations that can aid understanding of the distribution of a data set.

- range of values
- most likely values
- symmetry
- unimodal / multimodal
- tail behaviour
not angles or areas.

Linear measures at judging human are good.

Pie charts.
Vanilla Cream Plus Blueberry.

Straight more.

Dot Chart

Bar Chart
Histogram - visual display of variability

Range of data.  59 -> 87

Reject obvious errors
(no one in class is 5 inches tall)

class intervals: 
59 - 86
66 -

<table>
<thead>
<tr>
<th>class intervals</th>
<th>#</th>
<th>%</th>
<th>% - per - inch</th>
</tr>
</thead>
<tbody>
<tr>
<td>59 - 60</td>
<td>2</td>
<td>36</td>
<td>2.6</td>
</tr>
<tr>
<td>60 - 64</td>
<td>13</td>
<td>23.6</td>
<td>23.6/4 = 5.9</td>
</tr>
<tr>
<td>64 - 72</td>
<td>34</td>
<td>61.8</td>
<td>618/8 = 7.7</td>
</tr>
<tr>
<td>72 - 76</td>
<td>3</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>76 - 80</td>
<td>2</td>
<td>3.6</td>
<td>0.9</td>
</tr>
<tr>
<td>80 - 88</td>
<td>1</td>
<td>1.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>

class intervals include the upper limit

---

59 60 61 62 63 64
Does this accurately represent the data?

No - there were the same # of people in each of the two regions.

How many more people were between 64-72 compared with 60-64?

Tend to look at areas when comparing the bars.
Density plot — Area represents \%.

What \% are between 70 and 72? 

\[ 2 \times 7.7 = 15.4\% \]

Much easier to do this on a density plot than on the previous plot.