Chapter 1

Collecting and Presenting Data

OUTLINE

- Why do we want to collect data?
- How do we collect data?
- Once we have some data how do we present them?

Definitions

- *Random Variables*
- *Observation*
- *Data*
- *Population*

Definitions - Continued

Qualitative Variables

- Non numeric

Quantitative Variables

- Numeric
- Natural ordering

Qualitative Data

- *Ordinal Data*

- *Categorical Data*

Discrete Data

- Sequential

- Distinct Values

Continuous Data

- Numeric

- Measured on a continuous scale

Sampling

- Representative

- Unbiased

Three forms of sampling techniques.

- Random Sampling
- Quasi-random Sampling
- Non-random Sampling

Simple Random Sampling

- Each member of the population has an equally likely chance of being picked
- Unbiased

However

- Difficult to know who the whole population is.
- Some elements of population difficult to reach, which adds cost and time.
- Possible to pick an unrepresentative sample.

Stratified Sampling

- Random sampling, where clearly defined groups exist.
- Ensures sample reflects groupings in the population.
- Sample selected randomly within groups.

However

- Difficult to get clear information about the composition of each group
- Still need to know the whole population.

Systematic Sampling

- Quasi-random method.
- Randomly pick the first observation to sample then pick every k^{th} member of the population after the original.
- Easy to implement especially in production settings.

However

- It is non-random
- If there is a pattern or distribution to the process can have biased samples.
- Only works well with structured populations.

Multi-stage Sampling

- Quasi-random method.
- Used when you have large geographical spread of the population.
- Divide into geographically distinct areas and choose one at random to sample from.
- Save time and expense over sampling from whole population.

However

- Samples can be biased if stages are not carefully chosen
- Good design essential to get a representative sample.

Quota Sampling

- Quasi-random method.
- Similar to stratified sampling.
- Divide the population into groups, decide on a quota to sample and stop when we have reached it.

However

- Identification of appropriate quotas is difficult.
- Bias can be introduced by the interviewer.

Cluster Sampling

- Non-random sampling.
- Similar to multi-stage sampling.
- Population split into geographical areas, or clusters.
- All members of the cluster are observed.
- Can be inexpensive to conduct, due to small size of cluster.

However

- Easy to introduce bias.
- Sample can easily be unrepresentative.

Judgemental Sampling

- Non-random sampling.
- Targets individuals who are thought to have the information required.
- Can be very focused.

However

- Easy to introduce bias.
- Sample can easily be unrepresentative.

Sample Size

Larger samples give more precise information.

Categorical Data Frequency Tables

Example

Student	Mode	Student	Mode	Student	Mode
1	Car	11	Walk	21	Walk
2	Walk	12	Walk	22	Metro
3	Car	13	Metro	23	Car
4	Walk	14	Bus	24	Car
5	Bus	15	Train	25	Car
6	Metro	16	Bike	26	Bus
7	Car	17	Bus	27	Car
8	Bike	18	Bike	28	Walk
9	Walk	19	Bike	29	Car
10	Car	20	Metro	30	Car

Categorical Data Frequency Tables

Mode	Frequency
Car	10
Walk	7
Bike	4
Bus	4
Metro	4
Train	1
Total	30

Categorical Data Frequency Tables

Mode	Frequency	Relative Frequency (%)
Car	10	33.3
Walk	7	23.4
Bike	4	13.3
Bus	4	13.3
Metro	4	13.3
Train	1	3.4
Total	30	100

Discrete Data Frequency Tables

Another example:

Date	Cars Sold	Date	Cars Sold
01/07/04	9	08/07/04	10
02/07/04	8	09/07/04	5
03/07/04	6	10/07/04	8
04/07/04	7	11/07/04	4
05/07/04	7	12/07/04	6
06/07/04	10	13/07/04	8
07/07/04	11	14/07/04	9

Continuous Data Frequency Tables

- Identify the maximum and minimum observations.
- The class interval should be a convenient number.
- Want a reasonable number of classes.
- Decide on class interval (range of each group) e.g. $20 \leq obs < 30$

Continuous Data Frequency Tables

Example

214.8412	220.6484	216.7294	195.1217	211.4795
195.8980	201.1724	185.8529	183.4600	178.8625
196.3321	199.7596	206.7053	203.8093	203.1321
200.8080	201.3215	205.6930	181.6718	201.7461
180.2062	193.3125	188.2127	199.9597	204.7813
198.3838	193.1742	204.0352	197.2206	193.5201
205.5048	217.5945	208.8684	197.7658	212.3491
209.9000	197.6215	204.9101	203.1654	192.9706
208.9901	202.0090	195.0241	192.7098	219.8277
208.8920	200.7965	191.9784	188.8587	206.8912