CSC 439 - 3Units
Statistical Processing Systems

Professor Benjamin Aribisala

Department of Computer Science

Lagos State University

Course Outline

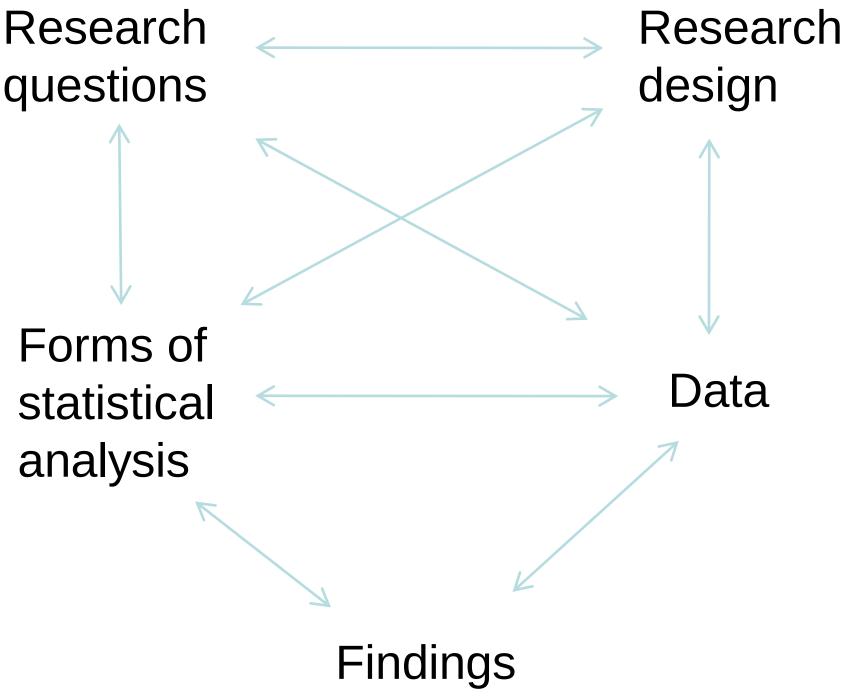
- Design and Analysis of Sampling Surveys
- Statistical Data Compression
- 2D, 3D Frequency Table, Points and Interval Estimation;
- Test of Significance
- Test of Hypothesis
- Analysis of Categorical Data
- Model Validation
- Chi-Square-Test
- Probability Plots.

Assessment Method

- CA = 30%
- Exam = 70%
- CA – Project, Assignment and Tests
- Practical sessions – Google Form, Python or R

Is research design a step on the way from research questions to findings?

(Not necessarily…)



Research questions

Research design

Forms of statistical analysis

Data

Findings

Forms of quantitative data ‘collection’

‘Controlled’ comparisons	Experiments / quasi-experiments (not common in some disciplines)
‘Talking’ with people	Surveys
Observing	Structured observation
Using existing quantitative information: I	Secondary analysis of existing ‘survey’ (micro-)data
Using existing quantitative information: II	Analysis of existing published quantifiable data
Looking at ‘texts’ (books, films, web pages, adverts…)	(Some forms of) content analysis

What do social researchers lose by not carrying out experiments?

- In an experiment, a sample of people could be allocated at *random* to two groups, one of which could be subjected to a *treatment*, e.g. unemployment, and then the frequency of some *outcome*, e.g. marital dissolution, could be compared between the *treatment* and *control* groups.

- However, in reality, for practical (and ethical) reasons, one can only *observe* whether unemployed people are more likely to experience marital dissolution than other people.

What do social researchers lose… (continued)

- But different rates of marital dissolution in these two groups might alternatively be attributable to systematic differences between the two groups (*selection effects*), e.g. in relation to social class or age at marriage (which are themselves related!).
- It is thus important to *control* for other factors that may have induced an observed relationship. *Controlling* for such factors inevitably involves the use of a *multivariate analysis*…

(A multivariate analysis involves three or more *variables*, in this case four relating to unemployment, marital dissolution, class and age at marriage).

Statistics without surveys…

- Some of the approaches on an earlier slide involve **unobtrusive methods**, i.e. they tend not to involve the (artificial, potentially power-laden, and much criticised) interactions found in survey interviews.
- However, all involve the operationalisation of concepts and coding of data (whether directly or indirectly), and may also involve decisions in relation to sampling, and as such are not immune from criticisms aimed at these processes, and the subjectivity involved therein.

Sample and population

- A range of statistical analyses of a **sample** can be carried out, including **descriptive** analyses.
- However, the topic of interest/research question is typically focused on some feature of the **population** (e.g. whether, on average, women *in the UK* earn more or less than men – as opposed to whether the women *in the sample* in question earn more on average than the men).
- Therefore statistical analyses usually involve the use of techniques for making **inferences** from a sample to the corresponding population.

Sampling error

Most pieces of quantitative social research rely on samples, but the most carefully designed sample will never provide a perfect representation of the population from which it was selected.

There will always be some sampling error!

However, the term is used in a specific way in quantitative social research. If the sampling approach used, or patterns of non-response, result in some degree of **bias** in the results, this is labelled as **non-sampling error**, as it reflects a deficiency in the *sampling approach*, rather being an inevitable consequence of examining a *sample*.

Sampling error or bias?

- A **parameter** is a quantity relating to a given variable for a population (e.g. the average (mean) adult income in the UK).
- When researchers generalize from a sample they use sample observations to **estimate** population parameters.
- The **sampling error** for a given sample design is the degree of error that is to be expected in making these estimations.
- So the **parameter estimates** generated by quantitative research are equal to the population parameters, plus a certain amount of **sampling error**, plus any **bias** arising from the data 'collection' process.

Time as a key dimension in quantitative research

- **Cross-sectional studies**

  Observations of a sample or ‘cross-section’ of a population (or of other phenomena) are made at one point in time – most surveys are cross-sectional.
  → This leads to a common criticism of survey research: that it is ahistorical/unsuited to the examination of social processes.

- **Longitudinal studies**

  These permit observations of the same population or phenomena over an extended period of time.
  → These enable analysis of change. They may also facilitate more credible assertions relating to causality.

Types of longitudinal study

1. **Trend studies** – these examine change within a population over time (e.g. the Census).
2. **Cohort studies** – these examine over time specific subpopulations or cohorts (often, although not necessarily, the same individuals) e.g. a study might interview people aged 30 in 1970, 40 in 1980, 50 in 1990 and 60 in 2000.
3. **Panel study** – These examine the same set of people each time (e.g. interview the same sample of (potential) voters every month during an election campaign.

Survey design issues

- Form of implementation
- Thinking about who or what you’re researching: case, population, sample
- Non-probability samples
- Probability samples (‘Random samples’)
- Sample size

Some general points…

- Typically, individuals are the ‘units of analysis’. (This is not always the case though: for example in a survey of schools)
- Individuals, referred to as respondents, provide data by responding to questions.
- The ‘research instrument’ used to gather data is often referred to as a questionnaire.
- Questionnaires/‘Interview schedules’:
  - collect standardised information.
  - are used to elicit information to be used in analyses.

Three types of surveys:

1. Self-administered questionnaires

   Including:

   1. Mail(ed) surveys (or e-mail surveys)
   2. Web-based surveys
   3. Group surveys (e.g. in a classroom)
2. Interview surveys (‘face-to-face’: including CAP interviewing)
3. Telephone surveys (including CAT interviewing)

Method	Advantages	Disadvantages	Tips to Remember
Self-completion	•Cheap •Cover wide area •Anonymity protected •Interviewer bias doesn't interfere •People can take their time	•Low response rate (and possible bias from this) •Questions need to be simple •No control over interpretation •No control over who fills it in •Slow	•*Simplify questions* •*Include covering letter* •*Include stamped addressed response envelope* •*Send a reminder*
Telephone survey	•Can do it all from one place •Can clarify answers •People may be relatively happy to talk on the phone •Relatively cheap •Quick	•People may not have home phones/be ex-directory •You may get wrong person or call at wrong time •May be a bias from whose name is listed/who's at home •Easy for people to break off •No context to interview	•*Because you rely totally on verbal communication – questions must be short and words easy to pronounce* •*Minimize number of response categories (so people can remember them)*
Face-to-face interview	•High response rate •High control of the interview situation •Ability to clarify responses	•Slow •Expensive •Interviewer presence may influence way questions are answered •If there is more than one interviewer, they may have different effects	•*Important that interviewer is non-threatening* •*Interviewer can clarify questions, but should be wary of elaborations that affect the content* •*Aim to ask questions in a clear, standardized way* •*If the list of possible responses is long, show them to the respondent for them to read while the question is read out*

Response rate

- You must keep track of the response rate, calculated as **the proportion of people who are selected to take part in the survey** (i.e. who are part of the ‘desired’ sample) **who actually participate**.
  For example, if you receive 75 questionnaires back from a sample of 100 people, your response rate is 75%.

  A more detailed example:

  – You are studying women over 50. You stop women in the street, ask their ages, and, if they qualify, you ask to interview them.

  – If you stop 30 women, but 20 are under 50 and only 10 over 50, your starting point (those qualified to take part) is thus 10.

  – If 5 of these are willing to talk to you, you have achieved a 50% response rate (5/10)

  – *Note: it is irrelevant that you originally stopped 30 women, hence your response rate is NOT 17% (5/30) – you ignore those people who do not qualify when calculating the response rate.*

Strengths of survey research

- Useful for describing the characteristics of a large population.
- Makes large samples feasible.
- Flexible - many questions can be asked on a given topic.
- Has a high degree of reliability (and replicability).
- Is a relatively transparent process.

Weaknesses of survey research

- Seldom deals with the *context* of social life.
- Inflexible – cannot be altered once it has begun (therefore poor for exploratory research).
- Subject to artificiality – the findings are a product of the respondents' consciousness that they are being studied.
- Sometimes weak in terms of validity.
- Can be poor at answering questions where the units of analysis are not individual people,
- Usually inappropriate for historical research.
- Can be particularly weak at gathering at certain sorts of information, e.g. about:
  - highly complex or 'expert' knowledge
  - people's past attitudes or behaviour
  - subconscious (especially macro-social) influences
  - shameful or stigmatized behaviour or attitudes (especially in the context of a face-to-face interview) – although survey research may nevertheless be able to achieve this in some circumstances.

Thinking about what you’re researching: Case, Population, Sample

Case: each empirical instance of what you’re researching

- So if you’re researching **celebrities who have been in trouble with the law** Pete Doherty would be a case, as would Kate Moss, Boy George, George Michael, Winona Ryder, OJ Simpson and Rachel Christie…
- If you were interested in **Fast Food companies** McDonalds would be a case, Burger King would be a case, as would Subway, Spud U Like, etc. …
- If you were interested in **users of a homeless shelter on a particular night**, each person who came to the shelter on the specified night would be a case.

Thinking about what you’re researching: Case, Population, Sample

- **Population** – all the theoretically-relevant cases (e.g. ‘Tottenham supporters’). This is also often referred to as the **target population**.
- This may differ from the **study population**, which is all of the theoretically-relevant cases which are actually available to be studied (e.g. ‘all Tottenham club members or season ticket holders’).

Sometimes you can study all possible cases

(the total population that you are interested in)

For example:

- *Post WW2 UK Prime Ministers*
- *Homeless people using a particular shelter on Christmas Day 2011*
- *National football teams in the 2010 World Cup*
- *Secondary schools in Coventry*

Often you cannot research the whole population

because it is too big and to do so would be too costly, too time consuming, or impossible.

For example, if your ‘population’ is:

- *Voters in the UK since WW2*
- *All the homeless people in the UK on Christmas Day 2011*
- *Club and National Football teams involved in cup competitions in 2012*
- *Secondary schools in the UK.*

On these occasions you need to **select some cases to study**.

Selecting cases from the total (study) population is called **sampling.**

How you sample depends (among other things) on some linked issues:

- What you are especially interested in (what you want to find out)
- The frequency with which what you are interested in occurs in the population
- The size/complexity of the population
- What research methods you are going to use
- How many cases you want (or have the resources and/or time) to study

Probability and non-probability sampling

Probability samples (‘Random samples’)

A probability sample has a mathematical relationship to the (study) population: we can work out mathematically what the likelihood (probability) is of the results found for the sample being within a given ‘distance’ of what would be found for the whole population (if we were able to examine the whole population!)

→ Such a sample allows us to make **inferences** about the population as a whole, based on the sample results.

Non-probability samples

→ Formally, these do not allow us to make inferences about the population as a whole.

However, there are often pragmatic reasons for their use, and, despite this lack of statistical legitimacy, inferential statistics are often generated (and published!)

Types of Non-probability sampling:

1. Reliance on available subjects:

- Literally choosing people because they are available (e.g. approaching the first five people you see outside the library)
- Only justified if less problematic sampling methods are not possible.
- Researchers must exercise considerable caution in generalizing from their data when this method is used.

Types of Non-probability sampling:

2. Purposive or judgmental sampling

- Selecting a sample based on knowledge of a population, its elements, and the purpose of the study. Selecting people who would be ‘good’ informants (individually/collectively).
- Used when field researchers are interested in studying cases that do not fit into regular patterns of attitudes and behaviours (i.e. when researching ‘deviance’).
- Relies totally on the researcher’s prior ability to determine ‘suitable’ subjects.

Types of
Non-probability sampling:

3. Snowball sampling

- Researcher collects data on members of the target population s/he can access, and uses them to help locate other members of the population.
- May be appropriate when members of a population are difficult to locate (and/or access).
- By definition, respondents who are located by snowball sampling will be connected to other respondents, thus respondents are more likely to share similarities with each other than with other members of the population.

Types of Non-probability sampling:

4. Quota sampling

- Begin with a matrix of the population (e.g. assuming it is 50% female and 9% minority ethnic, with a given age structure).
- Data is collected from people matching the defining characteristics of each cell within the matrix.
- Each cell is assigned a weight matching its proportion of the population (e.g. if you were going to sample 1,000 people, you would want 500 of them to be female, and hence 45 to be minority ethnic women).
- The data thus provide a representation of the population.
- However, the data may not represent the population well in terms of criteria that were not used to define the initial matrix.
- You cannot measure response rates.
- And, crucially, the selection process may be biased.

The logic of probability sampling

- **Representativeness:**

  A sample is **representative** of the population from which it is selected to the extent that it has the same aggregate characteristics (e.g. same percentage of women, of immigrants, of poor and rich people…)

- **EPSEM (Equal Probability of Selection Method):**

  Every member of the population has the same chance of being selected for the sample.

- **Random sampling:**

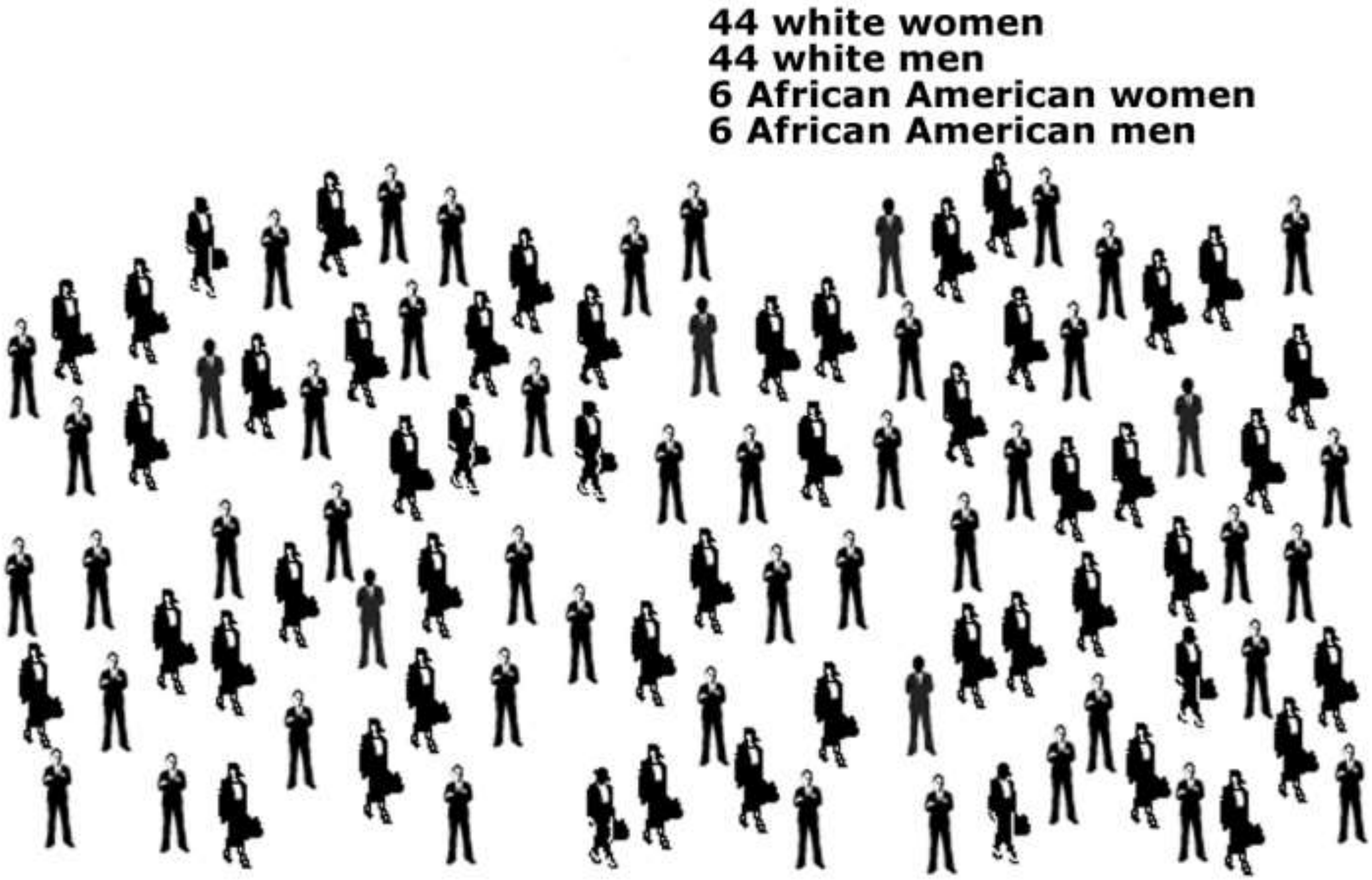
  Each element in the population has a known, non-zero chance of selection. ‘Tables’ or ‘lists’ of random numbers are often used (in print form or generated by a computer, e.g. in SPSS).

- **Sampling frame:**

  A list of every element/case in the population from which a probability sample can be selected.

  In practice, sampling frames may not include every element. It is the researcher’s job to assess the extent (and nature) of any omissions and, if possible, to correct them.

A Population of 100



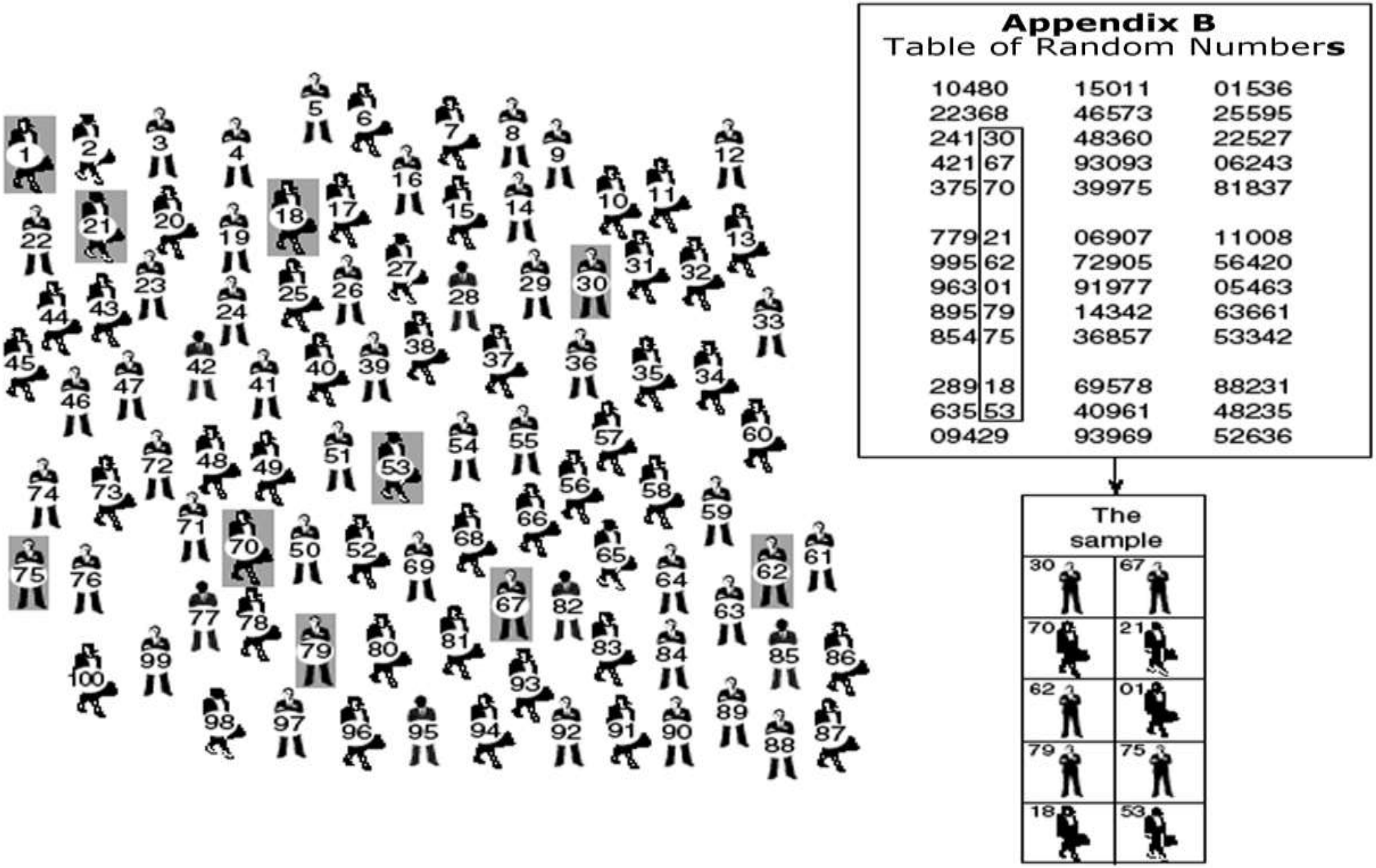
44 white women
44 white men
6 African American women
6 African American men

Types of probability sampling:

1. **Simple Random Sample**

- Feasible only with the simplest sort of sampling frame (a comprehensive one).
- The researcher enumerates the sampling frame, and randomly selects people.
- Despite being the ‘purist’ type of random sample, in practice it is rarely used.

A Simple Random Sample



Appendix B
Table of Random Numbers

10480	15011	01536
22368	46573	25595
24130	48360	22527
42167	93093	06243
37570	39975	81837
77921	06907	11008
99562	72905	56420
96301	91977	05463
89579	14342	63661
85475	36857	53342
28918	69578	88231
63553	40961	48235
09429	93969	52636

Types of probability sampling:

2. Systematic Random Sample

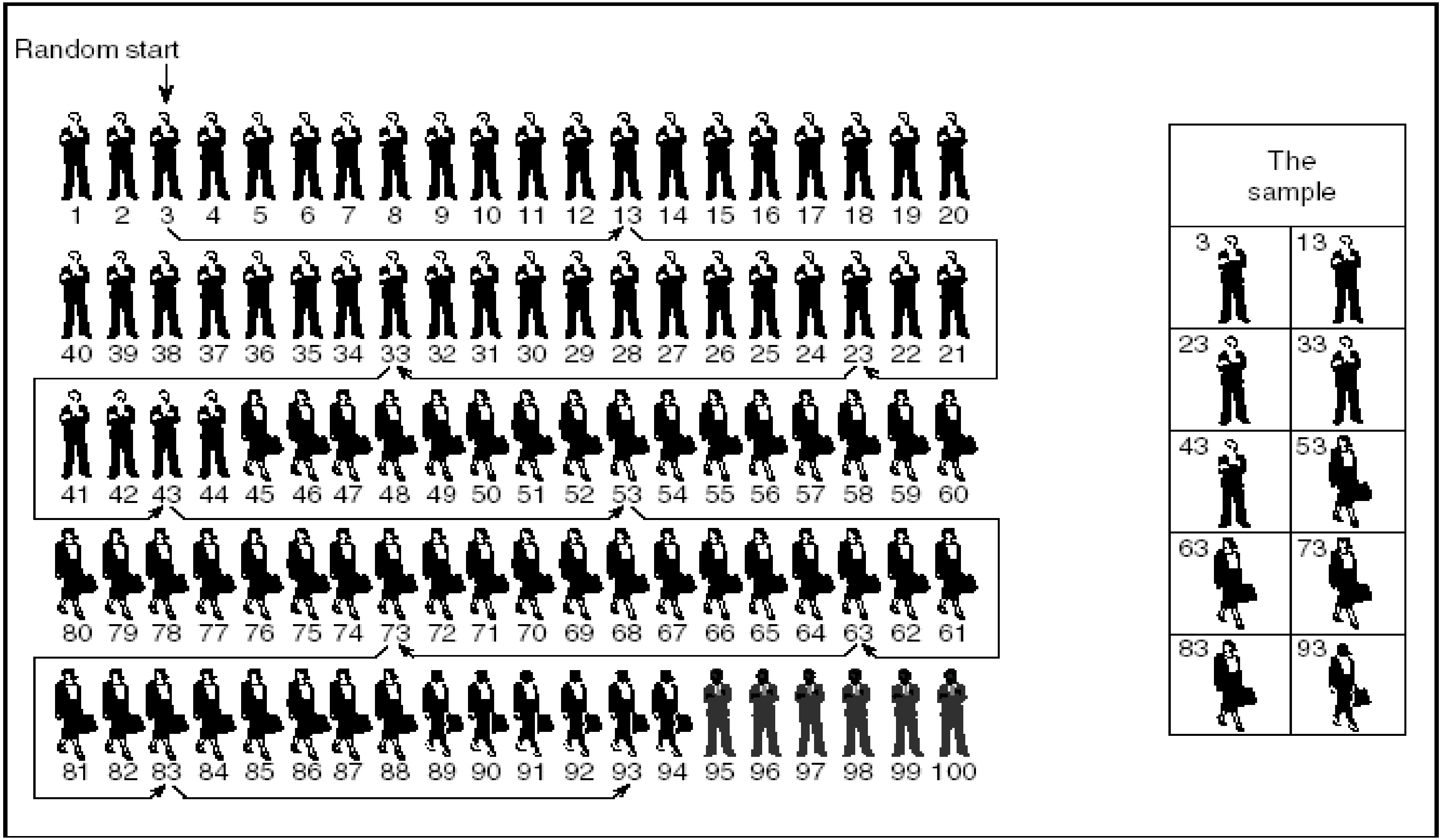
- Uses a random starting point, with every *k*th element selected (e.g. if you wanted to select 1,000 people out of 10,000 you'd select every 10th person: such as the 3rd, 13th, 23rd…).
- The arrangement of cases in the list can affect representativeness (e.g. if *k* is even, when sampling pages from a book with chapters starting on odd-numbered pages).

Types of probability sampling:

3. Stratified sampling

- Rather than selecting a sample from the overall population, the researcher selects cases from homogeneous subsets of the population (e.g. random sampling from a set of undergraduates, and from a set of postgraduates).
- This ensures that key sub-populations are represented adequately within the sample.
- A greater degree of representativeness in the results thus tends to be achieved, since the (typical) quantity of sampling error is reduced.

A Stratified, Systematic Sample with a Random Start



Types of probability sampling:

4. Multi-stage Sampling

- This is often used when it is not possible or practical to create a list containing all the elements within the target population.
- It involves the repetition of two basic steps: creating lists of sampling units and sampling from them.
- It can be highly efficient but less accurate.

Example of Multi-stage Sampling

Sampling Coventry residents

1. Make a list of all neighbourhoods in Coventry
2. Randomly select (sample) 5 neighbourhoods
3. Make a list of all streets in each selected neighbourhood
4. Randomly select (sample) 2 streets in each neighbourhood
5. Make a list of all addresses on each selected street
6. Select every house/flat [‘Cluster’ sampling!]
7. Make a list of all residents in each selected house/flat
8. Randomly select (sample) one person to interview.

Types of probability sampling:

5. Probability Proportional to Size (PPS) sampling

- A sophisticated form of multi-stage sampling.
- It is used in many large-scale surveys.
- Sampling units are selected with a probability proportional to their size (e.g. in a survey where the **primary sampling units** (PSUs) were cities, a city 10 times larger than another would be 10 times more likely to be selected in the first stage of sampling).

Note

- The sampling strategies used in real projects often combine elements of multi-stage sampling and elements of stratification.
- See, for example, the discussion of Peter Townsend’s poverty survey on p120 of Buckingham and Saunders, 2004.
- See also Rafferty, A. 2009. Introduction to Complex Sample Design in UK Government Surveys for summaries of the sample designs of various major UK surveys http://www.esds.ac.uk/government/docs/complexsampledesign.doc

Sample size

The sample size that is needed depends upon:

- The heterogeneity of the population: the more heterogeneous, the bigger the sample needed
- The number of relevant sub-groups: the more sub-groups, the bigger the sample needed
- The frequency of a phenomenon that you are trying to detect: the closer to 50% (of the time) that it occurs, the bigger the sample needed
- How accurately you want your sample statistics to reflect the population: the greater accuracy that is required, the bigger the sample needed.
- How confident you want to be about your results!

Other considerations when you are thinking about sample size

- The response rate – if you think that a lot of people will not respond, you need to start off by sampling a larger number of people.
- Form of analysis – some forms of statistical analysis require a larger number of cases than others. If you plan on using one of these you will need to ensure that you've got enough cases.

Generally (given a choice): Bigger is better!

(hence the sample size often reflects costs/resources.)

Content Analysis

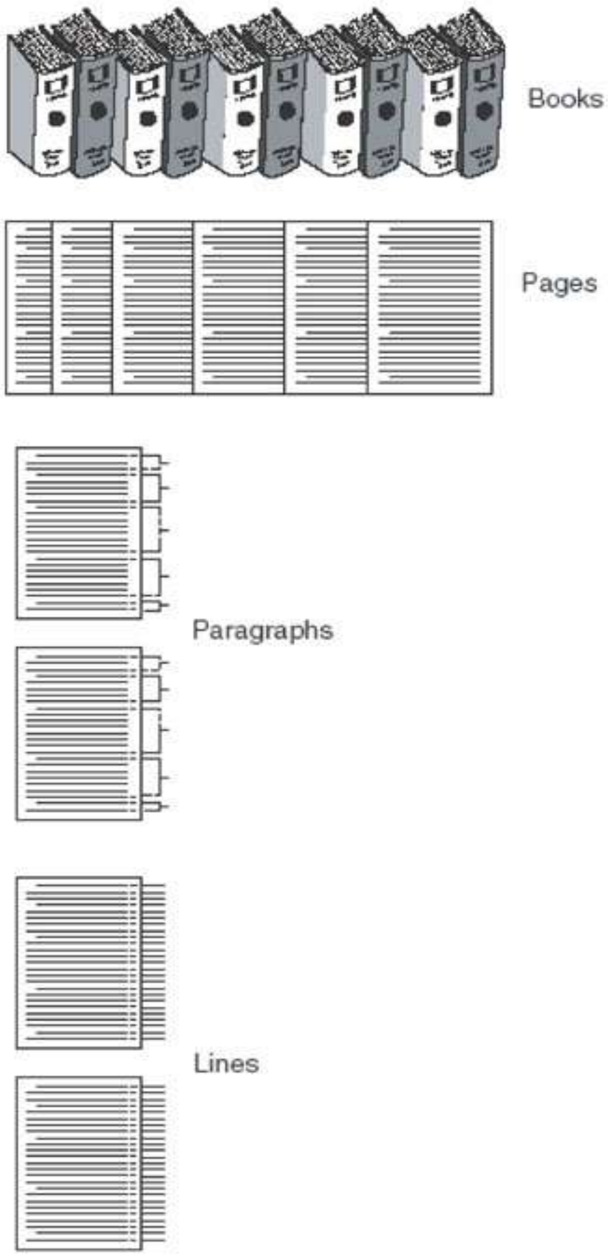
Method of transforming symbolic content of a document (such as words or images) from a qualitative unsystematic form into a quantitative systematic form.

See Bryman, 2008, Ch. 12 (online course extract)

Possible Units of Analysis for Content Analysis

…but a unit of analysis may also be:

a film,
a scene,
a TV episode,
a wall (containing graffiti),
a rubbish bin,
a politician's speech,
a web-site,
or a blog posting…



© 2005 Wadsworth - Thomson

Comparative-Historical Research

- Much comparative-historical research does not use statistics.
- However if you are looking at change over time or are comparing different countries or regions there are a large number of statistics that can be used:
- **Macro-level secondary statistics** – e.g. World Bank "development indicators" i.e. mortalitity rates; televisions per 1000 population; Literacy rates. Or "OECD Main Economic Indicators" – i.e. foreign direct investment; GDP; GNP… etc.

  [See the Library Statistics Workbook that is linked to the module web page; this is of value both in terms of accessing international data and statistical sources generally]
- **Primary statistics** – these are datasets that you construct for yourself from historical and comparative research. They may document anything from the strength and political composition of particular trade unions in a particular time and place; to land-holding patterns in different regions as described by local tax-records; to speeches made by Vice-Chancellors of UK universities at public forums over the last century… To conduct quantitative analysis of primary historical research it just needs to be systematically coded.

Sampling
Comparative-Historical Events

- If you are going to use comparative-historical data to create a dataset it is important to think about whether you have data from the entire population of events that you are interested in (i.e. every strike that occurred in the UK between 1990 and 2000), or whether you are focusing on a subset (thirty strikes that occurred in the UK between 1990 and 2000).
- If you present statistical information for a subset of events you are sampling and the same issues of occur as any other time that you sample data: your findings are only statistically generalisable if the sampling is random (or if each event has a known - typically equal - probability of selection into the subset).
- On the other hand, there are often substantive reasons to choose specific “important” events to be part of your subset (i.e. large-scale strikes that involved media campaigns). This is legitimate and statistics gleaned from these may be interesting and informative. However they are not statistically generalisable to all events (i.e. strikes generally) and so inferential statistics are not appropriate.

Observation(al) studies

- Observation is not just the preserve of qualitative methods. Quantitative methods can be applied where **structured** or **systematic observation** is carried out.
- Like qualitative observation studies (and surveys), this involves cross-sectional data (we can only observe the present).
- Unlike qualitative observation, structured or systematic observation is not inductive but requires the prior determination of what to observe (although this may be suggested by initial unstructured observations).

See Pole and Lampard, 2002, Ch. 4.

The observation schedule

- To produce quantitative data an observation schedule or coding scheme is required.
- This describes what is to be observed and how what is observed should be coded.
- For example, if I were observing in the Library Café and was interested in interactions between students and the staff working at the cash-registers I could code each student’s behaviour in the following way:

1. **No conversation, no eye contact, no smile**
2. **Eye contact and/or smile, no conversation**
3. **Conversation, only as required by the transaction**
4. **Conversation as required by the transaction and polite thanks.**
5. **Conversation that goes beyond transaction and polite thanks.**

The observation schedule

- The observations must be focused – and relevant to the research question
- The schedule (like closed questions in a questionnaire) should have categories that are mutually exclusive and exhaustive
- Recording should involve as little observer interpretation as possible – this is where reliability is diminished.

Sampling in Structured Observations

- It is important to be clear about the unit of analysis – are you sampling events/situations, interactions, or individuals?
- Sampling must consider the dimension of time in determining who, where, and when to make observations. It may sometimes be appropriate to sample at multiple time periods and in multiple sites.

Benefits and Drawbacks of Structured Observation I

- Like other ‘unobtrusive measures’ structured observation may avoid researcher contamination – enabling the study of people in their natural environment.
- Unlike surveys it does not depend on the negotiation of meaning between interviewer and interviewee (or the interviewee’s accurate representation of her behaviour).
- Unlike qualitative observation studies it can produce relatively reliable data and since observation (with a schedule) can be undertaken by more than one researcher, it enables large-scale data collection.

Benefits and Drawbacks of Structured Observation II

- However the researcher will only ‘see’ the predetermined categories of action that the schedule specifies. These may not be the categories of action that are relevant to participants.
- Since structured observation precludes questioning participants about their motives or opinions, it is wholly dependent on observing behaviour and on the ability of the researcher to appropriately assess this.
- It is ahistorical, in that it can only assess behaviour in the moment (unlike surveys which can ask, albeit imperfectly, about people’s pasts, or other methods such as content analysis, historical or secondary data analysis).