Masters Research Handbook

Lucia Rapanotti and Jon Hall

Branch: edit-stage4
Hash: 36a32d8dc6f2be2a3bb36b1048db3f15d74dd1ac

Contents

Stage 4: Generating and analysing data

You've now reached Stage 4, which means the end of your project is now in sight. In this stage you will be in the midst of your data generation and analysis, which is possibly the most exciting, yet demanding, part of your research: this is where you get your opportunity to make that contribution to knowledge.

This stage assumes that you have worked out most of your research design details and are now in a position to begin your data generation and analysis•.

With reference to our 5-stage framework, the activities which are in focus in Stage 4 are summarised in Table 0.1, which also provides some guidance for your interaction with your supervisor during this stage.

• If that's not the case, then, you should go back to Stage 3. You should also discuss your progress with your supervisor, revisiting your project timescale and risk.

Activity: Understanding the effort needed in this stage #1

Consider Table 0.1 carefully, paying particular attention to the entries in the 'Effort' column. Make a note of the activities which are most prominent in this stage and what their deliverables and learning outcomes are.

Discussion

Generating and analysing evidence will constitute by far your major effort in this stage (50% of study time): in particular, the framework assumes that you will have worked out the details of your research design in Stage 3, so you can focus on applying your data generation and analysis methods. You will also start to interpret you findings, an activity your will complete in Stage 5.

Note that your data analysis and interpretation may also prompt you to generate more data, including, perhaps, reviewing more academic literature or even re-thinking or adjusting your aim an objectives better to reflect your improving understanding. Therefore, so you should expect some iteration back to activities

Table 0.1: Research activities addressed in Stage 4 (20% of project length)

Research process activities	**Deliverables**	**Learning Outcomes: by the end of this stage you will:**	**Effort**	**Suggested focus of your interaction with your supervisor**
Identifying the research problem	Research problem statement, refined as needed	be able to assess and improve your research problem statement	1%	
Reviewing the literature	Substantial draft of your literature review, refined as needed	be able to assess and improve your current draft	1%	
Setting your aim and objectives	Aim and objectives, refined as needed	be able to assess and improve your aim, objectives and related tasks	2%	
Developing the research design	Research design description, refined as needed	be able to describe data generation and analysis procedures in detail	2%	Suitability of methods and procedures
Generating and analysing evidence	Raw data appropriately organised and stored; data summaries and outcomes of data analysis	know the difference between various sampling approaches; be able to organise and store your raw data; be able to apply appropriate data analysis methods; be able to present your data and evidence in a concise and effective way	50%	Appropriateness of data analysis and presentation
Interpreting and evaluating findings	Draft summary of findings from data/evidenced generated	be able to derive findings from your data analysis and critically assess them in relation to research aim and objectives	15%	Critical and logical thinking
Reflecting and reporting	Stage 4 report; draft abstract for your project	know the purpose and content of an abstract; be able to assess your research progress and write up a substantial report, including an abstract for your project	25%	Any further improvements required, particularly in relation to critical thinking and academic writing
Planning work and managing risk	Updated risk and work plan	be able to assess risk and draw a work plan	5%	Any major adjustment required to address deficiencies or manage risk

you have carried out in previous stages, and revision of things you have written.

0.1 Generating raw research data

Your *raw data* represent any data you generate as part of your research. Which data you generate, and how, is determined by the choices you have made in your research design, informed by your research aim and objectives. In this chapter, we look at key methods for generating data, by providing:

- design tips and techniques for each method;
- a basic workflow to get your data generation going;
- weaknesses exposed through the data generating technique;
- further sources to consult for more detail.

When we say "data generation", we don't necessarily mean that you will create a new-to-the-world "data"•. We simply mean new-to-your-masters-project data, which covers a multitude of sins, including:

- the creation of a brand new data set, which did not exist before your research project. This may be the result of data collected through new observations or measurements, a new survey, questionnaire or focus group, through selecting passages from documents, etc.
- the extension of a previously collected data set with new elements, derived from those that already exist, for instance adding the mean value of a collection of numerical data, or grouping together specific distinct data into new categories that you have created;
- the collection of previous data for reinterpretation, for instance if you are rerunning a previous experiment in order to confirm its results, or doing a meta-analysis of the literature in a particular area.

Our use of the term data generation includes all of these. Of course, in making a contribution to knowledge, you're going to have to do it regardless of the data generation method and so the data you generate as part of your research must allow you to conclude something new.

• Although this might indeed be the outcome of your data generation.

Related to data generation is the concept of "data source", which is the location from which your data originates. If you are re-using existing data sets, this may well be an archive or a digital repository. For new data sets you generate, this may well be the experimental or real-world setting of your own observations and measurements, or a population of interest from which you will derive a "sample" for further analysis. The latter underpins many data generation methods, so that we will consider sampling in some detail next.

0.1.1 Sampling: what, who (and how) to choose

Sampling is the process of selecting a subset• for further analysis from an infeasibly large population of interest.

Sampling assumes that there is a "sampling frame" as data source: a sub-set of the collection of the population of interest from which your sample is taken. In some sense, you have already experienced sampling as part of your literature review•. Unless you had infinite amounts of time – which you didn't – and infinite patience – which you might have – you could never be 100% certain that your literature search collected *all* relevant papers: the search space is infeasibly large (and not indexed particularly well). But you were systematic and achieved a practically good• coverage because of that.

More generally, sampling is used when we wish to study a population of interest, but this is infeasibly large or inaccessible for us to be able to study every single member of it. Instead, we choose a sample which is somewhat representative of the population characteristics, hoping that by studying the sample we can establish some properties or patterns of interest which can be assumed true of the population as a whole.

Broadly speaking, sampling can either be random or non-random.

In *random sampling*• some unbiased way of choosing the subset members from the population must be established upfront to inform sample collection, which is usually completed prior to any analysis. This is used particularly in quantitative research, and when generalisation of the results to the population is of primary importance, something enabled by the lack of bias in the sample selection process.

Instead, in *non-random sampling*•, the sample choice is based on the researcher's judgement and discretion, so that some element of bias may exist. Members can be added to the sample as the research progresses, interleaving data collection and analysis, until not more collection is possible or *saturation* is reached, that is collecting more data would not bring more relevant information. This kind of sampling is used particularly in qualitative research, where depth and richness of results are more important than the ability to generalise.

Random sampling techniques include:

simple random sampling where each member of the population has exactly the same chance to be se-

• We could have said "sample" but that would have been circular.

• You might remember the relatively complex procedures for recording search terms, discovered papers, their relationships, and your growing collection of notes on them.

• By *practically good*, we mean you found the most of the most important papers, some other papers, and didn't have to read *every single paper*. I.e., you found a *representative sample*.

• Random sampling is also called *probability sampling*.

• Non-random sampling is also called *non-probability sampling*.

lected. It has the advantage that it is easy to implement, and given the complete randomness of the sample, generalisation is fairly reliable. However, it can be time consuming if the population if very large, and may not lead to a representative sample if the population has large sub-groups, which may be over-represented in the sample, with minority groups being under-represented.

stratified random sampling where sub-groups of the population are identified based on common characteristics, the *strata*, and sampling is random across those strata. The strata are not mutually exclusive: for instance, the population may have sub-groups defined by gender, ethnicity and level of education, which may overlap. This approach overcomes the over/under representation problem of simple random sampling; however, deciding on the strata may be difficult and will also complicate data analysis.

cluster sampling where the population is divided up in naturally occurring separate clusters, and the sample is obtained by randomly selecting some clusters and then randomly selecting members of those clusters. It is more cost-efficient than the other two approaches, but can introduce bias if the selected clusters are not representative of the whole population, so that the over/under representation problem remains.

Non-random sampling techniques include:

purposive sampling in which participants are selected by the researcher based on particular characteristics, knowledge, or expertise they have. It is often used for small, rare or unique populations, and is particularly suited to studies which intend to be deep and narrow, and for which generalisation to the population is not the main concern. As the sample choice is made by the researcher, it is prone to bias. However, it also allows the researcher to involve participants who can provide insights into such rare or unique groups.

convenience sampling where participants are selected based on their availability or accessibility. This is quick and easy, but unlikely to produce a representative sample, so, once again, bias is an issue.

snowball sampling which relies on referral from previous participants to recruit new ones. This is an effective approach when a population is difficult to access or when the topic is sensitive or tabu. This too is unlikely to generate a representative sample, and is prone to bias. However, it is a way to gain access to members of a population which may be otherwise inaccessible.

In summary, when choosing a sample, you need to consider various factors, including the aim of your study, the kind of methods you are applying, and the level of access you may have. Trade-offs are likely

involved and you may not be able to obtain an ideal sample. Nevertheless, your sample will still be useful to your research, as long as you clearly explain and justify how it was obtained and what its limitations are.

Activity: Deep reading on sampling #2

Check back to your choice of research strategy. If you've chosen one which may require sampling, then you should go deeper into this topic to ensure you select the right kind of sampling for your study. We recommend you start by reading the following: **<empty citation>**

Guidance

The suggested reading is only a starting point. You should go deeper into the specific kind of sampling you are most likely to apply.

Activity: Your chosen sampling approach #3

Assuming your study requires you to perform some sampling, write down the sampling approach you are going to apple, with its justification in terms of your aim and objectives, and any trade-offs due to the practicality of accessing the sample. Record any possible weakness or limitation of your chosen approach, and how you will address them in your project.

Guidance

You can skip this activity if sampling is not required by your choice of research strategy.

0.1.2 Modern standards

Modern standards of research often require that your data be made available to other researchers so that your research can be verified or even rerun. In fact, it is increasingly the case that data sets are published and shared by entire research communities, often used as testbeds or benchmarks for new knowledge contributions. For instance, in medical applications of Machine Learning, in which new knowledge can be the fractional improvement of the performance of an AI algorithm to, say, diagnose a medical condition from images, not being able to share the data set you have used in your research can negate your knowledge contribution. Therefore, you should consider whether your data (or a sample of it) should appear as an appendix to your

dissertation, or even whether it should made available in its entirety to the wider research community, and how.

Modern standards of research also require you to comply with regulations on data privacy and protection•, so that as part of your data generation process you should also consider the need to anonymise data• without losing their research value for your project or the need to release commercial in confidence or otherwise sensitive data.

• We discussed GDPR in Section **??**.

• There may be a route to your degree in which your thesis is not published. Should you have any concerns about the release of data, please do find time to discuss this with your supervisor, balancing the needs for total anonymity with those than can be achieved through data anonymisation.

0.2 Data generation methods

In this section, we step through the most used data generation methods – those that were mentioned within the research strategies in Stage **??**, i.e., Interviews, Journalling, Observations, Questionnaires, Documents, Focus groups, Field work, Computational thinking, Mathematical thinking, and Statistical thinking.

cross-check with stage 3 if this is the full list

Most of these methods concern *empirical data*, that is data that is gathered through our five senses or from experience, and then used as the benchmark against which theories and advances in knowledge are made.

0.2.1 Observations

Activity: Do I need to know about Observations? #4

Check back to your chosen research strategy from Stage 3. Does it involve data generation using Observations? If so, read through this section and complete the activities.

Observations constitute one of the main ways in which empirical data are generated. In fact, according to **marvasti2014analysing**, "Observation is the foundation of science."

This methods requires the researcher to make their observations• of a phenomenon of interest in its setting. Such observations can be made directly, through our naked senses, or through instruments which enhance our sensory capabilities, such as a telescope, a microscope or the "a myriad of other ingenious inventions designed to make the invisible visible, the evanescent permanent, and the abstract concrete" **daston2011introduction**. Quantitative observations, say, the size or weight of an object, are usually referred to as 'measurements'.

• Hence the name...

Whole books have been written on observations as a research method•; we can give but a shallow

• See below for definitive sources you can use here.

introduction.

0.2.1.1 Observation tools

What to observe What can be observed is core to the characteristics of a phenomenon, and so it is the phenomena that you have identified as part of your research problem that you will be observing, whether they are interactions between atoms, between school children, or even your own actions, thoughts, and – perhaps even – biases. It may be that, in the course of your research, you will observe related or proxy phenomena, as needs must. Part of your observation skills will be to identify those related to and proxies for, to record those observations and to iterate those records into structures that capture the characteristics that are necessary to do the research on their back.

Observation types: natural, artificial, social Observations are versatile tools for almost any research domain, and your own domain will determine what sort of observations you will make. Observations range across *natural phenomena* – such as the different proportions of plant species that populate a train station wilderness garden – through the *artificial phenomena* – the way that buses drop off and pick up their passengers at a train station – to observations of *social phenomena* – the different ways in which a train station is used by commuters in the morning and the evening – as well as more complex combinations of each. Each will use different observations techniques and tools, and each with different constraints.

More on social observations Observations can be either *overt* or *covert*, depending on whether people are or not aware of being observed. Covert observations have the advantage that people's behaviour is not affected by their awareness of being observed, but, of course, they raise some important ethical and legal issues, such as the lack of informed consent or in relation to privacy and anonymity. If the judgement is that the phenomena observed do require privacy – perhaps you wish to observe commuters' use of restrooms in the train station or customers in a betting shop – then you must explicitly ask for permission – or change your research problem! Otherwise, in a public space, there may be no overriding expectation of privacy and observations can be done without explicit consent. Your university is likely to have strict regulations on the matter, or even prevent you from conducting covert observations as part of your research. In addition, as an observer of people, you can act as a *participant observer* – a researcher that interacts as a member of a community, in effect "living" alongside those you observe – you might be a commuter that uses the train station and so observe others in their use.

0.2.1.2 Workflow

Observations can be conducted on nearly any subject matter, and the kinds of observations you will do depend on your research question. Depending on the nature of the phenomena you are observing, you will prepare, run, and analyse your observations. When conducting observations, it is important to record only the events that are directly observable and to separate observations from interpretations to eliminate bias in your notes[•].

Paragraph adapted from AI

Data analysis was/is dealt with where?

• The right way to do this follows later in this Stage.

Preparing to observe To be able to observe you will need to:

- choose which phenomena: whether natural, artificial, or social
- if social: then choose which mode – either *participant* or *unobtrusive*.

In the case of the participant observer, where you are located in time and space will be determined by your participation. If you are a participant observer of train station usage, for instance, then you will make your observations whenever you use the train station. It may be that you make exaggerated use of the train station[•] to condense many months of participant observation into weeks or days – after all, your research project is time bounded – perhaps visiting ten times per day rather than just two. Keeping use to those of a participant is the key.

In the case of the other observations, you must determine the time and place at which those observations will be made.

• You could eat there, for instance, or use any shops that are colocated.

Unassisted observations if you are able to observe using only your senses – sight, smell, taste, touch, or feel – then you must be colocated in time and space with the phenomena you wish to observe.

Assisted observations many phenomena do not permit observation through the senses unassisted – for instance, the search for exoplanets[•] requires complex and delicate instruments which will be located on mountain tops, for instance. In this case, the availability of the equipment you will use will determine when and how you make your observations. In this case, you must plan to have and gain access to that equipment, and your other research tasks may have to fit around that access.

Records – created with the aid of technology, for instance, video recording of phenomena – may allow you to time-shift your observations to more convenient time. In addition, there may already be records made of the phenomena you wish to observe that are available to you[•].

• For instance, **jones2008exoplanets**. But don't let this exciting mission to explore strange new worlds; to seek out new life and new civilizations; to boldly go where no one has gone before distract you. Too much.

• In which case, `youtube` might be your friend.

Semi-assisted observations although some phenomena do not permit direct sense observations, increasingly the sense-prostheses needed to make observations is sufficiently portable to make colocated observations possible. Indeed, it may be that a smartphone is all you need to make excellent observations.

In any case, if you need equipment, you must ensure that you have access to it at a time and in a location that fits your plan•.

• Remember, plans never survive first contact with reality, so also plan to have a backup plan that you can use should your first plan to make your observations fail!

0.2.1.3 Other things to think about

Although you might see observation as being and seeing, observation is, by no means an easy way of generating research data. There are many issues that can arise, including:

- Observations can be influenced by the observer-observed relationship, overfamiliarity can, for instance, lead to observer bias such as in making of natural but undue inferences – see below
- likewise, maintaining analytic detachment is important in observational research to avoid bias.
- Different analysts, or different stages of analysis as your research progresses, may focus on different aspects of the same data. This can be a good thing should you analysis deepen due to understanding more about your observations, but may also lead to analysis drift or even "paralysis through analysis" in which no progress is made due to too much depth. To avoid this, keep a clear eye on the prize: your research goal, and set regular times at which you can reflect on progress.
- The analysis of observations depends on the researcher's chosen focus and philosophical and analytical framework. This is a natural dependency and can give great richness to observations, even those that are taken from the record. However, they can also lead to overthinking and, like the previous item, paralysis, as well as to prose that is obtuse and disengaging for the reader•.

• And examiner! Consider, also, that the examiner might not share your focus and that this difference might be sufficient for them to discount your work as without value. Not likely, but care is needed.

More on observer bias Bias in general is treated in Section **??**. Here we add something to that general description about observer bias that gives a tool for avoiding it, or at least being aware of it: the "double-entry notebook"•.

• See **driscoll2011introduction**; example on page 162.

Notebooks: avoiding observer bias The double entry notebook separates pure observation from judgemental observation, in which value judgements are made.

- pure observation: what can be determined from the observation alone, i.e., what can be determined without insight into the mindset of the person observed;
- judgemental observation: any additional data that is an inference by the observer – observing someone "waiting impatiently for the train door to open" is ascribing feelings to the person observe which, by their nature, are hidden from the observer.

Might need better term, to contrast with pure observation.

You may, of course, say that impatience is observable – and, indeed, it might be a valid inference from a statement such as "finally, the train has arrived!". The double–entry notebook allows such inferences into the mindset of the person observed – as well as other value judgements on their demeanour – so that another researcher reading your notes on observation can clearly differentiate direct observations from such inferences. You do not lose the ability to infer with a double-entry notebook.

Source **driscoll2011introduction**.

Activity: Deep dive into observations #5

To find out more about observations, take a look at these resources:
daston2011introduction;
marvasti2014analysing;
simpson2003using;
driscoll2011introduction;
angrosino2003observations;
sapsford1996data

0.2.2 Interviews

Activity: Do I need to know about interviews? #6

Check back to your chosen research strategy from Stage 3. Does it involve data generation using interviews? If so, read through this section and complete the activities.

Interviews are a versatile tool for generating data from participants by asking questions and recording detailed answers. Factors to consider in Interviewing including ensuring that you have chosen appropriately

[This section ADAPTED from `https://blog.scrintal.com/the-complete-guide-to-conducting-research-interviews-7a0027d02e0a`]

those that will take part, the questions you will ask and their structure, the tools that you need to take full value from the data generation, and how you will followup.

The personal approach that is a characteristics of interviews means that they are a great way of accessing a (group of) individuals' feelings, thoughts, ideas, and/or experiences, data that can be difficult to generate in other ways. They can also provide direction for new research by giving expert indications of where problems lie in a particular domain. As such, interviews can often be used as a way into a discipline, filling in useful background through personal experiences, and having access to otherwise difficult to access information.

Interviews do not need to be co-located – a video or audio link is all that is needed – but they do need to be in real-time, so that you must

0.2.2.1 Interview Tools

There are three main kinds of research interviews.

The structured interview The structure of a structured interview serves as a repeatable framework by which each participant can be asked the same questions in the same way. There is no scope for deviation from the structure, so that auxiliary questions and follow-ups are not used.

Structured interviews are the closest to being time bounded and predictable; if you only have 8 hours to conduct 24 interviews for instance, a structured interview would be the best way to achieve this.

Your skills as an interviewer will be tested by structured interviews: it is often difficult not to stray outside of the structure when an interesting answer is given, and you may have to cut a participant short if their answers overrun or diverge from the structure•.

• In our experience, this is a perennial problem, so don't underestimate the difficulties you will face as an interviewer.

The semi-structured interview The structure of s semi-structured interview serves to identify areas of interest to the researcher, with interesting responses being welcomed and followed up if appropriate. Interviewing a domain expert on your chosen topic would be well-served by the semi-structured interview as their expert knowledge could be probed with follow up questions.

The semi-structured interview does not naturally time-bound the interaction, and so – if you don't have unbounded amounts of time – you will have to balance the breadth of questions with the depth of responses.

The unstructured interview Although the interview may begin with the same question each time, there is no structure to constrain the route through the data that the participant wishes to take. The unstructured

interview requires the least domain knowledge of the interviewer as the direction can wholly be given by the participant, but retaining forward motion and focus can be a challenge.

0.2.2.2 Workflow

Your characteristics will determine whether you choose structure, semi-structured, or the unstructured form. Essentially, if you:

- already have a good level of domain knowledge then you can use on the less structured interview models, as you will be able to follow your participants' answers more easily and direct their comments towards your research areas[•]. If you don't already have a good level of domain knowledge, then you will be putting more effort into the design of the interview, so that you can simply capture your participants' responses to your – well-designed – questions.

  [•] Of course, you can always use structured interviews even if you do have domain knowledge – there's nothing to stop you.

- need deep insight then, again, you should use the less structured interview forms as these can allow you to probe your participants responses[•].

  [•] The less structured interview form you choose, the more domain knowledge you'll need. See the previous item.

- more here...

Your next step is to identify whom to interview – see the section on sampling above for guidance.

Additional guidance the number of participants in your interviews can not usually be predetermined: in the perfect case, you should interview until the responses you are receiving are "guessable"[•]. Of course, you will most likely have: i) limited time to complete your data generation; ii) a small population, so that you can interview everyone of interest; iii) other constraints???

[•] I.e., you no longer get novel answers to your questions, indicating that the topic has been covered.

Having decided on the participants, you should contact them. Your university may have strict guidelines on how to approach and work with interviews which you should investigate as a matter of urgency[•]. They're certain to contain issues around GDPR, informed consent, ethical guidelines, and the like.

[•] It's a surefire way of failing a research module, not following university requirements.

Activity: Regs	#7
Guidance	
TBD	

Before your interviews Prepare a draft of the questions you intend to ask. find a friendly test subject for your questions and do a dummy run of your interview. Use their feedback to improve your list of questions:

- did you put the participant at ease during the questionnaire: if not their nervousness might influence their ability to contribute;
- which questions worked well: i.e., they led to useful responses;
- which questions were confusing or led to unhelpful answers: in which case do you need only to reword them – perhaps with the help of your participant•, by having alternative versions of the questions, or by removing or replacing the question.
- (if time constrained) which questions overran: can you rephrased the question to be less "open"?
- (if structured) were you able to keep to your "script": how will you resist the temptation to problem more deeply? Or do you need to consider moving to a semi-structured or unstructured form?
- did you remember to record the answers in a useful form? Were you writing notes – if so did you capture everything of interest?• Did you audio-record the interview? Did the recording equipment work well? Is there another way of getting a transcript?
- are the questions in a logical order? Were they grouped by topic? Did they build responses in the most productive way?
- are there other questions you should have asked?
- other?

Repeat this with as many willing participants as you can until you're happy with the interview format•. Then run this form past your supervisor – they will be sure to have comments, perhaps from a more researcher perspective.

• "I would have asked it this way"...

• Longhand notes can be taken at 35 words per minute; spoken text is often as fast as 120 words per minute.

• Or until you run out of willing participants, or time! Make sure that you do not dip into your target population for these preliminaries.

0.2.2.3 Other things to think about

Where will you hold the interview you will need to ensure that you have an appropriately comfortable venue for your interview which, if to be colocated with your participant, access to comfort facilities. Public spaces – where you could share a coffee, for instance – may create a more immediate feeling of intimacy, and

so deeper responses, but background noise might interfere with your record keeping., so make sure to check the venue out at the appropriate time of day to ensure a recording device can handle any difficulties.

Protecting your data What would happen if your audio recording device went wrong? Would you have a backup or the interview? What if you forgot to turn it on?• Write a checklist of instructions for yourself to follow before and after each interview, so that you can be sure not to miss anything important. This checklist should include any permission issues, such as permission to record the interview.

• Oh so easy to do...

Being a good interviewer Try, to the extent possible, given the format, to allow your participant to govern the speed and direction of the interview. Allow them to talk in complete sentences without interruption, or have a good reason to interrupt. If you need to interrupt, apologise for doing so and tell them the reason why you have done so•. Be polite and encouraging, as your participant might be nervous.

• "I'm sorry to have to interrupt, but we only have 5 minutes left and ...".

At the end of the interview Say what will happen next. Perhaps share a sheet which describes the research you are doing•

• Or do this at the beginning, if you are willing to give the participant time to read it as part of the interview.

Activity: Deep dive into interviews #8

To find out more about interviews, take a look at these resources:
oates2008researching;
johannesson2014research;
secor2010social;
hays2003case;
mcclure2002common;
peoples2020write;
jorgensen2001grounded;
hycner1985some; **englander2012interview**; **ramsook2018methodological**;
robertson2002automated;
kielmann2012introduction

0.2.3 Journalling

This section currently adapted from **hayman2012journaling**.

Activity: Do I need to know about Journalling? #9

Check back to your chosen research strategy from Stage 3. Does it involve data generation using Journalling? If so, read through this section and complete the activities.

Journalling asks research participants to record and reflect on research-focussed experiences as "a way of thinking, understanding, and learning." (**hayman2012journaling**) One popular use of journalling is as a learning strategy for research participants, as so is often located in ethnographic research, or in data generation for grounded theory. The researcher is interested in the ability of journalling to expose elements and processes involved in deep learning, problem solving, critical thinking development, and reasoning.

0.2.3.1 Journalling tools

Journaling is an activity of recording personal thoughts, daily events, evolving insights, etc. As such the tools are basic and readily available, ranging from a hand-written diary• to a computer document.

• Of course, accessing hand-written notes will be more labour intensive than electronic ones.

Whatever the technology used, the main tool for journaling is, unsurprisingly, the journal as a record of the phenomena of interest: one or more journals can be used for different aspects of the journaling process. This should be organised to provide effective writing prompts, and ways of uncovering gaps in the journaler's perceptions of relevant phenomena. Techniques include asking questions, commenting on journal entries and, to be more engaging• using colours, stickers, and other distinguished comments.

• Perhaps for younger journalers...

The use of journals by the research and by the journaler might be different:

- the researcher will be looking for research data
- the journaler might use the journal for reflection on practice, and development as experience develops

The analysis of journaling involves thematic analysis and, perhaps, multiple rounds of coding (see the entry for documents **??**).

0.2.3.2 Workflow

Journaling or *self-reflective writing* can help research participants to reconstruct their thoughts from an activity or situation, indicating how they develop over time.

Adapted from **giguere2012self-reflective**.

The structure of a self-reflective portfolio should include

- goals of the reflection: established once at the beginning of the journaling exercise
- journaling entries contributed by participants: the use of these is to help the participant reconstruct their thoughts and development from the research exercise. To ensure that the researcher goal of the journal is approached, the researcher should develop journaling prompts. They can be broader or narrower in scope, and guide the focus of the participant onto journal entries that will be useful for the research goals.

  There are three identified challenges for journalling (**hayman2012journaling**):

  1. poor participation: as a researcher, you want to collect as rich a journal from each participant as possible. To assist in this goal, any guidance provided should be sincere, nonjudgmental, and conversational.
  2. feeling exposed: a participant may feel vulnerable or anxious documenting "sometimes intimate details of their lives" or by a perceived lack of writing skills.

     One method of allaying fears associated with journaling is to provide clear instructions, identify specific objectives, and provide ongoing support and guidance **taylor2006research**.
  3. staying on track: Journaling "momentum" of the participant should be monitored, if possible, during the research process. A participant that becomes distracted or demotivated might be helped to refocus their efforts through restatement of the research goal, with prompts repeated as necessary.

0.2.3.3 Other things to think about

Activity: Deep dive into journalling #10

To find out more about journalling, take a look at these resources:
kadarisman2017classroom;
burns2009action;
james2005journaling;
peoples2020write;
feinblum2016journaling;
hayman2012journaling;
mcgrath202115;

taylor2006research;
ovens2020weaving;
giguere2012self-reflective;
bacon2014journaling

0.2.4 Questionnaires

Activity: Do I need to know about questionnaires? #11

Check back to your chosen research strategy from Stage 3. Does it involve data generation using questionnaires? If so, read through this section and complete the activities.

Questionnaires• are versatile tools for generating data from participants by asking questions•. Questionnaires allow a researcher to characterise a population of interest by collecting participants' answers about their attitudes, preferences, opinions, behaviours, etc. You might use a questionnaire as a way of collecting statistically significant responses from a population sample, but there are other uses as well.

If you do use a questionnaire, its thoughtful design is of critical importance. Otherwise, you might be asking your (willing) sample to spend a considerable amount of their valuable time answering questions the content of which are not helpful for your research. As they might not be so willing to help a second time, getting the questions right• the first time is important.

Administering questionnaires are nowhere near as difficult as they used to be as the number of online resources for doing so increases. And, probably because of this, there are plenty of resources in the literature and online to help you design your questions.Their descriptions can be a little technical, however, so the following glossary and other tough points might help you engage with them better.

• Questionnaires are just one in a rich collection of *survey tools*, others of which are described below.

• There's a hint in the name – *question*naire – although why two "n"s; does no millionaire, billionaire, or debonaire use them?

Why population here?

• Often called *questionnaire design*, although this conjures up glossy format and whizzy web-pages which is of secondary importance. Unless, your questionnaire is about the design of questionnaires, of course.

Essential questions the smallest possible set of questions you absolutely need to ask to address your research aim and objectives. While using several questions will give you richer data sets, long questionnaires tend to put people off, so that fewer people may be willing to participate.

Profiling questions questions tat ensure your respondents match specific characteristics you are interested in: say, you are studying the usability of a new product, then you will need to know the extent your

respondents have engaged with that product. This is particularly the case if you are running a large survey and don't know who is going to respond.

Demographic questions often used so that you can then compare answers across different sub-groups, say, based on gender, age or ethnicity, etc.

Language questions should be clear and plain, and you should avoid jargon and idioms, to ensure your participants understand what you are asking, particularly if not native speakers. Your questions should also be objectives, that is you should avoid any judgemental term or tone, which may lead participants to answer in a particular way, or make assumptions about your respondents' habits or behaviours: for instance, asking participants what they eat for breakfast, assumes they all take breakfast, which may not be the case.

Double-barrelled questions also termed "compound", these are questions that ask more than one thing, while only allowing one answer. These should be avoided as it would be difficult, if not impossible, to establish in your analysis which part of the question each participant has answered. Instead, you should split the question into separate questions each addressing a specific thing.

Response options questions are broadly divided into *closed* and *open*-ended. Close questions restrict the possible responses to a set of given choices, while open questions allow respondents to use their own words freely to answer the question. If you use closed questions you should ensure that the possible answers cover all possible options• and exclusive, in the sense they don't overlap. Open questions can lead to richer answers, but you must ensure they are sufficiently constrained so that the answers don't end up being without value by being, for instance, vague or off topic.

• Or, at least those you're interested in.

Scales if your questions require participants to estimate or measure something, you need to worry about both validity and reliability when setting up the scales for possible answers. Validity means that the chosen scale should allow respondents to measure something accurately; while reliability means that, under the same conditions, respondents will be able to measure something consistently.

This might need rephrasing, as questionnaire weakness?

Question grouping, ordering and flow related questions• should be grouped together, and the flow between groups of questions should be logical. Question order in each group also matters: as a rule of thumb, simpler questions should precede more complex ones.

• For instance, those intended to establish a demographic of respondents.

0.2.4.1 Tools for creating (maintaining, and analysing) questionnaires

While you can design your questionnaires from scratch using your word processor, there are plenty of specialised digital tools, many of which are free, that can make it a lot easier•. They usually come with: templates and pre-defined question types that you can customise for your study; statistical analysis and data visualisation features that you can apply to the data you have collected; export functions that allow you to save the data to a spreadsheet for further analysis. Overall, if you need to develop questionnaires for your research, they can really help you speed up the process, so that it's well-worth the investment of time in climbing their learning curve.

• Examples include: add list here and URLs.

0.2.4.2 A simple questionnaire design workflow

More here?

Following the guidelines above, for the order and content of questions, it's very easy to complete a first draft of a questionnaire. Unless you have many years of experience in questionnaire design, your first draft will be far from suitable. Indeed, releasing your first draft without further thought may lead to you not only generating no useful data from it, but also p-ing off your audience sufficiently that they are not willing even to look at your second version.

So, once you have a first draft of your questionnaire, you should test it and refine it.

Early testing can be done by asking a friend, a family member•, or a colleague to work through the questions, provide their answers and any other feedback they might have. This will give you early indications of problems with your questionnaire•. Later in the process of designing it, however, you should take expert advice including, of course, that of your supervisor, to get to the final agreed form.

• Probably, but not always a friend:)

In addition, you could pilot your questionnaire on a small number of respondents first, then revise it as necessary before using it more widely.

In all cases, you are looking for evidence that there are issues, including confusion – which may point to a lack of clarity in the questions – or hesitation – which may point to a poor choice of response options or to inappropriate scales – or disengagement – which may point to too many questions being asked – may have occurred.

Be sure to loop back to those that have helped you to check that you have addressed their comments.

• Although it's sometimes difficult, you'll make more progress and quicker if you think of the questionnaire as imperfect, rather than you. You can then apply comments – even if they are negative – to the questionnaire rather than having a personal emotional reaction to them. For each comment, make sure you understand how it can be addressed in your questionnaire. This last tip also means that you can welcome (but ignore) comments that can't be addressed.

How can you check this with a non-colocated sample? Does this not need to go earlier?

ADD Extra reading

Activity: Improving your questionnaire design #12

Assuming your study requires you to use a questionnaire to collect data, consider each of the points above in relation to your draft questionnaire, making improvements whenever required.

Guidance

You can skip this activity if questionnaires are not relevant to your project. Reflecting on each of the points will help you avoid common mistakes and improve your questionnaire design.

Activity: Deep dive into questionnaires #13

To find out more about questionnaires, take a look at these resources:
oates2008researching;
hays2003case;
burns2009action; **mcclure2002common**;
najafi2016observation;
robertson2002automated;
kielmann2012introduction

0.2.5 Documents

Activity: Do I need to know about documents? #14

Check back to your chosen research strategy from Stage 3. Does it involve data generation using documents? If so, read through this section and complete the activities.

As a researcher, documents – in the form of academic articles – will already occupy a large proportion of your time/brain/computer. Your collection of academic papers could currently be as many as 50 or more. You will, therefore, already have good experience of interacting with documents and those interactions may well have been productive.

Other documents can also be used as primary research vehicles when used a part of a research strategy. In this sense, documents• are written records or figures produced as a by-product of the operation of an organisation, or part thereof. They will typically consist of literary, textual, or visual devices that enable organisational information to be shared and stories to be presented. They are social artefacts created for a particular purpose, crafted according to social conventions that apply at the time and place of writing, and can be used directly for their factual content or indirectly to uncover what they reveal about the writer or as examples of a specific form of presentation.

• Main sources: **coffey2014analysing**; **finnegan2006using**.

Documents can be in various forms such as words/text, pictures, or audio, and are stored on physical carriers like paper, stone, or microfiche. They have narrative structure, imposed by the writer, have cultural ways of telling, in a particular style, structure, and language. They also employ visual signs, literary devices, and symbols to present and display meaning. Documents are rarely produced and read in isolation and are often related to other texts, allowing for exploration of relationships and meanings within a text and in relation to other documents.

0.2.5.1 Document tools

The way in which you interact with documents will lead you more or less quickly to success of your document review at the end of which you should have a good feeling of the relationship between documents, their formal properties, and rhetorical features.

Researchers use documentary sources to collect and analyse data. To do this, a researcher will interact with them either *directly* or *indirectly*:

direct in which a document is analysed for its factual content: the data it describes;

indirect in which a document is analysed for the data it contains

The difference between these two modes may not be clear. An example might help. This is a copy of a passage from **fynes1873the-miners**:

Quote

Miner :– I believe you have something like 150 collieries to inspect?
Mr. Dunn :– Yes.
Miner :– Twenty-eight in Cumberland?
Mr. Dunn :– Yes.

> Miner :– Do you think you are able to inspect all these?
> Mr. Dunn :– Well, the Government thinks I am able, you know.
> Another Miner :– Were you satisfied with the one shaft at this colliery, if so there is an end to the matter; if not, what steps did you take to remedy the defect? Did you apply to the Secretary of State, showing him that it was defective?
> Mr. Dunn :– At this very moment there are three of the largest collieries in Northumberland – Seaton Delaval, North Seaton, and Newsham – managed by the most talented men in Northumberland, all with single shafts. Now, what would you have me to do? Do you think it is my duty to call in question the management of these pits?
> Miner :– Am I to understand this is an answer to my question?
> Mr. Dunn :– Well, I am not so well satisfied as if they had two, but I have not the power to alter it.

Here, a direct reading could be to identify collieries in which a single shaft existed at that time. An indirect reading could be to explore social relationships in the 18th century.

In more detail, indirect and direct reading use the following techniques to classify and conceptualise documents for qualitative research:

- Studying the form, content, and function of documents
- Analysing the ways in which documents justify decisions, display hierarchies, and exercise agency
- Distinguishing between primary sources (documents providing raw data) and secondary sources (documents providing information about primary sources)

0.2.5.2 Workflow

As we have already mentioned, direct interaction with documents is already something with which you are[•] familiar. However, according to **coffey2014analysing**[•], a good place to start in analysing documents is with the realisation that documents are "socially defined, produced, and consumed" so that, alongside the content, the processes by which a document is produced and is intended to be consumed, i.e., by which it becomes an "accomplishment" for some individual or organisation, often contain useful data.

In this case, there are many approaches possible from the simplest – counting instances, for instance, of words, phrases, or other elements[•] – and indexing and coding document elements to identify thematic

[•] Or, should be!

[•] Freely available from **coffey2014analysing**.

[•] **mckenzie1959shakespearian**, for instance, counts commas – as well as other punctuation – in Shakespear's *first Folio* to help identify which compositors were responsible for which parts of it.

content and identify patterns, ala **schreier2014qualitative**.

schreier2014qualitative's approach, a type of the more formal *qualitative content analysis*, involves the selection of material from a wide range of source with adequate coverage for the topic of interest. As such it can result ing large amounts of material to analyse and so an element of formal in the process is mandated. An important step in this is the building of a *coding frame*, i.e., a structure consisting of *main category* and two *sub-categories* (**schreier2014qualitative**):

- a main category[•] being those aspects of the material that interest the researcher,
- a subcategory being what is said in the material with respect to a main categories.

• Coding frames with more than 40 main categories are known to be difficult for one person to handle at one time.

Material selection It is important to choose material that covers all types of data and data sources in the proposed study. The material should be selected in a way that allows the majority of categories in the coding frame to be applied during the analysis. Once material has been selected, **schreier2014qualitative** recommends breaking down the material into smaller "chunks" and build the coding frame for one *chunk* at a time.

The process described by Schreier is circular and will need linearising to make better sense

Pilot phase is where the current draft of the coding frames categories and subcategories are tested and refined. After selecting and preparing an appropriate subset of the selected texts[•], trials of the encoding.

• This being one that will populate all categories and allow exploration of the subcategories.

Coding for keeps once you're convinced that your coding frame is stable, you then apply it to *all* documents. This will involve:

- paying attention to how documents are constructed as distinctive artefacts, their structure, vocabulary, level of formality, etc. (**coffey2014analysing**)
- analysing their form and content, considering how they were produced and how they were – and how they were intended to be – consumed
- considering the relationships *between* documents, highlighting dimensions of similarity, comparison, contrast, and difference by tracing how texts refer to other texts, sharing conventional formats, and constructing a uniform style[•]
- exploring the ways in which they function and are used in everyday life and social context
- examine their formal properties, including their linguistic registers, and rhetorical features.

• This scaled *Intertextuality* in the literature (**coffey2014analysing**); it's akin to how you went about analysing the academic literature as part of your literature survey.

Presenting your findings can be as simple as delivering the statistics that you have collected as the basis for further analysis. Alternatively, a completed coding frame with an accompanying narrative and example quotes as needed could be the output. Like other data, your coding frame can also be the starting point of further data generation and analysis, "examining the results [...] for patterns and co-occurrences of selected categories." (**schreier2014qualitative**)

0.2.5.3 Other things to think about

When analysing documents, it is important to consider the form and function of the documents themselves. Instead of focusing on whether a document is 'true' or 'valid' evidence, one should examine the formal properties and rhetorical features of texts. Documents should be analysed in relation to their production and consumption, as well as their intertextuality and relationships with other documents. It is also crucial to understand how documents function in everyday life and social contexts.

AI

Activity: Deep dive into documents #15

To find out more about documents, take a look at these resources:
coffey2014analysing;
schreier2014qualitative

0.2.6 Focus groups

Focus groups are a qualitative research method that allows researchers to generate detailed and valid data by engaging participants in interactive discussions. These discussions help in understanding complex phenomena and generating new hypotheses for further study or practice. The focus group methodology involves group composition, managing group discussions, and analysing results. The focus group's advantage lies in its ability to quickly identify a full range of perspectives held by participants and to encourage interaction among participants to expand on their contributions.

Adapted from AI summaries

Focus groups combine interviewing, participant observation, and group interaction, making them useful for uncovering data and ideas that may not come up in one-on-one questioning. As such, focus group participants should share intersectionality.

Focus groups interactions are made by a moderator that may or may not be the researcher themselves. Key qualities of a moderator include empathy, positive regard, ability to create a purposeful introduction, and effective use of pauses and probes.

To optimise data from focus groups, careful attention must be paid to the composition and number of groups, selection and training of the moderator, and development of the question route. The moderator plays a crucial role in facilitating group processes, establishing a permissive environment, maintaining focus, controlling participant interactions.

Activity: Do I need to know about focus groups? #16

Check back to your chosen research strategy from Stage 3. Does it involve data generation using focus groups? If so, read through this section and complete the activities.

0.2.6.1 Focus group tools

[STOLEN from wikipedia:
Variants of focus groups include:

- Two-way focus group - one focus group watches another focus group and discusses the observed interactions and conclusion
- Dual moderator focus group - one moderator ensures the session progresses smoothly, while another ensures that all the topics are covered
- Dueling moderator focus group (fencing-moderator) - two moderators deliberately take opposite sides on the issue under discussion
- Respondent moderator focus group - one and only one of the respondents is asked to act as the moderator temporarily
- Client participant focus groups - one or more client representatives participate in the discussion, either covertly or overtly
- Mini focus groups - groups are composed of four or five members rather than 6 to 12

- Teleconference focus groups - telephone network is used
- Creativity groups
- Band obsessive group
- Online focus groups - computers connected via the internet are used
- Phone/ web focus groups - live group conducted over the phone and online with 6 to 8 participants.

]

0.2.6.2 Workflow

Preparation The purpose of a focus group is to obtain data regarding ideas, attitudes, understanding, and perceptions on a specific topic, choosing participants that can contribute to these purposes is an important part of identifying participants. In particular, participants should be selected based on their experience and interest in the topic, rather than through random selection. Although the potential range of participants might be limited by context – you may need them to be selected from a small organisational group – the choice should be made so that they come from as diverse range of backgrounds, views, and experiences as possible. Focus groups should not, however, be chosen based on individuals suggested by fellow group members.

Having said that, the ideal number of participants in a focus group depends on the type of questions and the characteristics of the group, usually from between 8 to 12 participants•, although other sizes work too, the smallest useful size being 4 to 8. Anticipating subject loss, researchers should over-recruit participants by approximately 25%.

• Is a 12-person jury simply a focus group?

Focus groups are especially useful when the subject comprises a number of variables and psychosocial phenomena.

The Moderator Focus groups are facilitated by a moderator who guides the discussion. Being a moderator is a skilled role; if you do not have access to a skilled moderator, then accessing moderator training for yourself might be desirable•

• There are videos purporting to guide moderators on youtube.

It might be necessary to run more than one focus group, then they should be held in a series so that trends across different groups can be identified.

Environment Placing participants within an uncomfortable environment is likely to lead to negative outcomes. Given that a focus group might last for an extended period, time comfort should also be considered, with access to comfort facilities, etc, and an explicit timetable which includes breaks.

Defining the purpose of the focus group - Plan discussion topics

0.2.6.3 Other things to think about

Activity: Deep dive into focus groups #17

To find out more about focus groups, take a look at these resources:
powell1996focussmithson2000usingplummer2008focus

0.2.7 Field work

According to **wolcott2005art**•, fieldwork is a form of enquiry in which one is immersed personally in the ongoing social activities of some individual or group for the purposes of research. It involves personal involvement to achieve a level of understanding that will be shared with others. Fieldwork can include participant observation, interviewing, and archival research, all guided by sound human judgment, artistic decision, courtesy, and common sense. fieldwork encompasses everything the researcher will do in a study located in some field, i.e., some research context.

However, being in the field is not sufficient for fieldwork to be successful, although the physical location of the field does not need to be distant from the researcher's usual haunts. One could complete fieldwork on your university campus, or even your own garden, as well as in Iceland•.

Activity: Do I need to know about field work? #18

Check back to your chosen research strategy from Stage 3. Does it involve data generation using field work? If so, read through this section and complete the activities.

• **wolcott2005art**'s book is both detailed and entertaining. Reading it gives on a feeling that spending time in fieldwork with "Harry" would have been an education in itself. The book includes the importance of laundry to fieldwork, for instance, with experiences of fieldwork in a Canadian Indian reserve to illustrate.

AI summary

• Or on the moon – as some extremely lucky "researchers" have been able to do.

0.2.7.1 Field work tools

wolcott2005art offers the following tools needed by the researcher for successful fieldwork:

- Structured interviewing techniques: free listing, pile sorts, ratings, and rankings
- Use of computer software: questionnaire construction, data analysis, and text management
- Direct-observation data: coding behaviour, time allocation
- Social-network analysis: map structure, detect cliques, measure centrality, autocorrelation
- Detecting and visualising relationships among variables: measures of similarity and distance, multidimensional scaling, cluster analysis, correspondence analysis, and log-linear modelling

You'll imagine that records – primarily written (or typed), but also photographic or video, sound, etc – of fieldwork constitute the main raw data generated. These records – field notes – can have a number of forms:

From wikipedia, (**enwiki:1211911661**): better source needed?

Jot Notes Key words or phrases are written down while in the field.

Field Notes Proper A description of the physical context and the people involved, including their behaviour and non-verbal communication.

Methodological Notes New ideas that the researcher has on how to carry out the research project.

Journals and Diaries These notes• record the ethnographer's personal reactions, frustrations, and assessments of life and work in the field.

• Journals are covered in more detail below.

More here

0.2.7.2 Workflow

The quality of results obtained from field research depends on the data generated in the field. The data in turn, depend upon the field worker, their level of involvement, and ability to see and visualize things that other individuals visiting the area of study may fail to notice. The more open researchers are to new ideas, concepts, and things which they may not have seen in their own culture, the better will be the absorption of those ideas. Better grasping of such material means a better understanding of the forces of culture

operating in the area and the ways they modify the lives of the people under study. Social scientists (i.e. anthropologists, social psychologists, etc.) have always been taught to be free from ethnocentrism (i.e. the belief in the superiority of one's own ethnic group), when conducting any type of field research.

When humans themselves are the subject of study, protocols must be devised to reduce the risk of observer bias and the acquisition of too theoretical or idealized explanations of the workings of a culture. Participant observation, data collection, and survey research are examples of field research methods, in contrast to what is often called experimental or lab research.

0.2.7.3 Other things to think about

wolcott2005art, offering salutary advice, states that:

Quote

There may be discomfort and hardship aplenty connected with the experience, ranging from the distractions of diarrhoea or lost luggage to the despair of personal failure or lost hope, but the extent of one's suffering and sacrifice are not factored into judgments about the worth of the fieldwork as fieldwork.

so it may not all be plain sailing and, indeed, a risk assessment to an appropriate depth as well as the arrangement of insurance and other restorative things may be needed. If you have no budget for these, or are lacking the willingness to investigate and arrange them, then fieldwork will not be possible.

Fieldwork is an immersive data generation technique that can take substantial periods of time to complete periods as long as one year being not uncommon (**wolcott2005art**), sorter periods may require justification.

Activity: Deep dive into field work #19

To find out more about field work, take a look at these resources:
wolcott2005art;
randall2007fieldwork

0.2.8 Computational thinking

Clearly, part of computation thinking is programming a computer to achieve a goal.

However, **wing2006computational**[•], in her **wing2006computational** article that popularised Computational Thinking **wing2006computational** as a candidate curriculum antry, places programming in the broader context of computational thinking. If you are aiming at computational thinking as a way of focussing on your programming skills, it may be that you have misjudged this wider context.

wing2006computational says:

Computational thinking [...] has the following characteristics[•]:

Conceptualizing, not programming. Computer science is not [just] computer programming. Thinking like a computer scientist means more than being able to program a computer. It requires thinking at multiple levels of abstraction;

A way that humans, not computers, think. Computational thinking is a way humans solve problems; it is not trying to get humans to think like computers.

Complements and combines mathematical and engineering thinking. Computer science inherently draws on mathematical thinking, given that, like all sciences, its formal foundations rest on mathematics. Computer science inherently draws on engineering thinking, given that we build systems that interact with the real world.

Ideas, not artifacts. It's not just the software and hardware artifacts we produce that will be physically present everywhere and touch our lives all the time, it will be the computational concepts we use to approach and solve problems, manage our daily lives, and communicate and interact with other people.

Recognising the social context of computational thinking is, thus, critical to research in the area.

Activity: Do I need to know about computation thinking? #20

Check back to your chosen research strategy from Stage 3. Does it involve data generation using computation thinking? If so, read through this section and complete the activities.

The techniques generally associated with computational thinking are:

- Decomposition: breaking a problem into parts that are easier to solve;
- Pattern recognition & Generalisation: seeking significance from repeated structures within data

• **wing2006computational** is American, and so uses American spelling – -i*z*i-, art*i*fact, etc.

• **wing2006computational** includes other characteristics, of relevance to curricula but not so much to research.

- Abstraction: building higher level representations of things
- Algorithms: associating behaviours with computational structures
- Evaluation: called testing for software, but also has wider applicability when an algorithm is socialised, i.e., used within its final social context.

0.2.8.1 Computational thinking tools

Given the explosive growth in the use of computers over the past half century, you may not be surprised to hear that there are thousands of useful[•] computer tools to support computational thinking. Fortunately, for the vast majority of these, they support one or more computational models, which are the bases of computational thinking. Those computational models fit into the following categories

- computational mode: sequential, concurrent, agent based. Example tools: flowcharts, UML, Petri Nets,[•]
- computational paradigm: imperative, declarative. Examples: procedural, object oriented; logical, functional, reactive; etc.
- language: machine, low-level, high level, interpreted: machine code, C, Swift, Python.
- developmental process: lying on a continuum between plan-driven and micro-iterative
- requirements form: functional, non-functional;
- data structure: database, spreadsheet, json, etc.
- delivery platform: web, package, snippet, more here
- maintenance mechanism: GitHub,
- language libraries: React, Ruby on Rails, etc.[•]

[•] As well as some that are less than useful!

[•] Computational modes tend to be packaged for use, so – although the basis of a computational mode might be concurrent, you would not have to deal with the intricacies of modelling in, say, Petri nets. Unless, of course, your intention was to investigate their properties.

[•] Language libraries tend to come into and go out of date regularly as new paradigms of delivery are created and superseded.

0.2.8.2 Workflow

The main idea behind computational thinking is to link a wished-for behaviour to a computational structure. Here, we're thinking of the wished for behaviours of a client in context being translated to an algorithm, data model, or other computational element.

The ways in which the wished-for behaviour of a client in context are understood is through the discipline of requirements analysis. Although naive requirements analysis is possible for simple projects – asking a client what functional behaviours they need – it is more often driven by the explicit or implicit management or developmental risk, the risks being those of

- misunderstanding the context of deployment – where the final system will be installed
- misunderstanding the problem that is to be solved – what functional and non-functional behaviour is wished-for;
- misunderstanding the solution space – what technologies can be employed within the context and how they should integrate with human elements of the system
- misunderstanding of the change path by which any current system will become the new system
- misunderstanding the process needs for developing and delivering the system
- [more here]

0.2.8.3 Other things to think about

If you don't have experience with programming languages, then you can learn, but the learning curve can be very steep•.

Engaging with clients as stakeholders in a development can also be challenging, in that it makes great demands on both parties' time – there can be a great deal of iteration needed to get to a form where the client needs can be satisfied. Add to this validation of any delivered product by the client and a need for good project management skills come to the fore.

Often, developmental research projects constitute the research component of professional body-accredited degrees. Such degrees add requirements for good development practice including, for instance, documentation of the development process and its inputs, texting regimes, etc. These requirements can add substantially to the basis task of creating a solution, but aim to substantiate real-world problem solving.

• This video will give you some idea of what will be required: `https://www.youtube.com/watch?v=fLMZAHyrpyo`. Stephen Wolfram is a particularly interesting, if restless, speaker and a leader in this area.

But this isn't a great video, nor is any other that I've looked at, as they're created by computer scientists... They're also heavily advertise laden.

[More here]

Activity: Deep dive into computation thinking #21

To find out more about computation thinking, take a look at these resources:
angevine2017computational;
figueiredo2017improve;
lyon2020computational

0.2.9 Mathematical thinking

Mathematical thinking has had many centuries more than computational thinking to develop and the tools that exist as part of it are very stable. They are also much better explored and – due to the efforts of many great mathematicians, are more complex to apply to achieve their full potential.

Having said that, mathematical thinking has less of a basis in real-world problem solving, although mathematical techniques can apply to real-world problems, the difficulty being that closed form solutions, as mathematics tends to create•. Mathematics has limits of applicability, some of which are extremely subtle, even in relatively simple situations: complete closed form solutions in radicals do not exist for problems as simple as finding the roots to polynomial equations of order 5• or above; the general three body problem, for instance, being resistant to differential equation analysis; and others.

Alternatively, so called *numerical approximations* can be made to most problems, but these are often related to computational solutions and so might be better approached through computational thinking.

Another alternative that again takes us back to computational thinking, is the use of agent based simulations in which concurrently acting independent agents are used.

• A *closed form* solution is something like a set of differential equations that completely describe a – typically idealised – problem.

• So called *quintic equations* (**enwiki:1209364767**). A radical is a mathematical expression involving only the coefficients of the equation, and the basic arithmetic operations (addition, subtraction, multiplication, division, and taking the n^{th}-root).

More here.

Activity: Do I need to know about mathematical thinking? #22

Check back to your chosen research strategy from Stage 3. Does it involve data generation using mathematical thinking? If so, read through this section and complete the activities.

[More here]

0.2.9.1 Mathematical thinking tools

Differential equations; matrices; algebra; numbers theory; sets; functions; relations; logics•; topology; geometry; calculus; algebra; analysis;

• *Logics* is plural as there are many, depending on which area of mathematics you are using.

More here?

According to **burton1984mathematical**, there are four processes that are central to mathematical thinking:

Specialising exploring a problem through examples. Each example provides the opportunity for manipulating elements that are concrete, whether they are physical manifestations or ideas.

Conjecturing when enough such examples have been examined, you can conjecture about the relationships that connect them. Through conjecturing, underlying patterns are explored, expressed, and then substantiated.

Generalising if you are lucky enough to have found a pattern, then you might try to generalise it to creating order and meaning out of a – potentially, overwhelming – amount of data.

Convincing a generalisation must be tested until it is convincing to the reader – this is the basis of the knowledge contribution from mathematical thinking.

0.2.9.2 Workflow

Working from simple examples to more complex through an iterative workflow that establishes an appropriate mathematical approach. Building interpretations of a problem which are amenable to existing mathematical tools, or which suggest extensions to them. Developing and applying those tools to produce results. Iteration towards increasingly complete mathematical solutions, refinement of the problem to its important mathematics constituents. Formalisation and – if necessary – proof of results. Application.

0.2.9.3 Other things to think about

Mathematical thinking is arrived at through creative thinking and deep study of mathematical tools and techniques. The sophistication of mathematics often means that, either:

- a particular area of research has already been taken past the abilities of a masters-research-level mathematician;

- it is not amenable to (current) mathematical tools and techniques, and further creative[•] mathematical thinking will be necessary to progress.

Although neither of these characteristics are insuperable, they make timely contributions to knowledge through the application of mathematical approaches difficult[•]. It's worth moderating your expectations of what can be achieved in mathematical research at masters level – discussion with your supervisor of what their expectations are would be very worthwhile.

Activity: Understanding your supervisor's mathematical thinking expectations #23

Schedule some time with your supervisor to discuss what they hope will be achieved through your research project.

Mathematics abstracts from real-world complexity: in modelling traffic to improve flow through a complex junction, for instance, one would not necessarily consider the economy of individual cars, or the noise pollution created by a solution. This can reduce a real-world problem to a complexity that is approachable, but may also lead to non-solutions when applied back in the real world, for instance, leading to complaints from local home owners that noise pollution has risen through a solution.

Mathematics is a very tight community with very high publication standards. The language of mathematics is dense[•].

Because of this, many mathematical research projects at masters level are designed to provide a way into the mathematical literature. A supervisor will set a mathematical task that may already have been solved. The contribution a research student might make, then, is not a contribution to knowledge in the formal sense of extending mathematics, but the widening of the mathematical community to include another researcher whose skills have expanded to be able to make a novel restatement of a problem, for instance, and research further.

The study of the development of mathematical thinking, for instance, in schoolchildren is a very fruitful area with much still to be contributed. More here

Activity: Deep dive into mathematical thinking #24

To find out more about mathematical thinking, take a look at these resources:
stacey1982thinking

• And, most likely, deep and advanced, out of the box, out of this world, and further.

• What is often missing from the mathematical literature – or what isn't always visible to the new entrant – is the often vast timescales over which mathematical progress is made. Bertrand Russell and Alfred Whitehead spent over two decades of their professional lives in the creation of the three volume *Principia Mathematica*. A fourth volume – on geometry – was begun but never completed. In another example, the proof of Fermat's Last Theorem took 358 years to complete.

• A mathematical statement indicative of this complexity is: "Let q be x^5-x-1. Let G be its Galois group, which acts faithfully on the set of complex roots of q. Numbering the roots lets one identify G with a subgroup of the symmetric group $\mathcal{S}_5$. Since $q \bmod 2$ factors as $(x^2+x+1)(x^3+x^2+1)$ in $\mathbb{F}_2[x]$...".

0.2.10 Statistical thinking

Statistical thinking involves designing a study to collect data, analyse patterns in the data, and draw conclusions that go beyond the observed data. Sampling is key to being able to generalise results, while random assignment is key for cause-and-effect conclusions. Probability models help assess random variation and estimate margin of error.

Source: **chance2024statistical**.

Activity: Do I need to know about statistical thinking? #25

Check back to your chosen research strategy from Stage 3. Does it involve data generation using statistical thinking? If so, read through this section and complete the activities.

0.2.10.1 Workflow

0.2.10.2 Other things to think about

Add more definitive resources here

Activity: Deep dive into statistical thinking #26

To find out more about statistical thinking, take a look at these resources:
chance2024statistical

0.2.10.3 Sampling: what, who (and how) to choose

A core prelude to many of the data generating techniques introduced above is to choose the data source. You've already had experience of doing this when you conducted your literature search as part of Stage **??**•.

• You might remember the relatively complex procedures for recording search terms, discovered papers, their relationships, and your growing collection of notes on them.

More here

Unless you had infinite amounts of time – which you don't – and infinite patience – which you might have – you could never be 100% certain that your literature search collected *all* relevant papers the search space is infeasibly large (and not indexed particularly well). But you were systematic and achieved a practically good• coverage because of that.

• By *practically good*, we mean you found the most of the most important papers, some other papers, and didn't have to read *every single paper*. I.e., you found a *representative sample*.

More here?

Sampling is the process of selecting a subset• of an infeasibly large population of interest, and is used in the research strategies that work by estimating or predicting the properties of that population.

• We could have said "sample" but that would have

Sampling can either be random or non-random.

In *random sampling*[•] some unbiased way of choosing, before the fact, subset members from the population, while in *non-random sampling*[•], the choice is based on a researcher's judgement and discretion and can be added to as the research progresses. The lack of bias in the former means that results tend to generalise from the sample to the population. The potential for bias in the latter means that results may not generalise, but things of interest, of depth, and of richness can be followed as they are discovered. As a result, the former is used more in quantitative research, and the latter in qualitative research.

- Reword to separate the two?

Random sampling techniques include:

simple random sampling where each member of the population has exactly the same chance to be selected. It is easy and efficient to implement, and given the complete randomness of the sample, generalisation is fairly reliable. However, if the population has large sub-groups, these may be over-represented in the sample, with minority groups being under-represented.

- Add strengths and weaknesses.

stratified random sampling where sub-groups of the population are identified based on common characteristics, the *strata*, and sampling is random across those strata. The strata are not mutually exclusive: for instance, the population may have sub-groups defined by gender, ethnicity and level of education, which may overlap. This approach overcomes the over/under representation problem of simple random sampling.

- Add strengths and weaknesses.

cluster sampling where the population is divided up in naturally occurring separate clusters, and the sample is obtained by randomly selecting some clusters and then randomly selecting members of those clusters. It is more cost-efficient than the other two approaches, but can introduce bias if the selected clusters are not representative of the whole population, so that the over/under representation problem remains.

- Add strengths and weaknesses.

Non-random sampling techniques include:

purposive sampling in which participants are selected by the researcher based on particular characteristics, knowledge, or expertise they have. It is often used for small populations, especially rare populations which may otherwise be difficult to access. Purposive sampling is particularly suited to studies which intend to be deep and narrow, and for which subsequent generalisation back to the parent population is not a concern. As the sample choice is made by the researcher, it is prone to bias.

- Clarify
- Add other strengths and weaknesses, or perhaps the weakness is with the class rathe than the instance?

convenience sampling where participants are selected based on their availability or accessibility. This is quick and easy, but unlikely to produce a representative sample, so, once again, bias is an issue.

- Add other strengths and weaknesses.

• Random sampling is also called *probability sampling*.

• Non-random sampling is also called *non-probability sampling*.

snowball sampling which relies on referral from previous participants to recruit new ones. This is an effective approach when a population is difficult to access or when the topic is sensitive or tabu. This too is unlikely to generate a representative sample, and is prone to bias.

Add other strengths and weaknesses.

Activity: Deep reading on Non-Random Sampling #27

Check back to your choice of research strategy. If you've chosen one that uses non-random sampling, then you should read the following sources for more details **<empty citation>**

In summary, when choosing a sample, you need to consider various factors, including the aim of your study, the kind of methods you are applying, and the level of access you may have. Trade-offs are likely involved and you may not be able to obtain an ideal sample. Nevertheless, your sample will still be useful to your research, as long as you clearly explain and justify how it was obtained and what its limitations are.

Activity: Choosing your sampling approach #28

Assuming your study requires you to perform some sampling, write down the approach you are going to take, with its justification in terms of what is needed to address your aim and objectives, and any trade-offs due to the practicality of accessing the sample. Record any possible weakness or limitation of your chosen approach.

Guidance

You can skip this activity if sampling is not indicated by your choice of research strategy.

Add other core techniques here; which are there?

0.3 Managing your raw data

Your chosen research strategy may require you to generate data of many kinds and from many sources. The amount of data you collect could be a few ideas or a mound of documents, from a few to terabytes of data•.

• One project we know of collected ???

Before proceeding with your data analysis, you must ensure your data are properly organised and stored, so that you don't loose track of important information, and you can easily locate and refer back to appropriate data during your analysis and when writing up your research – you might need to include a representative

sample of your raw data in an appendix of your dissertation as evidence of your data generation, for instance, identifying which sample can take some time if done afterward while it can be immediate if done while.

Although techniques for doing so are outside of the scope of this book, your data storage should be secured against data loss either due to technical issues, such as computer failure, or due to a data breach, such as through hackers, at least to the standards required by law, any additional requirements made by your organisation, those of any participants, their organisations, and any other stakeholders•.

Activity: Deep dive into reflexivity #29

To find out more about reflexivity, take a look at these resources:

• Once upon a time, in a galaxy far, far away, data generation and storage used to be a *laissez-faire* thing. Today, your organisation can be fined vast amounts of money for any data misuse, so they tend to take it more seriously. If data loss were to happen, amongst other things, it'd probably means you'll fail your degree.

Can we point to needed safeguards rather than trying to be complete (and culpable) here? If so, revise below.

[More here]

0.4 Managing your raw data

Your chosen research strategy may require you to generate data of many kinds and from many sources. The amount of data you collect could be a few ideas or a mound of documents, from a few to terabytes of data•.

• One project we know of collected ???

Before proceeding with your data analysis, you must ensure your data are properly organised and stored, so that you don't loose track of important information, and you can easily locate and refer back to appropriate data during your analysis and when writing up your research – you might need to include a representative sample of your raw data in an appendix of your dissertation as evidence of your data generation, for instance, identifying which sample can take some time if done afterward while it can be immediate if done while.

Although techniques for doing so are outside of the scope of this book, your data storage should be secured against data loss either due to technical issues, such as computer failure, or due to a data breach, such as through hackers, at least to the standards required by law, any additional requirements made by your organisation, those of any participants, their organisations, and any other stakeholders•.

• Once upon a time, in a galaxy far, far away, data generation and storage used to be a *laissez-faire* thing. Today, your organisation can be fined vast amounts of money for any data misuse, so they tend to take it more seriously. If data loss were to happen, amongst other things, it'd probably means you'll fail your degree.

Can we point to needed safeguards rather than trying to be complete (and culpable) here? If so, revise below.

It is also important that you put your raw data in a form which is useable for analysis. Spreadsheets are particularly useful for this purpose, especially if your data is quantitative, so that this is a common way to organise and store raw data. In fact, most publicly available data sets used in research and beyond are stored as spreadsheet files: if you are going to use one such data set, then your raw data are likely already organised for you!

Spreadsheets organise data in rows and columns, so that you can easily enter your raw data using rows for your observations/measurements and columns for your variables. As we will see later on, spreadsheets

come a wide range of functionalities for data manipulation and for some level of data analysis. They are also easily extensible, so that you can grow your data sets incrementally.

Activity: Organising and storing your raw data #30

Consider the data and evidence you have collected or are planning to collect. List the actions you have taken/will need to take in relation to:

- organising your raw data
- storing and backing up your data
- protecting personal data

Make sure you complete those actions before moving on to your data analysis.

With your raw data properly organised and stored, you can now proceed to your data analysis.

End of "revise below"

****

0.5 Common analysis methods

Your choice of data analysis methods is part of your research design, and relates to the kind of data and evidence you have generated, and what you are trying to achieve, that is your aim and objectives.

This section provides an introduction to some common analysis methods. It is far from complete and does not go very deeply into the details of each method: entire books have been written on any of them! By studying this section, you won't become an expert in any of these methods, but you will have gained enough understanding to be able to make a judicious selection for your project. After that, you should review the related specialised literature to help you apply your chosen methods appropriately. You should also talk regularly to your supervisor for further guidance.

0.5.1 Using tables to analyse data

The following kinds of tables are used extensively in research and often found in dissertations.

Pivot tables Pivot tables can be used to summarise, sort, filter, re-organise or group data organised in rows and columns, and perform calculations on them, such as counting, generating totals or averages, and much more. Pivot tables are both powerful and versatile•, and one of the most widespread tools for data analysis.

You can generate a pivot table from any data set organised in rows and columns, regardless of whether the values are quantitative or qualitative: all common spreadsheet applications• include this function.

figure 0.1 gives an example: these are the first few raws of a data set related to the US housing market•. The dataset contains over 9,316 entries, each corresponding to a distinct property. Each property is characterised by a number of attributes: size in square feet, number of bedrooms and bathrooms, type of neighbourhood, the year it was built and its market price in US dollars. As you can see, this table includes both numerical and categorical variables.

- In fact, they are so versatile that we'll only be able to provide few illustrative examples. Much, much more can be found online!
- From MS Excel to Apple Numbers to Google Sheets.
- It was taken from one of Kaggle's free datasets, the housing price dataset. Kaggle is possibly the largest and best known online community for data science and machine learning.

ID	SquareFeet	Bedrooms	Bathrooms	Neighborhood	YearBuilt	Price
1	2126	4	1	Rural	1969	US$ 215,355.28
2	2459	3	2	Rural	1980	US$ 195,014.22
3	1860	2	1	Suburb	1970	US$ 306,891.01
4	2294	2	1	Urban	1996	US$ 206,786.79
5	2130	5	2	Suburb	2001	US$ 272,436.24
6	2095	2	3	Suburb	2020	US$ 198,208.80
7	2724	2	1	Suburb	1993	US$ 343,429.32
8	2044	4	3	Rural	1957	US$ 184,992.32
9	2638	4	3	Urban	1959	US$ 377,998.59

Figure 0.1: first few rows of the example dataset

Pivot tables can be used to summarise such data to answer certain questions. For instance, we may be interested in the average house price by neighbourhood and number of bedrooms, which would result in the pivot table in figure 0.2, which gives the average price of each combination. The 'grand totals' in the table are also averages, by row and by column.

Neighborhood	Rural	Suburb	Urban	Grand Total
Bedrooms	Price (Average)			
2	US$ 218,323.92	US$ 216,300.13	US$ 220,050.01	US$ 218,230.99
3	US$ 219,053.37	US$ 220,397.86	US$ 223,737.67	US$ 221,057.88
4	US$ 227,774.55	US$ 224,609.50	US$ 230,086.56	US$ 227,473.37
5	US$ 231,112.60	US$ 231,776.73	US$ 234,894.98	US$ 232,595.48
Grand Total	US$224096.13	US$223234.19	US$227166.20	US$224827.33

Figure 0.2: Pivot table of average property prices by neighbourhood and number of bedrooms

Alternatively, we may be interested in finding out how many properties of each kind have been built in each neighbourhood. In this case the pivot table would look like that in figure 0.3. The grand totals in this case are counts. Note how we have added combinations of bedroom and bathroom numbers to characterise each type of property.

These are just but two examples of questions about the data you can address by using pivot tables, out of a vast range of the possibilities. If your data are organised in tables, then it is well worth spending some time becoming familiar with pivot tables.

Activity: Pivot tables in Excel #31

Download the housing price data set from Kaggle and re-create the pivot tables in our example. Come up with other questions you could ask of the data and generate related pivot tables.

Guidance

Feel free to use your preferred spreadsheet application for this activity, as long as it supports pivot tables – most do.
You may have to register with Kaggle to gain access to the data set.

	Neighborhood	Rural	Suburb	Urban	Grand Total
Bedrooms	Bathrooms	ID (Sum)			
▼ 2	1	180	220	261	661
	2	129	214	283	626
	3	155	45	75	275
2 Total		464	479	619	1562
▼ 3	1	96	130	129	355
	2	397	149	245	791
	3	127	139	889	1155
3 Total		620	418	1263	2301
▼ 4	1	142	306	228	676
	2	453	379	410	1242
	3	315	175	194	684
4 Total		910	860	832	2602
▼ 5	1	325	282	378	985
	2	438	127	10	575
	3	534	367	390	1291
5 Total		1297	776	778	2851
Grand Total		3291	2533	3492	9316

Figure 0.3: Pivot table of property counts by neighbourhood and number of bedrooms/bathrooms

The Excel Help facility and documentation provides all the info you need to create a pivot table. However, you could also browse some of the very many freely available online resources and tutorials on this topic.

Frequency and contingency tables Frequency tables are used to summarise the frequency (or count) of values taken by a categorical variables in a data set. For instance, after studying a degree, a student's outcome may be classed as distinction, merit, pass or fail. A frequency table can then be used to summarise the frequency of each class of outcome for a particular students' cohort, as shown in Table 0.2.

Table 0.2: Example of frequency table

	Distinction	**Merit**	**Pass**	**Fail**
Outcome	12	26	42	5

Contingency tables• are a form of frequency tables used to tabulate the value frequencies of two categorical variables. For instance, following from our previous example, we may like to tabulate the outcome value frequencies in the cohort against gender, as shown in Table 0.3.

• Also known as *cross-tabulation* tables.

Table 0.3: Example of contingency table

Outcome by Gender	**Distinction**	**Merit**	**Pass**	**Fail**
Female	7	12	21	2
Male	5	13	19	3
Other	0	1	2	0
Totals	12	26	42	5

Contingency tables are frequently used to summarise and analyse data collected in survey research, and are a key tool in statistical analysis.

Both frequency and contingency tables can be generated as pivot tables in a spreadsheet. In fact, the table in figure 0.3 is a contingency table.

0.5.2 Statistical analysis

Statistical analysis in an umbrella terms for a set of methods which can be applied to numerical and categorical data. More precisely, in statistics data types are classified as:

- scalar, which includes all measurements and counts; with reference to the types in Section**??**, these are all numerical data, continuous, discrete, interval and ratio data.
- categorical, both ordinal and nominal.

There are two broad categories of statistical methods:

- descriptive statistics, whose aim is to describe data; and
- inferential statistics, whose aim is to make predictions from data.

We briefly consider each in what follows.

0.5.2.1 Descriptive statistics

These are used to describe various attributes of a data set. The basics are:

- count, to establish how many entries there are in the data set
- centrality, to establish the 'centre' of the data set. Three measures are commonly used: the *mean*, which provides the average value of the data set; the *median*, which provides its mid point•; and the *mode*, which indicates the value that occurs most frequently, if any•.
- dispersion, to establish the spread of the data in the data set. Range and standard deviation are two common measures. The *range* is the difference between smallest (minimum) and largest (maximum) values. The *standard deviation* is based on a mathematical formula which considers the distance of each value in the data set from the mean. It is not essential for you to know such formula, which is automatically computed by spreadsheets and statistical software•. The larger the standard deviation, the greater the dispersion.

• Remember that quantitative data can be ordered.

• There is no mode if no value is repeated in the data set.

• Of course, you can always look it up in the literature...

- skewness, to establish how symmetrically distributed the values in the data set are in relation to the centre. In the case of perfect symmetry, skewness is equal to zero, and mean and median are equal. When asymmetric, mean and median are different and the distribution may be either right (mean smaller than median, and negative skewness) or left (mean greater than median, and positive skewness) skewed. A perfectly symmetric distribution is usually referred to as a *normal distribution* or *bell curve*, from the shape of the line that can be obtained by plotting the data on a chart•.

• This oversimplifies the topic in order to give some intuition in case you have not come across these terms before. A lot more should be said about the normal distribution and its pivotal role in statistics!

Not all descriptive statistics apply to categorical data. In particular, the mode is used as the main measure of centrality for nominal data, while the median is used for ordinal data which are not numeric.

These are lots of definitions to digest, particularly if you haven't encountered these terms before! The following activity should help.

Activity: Descriptive statistics in Excel #32

Assume you have measured the weight in grams of each apple in a basket, obtaining the following numbers: 105, 120, 122, 125, 127, 128, 129, 130, 132, 133, 135, 135, 138, 140, 128. Enter these data in an Excel sheet and use its built-in data analysis function to generate the related descriptive statistics.

Guidance

In the current version of Excel, you can access this function from the Data tab, by pressing the Data Analysis button. If you find it difficult to locate this function, you should refer to the documentation or to some of the many tutorials on this topic which are freely available online.

Discussion

You should have obtained the following values:

Attribute	**Value**
mean	128.47
median	129
mode	128
standard deviation	8.55
skewness	-1.4
range	35
minimum	105
maximum	140
count	15

There are 15 values in this data set, with range 35 (the difference between maximum and minimum values). In terms of centrality, the mean (128.47) is slightly smaller than the median (129), and Excel reports a mode at 128. In reality, if you look at the data you will see that there are two modes in this data set[a], 128 and 135, but Excel only returns the first encountered! In terms of dispersion, the standard deviation is telling us that most apple weights are within 8.55 grams of the mean (below or above), so the apple weights are similar in the apple baskets. Note that the skewness is negative, which is consistent with the mean being smaller than median, so the data distribution is right skewed.

[a] Statisticians call this *bi-modal*.

In your dissertation, you can easily present such descriptive statistics as a table, possibly adapting that automatically generated by your spreadsheet.

In addition, charts can be used to visualise the data and examine their descriptive statistics.

With scalar data, like in our example, you can use a *histogram*. The one in figure 0.4 uses the apple weights from the previous activity: on the horizontal axis, we have the distinct weights, and on the vertical axis, their frequencies, that is how many times each weight appears in the data set. Given the values you have obtained for the data set descriptive statistics, you can easily locate on the chart min and max values, and mean, median and mode. In this case, the two 'peaks' correspond to the two modes we mentioned in the activity. You can also check that most of the values are within 8.55 grams from the mean, either way: the only values left out are 105 (to the left) and 138 and 140 (to the right). Skewness is not obviously notable on this chart, so that we will use a different chart for that purpose.

Before we do that, however, it is worth noticing that given our small data set of discrete values, we have

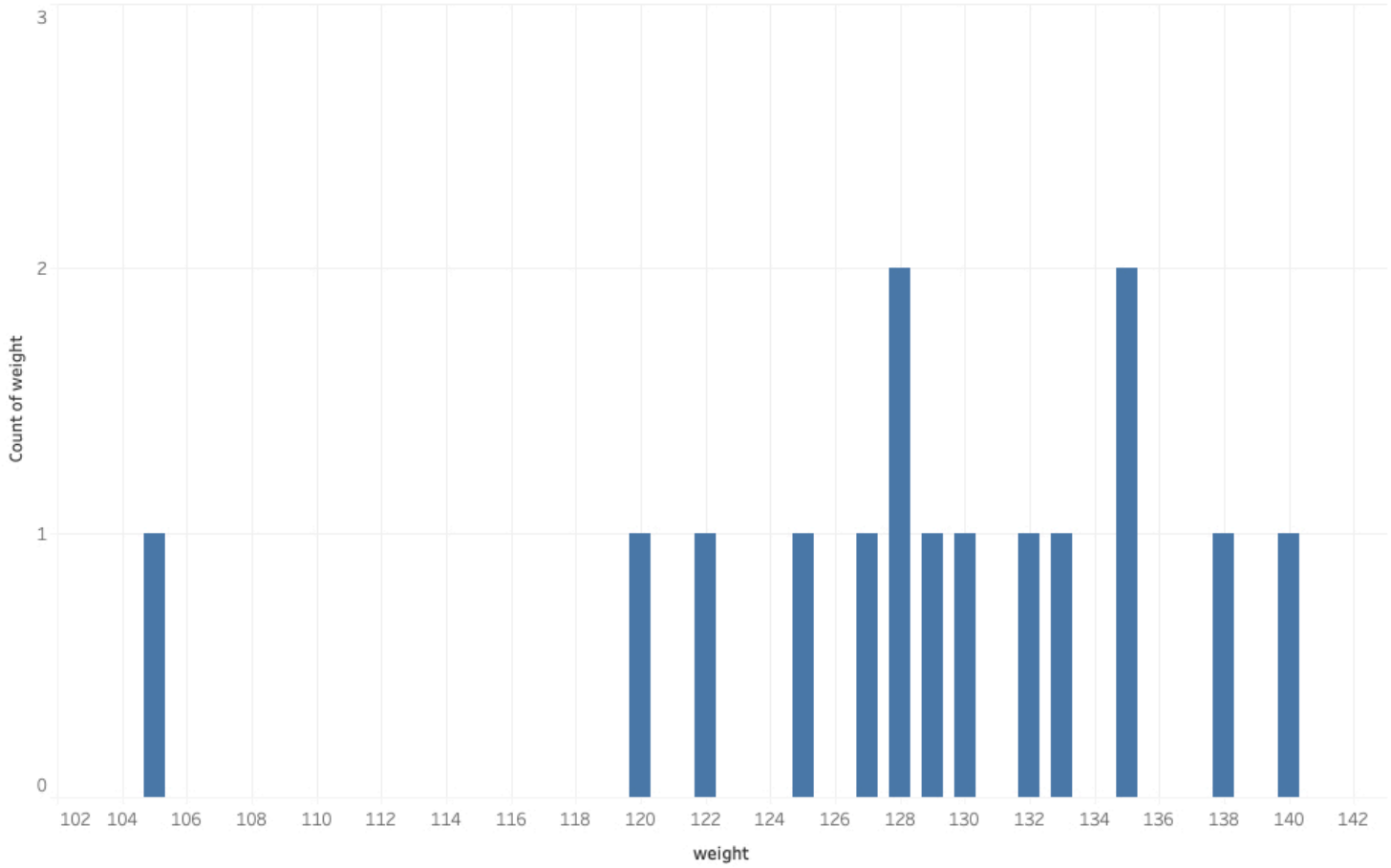

Figure 0.4: Histogram for the apple weights (bin size = 1)

used a histogram with 'bin' size equal to one, which allows us to plot each individual apple weight. A *bin* in a histogram is essentially a way to group a number of values, with bin size establishing the spread of each bin. Frequencies are then calculated by bin. Grouping values in bins is necessary with large data sets and/or with continuous data. figure 0.5 illustrates a histogram for our example, in which the bin size is 5: that is, each bin spans a set of 5 possible values.

In order to visualise both spread and skewness, a useful chart is the boxplot, illustrated in figure 0.6 for our example. This is made of a 'box' around the median of the data, and some 'whiskers' on each side of the box•. It is obtained by dividing up the data into quartiles, each containing a quarter (or 25%) of the data, with the median in the centre. The box includes the two quartiles on each side of the median, which, together, account for half of the values in the data set. The whiskers account for the two other quartiles, with a caveat: if there are very extreme values, these are treated as possible outliers and left out of the whiskers. This is, in fact, the case in our example where value 105 is treated as an outlier in the chart: it is a dot on its own, not included in the left whisker. The whisker length provides an indication of spread: the longer the whiskers, the more spread out the data. Instead, the position of the median in relation to the extreme of the box provides an indication of skweness: in our example the median is further away from to the right edge (just!), indicating that the data distribution is slightly right-skewed (consistent with the negative skewness value in the descriptive statistics).

• Which is why this chart is also called a *box and whiskers* plot.

To be more precise, the relation between a boxplot and its underlying statistical features is illustrated in figure 0.7. The two quartiles around the median represent the interquartile range (IQR) of the data set. The whisker lengths, calculated based on the formulae in the figure, allows the identification of lower and upper bounds beyond which values are seen as extreme and represented separately as outliers. An outlier, therefore, is just a value which is distant from most of the other values in the data set: it may point to an error, which should be corrected, or an anomaly, which may require further investigation, but that's not necessarily the case. However, it's good practice to investigate all outliers to understand why they have occurred.

Activity: Charts in Excel #33

Go back to your Excel sheet from the previous activities and generate charts similar to those in the figures above.

Guidance

In the current version of Excel, you can generate these charts from the Insert tab, by choosing from

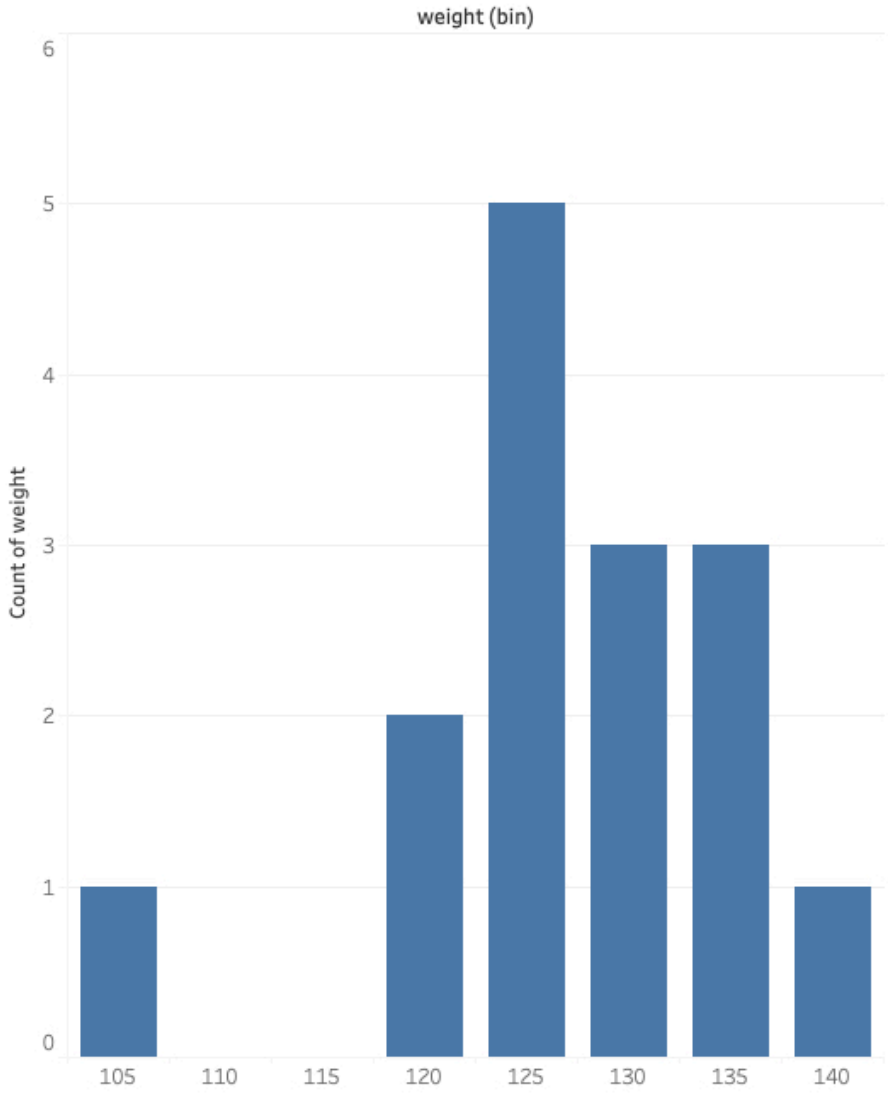

Figure 0.5: Histogram for the apple weights (bin size = 5)

Figure 0.6: Boxplot for the apple weights

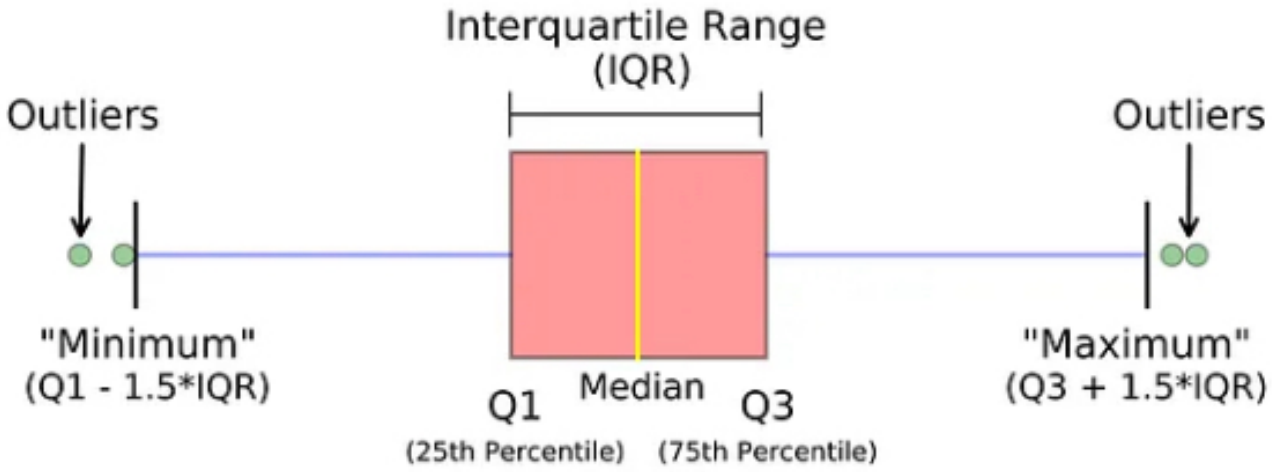

Figure 0.7: The features of a boxplot — LR to redraw as taken from the web

the Statistical charts menu. If you find it difficult to locate this function, you should refer to the documentation or check some of the many tutorials on this topic which are freely available online.

Table 0.4 summarises useful charts that can be applied to visualise your data and their descriptive statistics: it includes charts we have not used in our examples, but which are very common, so that you can find plenty of study materials online should you wish to look them up and use them.

Calculating descriptive statistics and visualising data in appropriate charts, should be the first step in your statistical data analysis, as these provide useful summaries and visualisations of key properties of your

Table 0.4: Common charts to visualise data sets and their descriptive statistics

Chart	Variable(s)	Purpose
bar chart	one categorical	to visualise counts/frequencies/proportions/percentages
staked bar chart	two categorical	to compare counts/frequencies/proportions/percentages between two groups
histogram	one scalar	to visualise distribution, including centrality, dispersion and skewness
scatter diagram	two scalar	to visualise relationships and possible outliers
boxplot	one scale or one categorical	to visualise spread, skewness, median, IQR and possible outliers
line chart	one scalar by time	to visualise change over time

data set. And, as you have found out in the activities, you do not need to be a statistician to be able to generate them!

Descriptive statistics may also help you identify errors or anomalies in the data, and can inform possible follow-up analysis, including inferential statistical analysis. Depending on your research aim and objectives, they could also be all you need in your project.

If you have collected scalar or categorical data, it is time for you to have a go at analysing them using descriptive statistics and charts.

Activity: Applying descriptive statistics to your data #34

Calculate descriptive statistics for your data set, and generate appropriate charts.

Guidance

MS Excel is relatively straightforward to use for this purpose, but feel free to use other tools you may be already familiar with, including statistical or data analytics packages. Whichever tool you use, you should ensure it supports the functionalities we have discussed in this section.

0.5.2.2 Inferential statistics

Inferential statistics relies on the concepts of population and sample: the *population* is the entire group you are interested in studying – say, all UK voters in a general election; while the *sample* is the portion or subset of that group you have access to in your research. Then the aim of inferential statistics is to establish

whether patterns or effects you have observed in the sample can be generalised to, i.e., inferred for, the whole population, or whether they are the result of chance. In inferential statistics this is achieved through statistical tests.

A statistical test tells you whether the proposition[•] you wish to test on your sample is likely to be true in the population under study. For this to work, your sample must be representative of the population[•].

A statistical test returns a measure of *statistical significance*, which is used to provide evidence (or otherwise) that the pattern or effect you see in your sample is also likely to exist in the population, and is not just the effect of chance[•]. As a corollary, if your sample is very large, almost all effects observed in the sample will be likely present in the population; vice-versa, if your sample is very small, most effects observed in the sample are unlikely to be present in the population, unless they are really very large. As a rule of thumb, most tests require a sample size of at least 30 observations, but more precise sample size estimates can be made based on population size and expected significance level[•].

Each statistical test comprises the following elements.

- **Hypotheses** There are two, *null* and *alternative* hypotheses. Inferential statistics assumes you can't prove something to be true, but you can disprove something by finding an exception. Here is a classic example: you can't prove that all swans are white, but you can disprove they are by finding a black swan! So, you must set the null hypothesis to what you want to disprove about the population, with the alternative hypothesis being what you are really interested in finding out. So, the null hypothesis is usually a statement of no pattern/effect in the population.

- **Significance** This is the level of statistical significance for the test. It's known as the *alpha* (α) value from the Greek name of the mathematical variable used to express it. Most tests are run with $\alpha = 0.05$, which gives a 5% probability that we may infer that the null hypothesis is disproved while in actually it is correct[•].

- **Sample(s)** You need to have one or more (representative) samples of the population of interest on which to perform the test. Multiple samples are used in some tests, typically to compare specific statistics in different groups within the population or changes within a group over time or after an intervention of interest, say treating patients with a new pharmacological drug.

- **p-value** This is the probability calculated for your test by your statistical package, and which is used to decide the outcome of the test.

• You can think of a proposition as an educated guess you have made based on some observations, but that has yet to be supported by evidence.

• We discussed sampling in Section **??**

• Contrary to the common language meaning of 'significance' as big or important, statistical significance only indicates that the effect is likely to exist in the population, where it may well be small or unimportant!

• Formulae for the ideal sample size are easily found in the literature and online.

• This is called a Type I error in Statistics.

- **Decision** This is based on the p-value in relation to the α value: if the p-value is less than the α value, then the null hypothesis is *rejected*, i.e. disproved, which means your alternative hypothesis that there is an effect in the population is supported by statistical evidence.

There are very many statistical tests to choose from, depending on the kind of data you have and their distribution, the purpose of your analysis and the number of samples involved.

Statistical tests are applicable to both scalar and categorical data and can be used to compare values of specific statistics or to establish statistical relationships between variables, specifically:

- an *association* between variables means that one variable can be used to provide some information about the other
- a *correlation* is a particular type of association such that the two associated variables always change together, for instance they both increase or decrease at the same time, or when one increases the other always decreases.

Statistical tests can be used to estimate the strength of an association (i.e., the extent changes in one correspond to changes in the other) and its direction (whether the variable changes are in the same or opposite way).

We will not detail all possible statistical tests in this introductory section — once again, entire books have been written about them! Instead, we provide Tables 0.5 and 0.6 as summaries of the most common tests that you can then follow up in the literature, should you wish to apply any in your research.

Even if these tests are only a sub-set of all statistical tests available, there is a lot to digest. The next activity should help you use these table to choose an appropriate test.

Activity: Choosing an appropriate test #35

Consider the following scenarios: for each, use the information in the tables to decide which test to apply and what the null hypothesis should be. For each, write down your reasoning, choice and null hypothesis.

- *Scenario 1* to investigate the amount of sugar contained in baby food of a particular brand against a recommended threshold, from a sample of 30 products of that brand.
- *Scenario 2* to investigate the number of products per hour of two manufacturing machines in the same plant, by observing the two machines' output over 24 hours.

Table 0.5: Common statistical tests for comparison. *Parametric* tests apply to normally distributed data (see Section0.5.2.1), while *non parametric* tests to skewed distributions.

Purpose	**Variables**	**Example**	**Parametric**	**Non parametric**	**Notes**
to compare the sample mean against a specific value	one scalar	to investigate whether AA batteries of a particular brand have the claimed lifespan	one sample t-test	n/a	
to compare the sample proportion against a specific value	one categorical	to investigate the proportion of people who voted for a particular party in a city against that for the whole country	one sample z-test	n/a	
to compare the means of two independent samples	scalar	to compare the mean scores (dependent) of students studying the same subject with two different teaching approaches (explanatory)	independent t-test	Mann-Whitney test/Wilcoxon rank sum	two samples are *independent* when there is no reason to believe that observations in one sample are influenced or determined by those in the other
to compare the means of three or more independent samples	scalar dependent; nominal explanatory	to compare the mean scores (dependent variable) of students studying the same subject with three or more different teaching approaches (explanatory variable)	one-way ANOVA	Kruskal-Wallis test	
to compare the average difference between paired samples against a particular value	scalar dependent; time or condition as explanatory	to compare the blood pressure readings (dependent vraiable) of a group of people before and after exercising (explanatory variable)	paired t-test	Wilcoxon signed rank test	in paired samples each data point in one sample is uniquely matched to a data point in the other sample; this happens, for instance, when we measure a factor before and after an intervention, or take different readings for the same group of individuals. Because of this, paired samples are not independent.

Table 0.6: Common statistical tests for association. *Parametric* tests apply to normally distributed data, while *non parametric* tests to skewed distributions.

Purpose	**Variables**	**Example**	**Parametric**	**Non parametric**	**Notes**
to investigate correlation between two continuous variables	scalar dependent and explanatory	to investigate the relation between blood pressure (dependent) and age (explanatory)	Pearson's Correlation Coefficient	Spearman's Correlation Coefficient	
to investigate association between two categorical variables	categorical dependent and explanatory	to find out if there are gender (categorical) differences in the choice of modes of transport (categorical) in a city	chi-squared	n/a	
to investigate association between two categorical variables when the sample is small	categorical dependent and explanatory	to find out if there are gender (categorical) differences in the choice of modes of transport (categorical) in a city	fisher's Exact test	n/a	the sample size n should be less than 20
to predict the value of one variable from that of one or more other variables	scalar dependent and any kind of explanatory	to predict house prices (dependent) based on location (explanatory, categorical) and number of bedrooms (explanatory, scalar)	linear regression	n/a	linear regression relies on associations between dependent and explanatory variables
to predict the value of a binary variable from that of two or more other variables	binary categorical dependent and any kind of explanatory	to predict whether a customer is likely or not to purchase a certain product (dependent) based on previous purchased products (explanatory, categorical) and average annual spent (explanatory, scalar)	logistic regression	n/a	a binary variable has only two possible values, so that logistic regression calculates the probability of each value based on the values of the explanatory variables. Because of this logistic regression can be used as a classification method

- *Scenario 3* to investigate the effect of temperature on the consumption of ice cream in a particular city over 12 months.
- *Scenario 4* to investigate whether taste in chocolate types, say white vs milk vs dark, is related to gender in particular country.

Guidance

To simplify things, always assume normal distributions.

Discussion

Assuming normal distributions, for each scenario, we have considered:

- the kind of data
- number of samples and their size
- purpose of the investigation

This is what we have concluded:

- *Scenario 1* scalar variable (amount of sugar); one sample of 30 products; to compare the sample mean against the recommended threshold. The test to use is a t-test with null hypothesis that the sample mean is above the threshold.
- *Scenario 2* scalar dependent (number of products per hour) and categorical explanatory (which machine); two samples (one per machine over the time span); to compare the means of products per hours for the two machines; there is no reason to think that the working of one machine may influence that of the other. The test to use is an independent t-test with null hypothesis that the two sample means are different.
- *Scenario 3* scalar dependent (level of ice cream consumption) and scalar explanatory (temperature); one sample over the period; to investigate any relationship between the two variables. The test to apply is Pearson's correlation with null hypothesis that there is no association between the two variables. If, in addition, we wanted to make predictions on ice cream consumption based on temperature, then we could also apply linear regression.

- *Scenario 4* both dependent (chocolate taste) and explanatory (gender) are categorical; one sample from the country; to investigate association. The test to use is a Chi-squared with null hypothesis that gender has no association with chocolate taste.

0.5.3 Quantitative analysis resources

Add here

0.5.4 Qualitative analysis

As indicated in Section **??**, common methods for qualitative data analysis include content, thematic, narrative or discourse analysis. While their goal may be different, they all apply the initial step of *coding*, which we discuss next.

0.5.4.1 Coding qualitative data

A *code* is essential a label which describes an extract from qualitative data set, with *coding* the process of creating and assigning codes to categorise those extracts.

Coding is important and it helps you ensure that your analysis is systematic, and the codes will help you explore themes and patterns in the data. However, codes are not themes: they are just labels used to group similar types of data, developed to support your follow-up analysis.

There are two main approaches to coding. In *deductive coding*, the codes are decided upfront, before looking at the data, and may be based on your research problem phenomena, or may have emerged from your literature review, including codes possibly used in previous studies. In *inductive* coding, the codes emerge from the data and are not pre-defined. Deducting and inductive coding can also be combined by starting with a set of pre-defined codes then adding new codes as you review the data.

Whichever your approach, you should follow a multi-pass coding process. The first pass should consist of going through the whole data set in order to establish which codes to use. In the second pass, and any subsequent ones, you should apply the codes to the data bit by bit, say by line by line in a text, or frame by frame in a video, etc. In the second pass and subsequent passes, the initial codes are reviewed and may become more or less detailed.

There are various ways to choose codes. For instance, *in vivo* coding uses the exact language which occurs in the data: this is used, in particular, for participants' speech, especially when different languages are used. On the other hand, *descriptive*• coding uses words which encapsulate a general idea, such as 'sport' or 'running': this is particularly useful for non textual data, like images or videos.

• This is a very common approach, although there are others which you can research in the literature.

Whichever codes you end up with, you should ensure they are properly defined, so that their are unambiguous and can be applied consistently. You should use a *codebook* for this purpose, which lists all the codes and their intended meaning, and that you can revisit and refine throughout the coding process.

The last step before detailed analysis is *code categorisation*, which is the process of reviewing what you have coded and organise it into categories. For instance, from codes such as 'football', 'tennis' and 'rugby' you may define a category 'sports'. In this way, you both organise your data and establish connections between codes and coded information.

Both coding and categorisation are iterative processes which carry on until you reach saturation, that is no more is gained from further coding or categorisation. At this point, you can proceed with your chosen analysis method, whether content, thematic, narrative, discourse analysis or other, in order to identify patterns and themes, and provide your own interpretation of the data.

Coding and categorising are time consuming tasks, particularly if you have a large amount of text to code. In most research, coding data by hand is impractical and you should at least make use of a word processor, perhaps using colours and comments to code fragments of your text. Better still, you could make use of a bespoke qualitative data coding tool: many such tools are now available, some of which can also automate coding and categorisation to some extent.

Add URLs in the following question

Activity: Investigating tools for qualitative data coding #36

Conduct a web search on tools which support qualitative data coding. List up to four which appear most commonly used. For each, indicate which coding features it offers and the extent it is freely available for students' research projects.

Discussion

Qualitative analysis tools are growing and changing rapidly, particularly due to the integration and exploitation of AI capabilities.
At the time of writing this book, the most used commercial products include NVivo, ATLAS.it and MAXQDEThey all provide support for coding, with more or less extensive automation, alongside various other features such as data visualisation, statistical analysis, automatic transcripts generation from audio

and video files, to name just a few. These commercial products are quite sophisticated with a steep learning curve and are usually quite expensive. They are also geared towards large research efforts, possibly by teams of researchers.
An increasing number of lighter, free products are also available. These include, for instance, Taguette, which supports manual coding and is both open source and free to use, or QDE Miner Lite, which is a free limited version of its full commercial release, and also supports manual coding. Such free products may be sufficient for Masters level research projects.
You may have found other similar tools.

0.5.4.2 Presenting qualitative data

While quantitative data can be summarised and presented using tables and charts, the same does not necessarily apply to qualitative data, which, due to their heterogeneous nature, cannot be easily set out in a standard manner.

Conveying the depth and richness of qualitative data in a succinct way is challenging, so that both selectivity and creativity are needed in presenting the data.

For textual data, like interview transcripts, verbatim quotations are often used to illustrate specific themes or points, or support certain conclusions. However, an excessive use of quotations will result in overlong accounts of the work, which may be difficult to follow or even obscure the main findings. Therefore it is important to select quotations which are particularly representative or poignant, avoiding verbose details that can be succinctly presented in the narrative around those quotations.

Diagrams, schematics or drawings can also be used effectively and imaginatively to present qualitative data and their analysis. Data visualisation is, in fact, a discipline in its own right•, and some visualisation techniques can be applied to qualitative data.

• Edward Tufte is one of the most influential figures in this field. His books provide compelling examples on how to use visualisation to present and analyse highly complex data.

Activity: Visualisation techniques for qualitative data #37

Conduct a web search on techniques for visualising qualitative data. List the techniques you have found and what they are used for.

Discussion

You may have encountered some or all of the following techniques:

- diagrams and schematics, to convey complex processes or structures
- graphic timelines, to summarise key events and their order
- word clouds, to summarise emerging themes or concepts from text, and their relative frequencies
- mind maps, to visualise how different ideas relate or contribute to a central concept or topic
- heat maps, to highlight trends or differences in tabulated data
- icons, alongside brief descriptions, to represent and quickly identify specific concepts
- bespoke drawings, for data which cannot be easily visualised using other standard techniques
- pie charts and bar charts, to summarise proportions and counts – which are actually quantitative, but may be the result of qualitative data analysis – of categorical data.

0.6 Writing up your analysis

In writing up your data analysis in your stage report or dissertion, you will need to decide:

- how to summarise your data and evidence. This will depend on their nature, and you will need to ensure that your summaries are appropriate to convey the essence of the evidence you have generated. In the previous sections, you have considered ways in which quantitative and qualitative data can be summarised using tables and visualisations. It may also be necessary for you to include sample raw data in an appendix.
- how to report findings. Your findings are your conclusions from your data analysis and should be reported as academic arguments which rely on the evidence you have generated.

- how to structure your narrative. Depending on your chosen research strategies and methods, different structures are possible. For instance, you may choose to start with a section which summarises all your evidence followed by one in which you analyse it, which may work well, for instance, for survey research. Alternatively, you could have separate sections each including a summary and analysis of a sub-set of your evidence: this may be appropriate for mixed methods research, with each section dealing with a different kind of data, or for design science research, with each section addressing a different design cycle. Whatever you choose, it is important that your report is effective in presenting your evidence and findings in a clear, rigorous and logical manner.

Activity: Writing up your analysis #38

Consider the data you have collected and analysed so far. Note down how you are going to address each of the points above in your report. Write an outline of your analysis section.

Guidance

A good starting point is to consider how other researchers report their data analysis and findings. To this end, go back to some of the articles you have reviewed and consider their data analysis section and any related discussion. Ensure you select articles that apply similar collection and analysis methods to those in your research design, or deal with similar types of data.

0.7 Interpreting and evaluating data

Having genearted and analysed a certain amount of data and evidence, it is time for you to start interpreting your findings in relation to your aim and objectives, and generally evaluate them in terms of their contribution to knowledge and possible limitations. This is a process you will repeat and complete in Stage 5, the concluding stage of your project, ending with your dissertation submission.

Interpreting your findings signifies addressing the following questions:

- What conclusions have you drawn from your data analysis?
- How do they relate to your aim and objectives?

- How do they relate to what you know from the literature?
- How do they relate to professional practice? (if applicable)
- Which new knowledge do they contribute?
- What do they fail to achieve?

Activity: Interpreting and evaluating your findings #39

Consider your data analysis and based on it, address each of the above questions. Write down your responses, ensuring your arguments are well-formed, with explicit reference to evidence.

Guidance

Your interpretation and evaluation of findings will be, of course, limited by the data/evidence you have generated and analysed up to this point. You will revisit and expand this work in Stage 5 in order to complete your project.

0.8 Drafting an abstract for your project

An abstract is a common way to summarise academic research. Abstracts are an integral parts of all published academic articles – you will have encountered many abstracts while reviewing the literature. They are also very common in academic dissertations, therefore it is highly likely you will be required to include one at the beginning of yours.

An abstract provides a short summary of the whole research written for a specialist audience, that is you can assume that the reader has good knowledge of the topic and field of study. It should be a stand-alone item, so that it can be understood without reference to any other part of your dissertation.

Its content should convey succinctly the research problem, how and where it arises and its significance, the research aim and research design, key results obtained by the research, their evaluation and their implications for further research or professional practice.

Writing an abstract for your research is a good exercise, even if one is not needed for your dissertation, as it gives you an opportunity to write a logical argument that connects all key elements of your research.

This can help you check that all the pieces fit together in a coherent manner. It is also something you can share with your supervisor and critical friends to communicate succinctly the essence of what you have done and achieved.

Activity: Drafting your abstract #40

Write a draft abstract for your project, which should reflect your research progress to date.

Guidance

You should go back to some of the articles you have reviewed to consider the content and structure of their abstract. Choose a structure which may fit your project and write up your draft abstract accordingly. As your research is yet to be completed, you will not be able to write up the full abstract, but you should end up with a draft that you can easily complete by the end of your project.

0.9 Reflecting and reporting in Stage 4

It's time to write your Stage 4 report. As in the previous stages, before you do, it is worth reflecting on your work and learning in this stage.

Activity: Reflecting on your learning and practice #41

As you did at the end of the previous stages, in this activity you are asked to stand back and reflect deeply on what you have leant and done, the wider context of your work and your own attitude to it. Specifically, you are asked to think deeply about each of the following:

- your study this far
- the way you work
- the context of your research
- your feelings about your project

You should also think of any significant changes with respect to your reflection in the previous stages

Guidance

You should be accomplished at reflection by now. However, should you need to, you can refer back to the guidance to this activity in Stage 1, Section **??**.

Your end-of-Stage 4 report will help you consolidate your work so far, adding yet another increment toward your full dissertation. We recommend you follow the guidance in Table 0.7 to write your report.

At the end of Stage 4, you should complete a report, extending that of Stage 3 and covering the work you have carried on in this stage. Its recommended structure and content are indicated in Table 0.7: much of the content should be carried forward from the previous stage.

Activity: Writing and assessing your report for Stage 4 #42

Using your word processor of choice, revise and expand your Stage 3 report by applying the structure and guidance in Table 0.7.
Assess your report by applying the criteria in Table 0.8. Revise and iterate until you are ready to move on.

Guidance

In completing your report, you should make good use of notes and summaries your wrote as part of the activities in this chapter. In evaluating your report, for each criteria, you should consider the related prompts, write down any further work needed for your next stage, and update your work plan and risk assessment table accordingly.

0.10 Takeaways

- Sampling is the process of selecting a sample from the population of interest, and is required in many research strategies. Many different approaches to sampling exist, depending on the nature and aim of your research.
- Questionnaires are common tools for data generations. Good questionnaire design relies on a wide range of considerations (see Section 0.2.4).

Table 0.7: Report structure and guidance

Report template	**Guidance**
Proposed title	Your title should continue to capture succinctly your research problem and aim. *It is likely this is the same as, or very similar to, that in Stage 3*
Abstract	You should include your draft abstract providing a succint account of your research to date
Sect 1 - Introduction 1.1 Background to the research 1.2 Justification for the research 1.3 fitness of the research	This section should continue to provide an introduction to your research topic in its wider context (as background) and your justification of why the research is worth pursuing. Its purpose is to introduce and justify your intended research in overview, before entering the detailed work of the subsequent sections. It should be well argued and supported by appropriate citations. In this section, you should also argue how the research fits within the scope of your qualification, and meets any other personal, professional or organisational criteria. *You may review this section from Stage 1 to reflect your growing understanding of the topic in context derived from your literature review.*
Sect 2 - Literature review 2.1 Review of existing relevant knowledge 2.2 Critical summary, including knowledge gap to be addressed by the research	Your review should provide a critical account of your in-depth engagement with the academic (and other) relevant literature, including identifying key trends, ideas and possible knowledge gaps. Most of your citations should point to academic articles. Your critical summary should highlight key insights from your review and provide a strong justification for your proposed research. Both coverage and depth of your review matter. You should ensure that your review is well structured, with a logical narrative flow and your arguments are well supported by evidence
Sect 3 - Research definition 3.1 Problem statement 3.2 Aim, objectives, tasks and deliverables 3.3 Knowledge contribution	You should ensure that your research problem is well articulated and appropriate for your course and your personal and professional circumstances, that your aim and objectives are consistent with research problem, that tasks and deliverables break down your objectives appropriately and are clearly related to your chosen research methods, and that the intended knowledge contribution of your research is clearly articulated
Sect 4 - Research design 4.1 Evidence and data 4.2 Research strategy and methods 4.3 Research procedures 4.4 Ethical, legal and EDI considerations	This section should demonstrated your critical engagement with all elements of research design, including a detailed account of the data and evidence needed in your research, the research methods and research strategies chosen, with justification, and applied within your project. Your account should be supported by a clear rationale and insights from the related literature, and appropriately justified in relation to your research problem, aim and objectives. It should also demonstrate your careful consideration of ethical and legal matters, and that your research complies with your

Table 0.8: Criteria for reviewing your research proposal

Criteria	Prompts
Completeness	Are all sections included and their content complete? What is missing?
Academic writing	Have you applied good academic writing practices throughout? Which main issues do you still have to address?
Logical structure and flow	Have you structured your writing appropriately to ensure a logical flow of arguments? Which restructuring may be needed?
Supporting evidence	Are your key arguments supported by appropriate references or other evidence? Which further evidence is needed?
Citation and reference style	Do all your citations and references comply with the required bibliographical style?
Avoiding plagiarism	Have you acknowledged the work of others and distinguished it from your own appropriately?
Grammar and spelling	Have you proof-read your report carefully to remove all typos and grammatical errors?

- When a large amount of raw data is collected, it is important to devise appropriate ways to store and organise them, paying particular attention to backing them up and protecting personal data.
- Tables are common ways to organise and present data, and a good starting point for data analysis. Pivot, frequency and contingency tables are commonly used in research.
- Descriptive statistics is used to describe data, with various attributes of data sets defined and calculated, such as centrality, dispersion or skewness. Charts are often used to visualise such attributes.
- Inferential statistics is used to make predictions from data, specifically to establish whether patterns or effects observed on sample data can be inferred for the whole population from which the sample was taken.
- Statistical tests are used to establish the statistical significance of observations on a sample in relation to the whole population. They are used both for comparing data to set values and to establish relationships between variables. Many statistical tests exist.
- Coding is the first step in qualitative analysis, and is the process of assigning labels to extracts from a qualitative data set to allow a systematic follow-up analysis. Different approaches to coding exist.

- Qualitative data are heterogeneous in nature, so that they cannot be easily set out in a standard manner. Many different, often bespoke, approaches to present and visualise qualitative data have been proposed in the literature.
- In writing up your data analysis you must decide how to summarise your data, how to report your findings and how to structure your narrative.
- Interpreting your findings means to indicate what you can conclude from the data, how that relate to your aim and objectives, and which new knowledge it contributes.
- An abstract is a short summary of your whole dissertation, written for a specialist audience as a stand-alone piece, that is understandable without reference to any other part of your dissertation.
- The template provided can help you structure your Stage 4 report.