Masters Research Handbook

Lucia Rapanotti and Jon Hall

February 22, 2024

Contents

Stage 4: Gathering 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 gathering and analysis, which is possibly the most exciting, yet demanding, part of your research: this is where you get an opportunity to make your original 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 5.3, 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 5.3 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

Gathering 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 5.3: 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
Gathering 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 gathered	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.

By the end of Stage 4 your data generation, analysis and interpretation should be on a solid ground, and consistent with your aim and objectives. Your research design description should also be close to its final form. Given the criticality of this stage, it is essential that you work very closely with your supervisor throughout.

5.8 Generating raw research data

Your *raw data* represent any data and evidence you generate as part of your research.

Which data you generate and how it is determined by the choices you have made in your research design, informed by your research aim and objectives. In this section, we look at key topics in generating data. This section provides

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

5.8.1 Data generation

Generating data, collecting data, ...

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

A core element of all data generating techniques is the choice of data sources. You've already had experience of doing this when you conducted your literature search as part of Stage **??**[*].

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.

Sampling is the process of selecting a subset[‡] of an infeasibly large population of interest, and is used in

[*]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

[†]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*.

the research strategies that work by estimating or predicting the properties of that population.

Update given Stage 3.

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.

*Random sampling is also called *probability sampling*.

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

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.

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 #2

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 #3

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?

5.8.3 Interviews

Activity: Do I need to know about interviews #4

Check back to your choice of research strategy. If you've chosen one that uses interviews. If so, read the next section and complete the activities.

[etc]

Activity: Deep dive into interviews #5

To find out more about interviews, take a look at these resources:
[C13]–oates2008researching"[p43]–johannesson2014research"[p194]–secor2010social"[p250]–hays2003case"–mcclure2002common"[p52]–pe
ramsook2018methodological"–robertson2002automated"[C4]–kielmann2012introduction";
Englander 2012

5.8.4 Journalling

Activity: Do I need to know about Journalling #6

Check back to your choice of research strategy. If you've chosen one that uses Journalling. If so, read the next section and complete the activities.

[etc]

Activity: Deep dive into journalling #7

To find out more about journalling, take a look at these resources:
[Ed]–kadarisman2017classroom"[]–burns2009action"[p55]–peoples2020write"[]–feinblum2016journaling"[]–hayman2012journaling"–mcgrat

5.8.5 Observations

Activity: Do I need to know about Observations #8

Check back to your choice of research strategy. If you've chosen one that uses Observations. If so, read the next section and complete the activities.

[etc]

5.8.6 Questionnaires

Activity: Do I need to know about questionnaires #9

Check back to your choice of research strategy. If you've chosen one that uses questionnaires. If so, read the next 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.

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.

*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 webpages which is of secondary importance. Unless, your questionnaire is about the design of questionnaires, of course.

§Check out the following resources: **<empty citation>**.

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.

5.8.6.1 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.

5.8.6.2 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 gathering 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 #10

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 #11

To find out more about questionnaires, take a look at these resources:
[“nopp C14]–oates2008researching”[p250]–hays2003case”–burns2009action;
mcclure2002common”–najafi2016observation”–robertson2002automated”[C6]–kielmann2012introduction”

5.8.7 Documents

Activity: Do I need to know about documents #12

Check back to your choice of research strategy. If you’ve chosen one that uses documents. If so, read the next section and complete the activities.

[More here]

Activity: Deep dive into documents #13

To find out more about documents, take a look at these resources:
<empty citation>

5.8.8 Focus groups

Activity: Do I need to know about focus groups #14

Check back to your choice of research strategy. If you've chosen one that uses focus groups. If so, read the next section and complete the activities.

[More here]

Activity: Deep dive into focus groups #15

To find out more about focus groups, take a look at these resources:
<empty citation>

5.8.9 Field work

Activity: Do I need to know about field work #16

Check back to your choice of research strategy. If you've chosen one that uses field work. If so, read the next section and complete the activities.

[More here]

Activity: Deep dive into field work #17

To find out more about field work, take a look at these resources:
<empty citation>

5.8.10 Computational thinking

Activity: Do I need to know about computation thinking #18

Check back to your choice of research strategy. If you've chosen one that uses computation thinking. If so, read the next section and complete the activities.

[More here]

Activity: Deep dive into computation thinking #19

To find out more about computation thinking, take a look at these resources:
<empty citation>

5.8.11 Mathematical thinking

Activity: Do I need to know about mathematical thinking #20

Check back to your choice of research strategy. If you've chosen one that uses mathematical thinking. If so, read the next section and complete the activities.

[More here]

Activity: Deep dive into mathematical thinking #21

To find out more about mathematical thinking, take a look at these resources:
<empty citation>

5.8.12 Statistical thinking

Activity: Do I need to know about statistical thinking #22

Check back to your choice of research strategy. If you've chosen one that uses statistical thinking. If so, read the next section and complete the activities.

[More here]

Activity: Deep dive into statistical thinking #23

To find out more about statistical thinking, take a look at these resources:
<empty citation>

5.8.13 Reflexivity

Activity: Do I need to know about reflexivity #24

Check back to your choice of research strategy. If you've chosen one that uses reflexivity. If so, read the next section and complete the activities.

[More here]

Activity: Deep dive into reflexivity #25

To find out more about reflexivity, take a look at these resources:
<empty citation>

5.9 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 #26

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"

****

5.10 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 gathered, 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.

5.10.1 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 5.4 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.

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 5.5, which gives the average price of each combination. The 'grand totals' in the table are also averages, by row and by column.

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 5.6. The grand totals in this

*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 5.4: First few rows of the example dataset

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

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 #27

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.
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 5.4.

Table 5.4: 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 5.5.

*Also known as *cross-tabulation* tables.

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

Table 5.5: 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 5.6 is a contingency table.

5.10.2 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 Section3.5.1, 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.

5.10.2.1 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.
- 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†.

*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...

†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 #28

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 5.7 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.

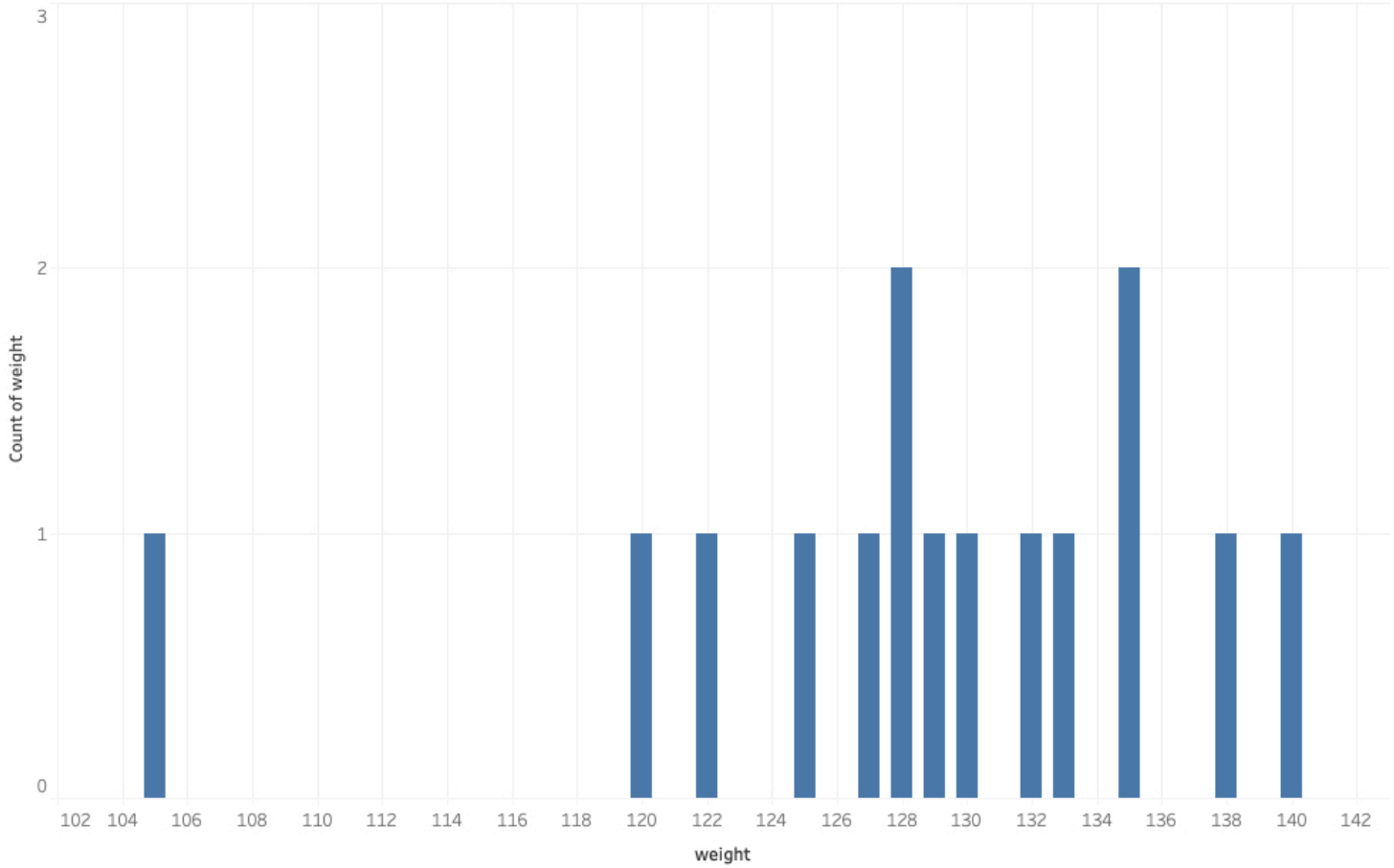

Figure 5.7: Histogram for the apple weights (bin size = 1)

Before we do that, however, it is worth noticing that given our small data set of discrete values, we have 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 5.8 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 5.9 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 5.10. 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 #29

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

Guidance

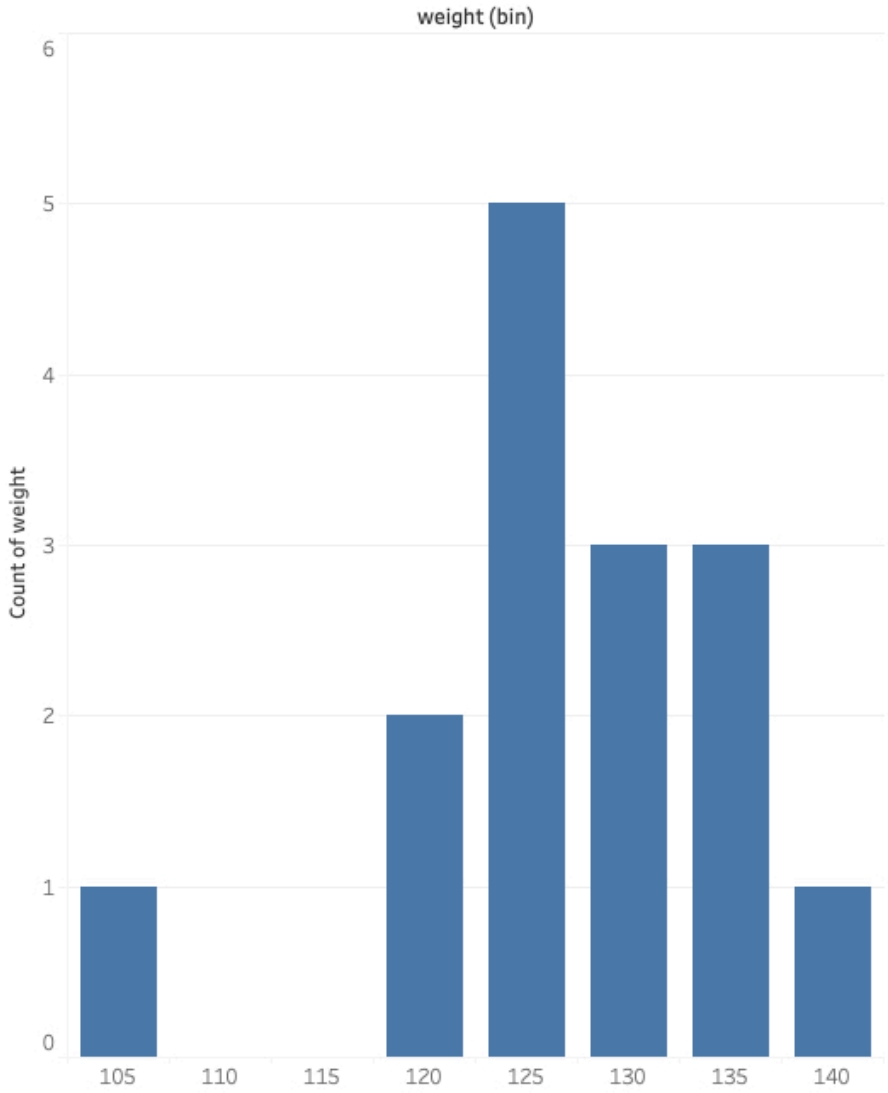

Figure 5.8: Histogram for the apple weights (bin size = 5)

Figure 5.9: Boxplot for the apple weights

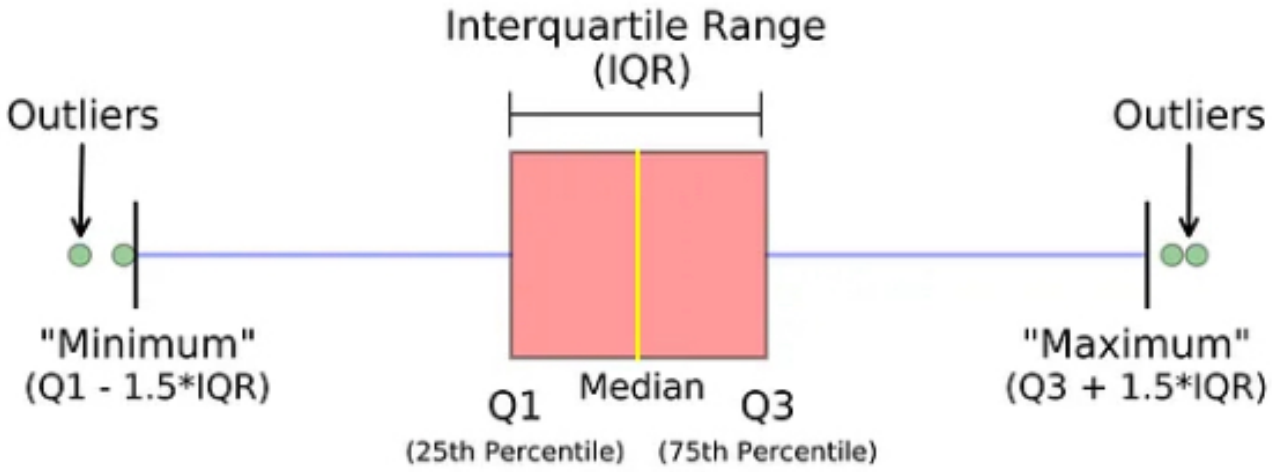

Figure 5.10: The features of a boxplot — LR to redraw as taken from the web

In the current version of Excel, you can generate these charts from the Insert tab, by choosing from 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 5.6 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

Table 5.6: 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

your statistical data analysis, as these provide useful summaries and visualisations of key properties of your 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 #30

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.

5.10.2.2 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 5.7 and 5.8 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 #31

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 5.7: Common statistical tests for comparison. *Parametric* tests apply to normally distributed data (see Section5.10.2.1), 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 5.8: 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.

5.10.3 Quantitative analysis resources

Add here

5.10.4 Qualitative analysis

As indicated in Section 4.2.1.2, 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.

5.10.4.1 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.

Activity: Investigating tools for qualitative data coding #32

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 MAXQDE. They 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.

5.10.4.2 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 #33

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[a] of categorical data.

[a]Which are actually quantitative, but may be the result of qualitative data analysis.

5.11 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 gathered.

- 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 #34

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.

5.12 Interpreting and evaluating data

Having gathered 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 #35

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 gathered and analysed up to this point. You will revisit and expand this work in Stage 5 in order to complete your project.

5.13 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 #36

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.

5.14 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 #37

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 3.7.

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 5.9 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 5.9: much of the content should be carried forward from the previous stage.

Activity: Writing and assessing your report for Stage 4 #38

Using your word processor of choice, revise and expand your Stage 3 report by applying the structure and guidance in Table 5.9.
Assess your report by applying the criteria in Table 5.10. 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.

5.15 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 5.8.6).

Table 5.9: 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 5.10: 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.

Bibliography

Englander, Magnus (2012). "The interview: Data collection in descriptive phenomenological human scientific research". In: *Journal of phenomenological psychology* 43.1, pp. 13–35 (cit. on p. 266).