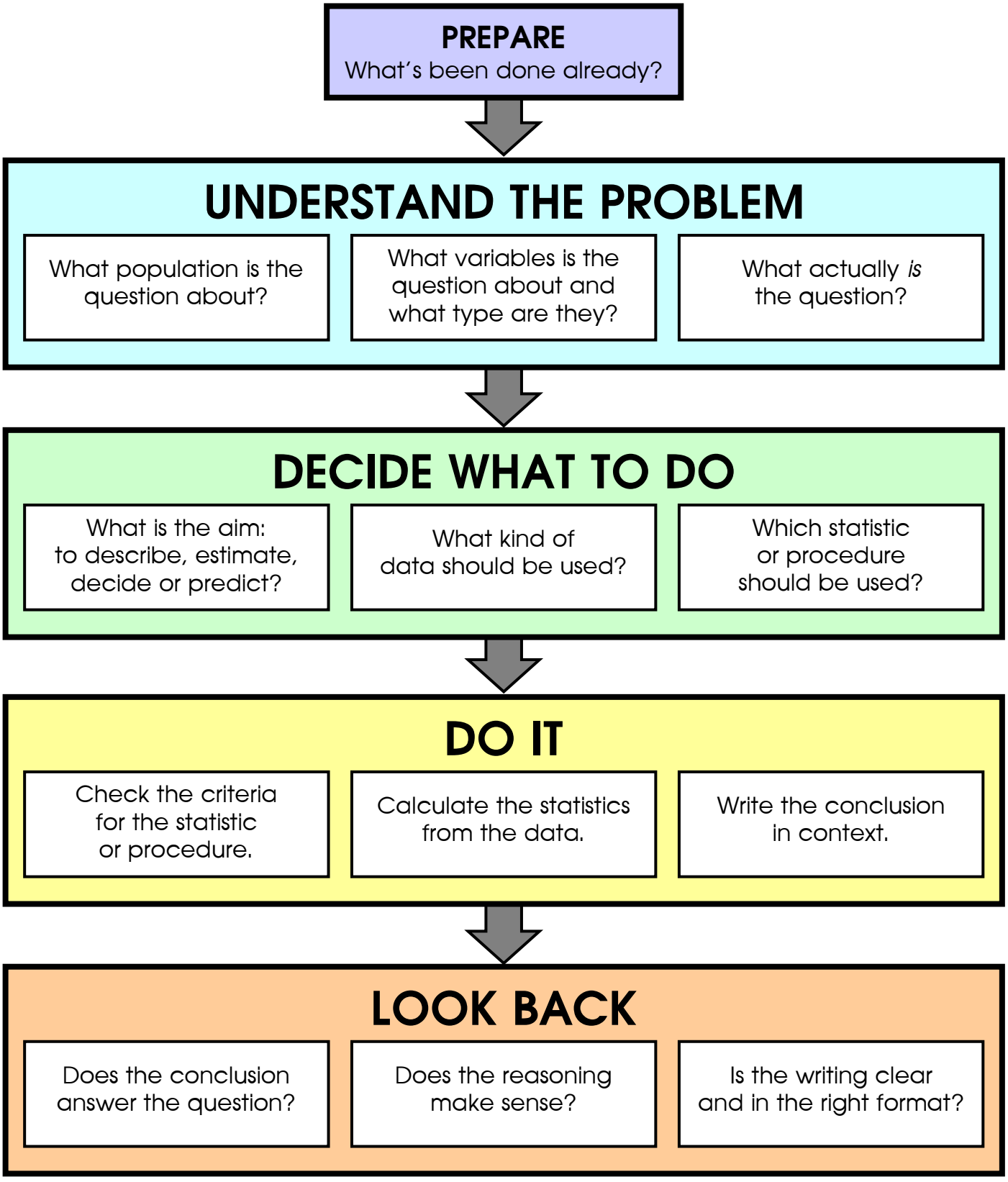
WORKING WITH STATISTICS



WORKING WITH STATISTICS

PREPARE
What's been done already?

Find all the relevant background information.
Describe the relevant theory behind the problem in your own words.
Why would someone be interested in asking the question?
What answer would you be expecting to get?

Find out if anyone has done this before.
Did they get any results that might be useful? (For example, they might have an estimate for a standard deviation.)
Did they run into any problems you should be aware of?
Did they use a method that looks particularly good?
Take into account that published articles may be inaccurate or have missing information.

UNDERSTAND THE PROBLEM
What population is the question about?

The **population** is whatever group of people/things your conclusion is going to be applied to.
Usually it's all people, or all things of a certain type.
Be specific! You need to be specific so you can be clear about exactly how to collect the data later, and how to state your conclusions when you have finished the analysis.

Example:
- All houses sold in South Australia in the year 2009.
- Males between the ages of 10 years and 18 years receiving three doses of 10mg of Immunisation X by injection at intervals of six months apart.
- All students visiting the Maths Drop-In Centre during Semester 1 2009.

UNDERSTAND THE PROBLEM
What variables is the question about and what type are they?

A **variable** is any piece of information you *record* about someone/something from your population.
Usually you'll be recording your variables on the people/things in some kind of sample.
You are likely to have many different variables of interest.
Be specific! You need to be specific so that you are clear about exactly how you are going to collect the data later.

Example:
- The price of the houses sold in South Australia in 2009.
- The real estate agent who sold the houses sold in South Australia in 2009.
- The ages of the males between the ages of 10 years and 18 years receiving three doses of 10mg of Immunisation X by injection at intervals of six months apart.
- Whether the males between the ages of 10 years and 18 years receiving three doses of 10mg of Immunisation X by injection at intervals of six months apart developed swelling at the injection site for the third immunisation.
- The amount of time from the third immunisation until the development of swelling for the males between the ages of 10 years and 18 years receiving three doses of 10mg of Immunisation X by injection at intervals of six months apart.
- The score out of 7 given by students visiting the Maths Drop-In Centre during Semester 1 2009 to indicate how strongly they agree with the statement "Visiting the Maths Drop-In Centre has made me more able to solve problems on my own".

Variables come in different types. You need to know what type of variables you have because the type of variable changes the kinds of questions you can ask and the kind of statistical analysis you can do.

In order to tell what type of variable you have, ask yourself the following questions:

- What actual values could my variable possibly take?
- What ways of comparing these values are meaningful?

There are two main types of variables:

- **Qualitative** (sometimes called **categorical**)**:**
  A qualitative variable has values that talk about what category something is in, or what qualities it has.
  Often the values are *words*, but sometimes they can be numbers too.
  There are two types of qualitative variable:
  - **Nominal:** The values for the variable are just different names for categories or qualities.
    There is no meaningful way to compare the different values – they just describe different categories.
    For example: whether swelling developed (values are "swelling", "no swelling");
    blood group (values are A, B, AB, O);
    colour (values are red, yellow, green …);
    group number (values are 1, 2, 3...).
  - **Ordinal:** The values for the variable describe different categories or qualities.
    It is meaningful to compare two different values and decide which is *more/better*.
    It is either not possible to subtract values to find the difference, or else this difference has no meaning.
    For example: how you like your steak (values are rare, medium rare, medium, well done);
    age categories (values could be 0-9 years, 10-19 years, 20-29 years, …);
    score out of 7 to indicate agreement (values are 1, 2, 3, 4, 5, 6, 7).
- **Quantitative** (sometimes called **numerical**)**:**
  Actual numbers counting or measuring something.
  It is possible and meaningful to subtract values to find the difference between them.
  There are two types of quantitative variables:
  - **Discrete:** The different possible values can be listed one at a time.
    Often discrete variables come about from *counting* things.
    For example: number of houses sold (values are 0, 1, 2, 3, 4, …);
    number of pencils in a pack of 5 that are broken (values are 0, 1, 2, 3, 4, 5);
    the fraction of pizza eaten (values are 1/8, 1/4, 3/8, 1/2, 5/8, 3/4, 7/8, 1).
  - **Continuous:** The different possible values cannot be listed one at a time, because if you pick any two values (no matter how close together they are), there's always another possible value between them.
    Often continuous variables come about from directly *measuring* things.
    For example: temperature (values could be any decimal at all, positive or negative);
    length (values could be any decimal at all, from 0cm upwards).

As well as thinking about these different types, you might be using the variables in different ways. You may want to use some variables to explain some others. For example, you may want to explain a person's heart rate using the time they just spent exercising (under stress); or you may want to explain the price of a house (at least partly) by the real estate agent selling it.
It's important to know which of your variables is which. That is, which of your variables is:

- **Response:** The variable you want to explain by looking at something else.
  In the above examples: the person's heart rate; the house price.
- **Explanatory:** The variables used to explain the changes in another one.
  In the above examples: the time spent exercising; the real estate agent.

Note that you may have several explanatory variables for each response variable. For example, the house price may also be explained using the suburb the house was in. There do exist statistical procedures to help you decide which explanatory variable is the most useful.

UNDERSTAND THE PROBLEM

What actually *is* the question?

Exactly what is it that you want to know?
Make sure you include your population and your variables.

Example:

- Was the price of houses sold in South Australia in 2009 higher for those sold by Real Estate Agent A as compared to the other real estate agents?
- What percentage of the males between the ages of 10 years and 18 years receiving three doses of 10mg of Immunisation X by injection at intervals of six months apart will develop swelling at the injection site for the third immunisation?
- What scores out of 7 did the students visiting the Maths Drop-In Centre during Semester 1 2009 give to indicate how strongly they agree with the statement "Visiting the Maths Drop-In Centre has made me more able to solve problems on my own" ?

DECIDE WHAT TO DO

What is the aim: to describe, estimate, decide or predict?

There are four possible purposes for your statistics:

- **To describe:** The aim is only to describe the data. You want an impression of what the data is like.
  For example: you want to draw a histogram showing the pain scores of heart attack patients.
  Statisticians call this **descriptive statistics**.
- **To estimation:** The aim is to estimate some value that describes the population from taking a sample.
  For example: you want to estimate the percentage of males between the ages of 10 years and 18 years receiving three doses of 10mg of Immunisation X by injection at intervals of six months apart will develop swelling at the injection site for the third immunisation.
  It's very common to use **confidence intervals** when estimating.
- **To decide:** The aim is to decide if something about the population is true or false.
  Usually it concerns some number that describes the population. (This number is called a **parameter**.)
  For example: you want to decide if the median price of houses sold in South Australia in 2009 higher for those sold by Real Estate Agent A as compared to the other real estate agents.
  Statisticians call this **hypothesis testing**.
- **To predict:** The aim is to use your data to predict something in the future. Often this involves creating a mathematical formula that uses some variables to predict others. Some people call this **modelling**.
  For example: you want to use a person's age to predict their chances of developing swelling after immunisation X.

In many cases, you have all three aims. However, for each aim there will be a particular question, or possibly several slightly different questions. Different types of aims will produce different types of questions.
If you have several aims, then start this "Working with Statistics" process again *for each aim* and *for each question.*

DECIDE WHAT TO DO

What kind of data should be used?

The kind of question you want to answer will tell you what kind of data to collect.
You may need to change your question to make it possible to actually answer the question using data and statistics.

The statistical procedure you choose later will tell you more about how to collect the data and how much to collect.
You may need to change the statistic or the data collection procedure so that they match.

How is the data going to be collected?
Are you going to use an:

- **Experiment:** You keep everything the same for your subjects in your sample and you impose some kind of treatment on them.
- **Observational Study:** You do nothing in particular to your subjects, but merely collect data on them.

Observational studies cannot give you information about the *cause* of things, and will not give you as powerful results.
There are many different types of experiment you could do for any given situation.

How are you going to choose your sample?
Will the subjects in your sample be representative of the population?
It would be best to have some form of **randomisation** if you can. Note that there are several different types of randomisation.

DECIDE WHAT TO DO

Which statistic or procedure should be used?

For example, do you want to do a t-test, draw a histogram, or calculate a confidence interval?
The question you want to answer, the data you collect and the way you collected it change the statistic or procedure you need to use.
You need to be able to justify why you chose this statistic or procedure.

DO IT

Check the criteria for the statistic or procedure.

Most statistical tests have criteria that must be met in order for them to work properly.
For example, many assume that the true population variable is normally distributed.
Check these criteria. This often involves drawing a graph, or perhaps performing a particular calculation.
If they aren't met, you may have to change the statistic or procedure you are using; or you may have to change the question you are asking; or you may have to collect new data.

DO IT

Calculate the statistics from the data.

Make sure you are doing the procedure you think you are doing, and using whatever technology you are using in the correct way.

DO IT

Write the conclusion in context.

Make sure you know how to interpret the results of your statistical test or procedure, especially the computer output.
Write your conclusion in terms of the original question and the population. Remember the details from your population and your question – you usually can't apply your conclusion to any larger population than the one in the question.
Make sure you say whether the results are important (as opposed to just significant).

LOOK BACK

Does the conclusion answer the question?

If it doesn't answer the question, then rewrite your conclusion.
Check that you chose the right statistical procedure and check that your sample matches your population.
In severe cases you may have to change the original question or even collect new data.

LOOK BACK

Does the reasoning make sense?

Have you told the reader the question? Have you told them how the data was collected and exactly what variables you recorded for each person/thing in your sample?
Is it obvious what statistical procedure you did?
Have you justified why you chose the statistical procedure, and why have the criteria been met?

LOOK BACK

Is the writing clear and in the appropriate format?

Have you included all the information?
Does the writing say what you think it says? Reading it aloud or asking someone else to read it will help answer this question.
Should the statistics be presented in a particular way? Try comparing to similar documents in your discipline area.

WORKING WITH STATISTIC*IANS*

PREPARE
Ask yourself some questions.

HAVE COURTESY

- Be patient.
- Give plenty of notice.
- Be grateful.

HELP THEM UNDERSTAND

- Let them ask questions.
- Give all the information.
- Use non-technical language.

GIVE RESPECT

- Let them think for themselves.
- Do your part.
- Trust their judgement.

LEARN SOMETHING

- Watch how they think.
- Make notes on the statistics procedures.
- Remember their questions for next time.

QUESTIONS ABOUT STATISTICAL ANALYSIS PROCEDURES

- What is the aim: to describe, estimate, decide or predict?
- What kind of questions can it answer?

- What kind of variables does it work on?
- How many variables does it work on at once?

- Do I need to collect data in a particular way?
- Do I need to record the data in a particular way?

- What criteria do I need to check?
- How do I check them?
- What can I do if the criteria aren't met?

- How do I calculate the statistic / do the procedure?
- What will the results look like?
- How do I interpret the results?