Australian Statistics Project Competition

www.amt.edu.au/statscomp

Teachers’ Pack 2012

Table of contents

1. Welcome! ....................................................................................................................................... 2
2. Prizes and awards ......................................................................................................................... 3
3. For more information ..................................................................................................................... 4
4. Competition rules (how to enter) .................................................................................................... 5
5. Guidelines for projects ................................................................................................................... 8
6. Making it easy for teachers .......................................................................................................... 10
7. What project could we do? .......................................................................................................... 10
8. A checklist for a good project ..................................................................................................... 12
9. Frequently asked questions ......................................................................................................... 13
10. Notes on statistical analysis ...................................................................................................... 15
1. Welcome!

Thank you for your interest in the Australian Statistics Project Competition. The Competition is designed to assist teachers to present lessons on data and statistics in a relevant and relaxed environment. The chance and data stream of school mathematics syllabi is particularly well suited for practical, project-based and team-based learning activities, and the Competition offers a framework and focus to support such learning activities.

In the past, teachers have found this competition easy to implement in the classroom while also being fun, informative and interactive for their students. To help you get started we have put together this Teachers’ Pack, which includes:

- The competition registration form;
- The project submission form;
- A list of the competition rules;
- A list of guidelines;
- A list of ideas for competition projects;
- Some useful resources;
- Answers to Frequently Asked Questions;
- Some notes on statistical analysis.

We trust that using this Teachers’ Pack may prove helpful and provide some interesting ideas for you and your class to help you get started.

About the competition

We learn by observing the patterns and associations in the world. Statistics provides efficient methods for collecting observations and identifying patterns and associations. Statistical results and interpretations inform decision making.

The Australian Statistics Project Competition encourages students to experience learning and decision making through the collection and analysis of data. Students will observe the importance of mathematics for identifying patterns and associations which form the basis for real-world learning and decision making. The overall process of developing, conducting and reporting the data-based project will encourage students’ creativity, planning, teamwork, accuracy, mathematics and computing skills, mathematical and statistical thinking, and clarity of communication. Participants will receive positive feedback through certificates, commendations for excellence and even cash prizes for state and national winners.

The competition is run by academics from seven Universities and administered by the Australian Mathematics Trust. The entry fee, including GST, is $19.80 per team of two or three students.

In contrast to a traditional mathematics competition the Australian Statistics Project Competition allows participants to experience the application of mathematics to real world problems while:

- choosing to investigate questions of personal interest;
- learning to work in a team;
- experiencing problem-based learning;
- taking time to develop their ideas; and
- applying all of their skills (creativity, design, management, record keeping, analysis, logic, interpretation, spelling and writing skills) in a mathematical context.
2. Prizes and awards

The competition is run by academics from seven Universities and administered by the Australian Mathematics Trust.

The entry fee, including GST, is $19.80 per team of two or three students.

Entries in each state/territory will be judged by judges appointed by the competition organisers in that state/territory.

The best entries from each state are submitted for National judging, and National winners chosen from these Projects. National winners are not eligible for State-level prizes.

Judging will be based on the judging criteria as detailed on the competition website and Section 8 in this handbook. The judges' decision is final.

Winners of major prizes will be announced on the Competition webpage in October and their schools will be notified by the Australian Mathematics Trust.

**National prizes ($240)**

There will be one National Prize in the Lower Secondary division and one National Prize in the Upper Secondary division. Teams winning a National Prize will receive $240 to be shared between team members. (National prize winners will be ineligible for State prizes.)

**State prizes ($120)**

There will be one State First Prize in the Lower Secondary division and one State First Prize in the Upper Secondary division in each of the seven regions. Teams winning a State Prize will receive $120 to be shared between team members.

**Highly Commended Certificates**

Each member of each team whose Project was highly commended by the judges will receive a certificate of commendation.

**Participation Certificates**

Each member of each participating team will receive a certificate.
3. For more information

**Competition web pages**
The web page for the Australian Statistics Project Competition is: [www.amt.edu.au/statscomp](http://www.amt.edu.au/statscomp)

This web site includes the contact details, registration and submission forms for each state, important information about the competition and it also points to other websites that may be of interest.

For further information please read the frequently-asked questions (Section 9 below).

**Contacts**
Submit questions regarding registrations, payments and prizes to:

Australian Mathematics Trust  
University of Canberra  
Locked Bag 1 Canberra GPO 2601  
Tel: +61 2 6201 5137  
Fax: +61 2 6201 5052  
Email: amc@amt.edu.au  
Web: [www.amt.edu.au](http://www.amt.edu.au)

Submit entries online or to your state coordinators. Submit any questions regarding statistics to your state coordinators.

**Queensland and Northern Territory**
Ms Taryn Swan  
University of Southern Queensland  
Springfield Campus  
PO Box 4196  
Springfield QLD 4300  
Email: taryn.swan@usq.edu.au  
Phone: +61 7 3470 4415

**New South Wales**
Dr Chris Reading  
SiMERR National Centre  
Education Building  
University of New England  
Armidale NSW 2351  
Email: statscomp@une.edu.au  
Phone: +61 2 6773 5060

**Australian Capital Territory and Southern NSW**
Dr Alice Richardson  
Faculty of Information Sciences and Engineering  
University of Canberra  
Locked Bag 1 Canberra GPO 2601  
Email: Alice.Richardson@canberra.edu.au  
Phone: +61 2 6201 2444

**Victoria**
Dr. Sue Finch  
Statistical Consulting Centre  
The University of Melbourne  
VIC 3010  
Email: s.finch@ms.unimelb.edu.au  
Phone: +61 3 8344 6994

**Tasmania**
Dr Des FitzGerald  
School of Mathematics and Physics  
University of Tasmania  
Private Bag 37  
Hobart TAS 7001  
Email: D.Fitzgerald@utas.edu.au  
Tel: +61 3 6226 2445

**South Australia**
Sue Middleton  
Mathematics  
University of Adelaide  
SA 5005  
Email: sue.middleton@adelaide.edu.au  
Phone: +61 8 8303 3696

**Western Australia**
Prof Ross Taplin  
School of Accounting  
Curtin University of Technology  
GPO Box U1987  
Perth WA 6845  
Email: R.Taplin@curtin.edu.au  
Phone: +61 8 9266 3033

**National coordinator**
Dr Peter Dunn  
Biostatistician  
Faculty of Science, Health and Education  
University of the Sunshine Coast  
Email: pdunn2@usc.edu.au  
Phone: +61 7 5456 5085
4. Competition rules (how to enter)

To enter the competition, the following rules should be followed. Teams should register (see Appendix A) by 22 June and submit their projects (see Appendix B) by 7 September. Note that there is a $19.80 entry fee per project. Schools may enter any number of projects. In overview, follow these steps:

1. Determine your division
2. Find a team and a teacher
3. Register your team/school.
4. Find a good project
5. Plan your project
6. Conduct your project
7. Prepare your presentation
8. Submit your project
9. Projects are judged

**Determine your division**

Please ensure you enter the correct division of the competition.

- **Lower secondary (Years 7-9):** all team members are in grades 7, 8 or 9 at high school.
- **Upper secondary (Years 10-12):** at least one team member is in grades 10, 11 or 12 at high school.

**Find a team and a teacher**

Statistics competition teams must normally be made up of a minimum of two students and a maximum of three students all from the same school.

All students in the team must be officially enrolled in the correct year level for the division entered at 22 June. Each team needs to be sponsored by a teacher in the same school as the students on 22 June.

The sponsoring teacher must take responsibility for certifying that the students have the school’s permission to enter the competition, that all team members are properly enrolled students at that school and that the team adheres to the competition rules.

**Register your team/school**

Teams planning to enter the competition are encouraged to register and pay their entry fee by 22 June 2012.

Registration involves completing the registration form (Appendix A) and forwarding it to the Australian Mathematics Trust.

The entry fee is $19.80 per team.

The Australian Mathematics Trust is a not-for-profit organisation. The entry fee is levied to cover the competition's administration, prizes and certificates.

**Find a good project**

A good project:
- answers a question which is interesting to you and others;
- answers a question which is simple and clear;
- looks at something which naturally varies (e.g. age, height, weight, opinions, tastes, behaviours, performance, time taken, weather, etc); and
- usually involves a comparison (between genders, between classes, between regions, between types, before against after, etc).

You can choose to collect your own data: your data collection could be based on one of the project ideas provided in Section 7 below or could be your own original idea.

You can choose to use data which already exists: your project could be based on one of the CensusAtSchool projects ideas or some other data set you find on the world-wide-web (you must acknowledge and reference your source). Projects should be achievable within a one month timeframe.
Plan your project

To make a project effective, a proper plan is essential. Ask yourself these questions:

What exactly is/are the question(s) of interest? That is, what exactly do you want to find out? Be clear in your objectives, and don't be too ambitious!

How can you obtain data to help answer the question of interest? That is, what data do you need to answer the question posed above? What data would be useful? How can this data be obtained? What other issues may affect what you actually measure?

Check your data is appropriate to the question
There is no point asking questions about green frogs if you can only find red frogs in your area.

Conduct your project

The team may obtain assistance and advice from resource material or resource people available to them.

The project should acknowledge all sources and assistance with appropriate footnotes and references on their final presentation.

Prepare your presentation

Summarise and check the data. Check that the data you have recorded are sensible, that the units of measurement are recorded (millimetres, centimetres, etc.). It is not unusual to find that there have been errors in recording data; try to find these errors, and either fix them if you can, or set them to missing.

Use statistics and graphs to analyse your data and answer your question.

Don't over-interpret your results. It would not be appropriate to make recommendations that would take millions of dollars to implement if you only have a small number of observations and trainee analysts. If you find something which you think could be interesting suggestions for further investigation may be appropriate.

Submissions must be electronic documents and adhere to the following guidelines

- Submissions must be in either pdf (.pdf files) or PowerPoint (.ppt files) format.
- Submitted files must be less than 2 MB in size.
- Presentations should be a maximum of eight slides.
- Presentations should be easy to read: The minimum font size should be 30 point.
- Don't overload your presentation with too much information; almost always, well-presented, simple presentations are better than presentations which try to do too much.
- Keep the layout and sequences logical.
- Use headings on each slide.
- Ensure any graphs and tables are clearly labelled, titled, uncluttered, easily read, and easily understood. Experience tells us that graphs are often poor, especially the default graphs produced by Excel.
- Remember the purpose of the presentation is to convey information as clearly and simply as possible: keep it simple, and keep the amount of text to a minimum. Too much text invites people to look at someone else's presentation instead.

The presentation should describe the purpose of the study, the methods used, the results obtained and the conclusions drawn. You should show how your study has answered your research question.

More useful information may be found at the website for the American Statistical Association's Poster and Project Competition at http://www.amstat.org/education/posterprojects/whatisastatposter.cfm

Submit your project

The school is responsible for submitting the project to the correct location. This will probably be organised by the supervising teacher.

Remember to submit by 7 September 2012.

To submit your entry in either PowerPoint (.ppt file) or pdf (.pdf file) format, email it to your State Coordinator and copy the email to: mail@amt.edu.au

In the email, name the project and students (with their school years).

Remember: Your file must not exceed 2 MB in size!
Projects are judged

The judges' decision is final.

Entries in each state/territory will be judged by judges appointed by the competition organisers in that state/territory.

The best entries from each state are submitted for National judging. National winners are not eligible for State-level prizes.

Judging will be based on the judging criteria as detailed on the competition website and Section 8 in this handbook.
5. Guidelines for projects

Most projects can be broken down into four broad steps, which can then be broken into smaller steps. The four main steps are:
1. Devise a question
2. Obtain the data
3. Analyse the data
4. Report and present the conclusions.

Devise a question

Data are used to answer questions. The question and the data need to match. There is no point trying to determine how far green frogs will jump if you only have data on red frogs’ jumps. If you plan to collect your own data, you should design the data collection around the question of interest. If you want to make use of existing data, you have to choose a question which is appropriate for the data. To properly plan a project, the following questions should be asked:

- **What exactly is/are the question(s) of interest?**
  It is important to know and understand what project you are doing, why you are doing it, and what you hope to learn from the project. Be clear in your objectives, and break it down into many smaller parts if you can. Don’t try to do too much all at the same time. Trying to do too much at once spoils some otherwise good projects.

- **What data did I use to answer the question of interest?**
  Finding an answer to your question is not the end of the process. You would normally want to tell others the answer – and they will be interested in seeing how you arrived at that answer. The role of your project is to show your question and answers to the the readers and to convince them that your conclusions are correct. One rule of thumb is to give sceptical readers sufficient information about your methods, so that they could run your study themselves, if they wanted to, to check your results.

Obtain the data

Carefully consider and document how you obtained your data.

Collect and compile the data

Collect your data carefully. Errors in data collection or data selection could invalidate your results. Look at the data you have and ask yourself whether it looks sensible. If not, double check for any errors. It is not unusual that there have been errors recording data; try to find these errors and either fix them if you can or set them to missing data.

Consider how you are going to manage the data

You may wish to use a computer at some stage to help prepare graphs or summaries. Just be careful to resist the temptation to use computer functions which you don’t really understand and can’t really explain on your project.

With smaller data sets, pen and paper analyses are feasible and completely acceptable. Just make sure you maintain accuracy and legibility.

Consider the completeness and adequacy of your data

Do you really have enough observations to answer the question of interest? Do you have all the measurements which may be relevant to the question of interest? Does the data set contain any missing data and, if so, how many. Identifying and acknowledging weaknesses in your data proves your understanding of the requirements of the statistical analyses to the reader.

Analyse the data

When analysing the data, you seek to identify the main patterns within the numbers you have collected.

Use graphs and statistics to analyse your data

Use some simple statistics to find out what your data is saying. You may use pictures and graphs (“a picture tells a thousand words”) such as histograms, bar charts, boxplots, scatterplots, etc and/or numeric statistics such as means, standard deviations, medians, two-way tables, etc. (See Section 10 for further details.)
Report and present the conclusions

When you report your findings, you must be clear and concise about what you did and how the data answers the question of interest.

Interpret and make conclusions
One of the goals of statistics is to gather data and turn it into information. Look at your data: what does it tell you? How does it answer your original question(s)? Be honest and open with any potential weaknesses – prove that you understand the strengths and weaknesses of what you have done. Suggest how you could do it better next time.

Avoid over-interpreting the results
Even if three students at your school have reported near misses when crossing a particular road this may not necessarily be sufficient justifications for recommending the building of a $2million pedestrian overpass. You are presenting no data on the effectiveness of pedestrian overpasses – you are only showing that some students are reporting a problem with the current crossing. This may warrant further investigation of how to improve the road. Keep your conclusions and recommendations in context with the results.

Check your presentation
You may like to read the Competition Checklist below for what to include on the presentation. Check for errors. In the past, good entries have been spoilt by incorrect wording, typing errors and calculation errors. In at least two of those cases, the entries missed out on prizes because of these simple errors.

Competition Checklist

- Have you registered?
- Does your entry clearly and concisely state the objective and research question?
- Are there clear statements of what data has been collected, and how?
- Is the collected data appropriate for answering the research question?
- Are the graphs/statistics appropriate for displaying and summarizing the data?
- Is the analysis appropriate?
- Are the conclusions clear, correct and appropriate?
- Is the repetition of an experiment to measure variation suitably acknowledged?
- Has the experiment been repeated enough times to give accurate conclusions?
- Is the file size less than 2MB?
- Is the discussion, and suggested improvements given?
- Is the minimum font size 30 point?
- Is the spelling and grammar correct and appropriate?
- Is the language clear, concise and error free?
- Is the project clearly and logically set out?
6. Making it easy for teachers

Connection to the syllabus

The Australian Statistics Project Competition has been designed to help teachers apply the syllabus. Our aim is that the competition should not be creating more work for the teacher; it should be helping the teacher implement the syllabus.

Issues considered during the design of the competition include:
- Syllabi usually call for students to be effective communicators. Creating a project is an excellent way of learning communication skills. The students must be able to communicate what they have done, the results they have obtained, and the questions they are answering. The students must write succinctly and clearly.
- Syllabi usually call for students to collect, analyse, explore and interpret data. The Australian Statistics Project Competition explicitly addresses this component
- Syllabi often encourage group work. The Schools Statistics Project Competition is a fantastic way for students to reap the benefits of teamwork.

The competition makes it easy to implement the syllabus as it provides the framework, ideas, details for specific projects and motivation to compete.

Time allocations

In past competitions, teachers have indicated that no more than four lessons are required for students to select a question, collect and analyse data and prepare their project completely in class. This figure, naturally, will vary from class to class, students to student. Because students work in teams, the work can be divided.

7. What project could we do?

Students can choose any project they wish that conforms to the aims of the competition. Some ideas are listed below; they might get you thinking of other ideas.

Students are encouraged to consider using data collected from (for example) a science class. You may have collected data in biology about the best conditions for growing bean sprouts. What a great idea for a statistics competition project!

Some ideas to get you started are given here. The competition website gives more details about some of these project ideas. Don’t feel restricted by the ideas listed; any topic can be chosen as long as it is statistical and sensible.

Examples requiring data collection

More detailed descriptions of these examples can be found by going to http://www.amt.edu.au/statscomp and selecting ‘project ideas’.

Splat that! (Physics)
Objective: To describe the accuracy and reliability of a simple catapult.

Ants for lunch (Biology)
Objective: To determine whether ants are more attracted to jam sandwiches or meat sandwiches.

Kids Bizz. (Psychology)
Objective: To explore the relationships between smoking, drinking, pocket money and parental marital status.

Tack toss. (Probability)
Objective: To determine the chance of a thumb tack landing point down.
Examples using CensusAtSchool data

More detailed descriptions of these examples can be found by going to [http://www.abs.gov.au](http://www.abs.gov.au) and selecting ‘Education’ then ‘CensusAtSchool’.

**Are Students Being Taken For A Ride?**
Students can investigate how students in Australia travelled to school in 2006 compared to 2004, using Excel to sort and analyse the data.

**Are males better drivers than females?**
Students compare the mean and median reaction times of males and females. Box plots can be used as a measure of spread. Excel functions can be used to construct a summary table and draw appropriate graphs.

**Water management**
Students collect and analyse CensusAtSchool data about the attitudes on the importance of environmental issues and water, in particular. They can use the tables provided to construct graphs to support their findings and conclusions

**Or be creative**

**Transport**
- How old are cars on the road?
- Cycling injuries: Helmets, do they make a difference?
- Who uses which forms of transport and why?

**Science**
- Does buttered bread always land buttered side down?
- How does the boiling point of water change with concentration of dissolved salt?
- What factors determine the coefficient of friction?
- Are extra-life batteries worth the extra expense?
- How do plant germination times vary under different conditions?
- How does the level of ant activity change with temperature?
- How fast do cups of water lose their heat?
- Do bananas keep better in the fridge wrapped in foil, cling wrap, or in a natural state?
- Does fruit ripen faster when placed in a paper bag with bananas? Does this work with all fruits, or just some?

**Psychology**
- Does reaction time vary with age?
- How does short term memory retention vary across various groups?

**Marketing**
- Do generic brands taste different?
- Can people distinguish amongst brands of tea/coffee/cola by taste alone?
- Are opinions affected by advertising?
- How knowledgeable are people about nutrition?

**Behaviour**
- How much do people eat?
- How much do teenagers drink?
- How do homework habits vary with grade/gender/type of school?

There are plenty of other options also. Don’t feel restricted by the suggestions given here; they are just ideas!
8. A checklist for a good project

Please see examples of good projects from previous years on the website www.amt.edu.au/statscomp

Judges will be looking at both the content of the project and the presentation. Before submitting the project, the following check list will ensure the judges will look favourably at your project.

Judging criteria

The judging criteria are given below. Please understand that the judges apply these criteria in an age appropriate way; that is, the standard required from a Year 11 student will be much higher than that required by a Year 8 student. Please read the Competition rules. Failure to adhere to these rules may result in disqualification from the Competition.

The projects will be judged based on the following three criteria, each equally weighted.

A. Planning and data collection

- Does the project clearly and concisely state the objective and research question?
- Are there clear statements of what data has been collected, and how? (If the CensusAtSchool Random Sampler is used, you must explain how you used it and exactly what you extracted. If the data were obtained from elsewhere, such as a webpage, you must acknowledge and reference the source.)
- Is the collected data appropriate for answering the research question?
- Is the research question creative, original and interesting?

B. Analysis

- Are any graphs/statistics appropriate for displaying and summarizing the data?
- Is the analysis appropriate?
- Are the conclusions clear, correct and appropriate?
- Is there repetition in the project to measure variation?
- Has the project been repeated enough times to give accurate conclusions?
- In the discussion, are suggested improvements given?

C. Presentation

- Is the presentation easily readable? This means:
  - Ensuring font sizes are large enough
  - Ensuring there is not too much text. A presentation is meant to convey information concisely. Too much text invites people to look at someone else’s presentation instead.
  - Using plain, simple fonts. Fancy fonts can be hard to read, though occasional use for headings may be suitable.
  - Ensuring graphs are clear, titled, uncluttered, and well explained.
- Is the spelling and grammar correct and appropriate?
- Is the language clear, concise and error free?
- Is the presentation clearly and logically set out?
- Is the design creative and original?

Further advice for making a good project

The following advice has been adapted from http://www.math.uow.edu.au/current/talk.html:

- Don’t think that making a project presentation is easy. The presentation of your project must be very good if people are to look at all of it.
- Do keep the material short, simple and easy to read. No-one is likely to spend more than about three minutes reading it.
- Think very clearly about what material you will display. Make sure that your type is not too small. (Do not use anything smaller than 16 pt font.) Don’t choose colours that are hard to see. Don’t use a dark background.
- Do check the English spelling and grammar in your presentation. This is good advice whether English is your native language or not: remember that errors in English language or grammar will show up very clearly in writing.
- Do get people to provide constructive criticism on the content of your presentation before you prepare the final version. Take note of that criticism.
- Do include a picture or graph if, by doing so, you can avoid even a few sentences. But remember that the display should be self-contained, with its own legend and title. The reader must be able to understand it very quickly.
9. Frequently asked questions

Is there any cost to enter the competition?

Yes, the entry fee is $19.80 per project. Projects should be prepared by teams of two or three students. Entry fees are collected by The Australian Mathematics Trust to cover the competition's administration, prizes and certificates. The Australian Mathematics Trust is a not-for-profit organisation.

Do we need to register?

Yes. Completing your registration by 22 June 2012 will guarantee your entry. Late registrations will be accepted wherever possible.

How long will it take to complete a project?

The time needed to complete a project will vary from class to class, and from student to student. In previous competitions, teachers have reported that projects could be completed in under four class lessons.

What are the benefits to students?

- It is a fun, interactive way of teaching the school curriculum.
- It will give them greater insight into mathematics and statistics as they plan, gather their own data, analyse their results, and create a presentation.
- The students gain experience in working as a team.
- Students enjoy it!
- There are many jobs requiring an understanding of statistics, and the competition gives practical experience in this field where there are many jobs.
- All participants receive certificates and are in the running for National and State prizes.

What are the benefits to teachers?

- It allows components of the school curriculum to be taught using a fun, different approach.
- It generates interest in, and demonstrates the relevance of, mathematics and science.
- It is a short and varied, yet fun and interesting activity.
- All students in the class can participate.

Can schools enter more than one entry?

Definitely! Everyone in your class can be part of a project entry.

How is it related to the CensusAtSchools program?

The competition is independent of but completely compatible with the Australian Bureau of Statistics’ CensusAtSchools program. The competition and the CensusAtSchools programs have almost identical aims. The competition provides students completing the CensusAtSchool program with a forum for displaying and testing what they have learnt. However, you do not have to participate in the competition to successfully complete the CensusAtSchools program and you do not have to participate in the CensusAtSchools program to enter the competition.

Do I have to do one of the suggested projects?

No! The suggestions are just that: suggestions. In the 2004 competition, for example, the winning entry used the Splat That! idea, while second and third place went to original ideas. It becomes a judgment call: some originality is desirable, but projects also need to be scientifically accurate and correct. Following the suggested projects too closely lacks originality but the more originality you insert into the project, the more chance of straying into error.

Do I have to use the CensusAtSchools data?

No. The competition does require you to analyse and interpret data. Whether you choose to pose a question which can be answered using the existing CensusAtSchools data set or whether you choose to pose a question which requires you to
collect your own data is entirely up to you. All entries will be judged on how well they describe the specific details of their data collection or data selection process.

Who can enter?

Any secondary students can enter. Students should work in teams of two or three, but team members do not have to be in the same year level as each other. Entries are invited from private schools, state schools or even home-schooled or distance education students (third place in 2003 was awarded to the Brisbane School of Distance Education).

Can individual students enter a project?

Usually, entries must come from a team of two or three. If you have special circumstances Contact the organizers for advice and clarification.

Do all students in a team need to be in the same year level?

No. However, the project will be judged according to the highest year level of the team members. (Usually, all team members are from the same school year level.)
10. Notes on statistical analysis

We learn through repeated observations

Every single child knows that if they drop a ball it will fall to the floor. They learn to expect a dropped ball to fall to the floor long before they ever hear of Isaac Newton or the law of gravity. Children (and indeed all people) learn about their world through repeated observation. Even very young children can be confident that a dropped ball will fall to the floor because they have seen it happen in a similar way countless times before.

When we want to answer new questions on how the world works, the best strategy is often to try to get lots and lots of observations related to that question. If we want to know if rainfall has increased on a particular town in the last 50 years we look to compare recorded (observations) of rainfall 50 years ago with current records. If we want to know how safe aeroplanes are we look to see how many crashes there were observed across a large number of flights. If we want to estimate how long a particular type of light bulb will run before it blows, we wouldn't just test a single bulb – that one may be unusual or faulty or different from normal in some way. We would test a large number of bulbs and see how long these bulbs normally last.

Repetition generates data

If we measure how long each light bulb lasts for we will end up with as many numbers as bulbs. Some bulbs will last a relatively long time, some will last a relatively short time but most observations will fall around the middle somewhere. The time that each bulb lasts will be a number and each bulb we test will generate a number. If we test 50 bulbs we will be able to record 50 numbers. These numbers, representing our repeated observations, are called data.

We are not limited to one observation per bulb. We may also wish to record some information on location of the bulb (eg 1 to represent inside and 2 to represent outside) to see if location affects lifetime of bulbs. In this case, we will have 2 data items for each bulb: one for lifetime and the other for location.

Graphs help show the patterns in data

The question of “how long does a light bulb of this type run before it blows” does not have a simple answer. Even if produced in the same factory and run in the same location, some bulbs will last a little longer than others. NO two items will ever be exactly identical. If we get a slightly different answer for each bulb we test, how can we have a single correct answer to the question “how long do bulbs of this type run for”? Giving 50 answers (one for each bulb) is not very helpful.

One good strategy is to draw a picture or graph of the 50 bulb’s observed lifetimes. (See below for discussion of various types of graphs.) A good graph will allow the reader to see how long most bulbs last and how much the lifetimes of bulbs vary between different bulbs. Such a graph will convey to the reader “Most of the light bulbs last about x weeks, but lifetime was seen to vary from y and z weeks”.

The best primary school students would get to this level of understanding.

We can also use numbers to summarise these patterns

A graph of data is useful as it allows us to immediately see the main characteristics of the data. For certain types of data we may be interested in the centre of the observed results, the spread of the observed results, the proportion of results falling into particular categories, etc. All of these main characteristics of the data can be reported as numbers. The centre of the observations can be reported numerically as a mean or median. The spread of observations can be reported as a standard deviation and the number of observations in any particular category can be reported as a percentage. (See below for further discussion on these numeric measures.)

An important advantage of a numeric (formula based) measures is that there is universal agreement on the value of that measure. Two people looking at a graph can quite legitimately have slightly different opinions of where the centre of the data lies. Anyone calculating a mean from a given data set should get exactly the same result – leaving no doubt or debate over the answer.
We often can’t measure everything – so we use a sample

It is not possible to measure the lifetime of every light bulb the factory produces – there won’t be any left to sell! It is not really possible to measure the average length of a particular species of cockroach – you would have to catch and measure every cockroach of that species in the world. How realistic is that?!

For practical reasons we often measure just a sub-group (or sample) of the individuals/items which are available for measurement. For example, we just test a sample of light bulbs produced by the factory or just measure some of the cockroaches of the species we are interested in.

A properly drawn sample can give us some information about the sample it was drawn from

We have already noted that no two individuals are ever identical. Therefore, when we are using a sample, each of the individuals/items which we are leaving out are different in some way to those which we are including. No member of the sample fully represents the excluded individuals/items. However, a properly drawn sample can provide us with results which are likely to be similar to the results which we would have obtained if we had measured every individual in the population.

Samples are used successfully every day. For example, when a chef is making a sauce she will take a little taste every now and again to check the flavour and consistency are correct. She doesn’t have to eat all the sauce to be sure it tastes nice – a teaspoonful is usually sufficient. The only condition on using a sample is that the sauce needs to be well stirred before taking the sample. If the sauce is burnt on the bottom and not stirred then tasting a sample from the top will not detect the problem.

Stirring the sauce is equivalent to randomising the selection of the sample. A sample consisting of individuals selected at random from the population of interest has the best chance of being similar to the population it was drawn from. If you sample only the easy to catch cockroaches, you may end up with a sample which is larger than the population of all cockroaches given bigger cockroaches are easier to catch.

The best lower secondary school entries should get to this level of understanding.

Things can go wrong in sampling, measurement and interpretation

Technically, samples should be randomly drawn from the whole population of interest. This minimises opportunities for biases. Bias occurs when the sample is systematically different from the population of interest. That is, if you wanted to know opinions of children your age on some particular issue, you should not just interview your friends. Your friends may all be girls – in which case the information you are collecting does not represent the opinions of all Year 8 and 9 children. The opinions of boys would not be represented.

Even if selected completely at random, a sample may not be representative of the population it was drawn from. Randomly select 10 people from your street and there is a chance that all 10 will be female. (That is, you have a 50:50 chance that the first person selected is female, then a 50:50 chance the second one selected is female, then a 50:50 chance the third person is female, ... there is a chance all may be female). Select 1000 at random from your town and there is a chance (albeit minute) that all 1000 are female. There must always be some uncertainty as to whether the results from the sample can be generalised to the wider population of interest.

Non-response can also make samples biased (ie different from the population they are drawn from). If all the 17 year olds on your survey refuse to answer a question, then the answers you obtain provide no information about the opinions of 17 year olds.

Measurements can go wrong too. Poorly posed questions can lead to misleading answers. For example “You agree with <issue> don’t you” will tend to encourage a ‘yes’ response no matter what the issue as people tend to find it easier to agree than to disagree. Measurements of how high a ball bounces will differ according to whether you are looking at the top, bottom or centre of the ball. Measurements of how long a light bulb lasts may only be accurate to the nearest day – it is unlikely that anyone will have the patience to watch the bulbs continuously; the observer may only check the bulb one per day.

The real relationship between shoe size and reading ability in children may be wrongly interpreted as “reading causes your feet to grow bigger”, whereas the real cause of the relationship
is a third, hidden variable: age (older children have bigger feet and better reading ability). Always consider the possibility that there may be other factors which you haven’t measured which could be affecting observed results.

It is always important to review all results asking whether there are any errors, uncertainties or potential weaknesses in sample, measurement and interpretation of these results. All statistical results should be accompanied by an appropriate discussion of the strengths and weaknesses of the methods used.

The best upper secondary school entries should get to this level of understanding.

We can use probability models to estimate the most likely characteristics of a population using results obtained from a sample.

Statistical inference is a group of mathematical techniques which are used to generalise information obtained from random samples back to the wider populations which we are actually interested in. For example, when conducting a survey of students in the playground we are not specifically interested in the answers from the 20 children who actually answer the survey. We are actually interested in what the answers from these 20 children can tell us about the all students in the school or all young people in the town.

The tools of statistical inference are beyond the scope of the current competition (they represent the majority of the content in university statistics courses). However, better high school students will consider how the sample of individuals that they have looked at relate to the wider population which they are really interested in.

No students are expected to apply formal statistical inference techniques in this competition.

Some tools for data analysis

Methods of statistical analysis may be divided into ‘Descriptive Analyses’ and ‘Statistical Inference’. Descriptive analyses can be conducted on data collected from whole populations and on data collected from samples. Statistical inference explores the mathematical relationship between samples and populations and will not be discussed here.

Descriptive analyses

Descriptive analyses summarise the main features of the data. Raw data is too complex and too detailed to be of any direct use – we need ways of extracting the key information from otherwise difficult to read data sets.

Example:

Here is a data set I collected from a class of university students:

<table>
<thead>
<tr>
<th>Student</th>
<th>Age in years</th>
<th>Gender</th>
<th>Faculty</th>
<th>Mark (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>Female</td>
<td>Business</td>
<td>79</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>Male</td>
<td>Business</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>Male</td>
<td>Business</td>
<td>62</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>Female</td>
<td>Business</td>
<td>59</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>Male</td>
<td>Engineering</td>
<td>58</td>
</tr>
<tr>
<td>6</td>
<td>33</td>
<td>Female</td>
<td>Sciences</td>
<td>75</td>
</tr>
<tr>
<td>7</td>
<td>35</td>
<td>Male</td>
<td>Sciences</td>
<td>71</td>
</tr>
<tr>
<td>8</td>
<td>29</td>
<td>Male</td>
<td>Business</td>
<td>82</td>
</tr>
<tr>
<td>9</td>
<td>41</td>
<td>Male</td>
<td>Engineering</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>Male</td>
<td>Business</td>
<td>69</td>
</tr>
<tr>
<td>11</td>
<td>24</td>
<td>Female</td>
<td>Business</td>
<td>56</td>
</tr>
<tr>
<td>12</td>
<td>31</td>
<td>Male</td>
<td>Business</td>
<td>72</td>
</tr>
<tr>
<td>13</td>
<td>24</td>
<td>Female</td>
<td>Business</td>
<td>80</td>
</tr>
<tr>
<td>14</td>
<td>31</td>
<td>Male</td>
<td>Engineering</td>
<td>87</td>
</tr>
<tr>
<td>15</td>
<td>21</td>
<td>Male</td>
<td>Business</td>
<td>87</td>
</tr>
<tr>
<td>16</td>
<td>25</td>
<td>Male</td>
<td>Business</td>
<td>83</td>
</tr>
<tr>
<td>17</td>
<td>24</td>
<td>Male</td>
<td>Engineering</td>
<td>73</td>
</tr>
<tr>
<td>18</td>
<td>23</td>
<td>Male</td>
<td>Business</td>
<td>90</td>
</tr>
<tr>
<td>19</td>
<td>31</td>
<td>Female</td>
<td>Business</td>
<td>78</td>
</tr>
</tbody>
</table>

There are 19 students in the class. Each row of data relates to one individual student. That is, the first student in the data set is a 30 year old female who is enrolled in the Business Faculty and received a mark of 79%. The columns contain ‘variables’. The variables are the characteristics of the individuals which I actually measured. The second column is the variable ‘Age in years’. I asked each of the 19 students their age and recorded it in this column. The third column records whether each of the students is a male or female, etc. Age, gender,
faculty enrolled in and marks vary from individual to individual (hence the name “variables”). The first column is not a variable; it is a label. Labels are not characteristics – the label tells me nothing about the individual’s abilities or attributes. Labels are just a tag I put on to that individual for my own convenience.

It can be difficult to see all the patterns in the raw data. You can see that there are 19 students in the class. You can see that there are more males than females and that the majority of the students are enrolled in the Business Faculty. If you look closely you can see that students’ ages range from 19 to 41 years. However, even with this small data set, it is very difficult to answer more complex questions like:

- What proportion of students are less than 30 years of age?
- Do older students get better marks?
- Do females get better marks?

There are a range of statistical tools available to help us summarise the important patterns in the data set. In order to select the right tool for the right job we first need to understand the difference between categorical and quantitative variables.

Quantitative variables are measured on a numeric scale. For example, age is measured in years and years is a numeric scale. You can do maths on numeric scales and get answers that make some sense. That is, if student 1 is 30 years of age and student 2 is 26 years of age then student 1 is 30-26=4 years older than student 2.

Categorical variables are measured on categorical scales. For example Gender is measured in categories (male or female). You can’t do maths on categorical scales. For example, Male+Female=? does not have a mathematical answer.

In the example above, age and mark are quantitative variables and gender and Faculty enrolled in are categorical variables.

Once you have figured out which of your variables are categorical and which are quantitative you can move on to selecting the appropriate statistical tool. The choice of which tool to use is summarised in the following tables:

### Statistical tools for summarising a single variable

<table>
<thead>
<tr>
<th></th>
<th>Categorical variables</th>
<th>Quantitative variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphical summaries</td>
<td>Bar chart</td>
<td>Histogram</td>
</tr>
<tr>
<td></td>
<td>Pie chart</td>
<td>Boxplot</td>
</tr>
<tr>
<td>Numerical summaries</td>
<td>Counts</td>
<td>Percentage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean &amp; standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Median &amp; quartiles</td>
</tr>
</tbody>
</table>

### Statistical tools for summarising the relationship between two variables

<table>
<thead>
<tr>
<th></th>
<th>Categorical and categorical</th>
<th>Categorical and quantitative</th>
<th>Quantitative and quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphical summaries</td>
<td>Complex bar charts</td>
<td>Side-by-side boxplots</td>
<td>Scatterplots</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multiple histograms</td>
<td></td>
</tr>
<tr>
<td>Numerical summaries</td>
<td>Frequency tables</td>
<td>(ANOVA)</td>
<td>(Correlation) (Regression)</td>
</tr>
</tbody>
</table>

Some of these statistical tools are described below.

**Frequency counts**

This is simply a count of the number of times each outcome occurs. For example a frequency count on gender in the example above is:

- 13 males and
- 6 females

Sometimes people prefer to present percentages rather than raw frequencies. In this example the the relative frequencies are:

- 68% males (ie 13/19 * 100% = 74%) and
- 32% females

If using relative frequencies, make sure you also mention the total number of individuals in your data set (the sample size). One out of ten and 100 out of 1000 both have the same relative frequency (10%) even though the study on 10 individuals provides far less complete evidence than the larger study.
Bar charts

Bar charts are graphical representations of the frequency counts. Examples of bar charts for gender are:

[Image of a bar chart showing gender distribution]

The categories (male and female in this case) are displayed on the horizontal axis, and the frequency counts or relative frequencies are given by the heights of the bars. All titles and scales are vital. Without the titles and scales, the reader will have difficulty in understanding the information presented. Notice the inclusion of the sample size (n=19) when using relative frequencies.

Pie charts

Pie charts are also constructed from frequency counts or relative frequencies. In this case, the whole sample size (19) is represented by the circle and the frequencies or relative frequencies are depicted as segments of the whole. For example:

[Image of a pie chart showing gender distribution]

Again, the titles and labels are just as important as the ‘pie’ itself.

Histograms

To create a histogram:

- divide the quantitative scale up into say 5 to 12 separate intervals
- count the frequency in each of these intervals
- draw a ‘bar chart’ of frequencies

There is an important difference between histograms and bar charts. Bar charts have gaps between the bars (representing the categorical scale) but histograms have no gaps between the bars.

I will summarise ‘Age in years’ using a histogram. As there is a relatively small sample size (n=19) five intervals, each 5 years wide will be sufficient in this case. I will use the intervals:

- 17.5 years to less than 22.5 years
- 22.5 years to less than 27.5 years
- 27.5 years to less than 32.5 years
- 32.5 years to less than 37.5 years
- 27.5 years to less than 42.5 years

These definitions look unusual but have a particular purpose – the centres of the five intervals are 20, 25, 30, 35 and 40 years respectively. It is important that each interval contains exactly the same number of years – each is 5 years wide in this case.

The frequency count for each interval is:

- interval centred on 20 years: 3 students
- interval centred on 25 years: 8 students
• interval centred on 30 years: 5 students
• interval centred on 35 years: 2 students
• interval centred on 40 years: 1 student

This information can be used to construct a histogram like the following.

![Histogram of age in years in the example data](image)

Notice that I have labelled the mid-point of each interval. Alternatively, you could label each interval with the range of values it contains (e.g., label the first interval 17.5 to 22.5 years). There are no gaps between the bars. The chart is fully labelled. Again, relative frequencies (percentages) could be used instead of actual counts.

**Mean and standard deviation**

Graphs contain more information, but there are times (particularly when it comes to statistical inference - below) when single number summaries of the data are more convenient. The mean and standard deviation are both single number summaries of the data. The mean is a measure of the centre of the data and the standard deviation is measure of the spread. (Looking at the histogram above, you can see that the centre of the data is probably somewhere between 25 and 30 years and students ages spread somewhere from say 17.5 years and 42.5 years).

The ‘mean’ is what people generally call the ‘average’. To work out the mean of age you add all the ages together and divide by the number of individuals you have data for. In this example the mean of age is:

\[
\bar{x} = \frac{30 + 26 + 21 + 19 + \ldots + 23 + 31}{19} = 27.26
\]

The standard deviation is, roughly, the average of how far each observation is from the mean. The higher the standard deviation, the more spread out the data.

Calculating the standard deviation is difficult. We recommend you use a scientific calculator or computer to help work it out. In this case the standard deviation of age is 5.516 years.

If you’re keen, here's how the standard deviation is done:

In this example, student 1 is 30 years of age. This is 30-27.26=2.74 years away from the mean of age. Student 2 is 26 years of age or 26-27.26=-1.26 years away from the mean. The fact that some distances from the mean are positive (above the mean) and some are negative (below the mean) causes problems. To make all distances positive we square them: so \((2.74)^2=7.5076\), \((-1.26)^2=1.5876\), etc. Once all of the distances have been squared and are all positive, they are added together to get a total squared distance score. To get an ‘average’ squared difference score we divide by one less than the sample size: i.e., divide by 18 in this case (it’s not easy to explain why we divided by 18 rather than 19 – we usually have to say ‘trust us on this’ to first year university students). Finally, to correct for the fact that we squared all the distances, the last step is to take the (positive) square root of the answer.
Median and quartiles

If you were to check back on the example data set you would find that 11 of the 19 students were younger than the mean age and only 8 were older. In this case the mean isn’t doing very well at describing the centre of the data. A different, and sometimes better, method for describing the centre of the data is to find the middle point or median.

To find the median, first sort the data from smallest to largest. For ‘Age in years’ this will give:

19, 21, 21, 23, 24, 24, 24, 25, 25, 25, 26, 29, 30, 31, 31, 31, 35, 41

The middle (median) is at 25 years, which is highlighted. There are nine data points below 25 and 9 data points above it. (If you have an even number of data points, the median is the point halfway between the middle two data points.)

The first quartile has one quarter of the data below it and three quarters above it. Perhaps the easiest way to find it is to find the median of the lower half of the data. That is:

19, 21, 21, 23, 24, 24, 24, 25, 25, 25, 26, 29, 30, 31, 31, 31, 35, 41

If we just look at the nine data points below the median, and the first quartile lies at the centre of those nine data points. In this case the highlighted 24 years has 4 data points below and 4 above.

The third quartile has three quarters of the data below and one quarter above. Using the equivalent method:

19, 21, 21, 23, 24, 24, 24, 25, 25, 25, 26, 29, 30, 31, 31, 33, 35, 41

The third quartile is 31 years of age.

The quartiles describe the spread of the middle 50% of the data.

Using the median and quartiles we can say that the centre of the age data is 25 years and the middle 50% of observations lie between 24 and 31 years.

Boxplot

Medians and quartiles are sometimes depicted graphically using a boxplot. Here is a boxplot of the age in years for the example data.

The dark line within the box represents the median (25 years). The bottom and top of the box show the first and third quartiles (24 and 31 years) and the whiskers extend up to the maximum age (41 years) and down to the minimum age (17 years).

Frequency tables

Suppose we wished to summarise the relationship between two categorical variables – let’s look at gender and faculty for example. To summarise this relationship we construct a table with one variable across the top and the second variable down the side as follows:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>8</td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
</tr>
</tbody>
</table>

It makes no difference which variable you put across the top and which one you put down the side. Using the table above we can easily see that there is a total of 19 and that 8 are male business students, 5 are female business students, etc. (Actually it appears there could be a relationship between faculty and gender in that Engineering has no females.)

Column (or row) percentages will often assist in highlighting relationships. For example:
Frequency table of gender by faculty

<table>
<thead>
<tr>
<th>Faculty</th>
<th>Business</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>62%</td>
<td>38%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Engineering</td>
<td>100%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Sciences</td>
<td>50%</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>68%</td>
<td>32%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Percentages are calculated for each row separately (such that each row adds to 100%). Now we can see that around one third (32%) of all students are female, but 0% of Engineering students are female. Females appear under-represented in Engineering.

Clipped or stacked bar charts

This is a clustered bar chart of two categorical variables:

The counts are taken from the frequency table. They may be converted to percentages and presented as relative frequencies.

Side-by-side boxplots and multiple histograms

When we wish to look at the relationship between a categorical variable and a quantitative variable we first separate the data into the categories. That is, if looking at the relationship between gender and marks, we would first separate marks into two groups – marks for males and marks for females. We then do the normal boxplot or histogram on marks within each group:

Notice that the boxplots or histograms have to be plotted on exactly the same scales (on both the vertical and horizontal axes) to allow direct comparison of results in the different categories.
Scatterplots

When looking at the relationship between two quantitative variables, put one variable on the horizontal axis and one on the vertical axis (for our purposes here, it doesn't really matter which variable is put where). Both the horizontal and vertical axes are numerical scales and each individual has their result plotted at the appropriate height on both axes. Here is a scatterplot of age against mark:

Consider the point at the top right of the graph. Looking straight down below this point on the horizontal axis, this student was 41 years of age. Looking straight across to the vertical axis, the student obtained a mark of 100%.

Looking at this scatterplot, there may be a general trend in the points from the bottom left to the top right (there are very few points at bottom right or top left). Such a trend would imply that older students tend to get better marks than younger students.