Statistics in Engineering and Computing

Why do statistics?

- To demonstrate that *differences* are *reliable*
- So if we re-ran an experiment, survey etc. that we would get the *same* or a *very similar* result
- Statistics is a way of predicting, in quantitative terms, how likely the experiment would be to give that same or very similar result
- …in terms of, e.g. which algorithm is faster, or which structure better resists earthquakes

Why do statistics?

There are other ways of ensuring results are reliable – some are in fact better, e.g.

- Mathematical proofs – common in theoretical computer science, this provides absolutely reproducable conclusions

Others are just different

- Re-testing – running a study multiple times and seeing if you get the same result, which is labour-intensive

Agenda for today

- Emphasis on *experimental* and other *empirical* testing, e.g. surveys
- Little discussion of maths *per se* – doing maths by hand is prone to error and would take too long
- There are many good sites and tools online and on your computer already – use those (or e.g. SPSS)
- More discussion of *what tests to do when* and what the basic concepts are
- …given with a few classic mistakes that I see

Choosing a Test

Choosing a test

- There are *many* statistical tests
- In various areas of computer science and engineering there are biases to particular tests
- …because they fit 'normal' problems in the area
- Beware of following these out of habit as people can (and do) regularly choose the wrong test

Choosing a Test

- What do other researchers do? *With similar data*
- What are we testing – a *measure* of a characteristic, or a *count* of a population
- What type of data is it? e.g. *continuous* or *discrete?*
- What is the *distribution* of the data?
- *How many* values are being tested?
- *How much* data is there?
- …and more

Step 1: Distribution Or Population?

- Distributions are when we repeatedly measure a characteristic of a thing, e.g.
  - Weight
  - Force
  - Speed
- Populations are when we measure what proportions or how many of a thing there are
  - Errors or faults (versus correct)
  - Successes versus failures
  - Which web browser do people use?

Step 2: What Sort of Measure?

- Continuous variables
  - i.e. fractions, decimals, "floating point"
- Discrete variables
  - i.e. ranks, ratings, integers
- There is some overlap – for larger numbers, we can teat discrete values as if they are continuous

Step 2: What sort of measure?

- But in general, ranks and rating numbers should be treated as discrete…
- We don't know the relative importance between being ranked 1st and 2nd – were they close or far apart?
- Are different people's ratings reliable and consistent?
- Also if all the range of integer numbers is small we seldom get a good fit to any *distribution*

Step 2: What sort of measure?

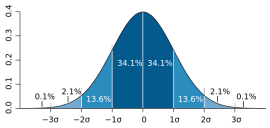
- There are two terms that help us chose the right sort of test if we are using a measure
- Ranks, ratings and small ranges of values are normally treated as *non-parametric*;
- Continuous values and large ranges of values are *parametric*

- Historically, many of our measures in engineering and science are *parametric*

Distributions

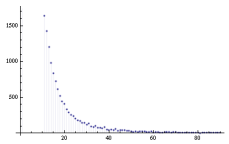
Step 3: Normal Distribution

- Also known as Gaussian distribution
- You may well have done basic statistics on this at school:



Step 3: Power-law Distribution

- Also call Zipfian distribution
  (and indeed a number of other names)



Step 3: Other distributions

- Uniform distribution – in theory, all values are equally likely; e.g. the numbers on a single die
- Bernoulli distribution – each outcome is a yes or a no (but the outcomes may not be equally likely)
- And many more – we will stick to a few

First choices

Choosing a Test

- We should now know if we're going to test:
  a) A population, to see if there are differences
  b) A measure, whether it is *parametric* or *non-parametric* and what *distribution* it is
- This should give us strong guidance for a test
- Let's take a couple of examples

Example: Cyber security

- We have two anti-virus products (AV-1, AV-2)
- Both were used to protect 100 computers each
- These were subjected to cyber attacks
- We know how many computers of AV-1 and AV-2 were 'hacked' and now have a virus
- Is this a population or a measurement problem?

Example: Racing Robots

- We have two robots that will be time tested several times:
  - One with wheels
  - One with legs
- Different race tracks and courses (but the same race tracks for each – each robot does every track once)
- Do we test a population or a measurement?
- Is the measurement parametric or non-parametric?
- What is the likely distribution?

A Common Test

A Common Parametric Test

Let's take a very common situation:

- Two competing methods to test
- Continuous, parametric measurements
- Normal distribution

- This can be tested by a very common method – Student's t-test

Student's t-test

Requires

- *Normal distribution*
- *Equal variance (i.e. standard deviation is the same)*
- *Only two sets of values*

This is a very common test, and you may well have seen it…

Student's t-test

- The test requires:
  1. The mean for the first method's data
  2. The standard deviation for the first method
  3. The mean for the second method
  4. The standard deviation for the second method
- This will produce a *t-value* and a *p-value*
- The t-value is specific to the t-test (remember this for a moment)
- The p-value is the probability, and is a general value

The p-value

- The p-value is a critical outcome for any test
- It is a number between 0 (good) and 1 (awful)
- It tells you how likely the result is to be just the outcome of random chance (rather than a real effect)
- If it is 1, it's *definitely* just luck
- If it is 0, there's *no chance* it's just luck

The p-value

- Typically, we accept values such as:
  - $p < 0.05$: one chance in 20 it is a "false positive"
    This is a typical value in general engineering
  - $p < 0.01$: one chance in 100 it is a "false positive"
    This is common in more sensitive areas of engineering and psychology
  - $p < 0.001$: one chance in 1000 it is a "false positive"
    More often used in medicine

Getting it Wrong

- Not checking the data is normally distributed
- Not checking for standard deviation being the same

Rachel is doing a study of how much people like the results of two different search engines. Each person gives search engine A and search engine B a score between 1 and 5. Rachel uses the t-test to check if A scores better than B

Getting it Wrong

- *Paul* is doing a study with 20 participants on a new technology to guide people who are navigating their way around a town using their mobile phone.
- Paul uses a t-test to check how many times people got lost using the new technology versus the old
- What might go wrong?

Refined Versions: Paired Tests

- *Paired tests* can be used when we use the same method on the same data, e.g.:
  - *A user tries to hit a target with a mouse or a touchpad*
  - *We strength test different structures with the same load*
  - *We measure the time taken with two algorithms searching the same dataset*
- This allows us to use more precise and sensitive tests in the statistics, e.g. a *paired t-test*
- Paired tests are more likely to give a positive result

Refined Versions: One-sided

- If we're pretty sure which way test is going to go, we can apply a one-sided test
  - e.g. "is an elephant heavier than an ant?"
- This (pretty much) increases your chance of getting a desirable p-value
- But should only be used when there is a strong argument to predict a better outcome from previous evidence

ANOVA

A t-test for more values

ANOVA

- ANOVA (ANalysis Of VAriance) is a derivative version that can test more than two value sets at once
- Note that it still requires parametric data
- Also still requires equal variances
- It is very, very often used badly or incorrectly, but it is commonly used

Why use ANOVA?

- You can do multiple t-tests to compare, e.g. three sets of data:
  - A vs B; B vs C; A vs C
- This is laborious, and ANOVA does it in one test
- There is however an even more important reason

Overtesting

- If we use t-test, then :
  - 4 methods need 6 tests
  - 5 methods need 10 tests
- If there's a 1/20 chance of a false positive, what is the risk of 10 tests producing one false positive?
- Almost 50%!
- One can use *Bonferroni corrections,* but ANOVA is simpler and better for novices to use

Testing Populations

Chi-Squared Test

- This is a test that can check if two populations differ significantly
- For example, we have two methods for giving heat feedback to users, and we have the following outcomes:

Method	Accurate	Inaccurate
Reinforcement	15	5
Discrete	9	11

Chi-Squared Test

- The test produces two results:
  - Test-specific number called χ (yep, that's a Greek Chi letter)
  - The general p-value
- Very simple test, but is very general purpose – you will find a lot of uses for it with populations

Two-way tests

- All the tests we've seen so-far are one way tests
- We only vary one attribute (variable) at a time
- But this isn't always possible or desirable
- E.g. we may have different test data; different sexes of participants; different structure types; etc.

A Two-Way Chi-Square

- Say we ran the study with heat interfaces, but now with 20 men and 20 women
- …women are known to have different heat sensitivity to men, so it's worth checking

Gender	Method	Accurate	Inaccurate
Male	Reinforcement	16	4
	Discrete	9	11
Female	Reinforcement	17	3
	Discrete	12	8

Reporting Tests

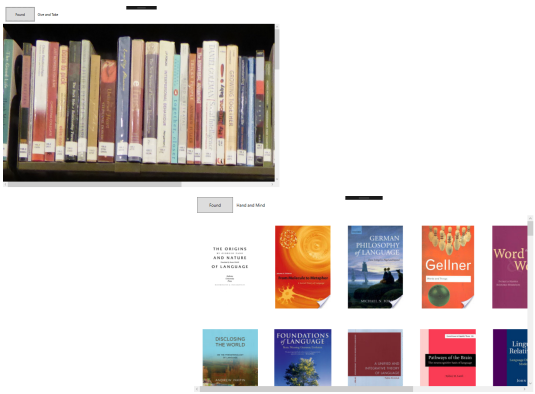
Reporting Tests

- You should always report both the p-value *and* the specific test-value for your test
- Also be clear what you did to check the distribution for any measurements
- And what you did to check any other requirements of the test, e.g. were the standard deviations similar for a t-test?

Reporting Tests

- Bad:
  - "We performed a t-test and produced $p<0.01$, so the test was successful"
- Good:
  - "We first checked if the standard distributions were similar; as they were 4.2 and 3.9, this verified that the requirement for similar variances was met. We then applied Student's t-test, and this produced $t=4.32$ ($n=20$); $p=0.0032$
- The 'n' here, by the way, is the size of the dataset, this is often reported as well, though it also varies a bit between tests

Example Study



Designing a Study

- How many people to use?
- …how much data to collect

- Always a problem!
- What have previous researchers done?
- …previous ones had used 12-24 participants

- And that was in my experience a viable number *if the differences are substantial*

Basic Tests

- If time to complete the task is normally distributed, then we can use t-test or ANOVA…
- Likewise any movement data
- …success or error rates, chi-squared test (proportion of success, for example)

- The previous researchers had used these tests in their work (c. 10-12 papers)

Problem: Oh dear, the stats

- If you're going to use parametric tests, you need to run a test that the data *is normally distributed*
- ….and our data wasn't!
- Was it our data? Did we do something wrong?
- Went back to the previous researchers – and their data had similar characteristics
- Low average (mean) values, and standard deviation greater than the mean… that's not normal!

What would you do?

- Re-run the experiment?
- Ask your supervisor?
- Phone a friend?
- Try a different test?

Find an expert

- Talked to a colleague who is a known statistics guru
- Asked the previous researchers what they had done (and for their reasoning)
- Ran different suggested tests on the data to see if we got different significance results (all $p<0.01$)
- Thankfully they all agreed
- Used the simplest standard test for reporting in the paper…but with a footnote about complications

Statistical Options

- Wilcoxon (non-parametric)
- Trimmed means
  (throw away the highest and lowest 10%)
- Log-linear (good fit test, but rare)
- Just use ANOVA anyway (allegedly "robust")

- Previous work had used ANOVA but not tested for normality…

The consequences

- The study was put in abeyance for a year…
- Then re-ran the test with other improvements and refinements, focusing on re-finding a target
- …not just finding it the first time

- The second test flowed from knowing how people behaved when searching for an idea
- ….rather than a specific book

Key Lesson

- Statistics isn't something you ever absolutely master
- You can't just add in at the end
  (or 24 hours before submitting!)
- Often you will need to learn a new method that matches your data
- Today is only an introduction

Summary

- This is only a brief introduction
- Choosing tests needs to be done with care
- Any maths is best done by computer
- But choosing the right test is something you need to do
- Sometimes previous researchers' methods *will not work* reliably on your data