Intelligent Control and Cognitive Systems brings you...

Just Enough About Statistics

Joanna Bryson and Will Lowe

Department of Computer Science
University of Bath
• “I’ve written an algorithm that does X”

• “I ran it on some data and it does better than the standard method”

• “So it’s better, right?”

• “Can I have a first?”
• “Kim and I designed a new user interface and asked Sandy to try it”

• “Sandy found it easier than the old one”

• “So it’s better, right?”

• “Can we share the best thesis prize?”
Actually, no.
Why not?

• Did Sandy drink a lot of coffee that morning? try harder for friends? work with similar interfaces before?

• Would your algorithm still be better on different data? in different network conditions? with different parameters?
• Observations are noisy and uncertain.

• They might not turn out the same way twice for all kinds of reasons.

• Statistics is about making inferences when there is noise and uncertainty.

• Where is uncertainty? Everywhere, except logic and pure mathematics.
• We use a *probability model* to express uncertainty about observations

• Divide what we observe into a *systematic part*, and a *random part*:

  • \( Y = f + \varepsilon \)

• Possible examples

  • \( f = 9.47 \)
  
  • \( f(X) = 3.81 + 2.82X \)
  
  • \( \varepsilon \sim \text{Normal}(0, 2) \)
Example

- \( Y = 11.34 \)

- **Systematic part (true value) is 9.46**

- **Random part (‘error’) adds 1.88**

- But we don’t know that yet.

- We want to know about the systematic part

- So we run experiments...
But if it’s just random...

• Worry 1: If it’s all just random, why take more observations?

• Worry 2: How can we know anything about \( \varepsilon \) when we don’t measure it?

• Answer to Worry 1: The Law of Large Numbers

• Answer to Worry 2: The Central Limit Theorem
Why more is better

• The Law of Large Numbers: “As the number of observations increase, the chance of being very wrong about the systematic part gets very small”

• Example: Suppose in reality:
  • $p(\text{success}) = p(Y=1) = 0.25$
  • $p(\text{failure}) = p(Y=0) = 0.75$

• Let’s graph the (frequency of successes) / (number of observations)
Law of Large Numbers

$p(Y=1)$

Trial
Noise and Normality

- The **Central Limit Theorem** (informal): “If $\varepsilon$ is the result of many smaller individual ‘errors’ then the more observations you have, the closer your observed averages are to being Normally distributed”
Noise and Normality

\[ p(\text{av}(Y)) \]

\[ \text{av}(Y) \]
Noise and Normality

- Remarkably, it doesn’t matter how the *actual* ‘errors’ are distributed (effects of coffee, network traffic etc.)

- **CLT** explains why we often assume normal distributions.
Noise is just signal you haven’t met yet.
Some applications

- What is the true accuracy of this classifier?
  - Point estimates and confidence intervals
- Is this interface easier to use than the others?
  - Experiments, Hypothesis Tests, and the Analysis of Variance
- What factors affect the performance of this application?
  - Regression
Confidence intervals

• How to estimate what you want to know?

• ‘Point estimates’ are usually sample averages (cf. law of large numbers).

• In papers you’ll hopefully see graphs with confidence intervals, sometimes called ‘error bars’.

• They express how confident we can be about the location of the true value.

• Conventionally you’ll see 95% intervals.
Number of Fights Involving Females

![Bar chart showing the number of female interactions across different conditions with confidence intervals.]

- a confidence interval
- a tighter confidence interval
Confidence intervals

- Confidence intervals come from the variance of the sample average in (theoretical) repeated trials.
- Typically the sample average variance is the sample variance divided by the number of observations.
- A 95% interval method comes with a guarantee: If we did this experiment again and again, and computed intervals, then only 5% of them would not contain the true value.
- 99% intervals are wider. (Why?)
Handy things to do with confidence intervals

• A rule of thumb:
  • If intervals overlap, then the difference in means is not statistically significant
  • If they don’t overlap, the difference is statistically significant

• Note: *only* a rule of thumb: check it!
Number of Fights Involving Females

← vs ↓ – a significant difference
Average Energy

Number of Extra Food Types Known

not significant↓

1  2  3  4  5  6  7
71 313 25 365 13 341 5 269 3 166 1 68 1 13
Confidence & Power

• An unintuitive consequence:
  • How certain you are (how small the standard error || narrow the confidence interval is) does not depend on the size of the population
  • Only the sample size and the variation within the population.

• It’s possible to survey 60,000,00 people with a sample of 1000, e.g. at election time...

• Determining the right sample size requires a power calculation based on effect size & variation.
Some applications

- What is the true accuracy of this classifier?
  - Point estimates and confidence intervals
- Is this interface easier to use than the others?
  - Experiments, Hypothesis Tests, and the Analysis of Variance
- What factors affect the performance of this application?
  - Regression
Experiments

• **Experiments** are designed to study causation – what *makes* your program run faster?

1. Pick subjects, factors, and design.

2. Randomize and control.

3. Analyze experimental data in an **ANOVA**.

   • $Y = \text{mean} + \psi + \varepsilon$

   • $\psi$ is how much $Y$ varies by *condition* (more on conditions later).
Hypothesis Testing

- **ANOVA** allows us to test for differences by seeing how well our data support:
  - **null hypothesis**: there is no difference \((a=0)\)
  - **alternative**: there are differences \((a \neq 0)\)

- These are only statistical hypotheses, so
  - **Cannot** say: “these are *definitely* different”.
  - **Can** say: “these appear significantly different \((p<.05)\)” or “we reject the null hypothesis at the .05 level”
When trying to test the hypothesis that a factor makes a difference we can make 2 kinds of mistake:

- Over-optimism (Type I error),
- Missed opportunity (Type II error).

‘p’ is the probability of thinking you’ve seen a difference when it’s really due to chance (Type I).

When the p value is small, either there’s a real difference, or something unlikely happened.

p values tell you nothing about Type II errors.
p and statistical significance

- When testing, we pick a measure of statistical significance level or p-value, typically .05
- This is the probability we are willing to risk of making an over-optimistic mistake
- Confidence intervals have the same interpretation: “the true value is within this interval” p<.05
- Either the true value is within this interval or something rather unlikely happened...
Statistics means never having to say you’re certain...
Experimental Design

- How many subjects do you need?
- Crossed designs and interactions
- Randomization and control
How many subjects?

- **Statistical power** is $1 - \beta$ (Type II error)

- $\beta$ = The probability of spotting an effect if it’s there

- Higher power experiments are better.

- Normally we fix a Type 1 error cutoff $\alpha$ (e.g. 0.05), then power depends on:
  
  - number of subjects
  
  - real size of effect (wait, we don’t know this!...)
    
    - importance of pilot studies...
Crossing and interactions

• Good designs get as much information out of as few subjects as possible.

• e.g. in crossed designs every subject gets every treatment

• have to randomise over ordering of conditions per subject to check for carry-over effects.

• Interactions: when the effect of one stimulus is different according to condition (e.g. sex)
Randomisation

- Control: divide subjects into conditions, e.g. sex, year of study; in which you expect effects to differ, (should we average effects in men and women?)

- Within a condition, these factors are fixed.

- Effects can also vary according to things we don’t know about. Can’t control these!

- Randomisation within condition approximately balances subjects with respect to these unobserved factors.
Control what you can, randomise the rest.
Some applications

• What is the true accuracy of this classifier?
  • Point estimates and confidence intervals

• Is this interface easier to use than the others?
  • Experiments, Hypothesis Tests, and the Analysis of Variance

• What factors affect the performance of this application?
  • Regression
Regression

- **Relationship between** performance \((Y)\), network load, time of day and cpu speed.

- \(Y = b_0 + b_1 \times \text{load} + b_2 \times \text{speed} + b_3 \times \text{time} + \varepsilon\)

- Assume we only care about network load. The rest are controls.
  - Get estimates and intervals for \(b_0, b_1, b_2, b_3\). Does the interval for \(b_1\) overlap 0?

- \(R^2\) measures how much of \(Y\)'s variance these factors (the model) explain.
Further reading...

• The web is full of free class notes and statistics tutorials e.g.
  • Ian Walker’s notes at Bath: http://staff.bath.ac.uk/pssiw/
  • http://davidmlane.com/hyperstat/
  • Statistics software packages from BUCS:
    • SPSS, Genstat, Minitab
  • But R is best, e.g. https://personality-project.org/r/