Just Enough About Statistics

Will Lowe
wlowe@latte.harvard.edu

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 MSc now?”

- “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?”

No you can’t

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?

Statistics

- 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 the uncertainty? *Everywhere*, except logic and pure mathematics

Statistical worldview

- We use *probability* to express uncertainty about observations
- Divide what we observe into a systematic part, and a random part:
- $Y = f + \epsilon$
  - $f = 9.47$
  - $f(X) = 2.82X + 3.81$
  - $\epsilon \sim \text{Normal}(0, 2)$

- Example: Y = 11.34
  - Systematic part (population value) is 9.46
  - Random part (personal 'error') adds 1.88
- We want to know about the systematic part

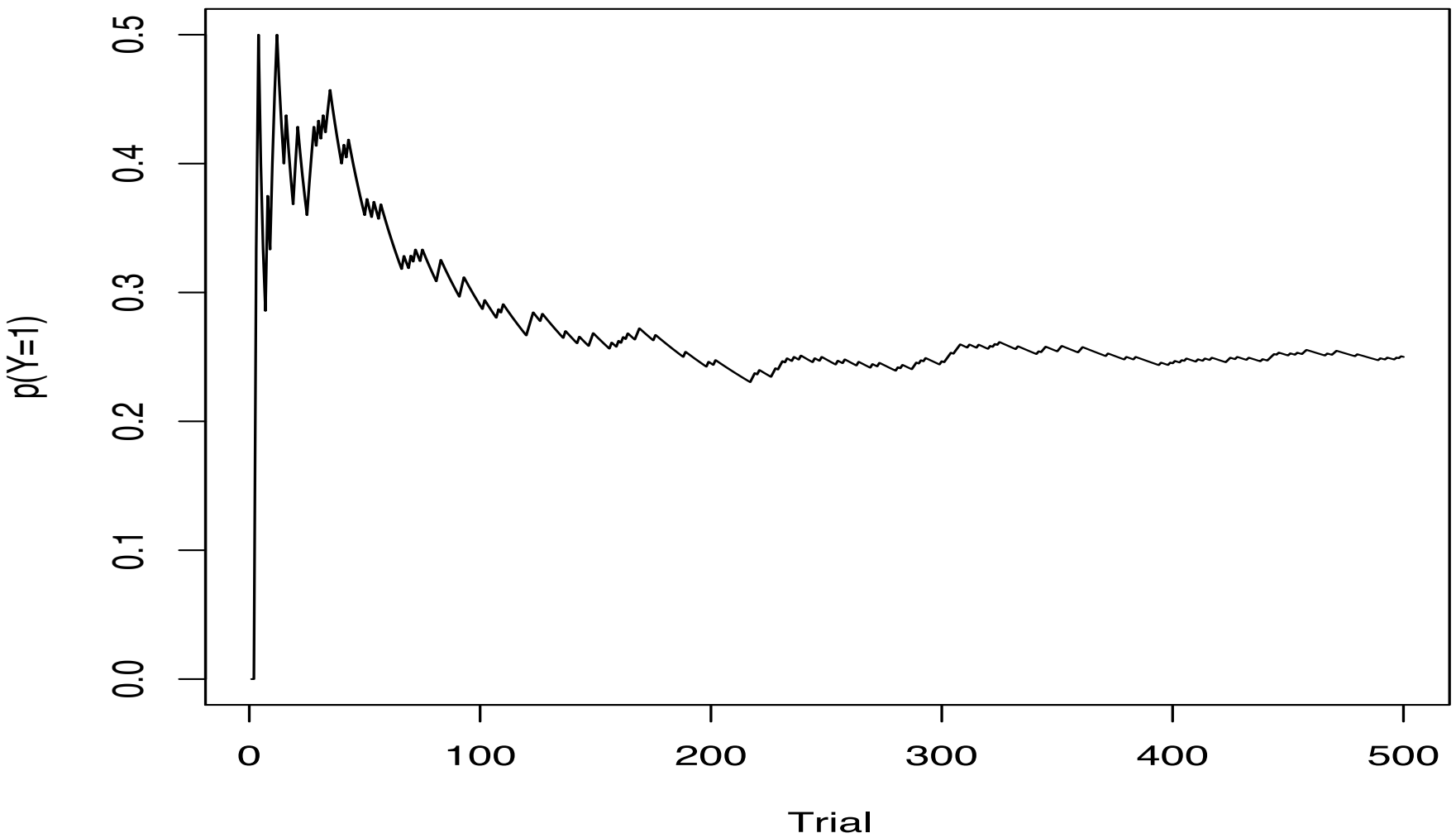
But if it's just random...

- Worry 1: If it's all just random, why take more observations?
- Worry 2: How can we know anything about ϵ 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 (informal): "As the number of observations increase, the chance of being *very* wrong gets *very* small"
- Example: p(Y=1)=0.25 and p(Y=0)=0.75
- Use frequency of Y=1 / number of observations

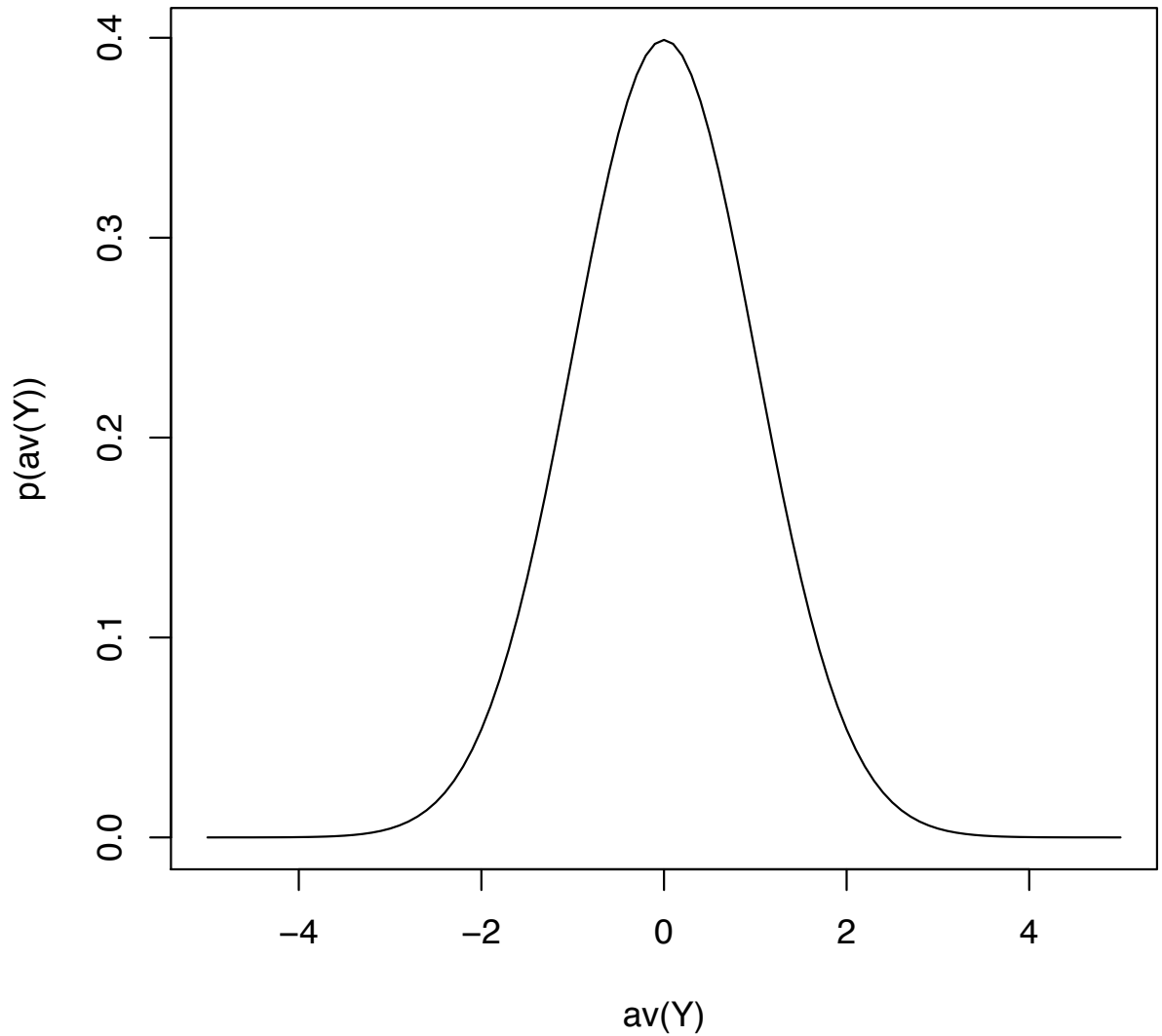
Law of Large Numbers



Normality

- The Central Limit Theorem (informal): "If ϵ is the result of many 'errors' then more observations you have, the closer your observation averages are to being Normally distributed"

Normality



Normality

- Remarkably, it doesn't matter how the actual 'errors' are distributed (coffee, network traffic etc.)

Replication and statistics

- Replication is essential for demonstrating things
- In statistics we always ask: "What would happen if we used this method again and again?"
- And we look for procedures that behave well in *repeated trials*.

Averages and intervals

- To see whether two conditions are different you:
  - Compute average performance in each condition
  - Compute confidence intervals

Confidence

- In papers you'll (hopefully) see graphs with error bars or *confidence intervals*
- They express how confident we can be about the location of the true value.
- Conventionally you'll see 95% intervals

Confidence

- The method we use to compute the interval has a *guarantee*: If we did this experiment again and again, and computed intervals, then 95% of them would contain the true value.
- Alternatively, the the proportion of intervals that *don't* contain the true value is .05

Interpreting intervals

- If intervals overlap, then the difference in means is not significant
- If they don't overlap, the difference is *statistically significant*
- 99% intervals are wider than 95% intervals
- Means can be significant at the .05 level, but not at the .01 level.
- How certain do you want to be?

In papers...

- In papers on experiments you'll also see:
  - "The new interface was superior, $F(1,19)=9.23$ $p<.05$"
  - This is conventionally how Analysis of Variance (ANOVA) tests are written
  - You only need to care about the 'p' part.
- What is 'p'?

p-values

- When trying to demonstrate things we can make 2 kinds of mistake:
  - *Over-optimism* (Type I error)
  - *Missed opportunity* (Type II error)
- 'p' is the probability of being over-optimistic (seeing a difference when there isn't really one)
- When p is small, either there's a real difference, or something *rather unlikely* happened.

Back to you and Kim...

- Get a several people to use the use the new interface, and several to use the old one (the more the better).
- Compute average performance for each group.
- Compute confidence intervals
- The new interface does better on average. Do the intervals overlap?
- No? Congratulations. You've demonstrated your interface is better!

Experimental Design

- How many subjects do you need?
- Crossed designs and interactions
- Randomization
- Control and repeated measures

Statistics means never having to say you're certain...