Just enough Statistics

Will Lowe

- I’ve written a program to categorize research articles based on my reading preferences
- My housemate used it a bit last night
- She said it was better than Winnow!
- So it’s better, right?
- Can I have an MSc. now?

- Sanjay and I designed a new search interface.
- We asked Jay to try it.
- He found it easier to use than than the library catalogue
- So it’s better, right?
- Can we *share* the best thesis prize?

- Sadly not.

Why not?

- Are you sure those were *representative* articles she tried?
- Is Winnow the right comparison?
- Does Jay usually prefer your style? He's a housemate, after all...
- Would your neighbours agree?
- Does the library catalogue interface suck?

Demonstration

- You need to show you've done a good job
  - Mathematicians *prove* it
  - The rest of us demonstrate it experimentally
- First consider inference problems in the abstract...

Inference

- When there is *no* uncertainty, use logic.
- When there *is* uncertainty, use probability.
- Statistics is about using probability to make rigorous and defensible inferences when there is noise and uncertainty.
- This lecture is about getting the *intuition* behind statistical, and experimental methods
- *Look up* the detail when you need it

Statistical view

- Observations are *noisy*, so our inferences from them are *uncertain*.
- *Lots* of different processes generate an observation. Divide them into
  - Systematic: what you are trying to measure (signal)
  - Random: everything else that gets in the way (noise)
- Task: Uncover the *systematic differences*

Statistical view

- Use a simple model to decompose search time *Y* into systematic and random parts, e.g.
  - $Y = s + e$
  - *e* is a noise distribution
  - *s* is the *true underlying difference* in search time between using your system and the library catalogue.
- We want to infer *s*.

Statistical view

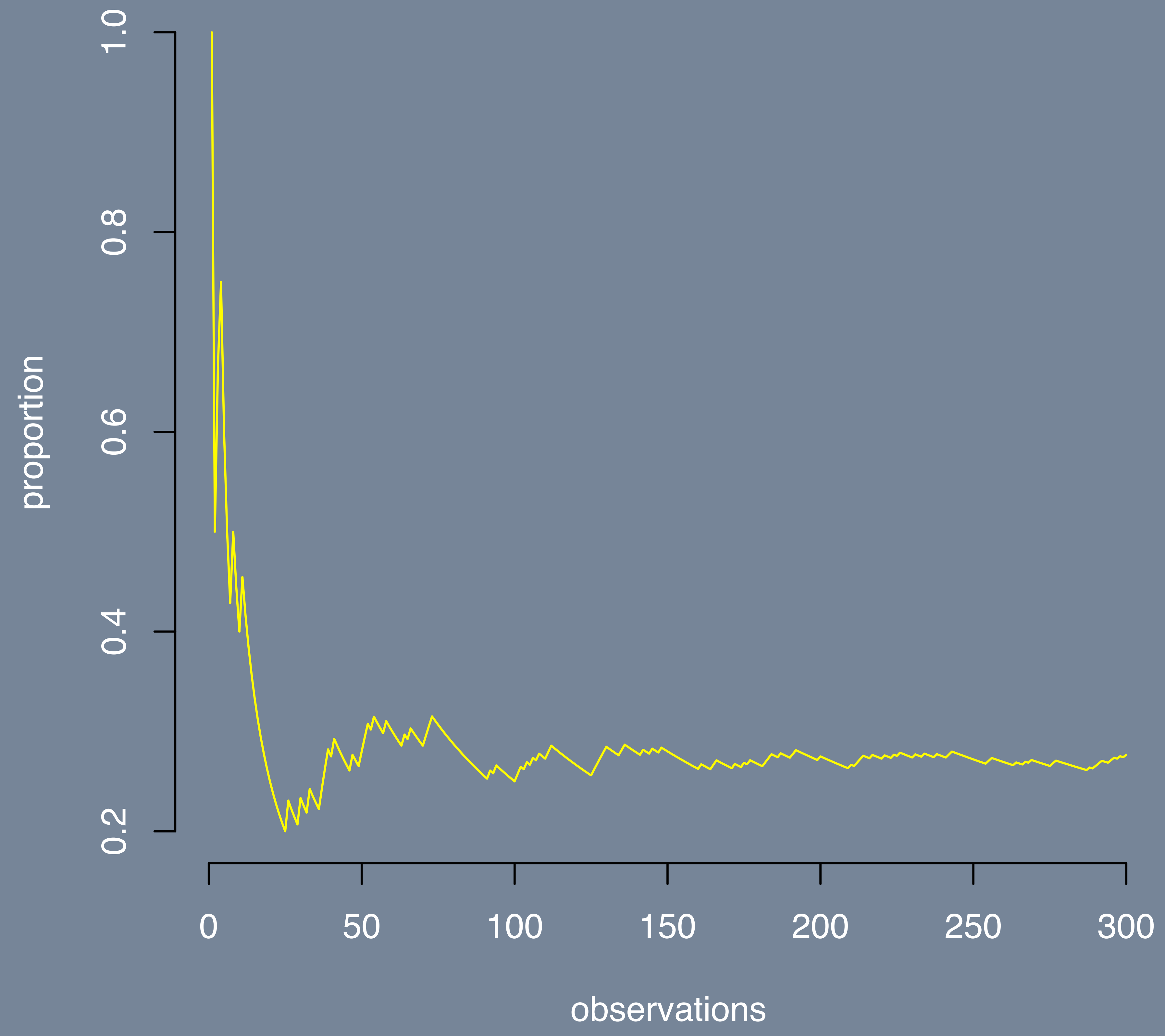
- Assume that this model describes observations on a population
  - university search users, general public, MSc candidates...
- Every time we make another observation
  - *e* is different (coffee, network traffic)
  - *s* is the same.
- *s* is the true or ‘population’ value

2 good questions

- If its all just random, why does taking *more* observations help?
- How can we know anything about *e* if we don't, or can't *measure* it?
- 2 good answers -
  - The Law of Large Numbers
  - The Central Limit Theorem

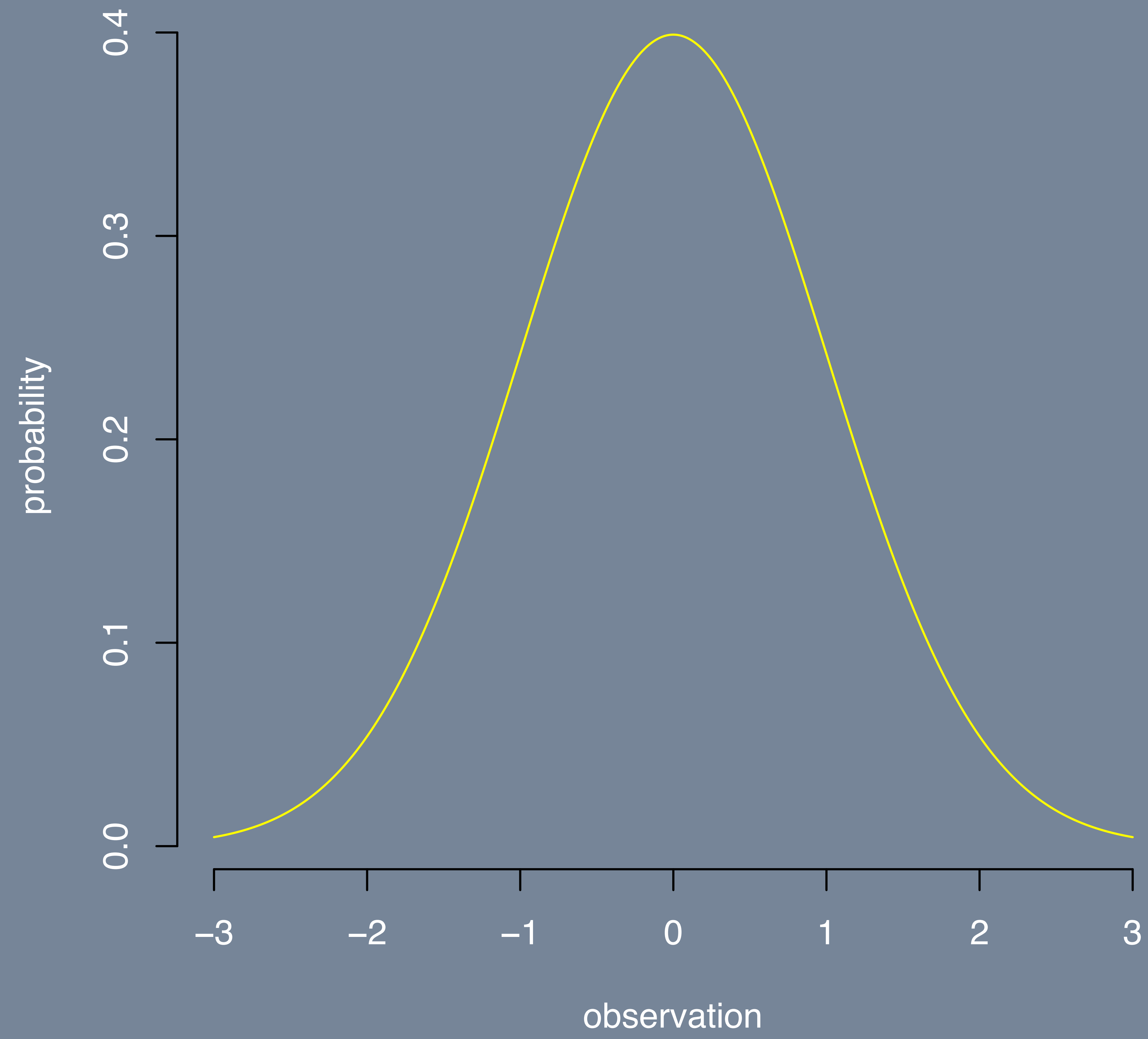
Large Numbers...

- As the number of observations increase, the chances of being *very wrong* (about the systematic part) get *very small*
- Simple example:
  - $p = Prob(Y{=}heads) = 0.25$
  - Estimate p using $h_{(i)}$ - the average number of heads seen after the *i*-th observation


1.0
0.8
0.6
0.4
0.2
proportion
0
50
100
150
200
250
300
observations

The Central Limit

- *If Y* is the result of many smaller individual noise sources, *then* the more observations you have, the closer the observations are to having a Normal Distribution.



0.0
0.1
0.2
0.3
0.4
probability
−3
−2
−1
0
1
2
3
observation

- Remarkably, *it does not matter* how the noise sources themselves are distributed
- This is fortunate: we usually have no idea how to mathematically characterize:
  - network lag
  - the effect of strong coffee
  - late nights reading about research methods…
- The CLT is why statistical models often assume Normally distributed noise.

Applications

- Statistical inferences divide into:
  - Estimation: what is the value of s ? What range of values would be plausible?
  - Testing: is $s > t$? How certain can we be of that?
- Estimation is usually used for *description*
- Testing is usually used for *demonstration*

- Estimation examples:
  - Point estimation
  - Confidence intervals (error bars)
- Testing examples:
  - Analysis of Variance (ANOVA)
  - Testing for Independence

Points…

- Point estimate for the true mean of N observations:
  - sample average: $\hat{Y} = \frac{1}{N}\sum^{N} Y_i$
- This estimate might be different next run
- This estimation *method* comes some mathematical guarantees. It is:
  - consistent
  - unbiased

…and Intervals

- The point estimate $\hat{Y}$ has a probability distribution of its own
- This distribution represents our uncertainty about the mean
  - wider distribution means less certainty
- The distribution width is called standard error

...and Intervals

- Choose an interval around $\hat{Y}$ that contains 95% of its (probable) values
  - e.g. +/- twice the standard error
- This interval construction method comes with mathematical guarantee:
  - *If* you construct the interval this way
  - *Then* it will contain the *true* mean 95% of the time (in repeated trials)
- Hence it is a 95% confidence interval

Points & Intervals

- 99% intervals are wider than 95% intervals
  - why?
- Conventionally, 95% intervals appear on graphs
- Rule of thumb for reading graphs:
  - *Overlapping intervals* mean that estimates are not reliably distinguishable, given your observations

Testing – ANOVA

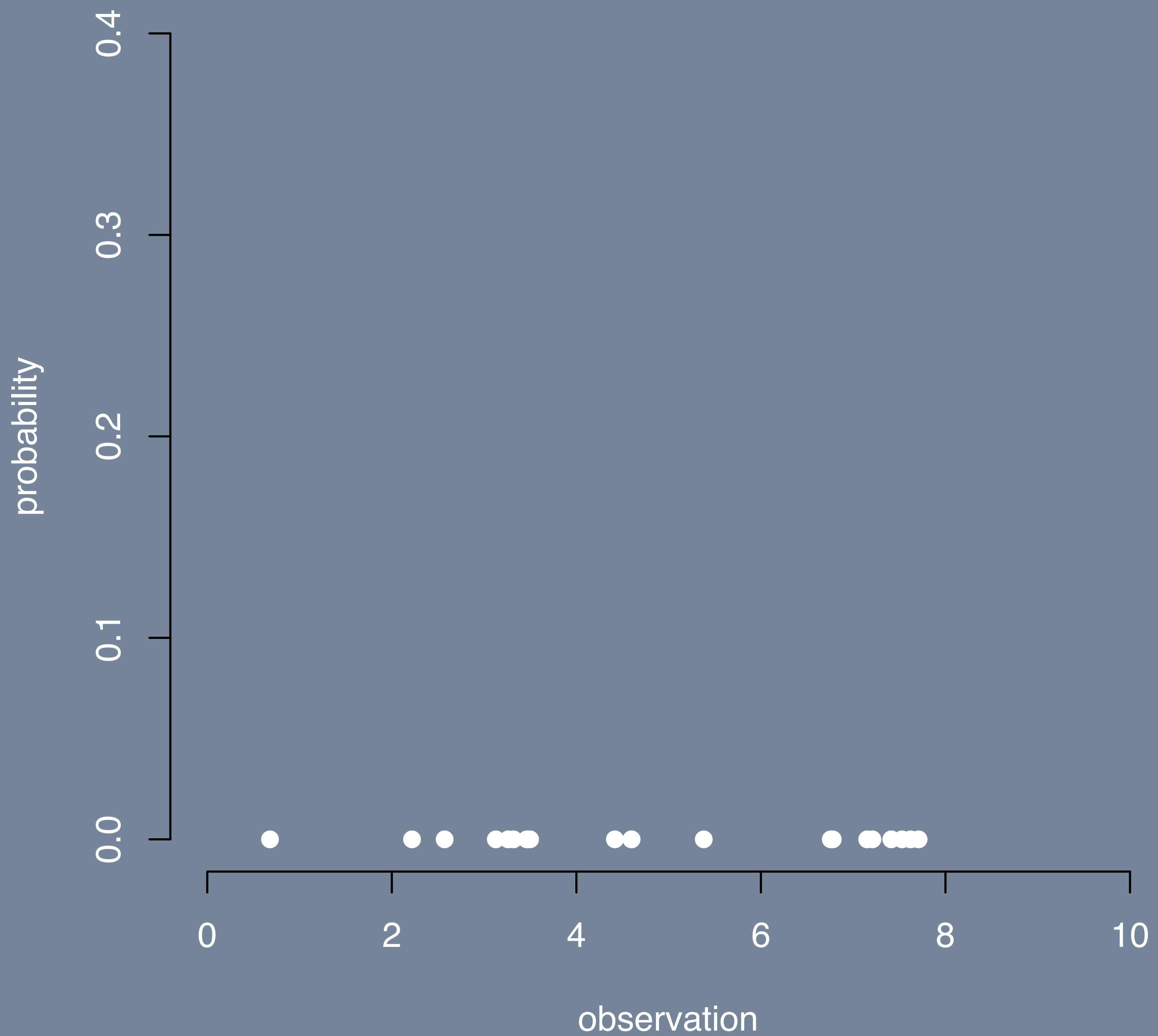
- Proper experimental demonstrations need a proper experimental *tests*
- e.g. Analysis of Variance (ANOVA)
- When you run an experiment there is observational variance
  - e.g. subjects search with two interfaces

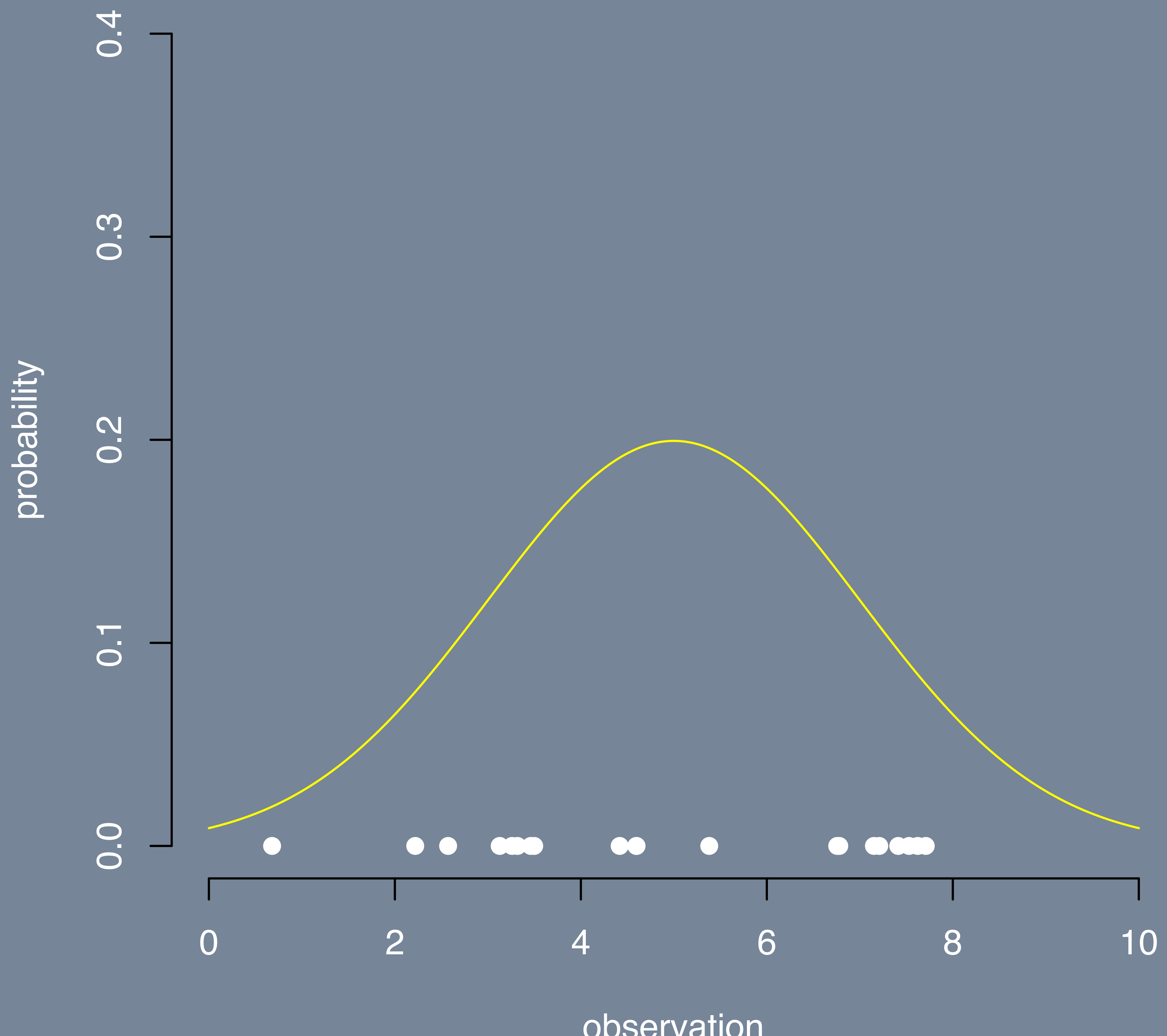
ANOVA

- Some of the variance is caused by *systematic* factors
  - e.g. one interface is just *better*
- Some of the variance is caused by *random* factors
  - e.g. Julie got bored in the middle

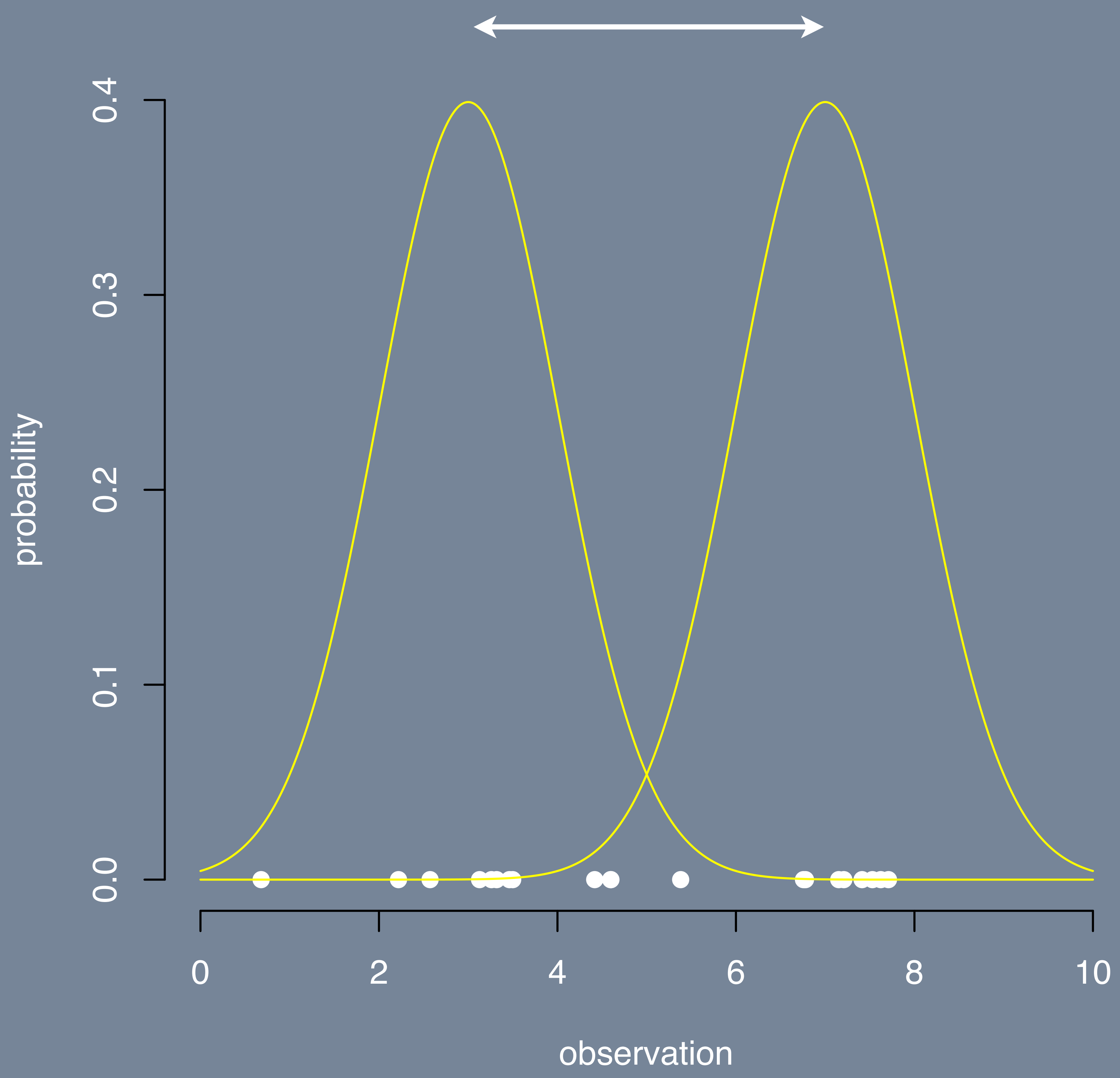
ANOVA

- ANOVA *analyses* the variance into two components by testing two competing hypotheses
  - H0: All variation is random
  - H1: Some variation is systematic
- Hypothesis 0 is sometimes called the *null hypothesis*


probability
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
observation



0.0
0.1
0.2
0.3
0.4
probability
0
2
4
6
8
10
observation



probability
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
observation

ANOVA

- ANOVA is a statistical test
  - Cannot say: “there is definitely a systematic cause for these observations”
  - Can say: “either there is a systematic cause for these observations, *or something unlikely happened*”
- Statistics means never having to say you’re certain…

ANOVA

- ANOVA is a *hypothesis test*
- There are two kinds of inference errors we can make:
  - Over-optimism (Type 1)
  - Missed opportunity (Type 2)
- Statisticians (and scientists, and engineers) are cautious: Type 1 errors are worse

Something unlikely

- You'll see ANOVA results in papers written like this:
  - "*s* and *t* are significantly different, $F(1,32)=13.01$ $p<.01$"
- p is the probability of inferring a systematic difference, when there isn't one
- We want p *small*, conventionally <.05

Experiments

- We can use experiments and ANOVA to test many hypotheses at once:
  - Test MSc students against the general public, *and*
  - Your interface against the library catalogue, …
- More efficient than separate experiments
- Reveals systematic interactions

Interfaces again

- How to demonstrate a superior interface:
  - Decide on your groups
    - MScs and the general public
  - Take a random sample from each
  - Decide on your comparison
    - New vs. library interface
  - Analyze the results...

Experiments

- Systematic differences in experiments are called ‘effects’
  - Main Effect: MScs are significantly faster than the general public
  - Main Effect: New interface is significantly faster than the library catalogue
  - Interaction Effect: MSc speed advantage *increases* with new interface

Resources

- Understanding what you’re doing:
  - Level 4 of the library
  - http://staff.bath.ac.uk/pssiw/
  - http://davidmlane.com/hyperstat/
- Doing it:
  - SPSS, from BUCS
  - R, from http://www.r-project.org/