ST221 Introduction to Statistics

Dr Rafael Moral

Dept of Mathematics and Statistics
Maynooth University

rafael.deandrademoral@mu.ie

Module objectives

- The objective of this module is to introduce students to descriptive and inferential Statistics.
- This course gives an introduction to the following topics:
  - Section 1: Probability theory – sample spaces, events, counting methods, conditional probability, independence, Bayes theorem
  - Section 2: Discrete random variables – the Bernoulli, binomial, and Poisson distributions
  - Section 3: Continuous random variables – the exponental and normal distributions
  - Section 4: Bivariate discrete random variables – covariance and correlation, independence
  - Section 5: Sampling distributions and Central Limit Theorem
  - Section 6: Inference – confidence intervals and hypothesis testing (one and two samples, paired samples)

Course organisation

Timetable

- Lectures will take place on campus and in person
- Lecture notes will be posted on Moodle prior to lectures
- Tutorials – on campus – weeks 4, 6, 8, 10
- **Please select a tutorial slot on Moodle**

Office hours

- Please message me on Teams or send an e-mail to schedule a call
- We can also meet in person in my office depending on the day of the week

Course organisation

Marks

- 10% – midterm (**6 November, in class**)
- 10% – four compulsory assignments (4×2.5%)
- 5% – two lab quizzes (2×2.5%)
- 10% – practice Moodle quizzes (2×5%, multiple attempts allowed)
- 65% – final exam (January; date TBC by Exams Office)

Please register for the Labs! Registration is currently open on Moodle.

Course organisation

- Notes will be a combination of handouts posted on Moodle and notes taken down during the lectures
- The notes on Moodle *will not be enough on their own* – **there are intentional blank spaces!**
- You must attend the lectures to complete the blank spaces in the notes
- At the end of each section we'll do a Summary Song, make sure to warm up your lovely singing voice before class

Recommended reading

- *Stats: Data & Models*, by De Veaux, Velleman and Bock

Statistics in your degree

- BSc Statistics is available as a Double Honours subject to students in 3rd and 4th year
  - BSc Statistics + Other
  - e.g. BSc Statistics + Applied Mathematics
  - e.g. BSc Statistics + Chemistry
  - e.g. BSc Statistics + Biology

Support

- Any problems please e-mail our departmental office `mathsstats@mu.ie`
- Maths Support Centre: `http://supportcentre.maths.nuim.ie/`

Why study Statistics?

- Statistics and Science go hand in hand...
- Scientists collect data, Statistics helps to get the best information from the data

Why study Statistics?

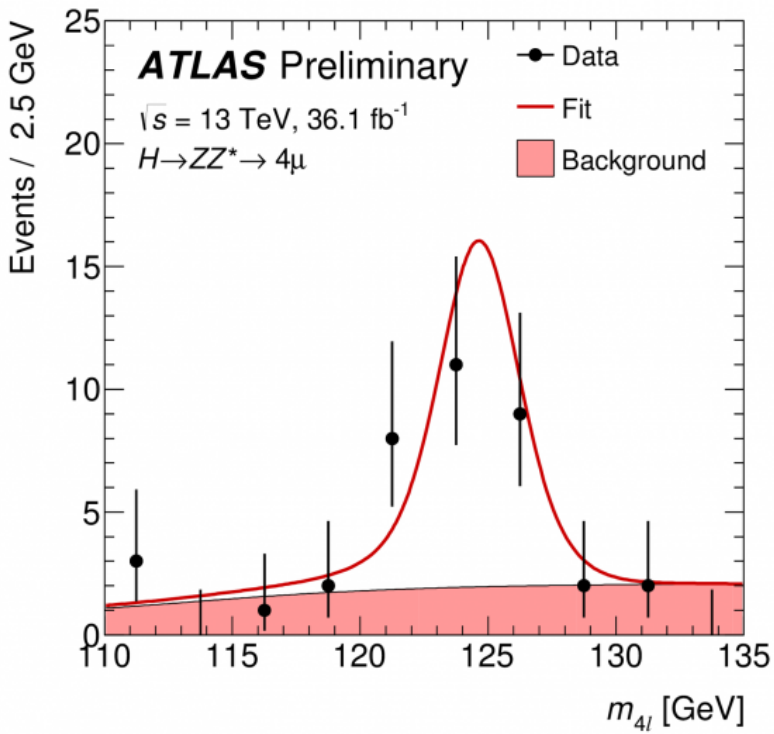
Why study Statistics?

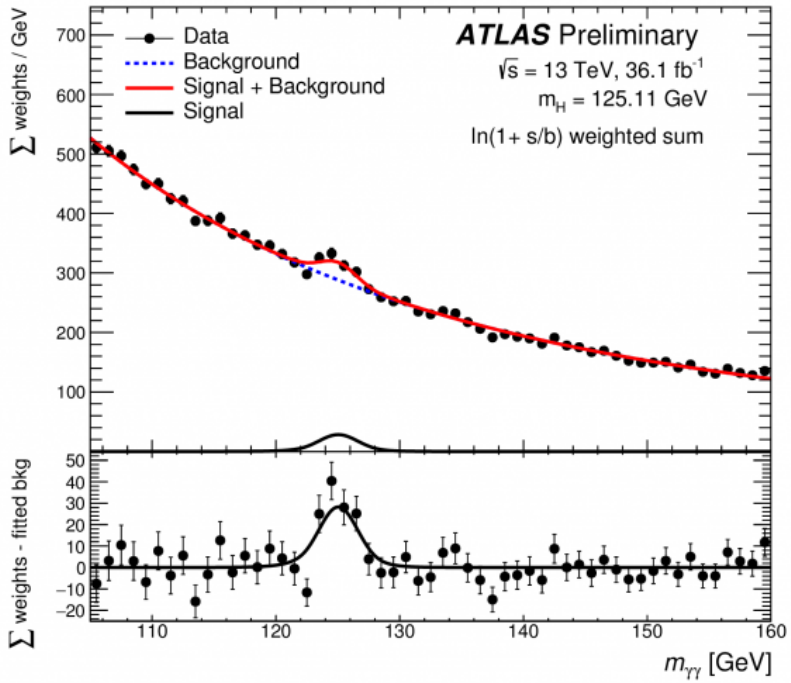
Why study Statistics?

Why study Statistics?

Why study Statistics?

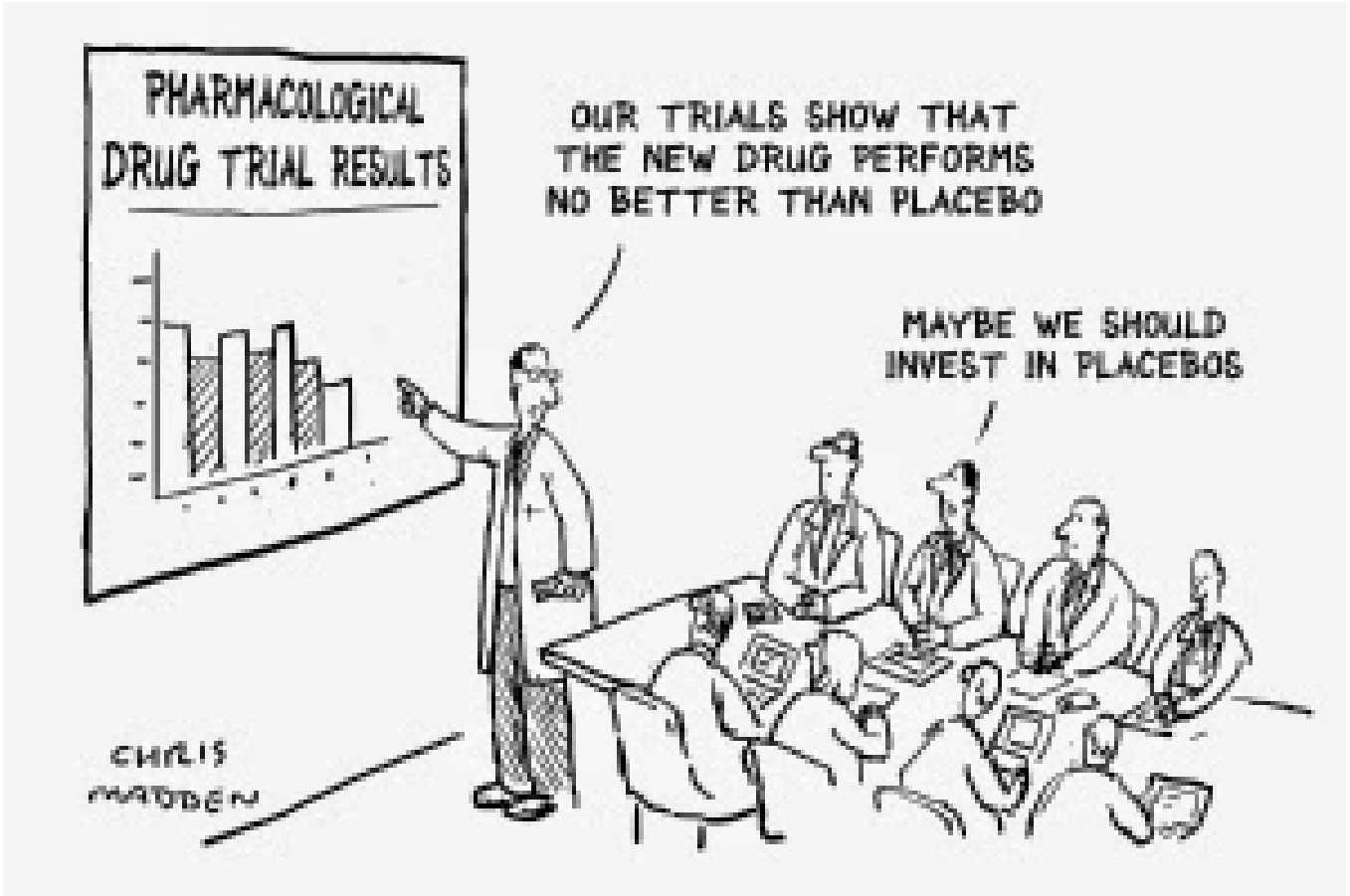
Scientifically interesting variability vs. background noise





Higgs Boson: `http://phdcomics.com/comics/archive.php?comicid=1489`

Scientifically interesting variability vs. background noise

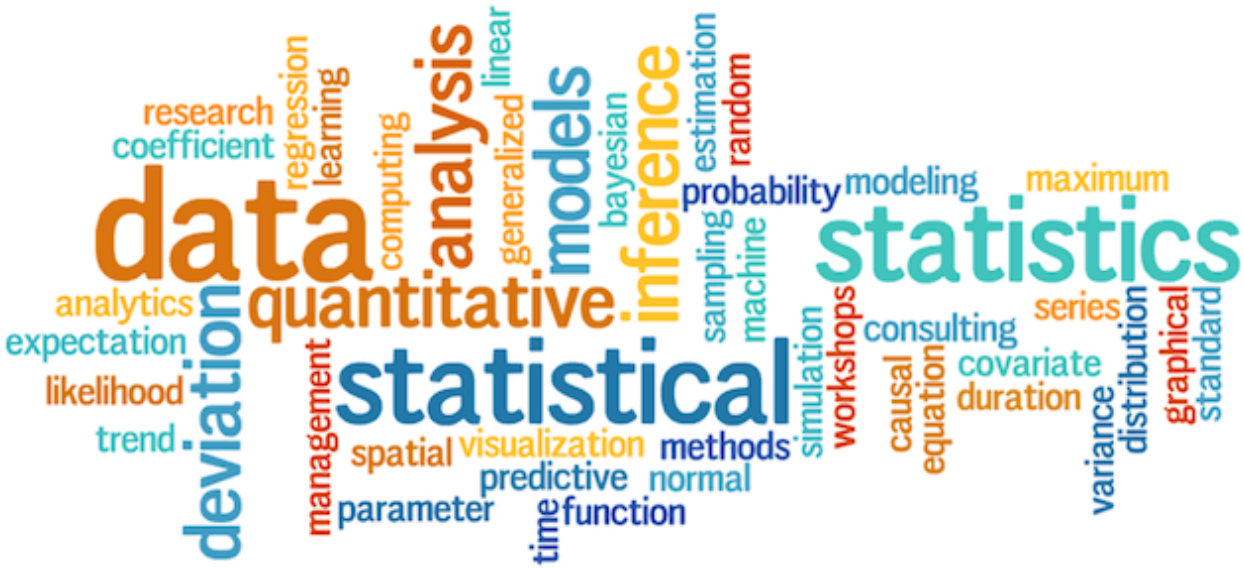


Statistics/Probability

- **Statistics** is about the collection, presentation, analysis and interpretation of data (information), usually in the presence of some inherent *variability*
- **Probability** is the mathematical framework for describing (modelling) *uncertainty*

Statistics

- The word **Statistics** comes from Latin – *statisticum* – which means "of the state", referring to analysis of data of a state/country
- With the development of Science, Probability Theory, and Informatics, Statistics acquired a scientific status, with applicability in virtually all areas of knowledge



A few definitions

- "Statistics is the science that teaches us what is the political context of every modern state in our world" (Hopper, 1770)
- "Statistics is the study of populations, variation and methods for data reduction" (Fisher, early 20th century)
- "Statistics are quantitative data affected, mostly, by a multiplicity of causes" (Yule and Kendall, 1958)
- "Statistics may be seen as a set of techniques that help decision making, when uncertainty conditions prevail" (Vieira, 1986)
- "Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data" (Merriam-Webster Dictionary)
- "Statistics is the science of learning from data" (David Cox, 2016)

A few definitions

- "Statistics is the science that teaches us what is the political context of every modern state in our world" (Hopper, 1770)
- "Statistics is the study of populations, variation and methods for data reduction" (Fisher, early 20th century)
- "Statistics are quantitative data affected, mostly, by a multiplicity of causes" (Yule and Kendall, 1958)
- "Statistics may be seen as a set of techniques that help decision making, when uncertainty conditions prevail" (Vieira, 1986)
- "Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data" (Merriam-Webster Dictionary)
- **"Statistics is the science of learning from data"** (David Cox, 2016)

The Birthday Problem

- Many people have shared birthdays within their families, or associates
- In his family (16 people), Rafael shares his birthday with his daughter and his auntie. Is this unusual?
- How large a group do we need for it to be more likely than not that there is a shared birthday?

The Birthday Problem – Analysis

- Let the birthdays of the n family members be $B_1, B_2, \ldots, B_n$
- Assuming that birthdays are uniformly spread, the probability that B_1 and B_2 are different is

$$\frac{364}{365}$$

The Birthday Problem – Analysis

- Let the birthdays of the n family members be $B_1, B_2, \ldots, B_n$
- Assuming that birthdays are uniformly spread, the probability that B_1 and B_2 are different is

$$\frac{364}{365}$$

- If we now consider B_3, the probability that this is different from the other two is

$$\frac{363}{365}$$

- Hence, the probability that all three are different is

$$\frac{364}{365} \times \frac{363}{365}$$

The Birthday Problem – Analysis

- Continuing in this way, the probability that all n are different is

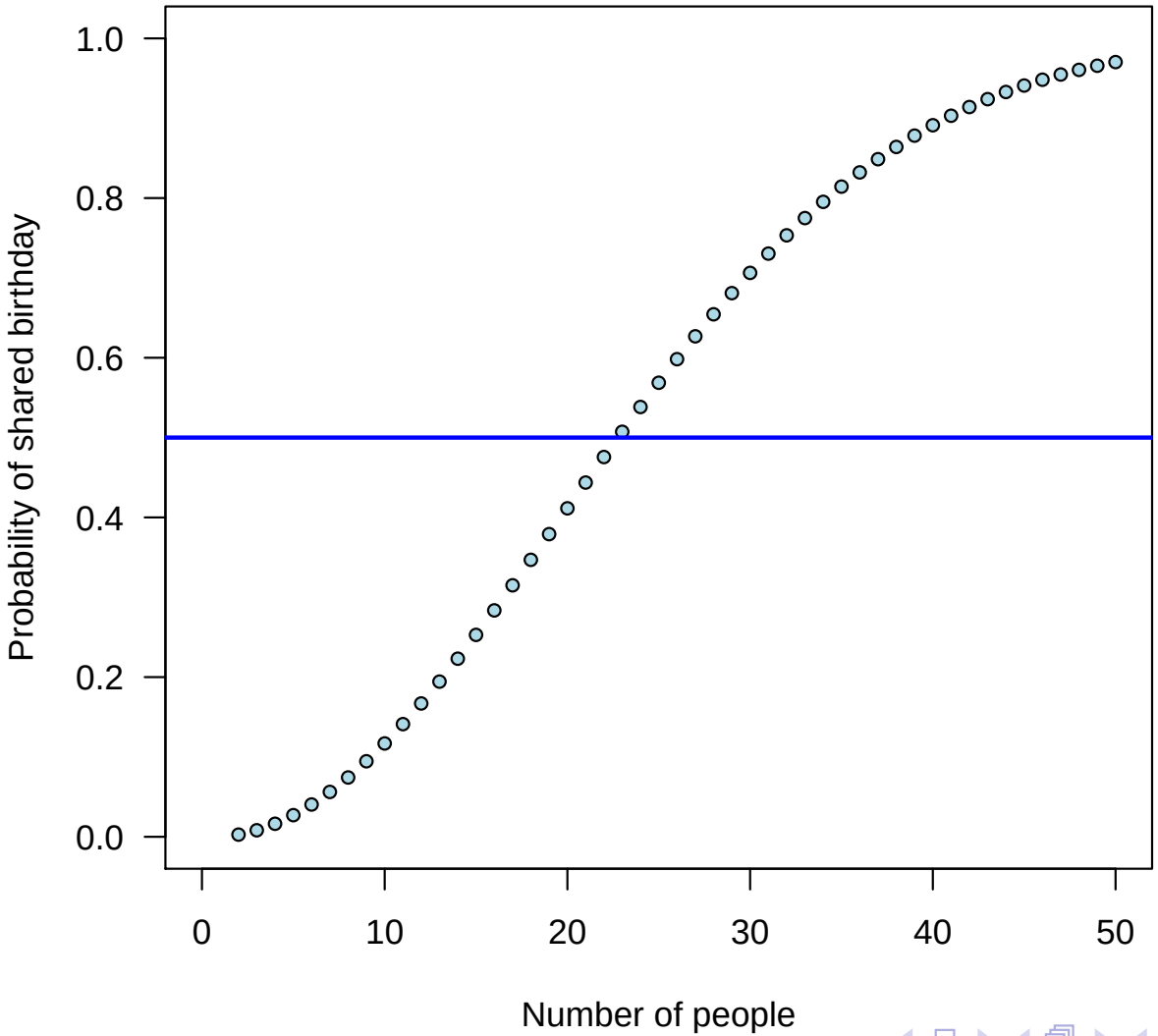
$$p_n = \frac{364}{365} \times \frac{363}{365} \times \ldots \times \frac{366 - n}{365}$$

- We have made some assumptions here about the ways in which probabilities multiply together to give the probabilities of combined events – we need to explore this further in the lectures
- The simplest way to proceed is to write a simple computer programme to tabulate p_n against n

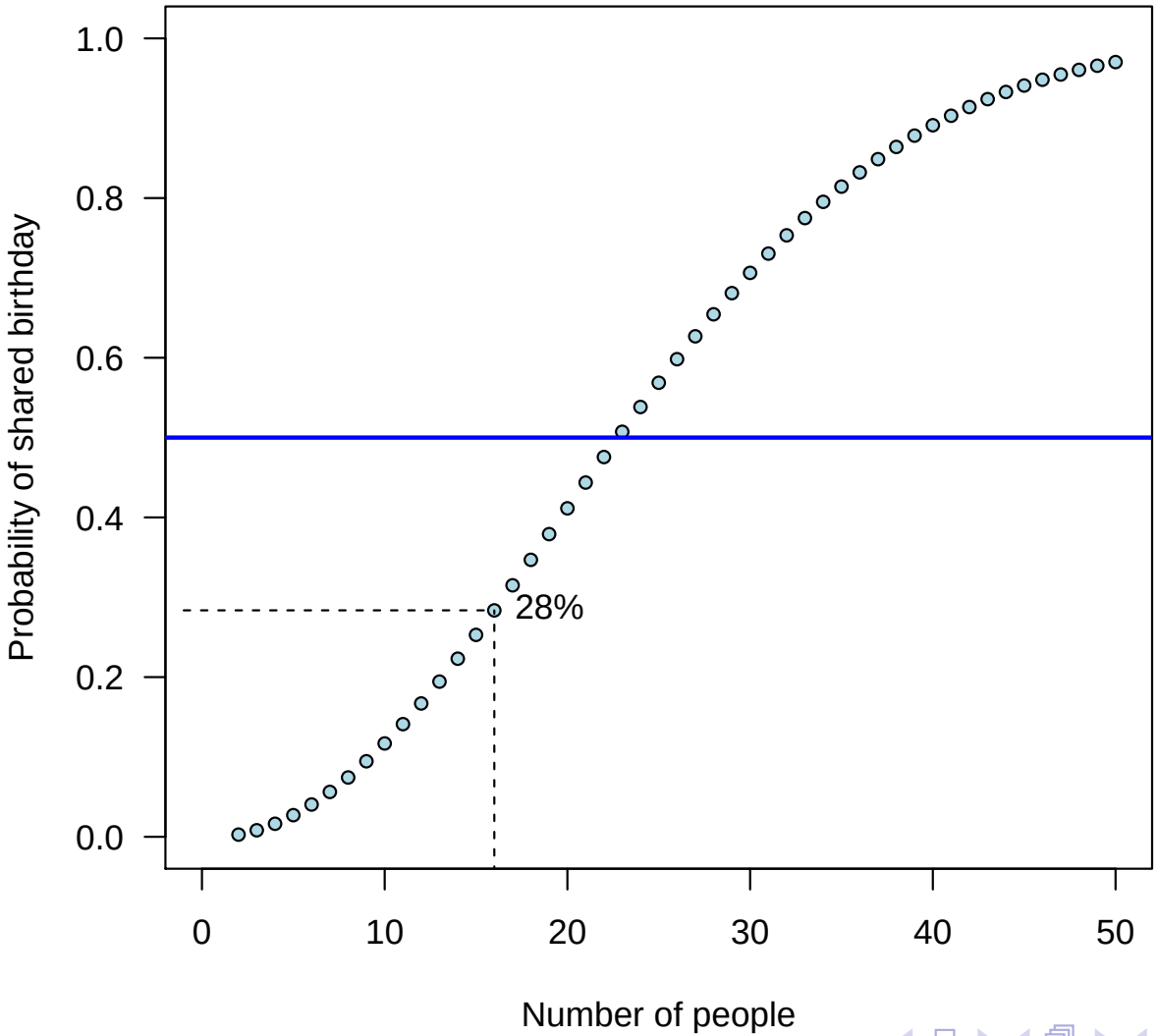
The Birthday Problem – Analysis

n	p_n	n	p_n
2	0.9973	13	0.8056
3	0.9918	14	0.7769
4	0.9863	15	0.7471
5	0.9729	16	0.7164
6	0.9595	17	0.6850
7	0.9438	18	0.6531
8	0.9257	19	0.6209
9	0.9054	20	0.5886
10	0.8831	21	0.5563
11	0.8589	22	0.5243
12	0.8330	23	0.4927

The Birthday Problem – Analysis



The Birthday Problem – Analysis



The Birthday Problem – Discussion

- We see from the table that only 23 birthdays are needed for there to be greater than a 50% probability of sharing
- The above analysis ignores leap years and makes two important assumptions
  1. that birthdays are spread evenly throughout the year
  2. that there is no pattern of birthdays within families
- Are these assumptions reasonable?
- Experiments with real data tend to show that the chances are considerably *greater* than 50%

Finally...

(a) You could spend millions on data collection, analytics, and advertising
or
(b)

Finally...

(a) You could spend millions on data collection, analytics, and advertising
or
(b)

Finally...