

The background of the slide features two white dice with black pips, resting on a red surface. The dice are positioned diagonally, with one in the foreground and another slightly behind it. A white rectangular box with a blue border is centered over the dice, containing the course title and instructor information.

# Physics 403

Probability, Statistics,  
and Exploration of Large Datasets

Segev BenZvi

Department of Physics and Astronomy  
University of Rochester

# Table of Contents

- 1 Course Overview
  - Texts and Materials
  - Software Overview
- 2 Inference and Probability in Science
  - The Meaning of Probability
  - Frequentist and Bayesian Statistics
- 3 Probability as an Extension of Classical Logic
  - Boolean Algebra

# Course Goals

## Use of Probability and Statistics to Interpret Experimental Data

### Probability

- ▶ definition, interpretation, Bayes' Theorem, random variables, probability distributions, expectation values, error propagation, ...

### Monte Carlo Methods

- ▶ random number generators, PDF transformation, acceptance-rejection

### Method of Maximum Likelihood

- ▶ likelihood, ML estimators, extended ML, binned ML, variance of ML

### Method of Least Squares

- ▶ goodness of fit, relation to maximum likelihood

# Course Goals (continued)

## Interval Estimation

- ▶ confidence intervals, lower and upper limits, significance, coverage

## Multivariate Techniques

- ▶ decision trees, boosting, machine learning, ...

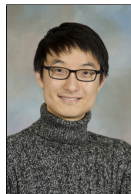
# Course Contacts

Segev BenZvi  
Instructor



- ▶ B&L 405
- ▶ (585) 276-7172
- ▶ [sybenzvi@pas.rochester.edu](mailto:sybenzvi@pas.rochester.edu)
- ▶ Office Hours: Tu 11-12, Th 2-3

Chang Rho  
TA



- ▶ B&L 477
- ▶ (585) 275-8556
- ▶ [crho2@ur.rochester.edu](mailto:crho2@ur.rochester.edu)
- ▶ Office Hours: W 2-3

# Course Format

## Lectures:

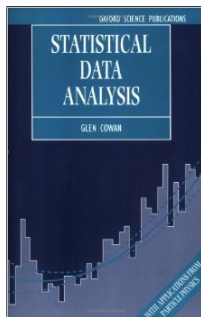
- ▶ M,W 10:25-11:40
- ▶ Lectures will be posted on BlackBoard

## Grading:

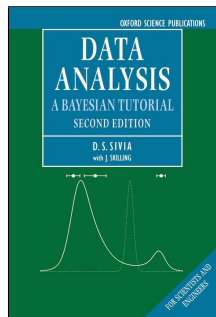
- ▶ 45% HW: bi-weekly problem sets
  - ▶ Theory and practical examples with a programming component
  - ▶ Due Friday @ 5 pm
- ▶ 20% midterm: in-class exam
- ▶ 10% lecture participation
- ▶ 25% final exam, format TBD:
  - ▶ Group project on current research, previous results, etc. In-class presentations lasting 15 minutes. Or...
  - ▶ Data analysis project; basically a take-home exam. Or...
  - ▶ In-class exam.

# Textbooks

There are two textbooks used in this course, both relatively concise:



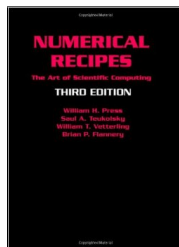
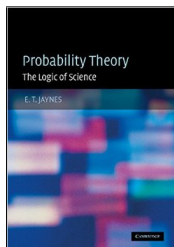
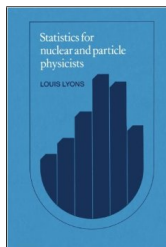
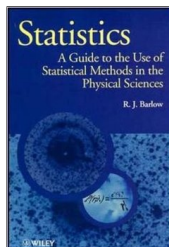
Cowan [1]: good introduction to “frequentist” statistics and data analysis as practiced in HEP



Sivia [2]: introduction to “Bayesian” statistics with applications and examples

# Additional Material

Supplementary material is on reserve in POA:



- ▶ *Barlow* [3], *Lyons* [4]: statistics for physicists with many examples. Focus on particle physics (both authors are in HEP) but generally applicable
- ▶ *Jaynes* [5]: Bayesian probability as an extension of logic
- ▶ *Numerical Recipes* [6]: book on algorithms and data analysis, easy to read; a great reference



# Homework

There will be 6 to 7 bi-weekly problem sets assigned during the semester:

- ▶ Solutions will be due every other **Friday at 5 pm**
- ▶ Some problems will require analytical solutions; others will have strong numerical, programming, and data analysis components
- ▶ You may informally discuss the problems with other students but your work must be your own
- ▶ Some problems could take time to solve. Don't put off the homework until the last minute
- ▶ **Show your work!** Problems that require programming and data analysis **must include printed source code** with your solutions or you won't get credit

# Software

This is not a software course, but you'll have to write basic programs (usually  $< 100$  lines) for data analysis. Useful software includes:

- ▶ **Python**: interpreted language with huge user community and excellent analysis and plotting packages (numpy, scipy, etc.)
- ▶ **ROOT**: C++ statistics package and plotting library from CERN, used in HEP. *Has python bindings!*
- ▶ **GSL**: GNU Scientific Library, very handy, written in C
- ▶ **Mathematica, R, ...**: use if you wish, but if you run into problems don't expect support from us

Typical **open-source computing environment**: Linux, Python

- ▶ See posted lecture on Linux and next week's lecture on programming

# Computing Methods Used in the Course

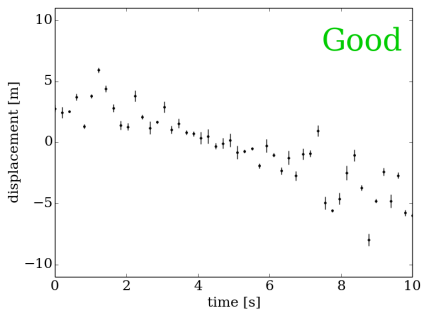
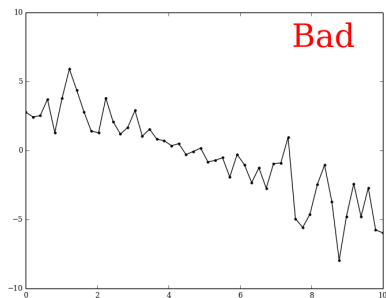
Data analysis will require you to become fluent with the following techniques:

- ▶ Random number generation and sampling
- ▶ Solving systems of linear equations (linear algebra)
- ▶ Matrix manipulation and inversion
- ▶ Numerical integration
- ▶ Minimization of linear and nonlinear functions
- ▶ Plotting: scatterplots, histograms, contour plots, corner plots

You'll need to understand how these methods work, but **don't reinvent the wheel**. I.e., don't write your own linear algebra library!

## A Word About Plotting

To receive credit, plots must be clearly readable, measurements should have error bars, and axes must be labeled (including units).



Beware of default plotting behavior like spurious lines between data points. Don't let yourself look like an amateur!

# Table of Contents

- 1 Course Overview
  - Texts and Materials
  - Software Overview
- 2 Inference and Probability in Science
  - The Meaning of Probability
  - Frequentist and Bayesian Statistics
- 3 Probability as an Extension of Classical Logic
  - Boolean Algebra

# Physics and Data Analysis

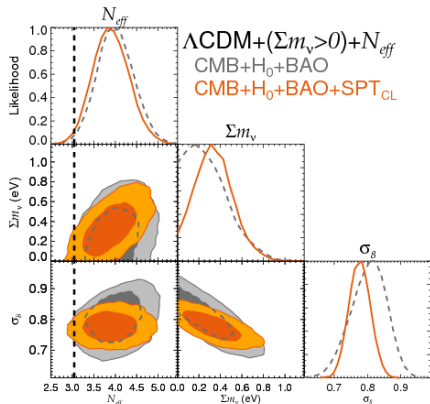
Two types of experimental inference:

**Hypothesis testing**, a test of a theory or model.

- ▶ How consistent are data with a given model?
- ▶ Which of several models best describes the data?

**Parameter estimation**, the measurement of a quantity.

- ▶ What is the best estimate of the parameter?
- ▶ How well did we make the measurement (uncertainty)?



Parameter estimation using CMB data [7].

# Deduction vs. Induction (Plausible Reasoning)

How does inference work in math and science?

**Deduction:** given a cause, we work out its effects.

## Example

If a fair coin is flipped 10 times, what is the chance that 7 times it will come up heads?

**Induction:** given a set of effects we identify plausible underlying causes.

## Example

If a coin was flipped 10 times and yielded 7 heads, what is the probability that it is a fair coin?

# Deduction vs. Induction (Plausible Reasoning)

How does inference work in math and science?

**Deduction:** given a cause, we work out its effects.

## Example

If a fair coin is flipped 10 times, what is the chance that 7 times it will come up heads?

**Induction:** given a set of effects we identify plausible underlying causes.

## Example

If a coin was flipped 10 times and yielded 7 heads, what is the probability that it is a fair coin?

Scientific inference works by **induction**. Given a set of measurements we make plausible inferences about their origin, using the language of **probability**.



# What is Probability?

There is not universal agreement on what *probability* means, and this does affect you. Three suggested interpretations:

# What is Probability?

There is not universal agreement on what *probability* means, and this does affect you. Three suggested interpretations:

1. **Propensity**. Ex.: an unbiased coin has a “propensity”  $1/2$  to land heads or tails. I.e.,  $1/2$  is a *property of the coin*.
2. **Degree of belief**. Ex.: your *belief* that a coin is unbiased means you assign probability  $1/2$  to the proposition “the coin will land heads.”
3. **Relative frequency**. Ex.: the relative frequency with which heads appears in a sequence of infinite tosses of an unbiased coin is  $1/2$ .

# What is Probability?

There is not universal agreement on what *probability* means, and this does affect you. Three suggested interpretations:

1. **Propensity**. Ex.: an unbiased coin has a “propensity”  $1/2$  to land heads or tails. I.e.,  $1/2$  is a *property of the coin*.
2. **Degree of belief**. Ex.: your *belief* that a coin is unbiased means you assign probability  $1/2$  to the proposition “the coin will land heads.”
3. **Relative frequency**. Ex.: the relative frequency with which heads appears in a sequence of infinite tosses of an unbiased coin is  $1/2$ .

Questions:

- ▶ Which of these seems most reasonable to you?
- ▶ Is one more “fundamental” than the others?
- ▶ How can scientists compare independent measurements and draw conclusions?

# The Meaning of Probability

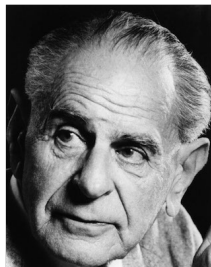
## Propensity

- ▶ Q: why does a given kind of experiment generate a given outcome at a persistent rate?
- ▶ A: probability is physical property, like particle decay in Quantum Mechanics.
- ▶ **Problem:** we don't observe the intrinsic probability, just the outcome.
- ▶ **Problem:** how to quantify ignorance?

## Example

I flip a coin and look at the result (tails).

- ▶ You: 50% probability the coin is tails.
- ▶ Me: the coin is tails w/ 100% probability.

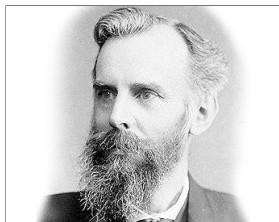


Karl Popper: "In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality [8]."

# The Meaning of Probability

## Relative Frequency, or “Frequentist Statistics”

- ▶ Probability & chance come from ignorance of physical causes, not intrinsic properties.
- ▶ We can't or don't usually need to care about causes, just the “objective” **consistency of long-run outcomes**.
- ▶ **Appeal**: we observe outcomes and relative frequencies all the time in controlled experiments!
- ▶ **Problem**: very awkward to talk about non-repeatable phenomena. Invoke “ensembles” of hypothetical identical systems.
- ▶ **Problem**: how to “objectively” identify an ensemble?



John Venn: “On the view here adopted we are concerned only with averages, or with the single event as deduced from an average and conceived to form one of a series [9].”

# The Meaning of Probability

Degree of Belief, or “Bayesian Statistics”

- ▶ Probability quantifies our knowledge or ignorance of a situation given “data” plus our *a priori* beliefs.
- ▶ Degree of belief is always subjective, but can/should be revised as we learn (a.k.a., “acquire new data”).
- ▶ **Appeal:** this is common sense, and describes scientific inference very well.
- ▶ **Problem:** how do we quantify our *a priori* beliefs, i.e., our beliefs before taking any data?
- ▶ **Problem:** if probability is subjective, how can we achieve consensus?



Laplace: “Probability theory is only common sense reduced to calculation.”

# Frequentist and Bayesian Statistics

The “propensity” view is not common but you will often encounter both frequentist and Bayesian approaches.

## Bayesian

- ▶ Data are known and fixed; calculate probability of a hypothesis given data.
- ▶ Fundamental; a model of *inductive reasoning*.
- ▶ Well-defined procedure for calculating a probability.
- ▶ **Must quantify prior knowledge of parameters/hypotheses, even if your “knowledge” is total ignorance. Can be HARD!**

## Frequentist

- ▶ Model parameters are unknown but fixed; calculate probability of data given a hypothesis.
- ▶ Under limited circumstances, guarantees experiments obey long-run relative frequencies.
- ▶ Uses “random variables” to model the outcome of unobserved data.
- ▶ **Ad hoc, “cookbook” approach to statistics.**

# Table of Contents

- 1 Course Overview
  - Texts and Materials
  - Software Overview
- 2 Inference and Probability in Science
  - The Meaning of Probability
  - Frequentist and Bayesian Statistics
- 3 Probability as an Extension of Classical Logic
  - Boolean Algebra



# Deductive Logic and Boolean Algebra

How to combine premises  $A$  and  $B$  which can be **true (1)** or **false (0)**:

- ▶  $A + B$ : logical disjunction (OR): e.g., “ $A$  or  $B$ ” is true
- ▶  $AB$ : logical conjunction (AND): e.g., “ $A$  and  $B$ ” are both true
- ▶  $\bar{A}$ : logical negation (NOT): e.g., “not  $A$ ” is true
- ▶  $\implies$  : implication: infer something from several premises (syllogism)

Comment:  $A \implies B$  (read “ $A$  implies  $B$ ”) does not mean either  $A$  or  $B$  is true. Instead, it means

$$A = AB.$$

If  $A$  is true then  $B$  must be true. If  $B$  is false then  $A$  must be false.

## Example

$A =$  “it is raining”,  $B =$  “it is cloudy”.

# Truth Tables

A handy visual trick to combine premises  $A$  and  $B$  which can be **true (1)** or **false (0)** is to use a truth table:

$A$	$B$	$A + B$	$AB$	$\overline{A + B}$	$\overline{AB}$
0	0	0	0	1	1
0	1	1	0	0	1
1	1	1	1	0	0
1	0	1	0	0	1

$\overline{A}$	$\overline{B}$	$\overline{A + B}$	$\overline{A} \overline{B}$
1	1	1	1
1	0	1	0
0	0	0	0
0	1	1	0

- ▶ Note:  $\overline{AB} = \overline{A} + \overline{B}$ : “not( $A$  and  $B$ ) = not( $A$ ) or not( $B$ )”
- ▶ Note:  $\overline{A} \overline{B} = \overline{A + B}$ : “not( $A$ ) and not( $B$ ) = not( $A$  or  $B$ )”

# Basic Logic Identities

Often Used in Programming and Digital Logic

$$\text{Idempotence : } \begin{cases} AA = A \\ A + A = A \end{cases}$$

$$\text{Commutativity : } \begin{cases} AB = BA \\ A + B = B + A \end{cases}$$

$$\text{Associativity : } \begin{cases} A(BC) = (AB)C = ABC \\ A + (B + C) = (A + B) + C = A + B + C \end{cases}$$

$$\text{Distributivity : } \begin{cases} A(B + C) = AB + AC \\ A + (BC) = (A + B)(A + C) \end{cases}$$

$$\text{Duality : } \begin{cases} \text{if } C = AB, \text{ then } \overline{C} = \overline{AB} = \overline{A} + \overline{B} \\ \text{if } D = A + B, \text{ then } \overline{D} = \overline{A + B} = \overline{A} \overline{B} \end{cases}$$

# Going from Boolean Premises to Statements of Plausibility

- ▶ Instead of deductive assertions like “if  $I$  is true, then  $A$  is true”, we are interested in quantifying  $P(A|I)$ , the *probability of  $A$  being true, given that  $I$  is true*.
- ▶ Comment 1: This is a **conditional probability**. The “|” means “given,” so all information to the right is assumed true.
- ▶ Comment 2: There is **no such thing as an absolute probability**. All probabilities are conditional at some level.
- ▶ Comment 3: Don't expect conditional relationships to commute.

## Example

Let  $A$  = “it is cloudy” and  $B$  = “it is raining.” In general,

$$P(A|B) \neq P(B|A).$$

# Requirements for Probabilistic Reasoning

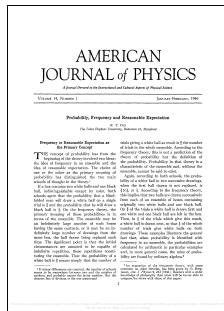
What properties are required for logical and consistent reasoning?

The following features are desired:

1. Degrees of plausibility are represented by **real numbers**;
2. **Common sense**: as data supporting a hypothesis accumulate, its plausibility increases continuously and monotonically;
3. **Consistency**: if there are two different ways to use the same information, both methods should give the same conclusion.

# Probability as a Generalization of Deductive Logic

- ▶ 1933: Kolmogorov describes probability using axiomatic set theory (see Cowan, Ch. 1 [1]).
- ▶ 1946: Combining common sense and consistency with Boolean algebra, Cox found that real numbers representing probabilities *must* obey Kolmogorov's axioms [10].
- ▶ “Subjective” probability is really an extension of Aristotelian deductive logic.
- ▶ Consistency guarantees that two observers with the same prior information and data **will assign the same probability** to an event [5].



R.T. Cox: “Employing the algebra of symbolic logic it is possible to derive the rules of probability... which appeal rather immediately to common sense [10].”

## Further Reading I

- [1] Glen Cowan. *Statistical Data Analysis*. New York: Oxford University Press, 1998.
- [2] D.S. Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. New York: Oxford University Press, 1998.
- [3] R.J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*. New York: Wiley, 1989.
- [4] Louis. Lyons. *Statistics for Nuclear and Particle Physicists*. New York: Cambridge University Press, 1986.
- [5] E.T. Jaynes. *Probability Theory: The Logic of Science*. New York: Cambridge University Press, 2003.
- [6] W. Press et al. *Numerical Recipes in C*. New York: Cambridge University Press, 1992. URL: <http://www.nr.com>.

## Further Reading II

- [7] B.A. Benson et al. "Cosmological Constraints from Sunyaev-Zel'dovich-Selected Clusters with X-ray Observations in the First 178 Square Degrees of the South Pole Telescope Survey". In: *Astrophys.J.* 763 (2013), p. 147. arXiv: [1112.5435](https://arxiv.org/abs/1112.5435) [astro-ph.CO].
- [8] Karl Popper. *The Logic of Scientific Discovery*. New York: Hutchinson and Co., 1959.
- [9] John Venn. *The Logic of Chance*. New York: Macmillan, 1888. URL: <https://archive.org/details/logicofchance029416mbp>.
- [10] R.T. Cox. "Probability, Frequency, and Reasonable Expectation". In: *Am. J. Phys.* 14 (1946), p. 1.