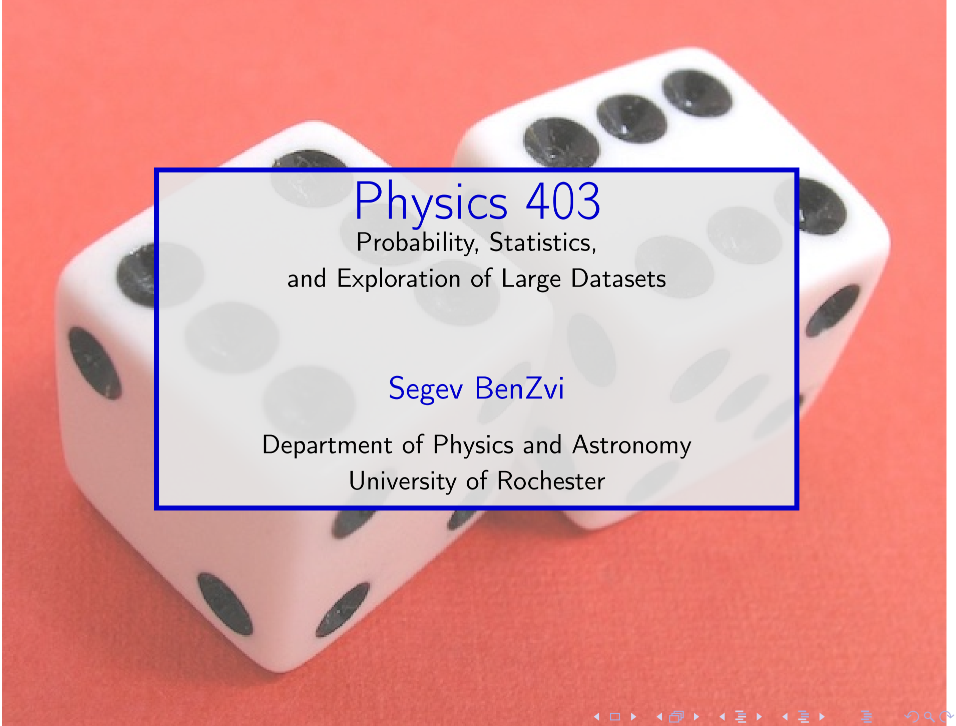


Physics 403

Probability, Statistics,
and Exploration of Large Datasets

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Table of Contents

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

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
- Office Hours: Th 2:00 - 3:00

Brian Coopersmith

TA

- B&L 373
- (585) 275-4001
- coopersmithbrian@gmail.com
- Office Hours: W 12:45 - 1:45

Course Format

Lectures:

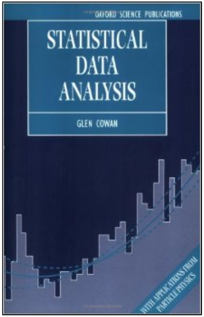
- M,W 9:30-10:50. Lectures posted on BlackBoard

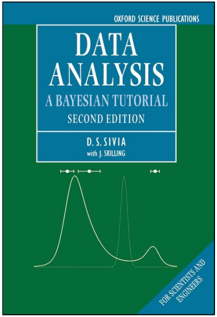
Grading:

- 45% HW, 25% final project, 20% midterm, 10% lecture participation
- Bi-weekly problem sets (theory and practical examples with a programming component), due Friday @ 5 pm
- Final project on a topic of your choice
  - Your current research, previous result, . . .
  - $\sim$ 20 minute presentations, starting last 3 lectures

Textbooks

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



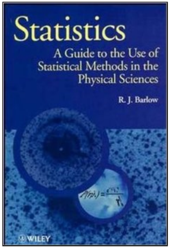


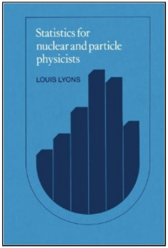
Cowan [1]: good introduction to "frequentist" statistics and data analysis as practiced in HEP

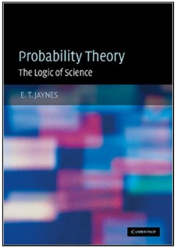
Sivia [2]: introduction to "Bayesian" statistics with applications and examples

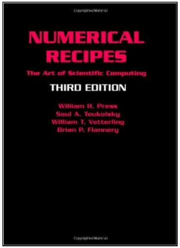
Additional Material

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

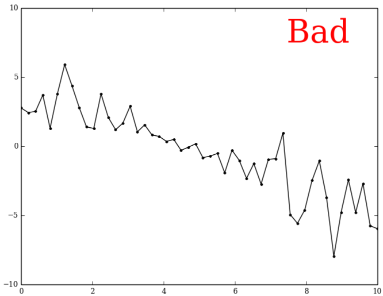
Data analysis will require you to become fluent with the following techniques:

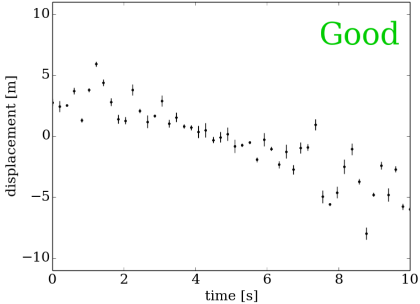
- Random number generation and sampling
- Solving systems of linear equations (linear algebra)
- Matrix manipulation and inversion
- Numerical integration
- Minimization of linear and nonlinear functions
- Plotting: scatterplots, histograms, contour plots, corner plots

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

Physics and Data Analysis

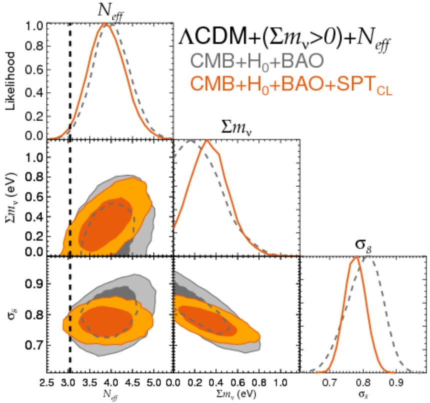
Two types of experimental inference:

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?

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:

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

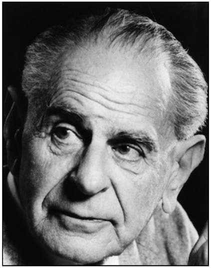
Propensity

- 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.

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

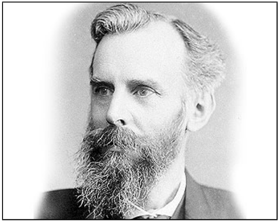


Karl Popper: *"In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality [8]."*

The Meaning of Probability

Relative Frequency, or "Frequentist Statistics"

- Probability & chance come from ignorance of physical causes, not intrinsic physical 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 generalization of *deductive logic*.
- 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

Deductive Logic and Boolean Algebra

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

- $\mathbf{A} + \mathbf{B}$: logical disjunction (OR): e.g., "A or B" is true
- $\mathbf{AB}$: logical conjunction (AND): e.g., "A and B" are both true
- $\overline{\mathbf{A}}$: 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

$\overline{A}$	$\overline{B}$	$\overline{A}+\overline{B}$	$\overline{A}\,\overline{B}$
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?

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 the desiderata of 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]."*

Basic Rules of Probability

1. **Representation**: truth: $P(A|I) = 1$; falsehood: $P(A|I) = 0$.
2. **Sum Rule**: $P(A|I) + P(\overline{A}|I) = 1$

Example

$A =$ "a coin toss gives tails." Clearly $P(A|I) + P(\overline{A}|I) = 1$.

3. **Product Rule**: $P(A, B|I) = P(A|B, I) \times P(B|I)$

Example

Two red marbles and one blue marble are in a bag. Two marbles are drawn from the bag *in sequence* and without replacement. What's the probability that both marbles are red?

Basic Rules of Probability

1. **Representation**: truth: $P(A|I) = 1$; falsehood: $P(A|I) = 0$.
2. **Sum Rule**: $P(A|I) + P(\overline{A}|I) = 1$

Example

$A =$ "a coin toss gives tails." Clearly $P(A|I) + P(\overline{A}|I) = 1$.

3. **Product Rule**: $P(A, B|I) = P(A|B, I) \times P(B|I)$

Example

Two red marbles and one blue marble are in a bag. Two marbles are drawn from the bag *in sequence* and without replacement. What's the probability that both marbles are red?

$$P(R|I) = 2/3$$
$$P(R, R|I) = P(R|I) \times P(R|R, I) = 2/3 \times 1/2 = 1/3$$

Further Properties of Probability Functions

Some additional important properties of probability functions can be derived using Boolean algebra and repeated applications of the sum and product rules. For example:

$$
\begin{aligned}
P(A|I) &= 1 - P(\overline{A}|I) \\
P(A+\overline{A}|I) &= 1 \\
P(A|I) &\in [0,1] \\
P(A+B|I) &= P(A|I) + P(B|I) - P(A,B|I)
\end{aligned}
$$

Furthermore, from the product rule, A and B are called **independent** if $P(A|B,I) = P(A|I)$ and $P(B|A,I) = P(B|I)$, so that

$$P(A,B|I) = P(A|I) \times P(B|I).$$

Example

We draw marbles from our bag but *replace them* after each draw.

Bayes' Theorem

The very important Bayes' Theorem can also be derived directly from the product rule:

$$P(A, B|I) = P(A|B, I) \times P(B|I)$$
$$P(B, A|I) = P(B|A, I) \times P(A|I)$$

Logically, $AB = BA$, so $P(A, B|I) = P(B, A|I)$. Therefore

$$\boxed{P(A|B, I) = \frac{P(B|A, I) \times P(A|I)}{P(B|I)}}$$

"The probability of A given B and I is equal to the probably of B given A times the probability of A irrespective of B, divided by the probability of B irrespective of A."

Bayes' Theorem and Inference

Replace A with *hypothesis* H and B with *data* D to see how Bayes' Theorem applies to model selection and parameter estimation:

- *A priori* probability of the hypothesis (**"prior"**)
- **"Likelihood"** of data given the hypothesis

$$P(H|D,I) = \frac{P(D|H,I) \times P(H|I)}{P(D|I)}$$

- **Posterior** probability
- **"Evidence"** or **"prior predictive"** of the data

Using Bayes' Theorem you can construct a probability for any hypothesis given an observation.

The Posterior Probability

The posterior probability $P(H|D, I)$ gives the probability that hypothesis H is true given the data D and background information I.

Example

You have some data $(\mathbf{x}, \mathbf{y})$ that appear to be linear. Your hypothesis H could be "the data were generated by a function $f(\mathbf{x}) = a\mathbf{x} + b$."
In this case, $P(H|D, I) = P(H|(\mathbf{x}, \mathbf{y}), I)$ gives the probability that the data were generated by $f(\mathbf{x})$.

In order to calculate $P(H|D, I)$, you need to quantify:

- The **likelihood** $P(D|H, I)$, which is usually quite easy.
- The **prior** $P(H|I)$, which is not always obvious.

Comment: in frequentist statistics priors are not calculated at all. Only the likelihood is used.

The Likelihood

$P(D|H,I)$: "what is the probability of observing D given H?"

Example

Using the example from the last slide, if the measurements $\mathbf{y} = \{y_i\}$ are *independent* and have Gaussian uncertainties of width σ, we would write

$$
\begin{aligned}
P(D|H,I) &= p(y_1|H,I) \times p(y_2|H,I) \times \ldots \times p(y_N|H,I) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y_i - (ax_i + b)}{\sigma}\right)^2\right\}, \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - (ax_i + b)}{\sigma}\right)^2\right\}.
\end{aligned}
$$

Note that the likelihood does NOT give us the probability that the data are linear; we already *assumed* $\mathbf{y} = a\mathbf{x} + b$ when constructing $P(D|H,I)$.

The Prior

Choosing the prior $P(H|I)$ is tricky. It could be:

- A known relative frequency from previous observations.
- A theoretical input with some given uncertainty.
- A noninformative probability density function that indicates our total ignorance (meaning of "noninformative" to be defined later).
- A personal opinion.

Example

Using the example from the previous two slides, $P(H|I)$ could be:

- Our prior belief in H that the data are linear;
- Our belief in the likely values of the model parameters a and b, with I corresponding to previous measurements or values motivated by theory.

As a rule, we want a prior that doesn't overly bias us against new discoveries in the data. Doing this correctly can be non-trivial.

Law of Total Probability

Marginalization: The Evidence Term in Bayes' Theorem

What is the meaning of the normalization or "evidence" term $P(D|I)$?

- Probability of the observation D, independent of the hypothesis H.
- H doesn't affect $P(D|I)$ – it is a *nuisance parameter* – so we marginalize it [2]:

$$\begin{aligned} P(D|I) &= P(D,H|I) + P(D,\overline{H}|I) \\ &= [P(D|H,I) \times P(H|I)] + [P(D|\overline{H},I) \times P(\overline{H}|I)]. \end{aligned}$$

  We express $P(D|I)$ in terms of the joint probability of D and the mutually exclusive hypotheses H and $\overline{H}$.
- **Justification**: logical negation, sum rule, and product rule.
- If there are M mutually exclusive (and exhaustive) hypotheses then

$$P(D|I) = \sum_{i=1}^{M} P(D|H_i,I) \times P(H_i|I), \qquad \text{with} \sum_{i=1}^{M} P(H_i|\ldots) = 1$$

Marginalization

Discrete and Continuous Cases

Relation between joint (multidimensional) and marginal (1D) probabilities:

- Discrete Case: for exclusive, exhaustive B_i,

$$P(A|I) = \sum_i P(A, B_i|I) = \sum_i P(A|B_i, I)P(B_i|I)$$

- Continuous case: $P(A|I)$ is now a probability density

$$P(A|I) = \int P(A, B|I)dB = \int P(A|B, I)P(B|I)dB$$

Can interpret this procedure in two ways:

1. Marginalization: get $P(A|I)$ from joint distribution $P(A, B|I)$.
2. Normalization of Bayes' Theorem: we don't know how to calculate $P(A|I)$ directly, so we expand it in a "basis" using the set of $\{B_i\}$.

Application of Bayes' Theorem

Example

We have 3 coins, two fair (F) and one completely biased (B) toward tails. We pick one coin and flip it 3 times, finding tails in all three tosses, i.e., $D = \{\text{T,T,T}\}$. What is the probability that we picked the biased coin?

Application of Bayes' Theorem

Example

We have 3 coins, two fair (F) and one completely biased (B) toward tails. We pick one coin and flip it 3 times, finding tails in all three tosses, i.e., $D = \{\text{T,T,T}\}$. What is the probability that we picked the biased coin?

$$\begin{aligned}
P(B|D,I) &= \frac{P(D|B,I)P(B,I)}{P(D,I)} \\
&= \frac{P(D|B,I)P(B,I)}{P(D|B,I)P(B,I) + P(D|F,I)P(F,I)} \\
&= \frac{1^3 \cdot (1/3)}{1^3 \cdot (1/3) + (1/2)^3 \cdot (2/3)} = \frac{1/3}{1/3 + 1/8 \cdot 2/3} \\
&= 4/5
\end{aligned}$$

Similarly, you can calculate that $P(F|D,I) = 1/5$, or just infer it from the sum rule because the fair and biased hypotheses are exclusive.

Summary

- **Sum Rule**:

$$P(A|I) + P(\overline{A}|I) = 1$$
$$\sum P(H_i|I) = 1 \quad \text{for exclusive } H_i$$

- **Product Rule**:

$$P(A, B|I) = P(A|B, I)P(B|I)$$

- **Bayes' Theorem**:

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)}$$

- **Law of Total Probability**:

$$P(A|I) = \sum_i P(A, B_i|I) = \sum_i P(A|B_i, I)P(B_i|I)$$

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 [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.