Introduction to Probability and Statistics

October 4, 2016

Course Description

Course history: came out of the new observational cosmology program at UCD
Philosopy: Pedantic but also practical
Books

- Wall & Jenkins, Practical Statistics for Astronomers
- Ivezic et al., Statistics, Data Mining, and Machine Learning in Astronomy
- Press et al., Numerical Recipes
- Bevington, Data Reduction and Error Analysis
- Taylor, An Introduction to Data Analysis
- etc.

The Scientific Method and Probabilities

Consider the scientific method (here, somewhat expanded):

1. Hypothesis
2. Data collection
3. Data reduction / calibration
4. Construction of a test statistic
    - **Statistic:** a quantity that summarizes the data, and depends *only* upon the data themselves
    - e.g., the mean
5. Decision / Hypothesis testing
    - Based on probability
    - e.g., How well do the data agree with the model?
    - e.g., Is this particular observation unusual?

Thus, any observational / experimental science is one of probabilities

Uncertainties: First thoughts

A measured or derived quantity is **useless** without an estimate of the uncertainty (error) associated with that measurement or derivation.

- Consider this quantity and its true value $g = 9.81$ m/s^2
- One measurement finds 9.83
- Another finds 10.1
- Which is more consistent with the true value? *We don't know just given the central values!*

A quantity such as 10.1 ± 0.4 *implies* a probability

Statistical vs. Systematic Uncertainties

Two types of uncertainties:

- Random (statistical): Randomly distributed around some central value
- Systematic: Can shift the whole distribution. Often much harder to track down or to realize that these errors are present

Another way to characterize these: **Accuracy vs. Precision**

And yet another way of thinking about it:

> *"If you are observing Arcturus when you are supposed to be observing Vega, the error will never average away."*

[Bulls eye figure from Taylor??]

Applicability to Astrophysics and Cosmology

Some "fun" aspects

- We cannot repeat or control the "experiment" – there is just one Universe
- We are often dealing with very small numbers

Some examples from astrophysics / cosmology

- **Detection of signals:** e.g., object detection in an image, CMB fluctuations, spectral features, etc.
- **Detection of correlations:** e.g., Hubble diagram, galaxy distributions, etc.
- **Tests of hypotheses:** e.g., isotropy of Universe, physical association of objects, etc.

Bayesian vs. Frequentist I

This is a bit of a preview, but there are two major views as to how to approach statistical inference: frequentist and Bayesian.

- Frequentist is the classical approach that is, or used to be, taught in many lab and statistics classes
- Bayes' theorem (more on this later), was seen as interesting but not that useful
- Now the Bayesian approach is seen as a very powerful approach to statistical inference.

What follows on the next few slides is my unsophisticated take on how they differ.

Bayesian vs. Frequentist II

Frequentist approach

- The underlying idea: there is one correct hypothesis
- Collect your data and then determine a statistic
- Use your hypothesized model to evaluate the probability of getting that particular data set from the model (the "likelihood").
- Then make a decision, e.g., detection or not.
- Often in astro, test the null hypothesis: i.e., what is the probability, given the underlying *assumed* distribution, of getting the observed data *by chance*

Bayesian vs. Frequentist III

An alternative take on the frequentist approach:

- Use the assumed distribution to get the probability that a given range around our single measurement contains the true value
- In practice, typically find the maximum likelihood and then use the chosen statistic to establish confidence regions

Bayesian vs. Frequentist IV

Bayesian approach

- Underlying idea: The data are unique and known, so use them to determine a probability distribution for the parameters of a model
- No statistic is calculated
- Instead, the statistic (e.g., the true mean of the distribution) is considered to be unknown and we use the data to determine the probability distribution of the statistic

The Basics

Probability: "A numerical formulation of our degree or intensity of belief"

Consider an event A. For example, A could be getting a 2 when you throw a 6-sided dice. The probability of that event occuring is $P(A)$

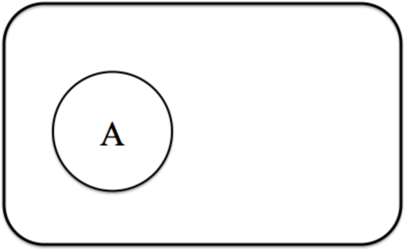
The Kolmogorov Axioms

1. $0 \leq P(A) \leq 1$
2. The "sure event" has $P(A) = 1$
3. If two events A and B are *exclusive*, then $P(A \text{ or } B) = P(A) + P(B)$

What does "exclusive" mean? Stay tuned...

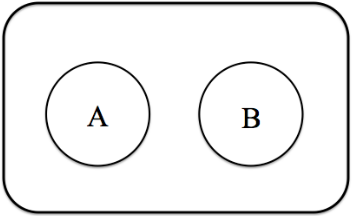
Probability as Set Diagrams I

- One way to think about probability is in terms of sets
- In this picture, $P(A)$ is the ratio of the area containing event A (the circle) to the area of the region containing all possible events (the rectangle)



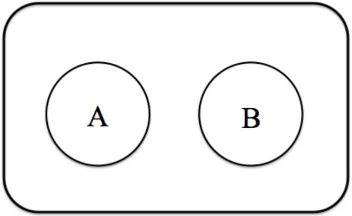
Probability as Set Diagrams II

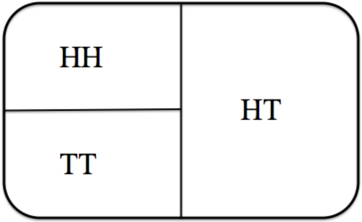
- In this formulation, what does "exclusive" mean?
- A and B are exclusive events if there is no overlap between them, as in the diagram.
- In other words $P(A \text{ and } B) = 0$
- One example: The possible results of two coin tosses.



Probability as Set Diagrams II

- In this formulation, what does "exclusive" mean?
- A and B are exclusive events if there is no overlap between them, as in the diagram.
- In other words $P(A \text{ and } B) = 0$
- One example: The possible results of two coin tosses.



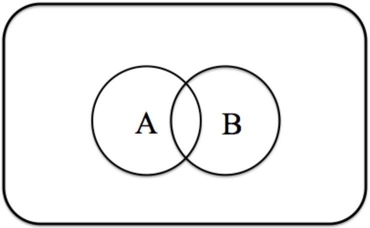


Independent Events I

- Events A and B are independent if $P(A)$ has no dependence on whether B has occured or not.
- Can independent events be exclusive?

Independent Events I

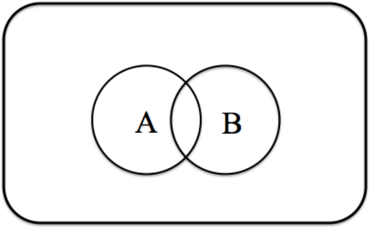
- Events A and B are independent if $P(A)$ has no dependence on whether B has occured or not.
- Can independent events be exclusive?
- No! In the exclusive case, if B has happened then you know that A **cannot** happen. That implies dependence.
- In the set diagram, the following *could* indicate independence as long as other conditions are met (see next slide)



Independent Events II

Conditions for independence

- Fraction of B that is also A is the same as the fraction of all possibilities that is A
- In other words [Area(A and B)]/Area(B) = Area(A)/Area(Total)
- $\Rightarrow P(A \text{ and } B)/P(B) = P(A)$
- $\Rightarrow P(A \text{ and } B) = P(A)P(B)$ if the events are independent



Generalizing Kolmogorov Axiom #3

- Remember, the axiom stated that if A and B are exclusive then $P(A \text{ or } B) = P(A) + P(B)$
- **General statement:**
  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- Avoids double-counting the area of intersection

Silly example:

- A: $1/3$ of kangaroos have blue eys
- B: $1/5$ of kangaroos are left-handed
- These two characteristics are independent
- What is $P(A \text{ or } B)$?

Conditional Probability

- In the previous discussion, we've talked about events happening once another event has happened.
- We describe the associated probability as the conditional probability
- e.g., $P(B|A)$ (say "probability of B given A") is the probability that B occurs *given that A has occurred first*.
- Thus, $P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B)$
- If A and B are independent then, e.g., $P(A|B) = P(A)$, since the occurance of A does not depend at all on B in this case.

Bayes' Theorem: the Basics

The derivation

- Start with statement from previous slide: $P(A)\,P(B|A) = P(B)\,P(A|B)$
- Rearrange to get **Bayes' Theorem**

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

Each term has an associated designation

- $P(B|A)$: the posterior probability
- $P(A|B)$: the likelihood
- $P(B)$: the prior probability
- $P(A)$: the evidence

Terms in Bayes' Theorem

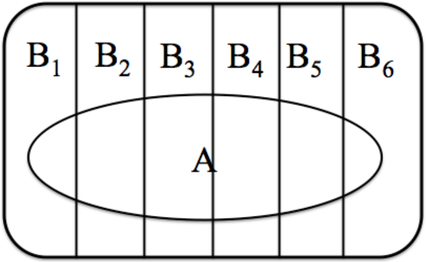
- The posterior is what you are trying to calculate
- The likelihood is $P(\text{data}|\text{model})$, which is the standard likelihood that is often calculated in frequentist analyses
- The prior reflects your expectations prior to conducting the experiment. You may have some previous information, e.g., from an earlier experiment, or be fairly ignorant about the possible values.
- What to use for a prior if you don't have a strong idea
    - A uniform or flat prior
    - The "Jeffries prior": uniform in log.
- The evidence is often just used as a constant of proportionality, but can powerfully be used to compare two models that are being fit to a given data set.

The Power of Bayes' Theorem

- The standard frequentist analysis calculates the likelihood: $P(\mathrm{data}|\mathrm{model})$
- Bayes' Theorem gives you, instead, what you really want, which is the probability distribution of the model parameters: $P(\mathrm{model}|\mathrm{data})$

$$\boxed{P(\mathrm{model}|\mathrm{data}) \;\propto\; P(\mathrm{data}|\mathrm{model})\, P(\mathrm{model})}$$

A more complex situation



- Consider a set of *exclusive* events, B_i, that span the space of possible outcomes, i.e., such that $\sum B_i = 1$.
- Then $P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \ldots$, since the B_i are exclusive
- $\Rightarrow \quad P(A) = \sum P(A|B_i)\, P(B_i)$

Marginalization

The summation over the B_i in the previous slide is called **marginalization**

- We often do this when we care about one parameter (e.g., A) and don't care about another "nuisance parameter" (e.g., B), so we sum over all possible values of B to just get an inference on A
- For example, B might represent some instrumental parameter (e.g., temperature of the detector) while A might be some cosmological parameter that is the goal of our experiment.
- In the case of continuous distributions, we marginalize by integrating. For example:

$$P(A) = \int_{-\infty}^{\infty} P(A|B)\, P(B)\, dB$$

A More General Form of Bayes' Theorem

- We may need to marginalize to get the evidence, e.g., $P(A)$ in Bayes' Theorem
- This is one reason why it can be hard to compute the evidence
- Consider a discrete case, where you care about one of the values of B:

$$\boxed{P(B_i|A) = \frac{P(A|B_i)\,P(B_i)}{\sum P(A|B_j)\,P(B_j)}}$$

- Let's do an example

Example: Test for a dread disease 1

Consider a test for, e.g., ebola

- The test is 98% accurate:
  - if you have ebola, then the test is positive 98% of the time
  - if you do not have ebola, then the test is negative 98% of the time
- $P(\text{ebola}) = 0.5\%$
- You test positive for ebola. Should you worry?a
- In other words, what is $P(\text{ebola}|\text{positive})$

Example: Test for a dread disease 2

Set up the problem

- Let A = a positive test result
- B_1 = you have ebola
- B_2 = you do not have ebola

$$P(B_1|A) = \frac{P(A|B_1)\,P(B_1)}{P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2)}$$

$$P(B_1|A) = \frac{(0.98)\,(0.005)}{(0.98)\,(0.005) + (0.02)\,(0.995)}$$

- In the end, $P(B_1|A) = 0.2$, so a 20% chance that a positive test means that you have ebola!
- **Moral:** For very rare events, you want a **very** accurate test. 98% doesn't cut it.