Philosophical Foundations

Manpreet Katari

Scientific Principles

- **Objectivity -** Avoid designing experiments that can be biased by confounding factors.
- **Realism -** Data collected should not be depended on the study. The study should be repeatable with similar conclusions.
- **Communalism -** Science grows as principles are challenged and refined by many researchers.

Scientific Methods

- **Observations** - thoughts that make us curious.
- **Testable hypothesis** - a possible explanation of the observation we have made
- **Performing the test** - collecting information to see if the hypothesis is correct.
- **Conclusion** - A decision is made whether the data supports the hypothesis.

Logic

- **Induction** - observations that lead to conclusions that are probably true.
  - NYU is a very selective university. (premise)
  - An NYU student is smart (conclusion)
- **Deduction** - general observations lead to a more specific conclusion. If the premise is true, the conclusion must be true.
  - 1440 is the average SAT score for an NYU student (premise)
  - 1440 is the 97th percentile for all students taking the SAT. (premise)
  - An average NYU student is in the 97th percentile. (conclusion)

Modus Tollens and affirming the consequent.

It is easier to reject the hypothesis when there is negative evidence. However it is not so easy to confirm the hypothesis.

If A is true then B is true

Data collected so far says B is not true then

A is not true

Modus tollens

If it rains, it is cloudy

It was not cloudy

It did not rain

If A is true then B is true

Data collected so far says B is true then

A is true

Modus tollens

If it rains, it is cloudy

It is cloudy

It rained

affirming the consequent

Variability and Uncertainty

It is impossible to prove a deductive argument because of variation.

There is variation even at the subatomic level

- *Heisenberg uncertainty principle*

Introduction to Probability

- Variable - stores values that can be changed
- Constant - value that does not change.

$f(t) = g(t)$

g is the gravitational constant ($9.8ms^{-2}$)

t and f(t) are variable because as t changes, f(t) will also change.

A random variable is a variable where the value is unknown until the experiment is done. It is usually defined by a probability. Outcome of a coin toss is a random variable with a probability of of 50% for heads.

Probabilistic Models

Deterministic models produce the same output when given the same input.

The speed of the ball when it is dropped from 10 m.

However, there are many factors that will affect the speed the ball and will cause slight variations. For this reason we often provide a variable to represent error.

$f(t) = gt + \varepsilon$

Majority of the measurements will have very small errors, where most values for the error term will be, or be very close to 0.

Set Theory

A set is a collection of objects, often defined in { }

A = { H,T }

To show that an element is part of a set you use $\in$ or $\notin$

H $\in$ A or J $\notin$ A

If A is a subset of B or A is not a subset of C

A $\subset$ B or A $\not\subset$ C

S is often the sample space (all possible values)

An empty set B = $\varnothing$

Probability

Frequentist interpretation - P(A) is the number of times A was observed in all trials.

P(H) = 0.5 - when you flip the coin, you get H 50% of the time.

Frequentist assume that there is a true value, but there may be none.

There is no prior knowledge used in the calculation of P(A)

Probability

Degrees of belief (bayesian)

P(A) the confidence that A is true. P(A) does not have to be fixed and it is not being determined, it’s just the confidence that A is true.

Prior information is expected to help augment the data.

A case for priors (example from book)

The gene for hemophilia is carried on the X chromosome. As a result, a woman will either express the disease (if she carries the disease allele on both of her X chromosomes) or be a carrier (if she carries the allele on a single X chromosome). Males on the other hand will always express hemophilia (if they carry the disease allele) because they have only one X chromosome. Consider now a woman named Mary whose mother is a carrier, and whose father did not have hemophilia. Because Mary received an X chromosome from each parent, she has a 50% chance (= probability 0.5) of carrying the gene. Suppose, now that Mary has two sons, each of whom is disease free. Surely, this new information should now be used to update the priors, and reduce the estimated probability that Mary is a carrier. Bayesian methods allow such calculations. *

Aho, Ken A.. Foundational and Applied Statistics for Biologists Using R (Page 26). Chapman and Hall/CRC.

Classical Probability

N = all possible outcomes

N(A) = outcomes resulting in A

$P(A) = N(A)/N$

$0 \leq P(A) \leq 1$

$P(A') = 1-P(A)$

$P(S) = 1$

$P(\varnothing) = P(S') = 0$

Disjoint

Two events cannot occur simultaneously they are mutually exclusive or disjointed.

If we want to know the probability of either outcome when the events are disjointed, we need to determine the **union**

$P(A \cup B) = P(A) + P(B)$

If the events are not mutually exclusive and we want to when they are both true we are looking for **intersect**

$P(A \cap B) = P(A) \times P(B)$

If the event are not mutually exclusive and we want to calculate the union it is

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ *# inclusion-exclusion principle*

Boole’s Inequality

If all events are disjoint than P(union of all A) = sum of P(A)

If they are not disjoint than P(union of all A) < sum of P(A)

$$P\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P(A_i).$$

The difference between A and B is P(A) but not in A ∩ B. This can also be written as

P(A ∩ B’) = P(A) - P(A ∩ B)

Independence and Bonferroni's Inequality

If the outcome of one event does not affect the outcome of another event they are independent of each other.

$P(A \cap B) = P(A)P(B)$

$$P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} P(A_i')$$

Probability of intersection of all

independent variables is greater than

1 - sum of not the independent events.

$$\geq \sum_{i=1}^{k} P(A_i) + (1-k),$$

Simpson’s diversity index and probability

The value can be used to understand how diverse and evenly distributes your data is. If you pick 2 times, what are the chances that they will be from different species/types of data?

$$D_1 = 1 - \sum_{i=1}^{s} p_i^2,$$

If there are two species, the probability of of picking the same species twice is p^2 . Since they are disjointed, the probability of the union of all species it the sum of the probabilities squared. So probability of getting two is simply 1 - that probability.

Conditional Probability

If A and B are not independent, we need to use conditional probability to calculate P(A ∩ B). P(A|B) is read probability of A given B.

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

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Odds

$$\Omega(A) = \frac{P(A)}{1 - P(A)} = \frac{P(A)}{P(A')}.$$

Odds is similar to probability but it compares probability of an event occurring to the probability that it doesn't rather than total events.

EXAMPLE 2.7 Cooper et al. (1993) studied the effectiveness of zidovudine (AZT) for the treatment of asymptomatic HIV+ patients. This study was of interest because while the effectiveness of AZT for preventing further loss of T-helper cells for patients with full-blown AIDS had been demonstrated in earlier studies, its effectiveness in cases with less severe disease development was unknown. Cooper et al. found that the disease progressed in 28% of asymptomatic HIV+ patients receiving a placebo, and in 16% of the patients receiving AZT. Based on these results, the odds of the disease progressing for placebo patients was 0.28/0.72 = 0.39, while the odds of the disease progressing for AZT patients was 0.16/0.84 = 019.

Aho, Ken A.. Foundational and Applied Statistics for Biologists Using R .

Odds Ratio and Relative Risk

Odds Ratio is the ratio of odds of two separate events.

Odd Ratio of placebo vs AZT patients is 0.39/0.19 = 2.04.

The odds of disease progression is 2x more in placebo.

Ratio of probabilities is the called relative risk.

The probability disease will progress in placebo is

.28/.16 = 1.75

Combinatorial Analysis

When an event include multiple selection from S, pay attention to:

1) Will the element be replaced to be selected.
2) Does the order of selection matter.

Multiplication principle (with replacement)

The total space of all possible outcomes N(S) is the multiplication of the number of times selection is occurring and the N of the selection.

EXAMPLE 2.8 A medical researcher is curious how survivorship from lymphoma can be enhanced by dosage combinations of two different drugs. She plans on randomly assigning one of three different dosages of drug 1 and one of two different dosages of drug 2 to subjects. How many distinct different treatments will she need to make?

EXAMPLE 2.9 The litter size for domestic dogs (Canis familiaris) generally ranges between 6 and 10 pups (HSUS 2013). How many different sorts of litters (combinations of male and female pups) are possible for a litter size of six? How about a litter size of 10?

Aho, Ken A.. Foundational and Applied Statistics for Biologists Using R (Page 36). Chapman and Hall/CRC. Kindle Edition.

Permutations (without replacement)

The number of permutations of any S of size n is n!

n! = n * (n-1) * (n-2) * … 1 # 0! = 1

If the selection is a subset of S

nPerm = n! / (n-r)! # where r is the number of S taken at a time.

How many ways to arrange 5 distinguishable items?

How many ways to select 2 out of the 5 items (without replacement)?

Combinations - without replacement & order doesn’t matter

Binomial coefficient

(n choose r)

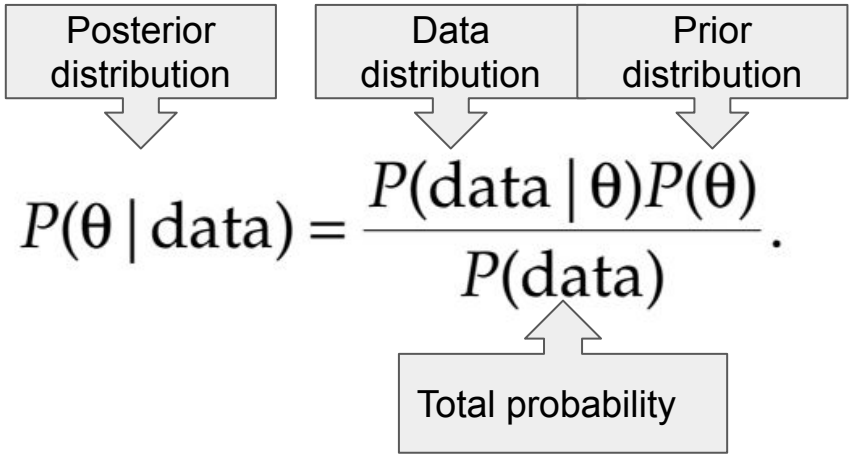
$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

How many ways to select 2 out of the 5 items

(without replacement & order doesn’t matter)?

Bayes Rule

The Bayes rule allows us to look at the alternative conditional probability.



$$P(\text{data}) = \sum_{k=1}^{c} P\left(\text{data} \mid \theta_k\right) P(\theta_k),$$

Bayes Example

EXAMPLE 2.10 AN INTRODUCTORY ILLUSTRATION From previous research, it is believed that three varieties of a species of oak (Quercus sp.) occur at equal frequencies on a landscape. Thus each variety has an equal probability, 1/3, of being randomly selected. Assume that variety one will always die when infected by a particular fungus, variety two will die with probability 0.5, and variety three will be unaffected by the infection. A tree is randomly chosen from the landscape, which has died from the fungal infection. What is the probability that it came from variety one?

Aho, Ken A.. Foundational and Applied Statistics for Biologists Using R (Page 39). Chapman and Hall/CRC. Kindle Edition.