Masters Bridge Program **Summer 2021**

Session 1: Basic concepts

Instructor: Bryan Liu *Contact: runjing_liu@berkeley.edu*

The materials for this workshop are adapted from Jim Pitman's undergraduate text on probability. A pdf can be found for free at

`https://link.springer.com/book/10.1007/978-1-4612-4374-8`

This first session comes from Chapter 1.

1.1 Probability spaces

The building blocks of probability theory are three ingredients:

- **An outcome space:** the set of possible outcomes, denoted Ω.
- **Events:** subsets of the outcome space, often denoted with a capital letter, e.g. $A \subseteq \Omega$.
- **Probability:** a function which maps events to a real number between 0 and 1. We use $\mathbb{P}$ to denote such a function, in which case $\mathbb{P}(A)$ represents the probability of event A.

Example 1.1 (Rolling two dice)**.** Two dice are rolled, and number on the top faces are recorded.

The *outcome space* are the pairs,

$$\Omega = \begin{Bmatrix} (1,1), & (1,2), & (1,3), & (1,4), & (1,5), & (1,6) \\ (2,1), & (2,2), & (2,3), & (2,4), & (2,5), & (2,6) \\ (3,1), & (3,2), & (3,3), & (3,4), & (3,5), & (3,6) \\ (4,1), & (4,2), & (4,3), & (4,4), & (4,5), & (4,6) \\ (5,1), & (5,2), & (5,3), & (5,4), & (5,5), & (5,6) \\ (6,1), & (6,2), & (6,3), & (6,4), & (6,5), & (6,6) \end{Bmatrix}.$$

Events are subsets of the outcome space. One such event might be "the sum of the two numbers is 5". This event is the subset

$$A = \big\{(1,4), \quad (2,3), \quad (3,2), \quad (4,1)\big\}$$

Assuming the die are fair, each pair in the event space is assigned to have probability $1/36$, and the probability of the event above is

$$\mathbb{P}(\text{the sum of the two numbers is 5}) = \frac{4}{36} = \frac{1}{9}$$

□

Example 1.2 (Particle decay)**.** We measure the time it takes for a radioactive particle to decay. The *outcome space* is $\Omega = \mathbb{R}^+$, the positive real line.

Events are subsets of the positive real line. For example, one event might be "it takes at least 3 hours for the particle to decay".

The probability of this event is given by

$$\mathbb{P}(\text{at least 3 hours for the particle to decay}) = \int_3^\infty \lambda \exp^{-\lambda t}\ dt,$$

where λ is a physical constant unique to the particle in study.

In general, an event A is of the form

$$A = \{t : t \geq a \text{ and } t \leq b\},$$

where t is the decay time, while a and b are real numbers. The probability of such events is

$$\mathbb{P}(A) = \int_a^b \lambda \exp^{-\lambda t}\ dt. \tag{1.1}$$

When the probability of decay times follow the formula in (1.1), we say that the decay time is *exponentially distributed*.

□

Example 1.3. (Stock prices) Let S_0 be the initial stock price, and S_1 be its price after one unit time. Consider the log ratio of stock prices,

$$\Delta \log S := \log \frac{S_1}{S_0}.$$

We model the log ratio as a random quantity. The outcome space is the real line. Events are subsets of the real line, for example, "the log ratio is greater than a but less than b", a and b real numbers. In math, an event A is the set

$$A = \{\Delta \log S \in [a, b]\}.$$

The probability of such events are modeled as

$$\mathbb{P}(A) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{1}{2\sigma^2}(x-\mu)^2}\ dx \tag{1.2}$$

Here, μ is the *exponential growth rate of the stock* (exponential because we are working with log-prices) and σ^2 is the *volatility* of the stock.

Random quantities with event probabilities given by (1.2) are called *normally distributed*. □

In the first two examples, the assigned probabilities are more or less well-motivated by our understanding of the laws of physics. In the last example, the assumption of a normal distribution is a modeling choice, which may or may not be accurate. An important job for the statistician is arguing that the modeling choice is appropriate, either with theoretical derivations or with empirical evidence.

1.2 Rules of probability

Probabilities must satisfy

- **non-negativity**: $\mathbb{P}(A) \geq 0$ for events A.
- **additivity**: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if A and B are disjoint.
- **total one**: $\mathbb{P}(\Omega) = 1$.

From these basic rules we can derive

- **The complement rule**: $\mathbb{P}(\text{not}A) = 1 - \mathbb{P}(A)$.
- **The difference rule**: if $A \subseteq B$, then $\mathbb{P}(B \text{ but not } A) = \mathbb{P}(B) - \mathbb{P}(A)$.
- **Inclusion-exclusion**: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

IN CLASS: proof by picture.

Next, we extend the inclusion-exclusion principle to derive Boole's inequality.

Exercise 1 (Boole's inequality)**.**

(a) Argue that for any events A, B,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

(b) Use induction to argue that for any set of events $B_1, ..., B_n$,

$$\mathbb{P}(\cup_{i=1}^n B_i) \leq \sum_{i=1}^n \mathbb{P}(\cup_{i=1}^n B_i) \tag{1.3}$$

Example 1.4 (Bonferroni's correction)**.** This simple inequality (1.3) is useful in multiple testing situations. Suppose we are running n trials (say to test the effect of n different drugs), and for each trial we set up a level-α hypothesis test. Let B_i be the event that we reject the i-th hypothesis. If our test is correctly set-up, then we must have $\mathbb{P}(B_i) = \alpha$ for all i, where the probability $\mathbb{P}$ is computed under the assumptions of the null hypothesis.

Notice however, that in general, we must have $\mathbb{P}(\cup_i B_i) \geq \alpha$, so the probability that we make at least one false rejection is greater than α.

To control the probability of making even one false rejection then, we design a test such that the probability of rejection for each test is instead α/n – that is, $\mathbb{P}(B_i) = \alpha/n$ for all i. In this case, Boole's inequality shows that

$$\mathbb{P}(\cup_{i=1}^n B_i) \leq \sum_{i=1}^n \mathbb{P}(\cup_{i=1}^n B_i) = \alpha.$$

This correction of α values by dividing by n is known as *Bonferroni's correction*. Notice that we did not need any assumptions on the nature of the events B_i; we only assumed we have correctly specified hypothesis tests, which makes this correction of general interest.

However, correcting an α-rejection level to an (α/n)-rejection level considerably decreases the power of the tests (i.e. it is harder to make any true rejections). When we know more about the nature of the events, e.g. their correlation structure, it is often possible to employ less stringent corrections.

□

1.3 Conditional probabilities

Given two events A and B, we can compute the probability that A occurs if we know that B occurs. This is known as a conditional probability, written $\mathbb{P}(A|B)$.

The general formula is given by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1.4}$$

IN CLASS: show by picture.

We can re-write (1.4) as

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B). \tag{1.5}$$

Events A and B are **independent** if $P(A|B) = P(A)$. For independent events, (1.5) becomes

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{1.6}$$

Exercise 2 (Sequential vs parallel components)**.** A system consists of two components C_1 and C_2. Suppose the failure probability of C_1 is 0.05, while the failure probability of C_2 is 0.02. The failure events are independent.

(a) Suppose the system is connected in series, so if one of the two component fails, the entire system fails. What is the probability that the system works?

(b) Suppose the system is connected in parallel, so as long as one of the component works, the system works. What is the probability that the system works?

Exercise 3 (Conditional probabilities)**.** Suppose there is a 40% chance of raining Monday. If it rains Monday, then the probability that it rains Tuesday is 75%. But if it does not Monday, then the probability that it rains on Tuesday falls to 15%. Calculate the following probabilities that on Monday and Tuesday,

(a) It rains at least once.

(b) It rains exactly once.

(c) It rains on Tuesday.

IN CLASS draw tree diagram. Using the above tree diagram, we can derive

The rule of average conditional probabilities:

$$\mathbb{P}(A) = \mathbb{P}(A|B_1)\mathbb{P}(B_1) + ... + \mathbb{P}(A|B_n)\mathbb{P}(B_n) \tag{1.7}$$

where $B_1, ..., B_n$ is a *partition* of Ω, meaning that the B_i are disjoint and that $\cup_i B_i = \Omega$.

Exercise 4 (Stratified sampling)**.** Three high schools have senior class of size 100, 400, and 500, respectively. Consider two schemes for selecting a student from among the three senior classes:

I Make a list of all 1000 seniors, and choose a student at random from this list.

II Pick one school at random, then pick a student at random from the senior class in that school.

Are these sampling schemes equivalent? Why or why not?

Consider a third scheme: where we pick school i with probability p_i (so $p_1 + p_2 + p_3 = 1$), and then pick a student at random from the senior class in that school. Find the probabilities which make this last scheme equivalent to scheme I.

1.4 Bayes rule

We can combine the conditional probability formula (1.4) and the rule of average conditional probabilities (1.7) to arrive at **Bayes rule**:

For a partition of all possible outcomes $B_1, ..., B_n$,

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A|B_1)\mathbb{P}(B_1) + ... + \mathbb{P}(A|B_n)\mathbb{P}(B_n)}. \tag{1.8}$$

Exercise 5 (Conditional probabilities continued)**.** Return to the conditional probabilities in Exercise 3. Suppose I was out of town Monday, so I don't know if it rained then. I come back on Tuesday,

and saw that it rained. Apply Bayes rule to compute the probability that it rained Monday. [1]

Exercise 6 (Identical or fraternal twins)**.** [2] A mother finds out via sonogram that she is going to have twin boys. She wants to know, what is the probability that her twins will be identical rather than fraternal? The doctor tells her that in the United States, one-third of twin births are identical, while two-thirds are fraternal. For identical twins, the siblings must be of the same sex; for fraternal births, each sibling has a 50/50 chance of being male or female, independently of each other. Apply Bayes rule to compute probability that the mother's twins will be fraternal.

Exercise 7 (Medical diagnosis)**.** A diagnostic test for a type of cancer has a *sensitivity* of 98% and a *specificity* of 90%. Here,

- **sensitivity** is the probability that a test returns a positive result if the individual has the disease.
- **specificity** is the probability that a test returns a negative result if the individual does not have the disease.

Suppose I walk into the doctor's office and get a test for this type of cancer. I test positive. Consider three scenarios:

(a) Going home, I Google that this cancer occurs in 0.2% of the adult population in the United States. Being a statistician, I use Bayes rule to compute the probability that I have this cancer. What is the probability?

(b) But wait. Upon talking with my doctor again, he tells me that in his experience, smokers are more likely than the general public to have this cancer. In fact, about 20% of smokers develop this cancer. My doctor is well-trained in probability, and she knows that I smoke. She uses Bayes rule to compute the probability that I have cancer. What is this probability?

(c) I go visit a geneticist who sequences my genome. He tells me that among people with similar genetic characteristics as mine, about 5% of people develop this cancer. The geneticist then uses Bayes rule to compute the probability that I have cancer. What is this probability?

Discussion

Bayes rule is a mathematical formula for turning observed data into inferences concerning unknown quantities. In the previous exercise, the observed data is that I have a positive diagnostic test. The unknown quantity that I would like to know is whether or not I have cancer. In story of the mother

[1] To a frequentist, this question would be nonsensical: Monday has already occurred, so either it rained or it did not rain; there is no sense in assigning a probability to this event. In the mind of a Bayesian however, the event "it rained on Monday" is unobserved – and therefore random. The probability assigned to this event quantifies the uncertainty in the occurrence of the unobserved event. The Bayesian will continue to view this event as random until more information is collected, for example, by asking a friend who was in town whether or not it rained on Monday.

[2] Example taken from Chapter 3 of `https://web.stanford.edu/~hastie/CASI/`.

with twins, the observed data is that her twins are both boys. The unknown quantity is whether the twins are identical or fraternal.

Crucially, the inferences made using Bayes rule depend on specifying a *prior probability*, that is, the background probability of the unknown event before seeing any data. In the twin example, the doctor can go into a hospital database of all births in the United States, and see what fraction of twins were fraternal, and what fraction were identical. In this case, these proportions were one-third and two-thirds, respectively.

However, as the medical diagnosis example tries to illustrate, this prior probability may not be straight-forward to specify. Which prior probability should I use, and which is correct? My Google search, my doctor, or my geneticist? The resulting conclusion about my cancer diagnosis depends what assumptions we make; and the assumptions we make depend on the available knowledge at hand.

More fundamentally, how can we interpret the probability, given that it changes based on the information at hand or based on expert opinion? The answer depends on how we define the *population of interest*.[3] Let us take scenario (c) in Exercise 7. Here, the *implicit theoretical population is the set of individuals with similar genetic information as me*. Suppose I sample 1000 individuals from this theoretical population, and test each individual. Then each individual would fall in one of four categories: (test positive, diseased); (test positive, not diseased); (test negative, diseased); (test negative, not diseased). Based on the probability model described in Exercise 7, the counts of individuals falling in each catagories would be approximately .

	Test positive	Test negative	Total
Diseased	49	1	50
Not diseased	95	855	950
Total	144	856	1000

If I am a member of this sample of 1000 individuals from the theoretical population, and I know that I test positive, then this places me in the first column of the table above. Examining the first column of the table, I see that $49/144$ of these positive individuals have the disease. I conclude then, that I have a $49/144 = .34$ myself of having this disease.

This demonstrates the frequency interpretation probability: if I sample a set of individuals from my theoretical population, then I would expect about 34% of these individuals who test positive will be actually diseased. Notice that this probability depends on the prevalence of this disease in the population – i.e. how I define my theoretical population.

Bayes rule is a provable mathematical fact about probabities and events. Going from mathematical principles to real data applications however, requires care in thinking about what probabilties, and our resulting inferences actually tell us about the worl.

[3] Defining the population is the first step in applied statistical analysis, argues Bin Yu in `https://www.stat.berkeley.edu/~binyu/ps/papers2018/AI+Stat18.pdf`