Stat 110

Notes by Edward Chen

October 2022

Preface

These are my notes to Harvard's Stat 110 class, taught by Joseph Blitzstein and Jessica Hwang. The class covers all the basics of probability–counting principles, probabilistic events, random variables, distributions, conditional probability, expectation, and Bayesian inference.

Compared to other notes, this should not be used as a whole subsititue for the class, rather as an indication of the more complex topics covered. Put harsher, I will likely spend less time note taking on things I already know and more on things I'm confused out. For clarity's sake, I am pretty comfortable with probability from my competition math background.

Lecture videos are freely available at

1 Chapter 1

Probability gives us a logical framework to view and anlyze uncertaintiy. It is the foundation and language for statistic ass well as the basis for topics such as Statistics, Physics, Biology, and Computer science.

A **sample space** S is the set of all theoretically possible outcomes.

Let D be the set of all coin flips with at least two consecutive heads. The sameple space, expressed as a set would be:

$$D = \cup_{j=1}^{9}(A_j \cap A_{j+1})$$

Some more set notation:
sameple space: S
s is a possible outcome: $s \in S$
A is an event: $A17S$
A implies B:

A and B are mutually exclusive: $A \cap B = \emptyset$

Example 1.4.10: Birthday Problem

Theorem 1.4.15 (Binomial coefficient formula). For $k \leq n$, we have:

$$\binom{n}{k} = \frac{n(n-1)...(n-k+1)}{k!} = \frac{n!}{(n-k)!k!}$$

Some proof methods covered in the chapter include complementary counting, stars and bars, and story proofs(1.5).

There are a lot of formulas with binomial coefficients:

$$\binom{n}{k} = \binom{n-k}{k}$$

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

$$n\binom{n-1}{k-1} = k\binom{n}{k}$$

$$\text{Vandermonde's} : \binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$$

Example 1.5.4(Partnerships). Let's prove

$$\frac{(2n)!}{2^n \cdot n!} = (2n-1)(2n-3)...3 \cdot 1$$

Story proof:: We will show that both sides count the number of ways to break $2n$ people into n partnerships. Take $2n$ people, and give them ID numbers from 1 to $2n$. We can form partnerships by lining up the people in some order and then saying the first two are a pair, the next two are a pair, etc. This overcounts by a factor of $n! \cdot 2^n$ since the order of pairs doesn't matter, nor does the order within each pair. Alternaitvely, count the number of possibilities by noting that there are $2n-1$ choices for the parnter of person 1, then $2n-3$ choices for person 2, and so on.

Definition 1.6.1 (General definition of probability). A *probability space* consists of a sample space S and a *probability function* P and returns a real number between 0 and 1, $P(A)$, where A is the event it takes in.

Unlike the naive definition, here we can have events with different probabilities.

The *frequentist* view of probability is that it represents a long-run frequency over a large number of repetitions of an experiment. The *Bayesian view* is that it represents a degree of belief about the event in question, so we can assign probabilities to hypotheses.

Inclusion-exclusion example: With a triple venn diagram, we can write:

$$P(A \cup B \cup C) = P(A)+P(B)+P(C)-P(A \cap B)-P(A \cap C)-P(B \cap C)+P(A \cup B \cup C)$$

Generally, we can write:

Example 1.6.4 (de Montmort's matching problem). Consider a shuffled deck of n cards, from 1 to n. You flip each one over one by one. What is the probability the ith index card you turn over has value i?

With PIE, we get

$$P(\cup_{i=1}^{n} A_i) = \frac{n}{n} - \frac{\binom{n}{2}}{n(n-1)} + \frac{\binom{n}{3}}{n(n-1)(n-2)} - ... + (-1)^{n+1} \cdot \frac{1}{n!}$$

$$= 1 - \frac{1}{2!} + \frac{1}{3!} - ... + (-1)^{n+1} \cdot \frac{1}{n!}$$

With large n, this approaches the Taylor series for $\frac{1}{e}$:

$$e^{-1} = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + ...$$

R: Please read section 1.8 for an introduction to R. R allows us to simulate and deal with large sets of data.

2 Chapter 2: Conditional Probability

Conditional Probability explores how we should update our probability when we receive new information.

Definition 2.2.1. If A and B are events with $P(B) > 0$, then conditional probability can be expressed as:

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

Example 2.2.5(Two Children): Martin Gardner posed the following puzzle:
Mr. Jones has two children. The older hild is a girl. What is the probability that both children are girls?
Mr. Smith has two children. At least one of them is aboy. What is the probability that both children are boys?

The intuition is that they should both be $\frac{1}{2}$, but the respesctive probabilities are actually $\frac{1}{2}$ and $\frac{1}{3}$.

Theorem 2.23 (Bayes' rule).

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

Definition 2.3.4 (Odds). The *odds* of an event A are

$$\text{odds(A)} = P(A)/P(A^c)$$

. Ex.: if $P(A) = 2/3$, we say the odds in favor of A are 2 to 1.

Theorem 2.3.5 (Odds form of Bayes' rule)

Example 2.4.5 (Unanimous agreement). The artaicle "Why too much evidence can be a bad thing" by Lisa Zyga says:

Under ancient Jewish law, if a suspect on trial was unanimously found guilty by all judges, then the suspe

This is because a systemic error occurs independet of the conviction, so there may be an unseen bias at play here.

Independence

Events A and B are *independent* if $P(A \cap B) = P(A)P(B)$.

For infinitely many events, we say that they are independent if every finite subset of the events is independent.

Definition 2.5.7 (Conditional independence). Events A and B are said to be *conditionally independent* given E if $P(A \cap B|E) = P(A|E)P(B|E)$.

> It is easy to make terrible blunders stemming from confusing independence and conditional independence. Two events can be conditionally independence given E, but not independent given E^c. Two events can be conditionally independent given E, but not independent. Two events can be independent, but not conditionally independent given E. Great care is needed

todo: 65-67

An important property of Bayes' rule is that is is *coherent*: for the same updates, no matter the update size, the final state should be the same.

todo: 68-75

Example 2.7.1 (Monty Hall):

Here the strategy is to use conditioning as a problem-solving tool. If we knew where the car actually is, it makes the probabilities a lot easier to deal with:

$P(\text{get car}) = P(\text{get car}|C_1) \cdot \frac{1}{3} + P(\text{get car}|C_2) \cdot \frac{1}{3} + P(\text{get car}|C_3) \cdot \frac{1}{3} = \frac{2}{3}$

That is why it is always better to switch because now you have a 2/3 success strategy.

Example 2.7.3 (Gambler's ruin):

The gambler's ruin problem is a problem where we have two players, one with n dollars and the other with n dollars. They play a game where they bet 1 dollar each. There is probability of p that the first player wins and gets 2 dollars, and otherwise the second player gets the 2 dollars. The game ends when one of the players has 0 dollars. What is the probability that the first player wins?

This can be solved by conditioning on the first step and is just a general version of the random walk problem. We have **absorbing state** at 0 and $2n$ and we have the relationship:

$$p_i = p \cdot p_{i+1} + (1-p) \cdot p_{i-1}$$

To solve, we start by guessing $p_i = x^i$. Then we get:

$$x^i = p \cdot x^{i+1} + (1-p) \cdot x^{i-1}$$

$$px^2 - x + (1-p) = 0$$

$$x = \frac{1 \pm \sqrt{1 - 4p(1-p)}}{2p}$$

3 Chapter 3: Random Variables

A **random variable** is more than "a variable that takes on random values." A better definition is that a random variable is a function that maps a sample

space to the real numbers.

Definition 3.1.2 (Bernoulli distribution). The **Bernoulli distribution** is the probability distribution of a random variable X that takes on the values 0 and 1 with probabilities p and $1-p$, respectively. We write X $Bern(n,p)$ as X is distributed as a Bernoulli function with n independent Bernoulli random variables that have probability p of being 1.

Interestingly: the binomial distribution is the same distribution of the sum of n independent Bernoulli random variables.

Todo: proposition 9.2

Definition 3.2.2 (Probability mass function). The **probability mass function** of a discrete random variable X is a function f_X that assigns to each value x of X the probability that X takes on the value x.

Definition 3.2.3 (Cumulative distribution function). The **cumulative distribution function** of a discrete random variable X is a function F_X that assigns to each value x of X the probability that X takes on a value less than or equal to x.

Hypergeometric distribution is the probability distribution of the number of successes in a sample of size n drawn with replacement from a finite population of size N containing exactly K successes.

Theorem 3.4.2 (Hypergeometric PMF): The probability mass function of a hypergeometric random variable X is given by:

Hypergeometric distributions are very similar to binomial distrbutions in that they take on n Bernoulli trials, but note that hypergeometric distributions are dependent as we are drawing without replacement.

img pg 130

Exercise 32

In Evan's history class, 10 out of the 100 key terms will be randomly selected to appear on the quiz. Evan must then choose 7 of them to define. What is the probability distrbution of the number of terms, X, Evan will know on the quiz?

We can use the hypergeometric distribution here. We have $N = 100$, $K = 10$, $n = 7$, and x is the number of terms Evan knows.

4 Chapter 4: Expectation

Here is a fundamental derivation of binomial expectation for $X \sim Binom(n,p)$:

$$EV(X) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = n\sum_{k=0}^{} \binom{n-1}{k-1}p^k q^{n-k} = np\sum_{j=0}^{} \binom{n-1}{k-1}p^j q^{n-1-j} = np$$

Geometric and Negative Binomial Distributions

Definition 4.1.1 (Geometric distribution). The **geometric distribution** is the probability distribution of the number of Bernoulli trials needed to get the first success.

Definition 4.1.2 (Negative binomial distribution). The **negative binomial distribution** is the probability distribution of the number of Bernoulli trials needed to get r successes.

Negative binomial expectation is just given by: $E(x) = E(x_1) + E(x_2) + ... + E(x_r) = r \cdot \frac{q}{p}$.

Poission Distribution

Poisson Distribution is a common distribution used for modeling the number of "successes" given a low-probability, high frequency event. The PMF is given a parameter λ by:

$$P(x = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

It is a valid PMF as $\sum_{k=0}^{\infty} P(x = k) = e^{-\lambda} \cdot e^{\lambda} = 1$. With the same derivation as above, we can show that the expected value of a Poisson distribution is λ.

Connection with binomial distribution: actually converges pretty well for large n, small p_j, but here events can also be "weakly dependent," and have different p_j's.

Poisson can be used for finding the approximate probability function of the birthday problem, where lets say we have n people and want to find probability of there is exactly k triplets of people with the same birthday. We know that $EV = \lambda = \binom{n}{3}(\frac{1}{365})^2$ so we can create a Poisson distrbution. Notice there is some weak dependence here.

5 Continous

We've learned about DPFs, but now we want can say similar things about **probability density functions**, or PDFs. In fact, to they are easily using

the Fundamental Theorem of Calculus in that to get a PDF is just to take the derivative of DPF's.

Variance

Definition 4.2.1 (Variance). The **variance** of a random variable X is defined as:

$$Var(X) = E[(X - EX)^2] = Ex^2 - (Ex)^2$$

Something more intereprable is standard deviation: $\sigma = \sqrt{Var(X)}$.

But how do we actually find Ex^2? We can use the Fundamental Theorem of Calculus to get:

$$Ex^2 = \int_{-\infty}^{\infty} x^2 f_X(x) dx$$

Generally, this is called the Law of the Unconscious Statistician(LOTUS):

$$E(g(x)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Uniform and Universality of the Uniform

Given being able to produce random variables from a uniform distribution, we can produce any other distribution.

The proof follows that if we let F be the CDF of X, then $F^{-1}(u) \sim F$ if $U \sim$ Unif(0,1). The intuition is that we plug a random variable into its own CDF.

Normal Distribution

The normal distribution is a very important distribution that is used to model many things. It is defined by two parameters, μ and σ^2. The PDF is given by:

$$f_X(x) = ce^{-\frac{z^2}{2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Central Limit Theorem: the sum of independent random variables tends to a normal distribution.

Main intution here is that we can actually integrate the definite integral of $ce^{-\frac{z^2}{2}}$ using polar coordinates to get $\frac{1}{\sqrt{2\pi}}$.