Statistics Lecture Notes

Simon Xiang

Lecture notes for Statistics (M 358 K). These notes were taken live in class (and so they may contain many errors). Source files: https://git.simonxiang.xyz/math_notes/files.html

Contents

1 Probability "review"

We "review" (learn) probability. INFORMALLY:

Definition 1.1. The "**probability**" of an outcome is the proportion of times it would occur if we observed the random process an infinite amount of times.

This fact that probabilities stabilize is called the **Law of Large Numbers**.

Definition 1.2. Two outcomes are "**disjoint**" or "**mutually exclusive**" if they cannot both happen.

For example, rolling a 1 and a 2 are disjoint outcomes, while rolling a 1 and rolling an odd number are not. For disjoint outcomes we can add to find their probabilities. We express this as follows:

Addition Rule of Disjoint Outcomes. If A_1, A_2 represent disjoint outcomes, then the probability that one of them occurs is given by

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

This naturally generalizes to the fact that if we have k disjoint outcomes $A_1, \cdots, A_k$, then

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

Often times data scientists do not work with individual outcomes but rather **events**, or sets of outcomes. The addition rule applies as well. What if A, B are not disjoint? Consider a deck of cards, then $P(\text{diamond or face card}) = 13/52 + 12/52 - 3/52$, since we double count diamond face cards. This leads to the general addition rule.

General Addition Rule. If A and B are events, then

$$P(A \cap B) = P(A) + P(B) - P(A \cap B).$$

Definition 1.3. A "**probability distribution** is a table of all disjoint outcomes and their associated probabilities. A probability distribution must satisfy the following:

(1) The outcomes listed must be disjoint.

(2) Each probability must be in between 0 and 1.

(3) The probabilities must total 1.

The set of all possible outcomes is called a **sample space** S. For rolling a die, $S = \{1, 2, 3, 4, 5, 6\}$. Let $D = \{2, 3\}$ represent the event where the outcome of rolling a die is 2 or 3. Then $D^c = \{1, 4, 5, 6\} = S \setminus D^c$. We have $A \cup A^c = S$, $A \cap A^c = \emptyset$, and $P(A \cup A^c) = 1$ by definition. This implies that $P(A) = 1 - P(A^c)$.

Definition 1.4. Two *processes* are "**independent**" if knowing the outcome of one provides no useful information about the other. We could also view this as saying that the realization of one does not affect the probability distribution of the other.

Example 1.1. Rolling two die is independent. Rolling one provides us no information about the other.

Multiplication Rule for Independent Processes. If A and B represent events from two different and independent processes, then

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

This generalizes to saying that

$$P\left(\bigcup_{1\le i\le k} A_k\right) = \prod_{1\le i\le k} P(A_i).$$

1.1 Conditional probability

Data set: ML algorithm classifies fashion or not, based off of the "truth". In this case the correctness of the prediction depends on the truth variable. We can explore these possibilities with a contingency table.

	`fashion`	`not`	Total
`pred-fashion`	197	22	219
`pred-not`	112	1491	1603
Total	309	1513	1822

Figure 1: Contigency table for ML fashion.

Let `mach-learn` denote the ML classifier (includes `pred-fashion` and `pred-not`), and let `truth` denote `fashion` or `not`. What if we want to know the chance the ML classifier correctly identifies something being about fashion? Then

$$P(\texttt{mach-learn is pred-fashion given truth is fashion}) = \frac{197}{309} = 0.638$$

On the same vein, what if we sample a photo from the data set and learn that the ML algorithm predicted incorrectly that a photo was not about fashion? Then

$$P(\texttt{truth is fashion given mach-learn is pred-not}) = \frac{112}{1603} = 0.070$$

The probabilities based on a single variable without any regard to other variables are **marginal probabilities**, in this case the row and column totals for each separate variable. For example, any probability based solely on `mach-learn` is a marginal probability.

$$P(\texttt{mach-learn is pred-fashion}) = \frac{219}{1822} = 0.12$$

A probability of outcomes for two or more variables is a **joint probability**.

$$P(\texttt{mach-learn is pred-fashion and truth is fashion}) = \frac{192}{1822} = 0.11$$

It is common to substitute a comma for "and". We use **table proportions** to summarize joint probabilities.

	`truth: fashion`	`truth: not`	Total
`mach-learn: pred-fashion`	0.1081	0.0121	0.1202
`mach-learn: pred-not`	0.0615	0.8183	0.8798

Figure 2: Probability table of the fashion data set.

From here we can make a probability distribution of the conditional probabilities. Conditional probability is the probability of an outcome given some condition to be true. For example, the probability that `truth` is `fashion` given `mach-learn` is `pred-fashion` is $\frac{197}{219} = 0.9$, an example of conditional probability (compute `truth` under the condition that `mach-learn` is `pred-fashion`). We can write "given" with the vertical line |.

Definition 1.5. The **conditional probability** of outcome A given condition B is given by

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

We saw a multiplication rule for independent processes. Here we see one for events that may not be independent.

General Multiplication Rule. If A and B represent two outcomes or events, then

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

It is useful to think of A as the outcome of interest and B as the condition.

Use tree diagrams. The end leaves give conditional probabilities, and multiplying gives unions of probabilities.

Sum of Conditional Probabilities. Let $A_1, \cdots, A_k$ represent all the disjoint outcomes for a variable or process. If B is an event, we have

$$P(A_1 \mid B) + \cdots + P(A_k \mid B) = 1.$$

We also have that when an event and its complement are conditioned on the same information, the rule for complements holds:

$$P(A \mid B) = 1 - P(A^c \mid B).$$

We are often given a conditional probability of the form $P(A \mid B)$ but would really like to know $P(B \mid A)$. Breast cancer setup; 0.35% of patients have BC, in 11% the text gives a false negative, in 7% gives a false positive. We want to find $P(BC \mid +) = P(BC \cap +)/P(+)$. We have $P(BC \mid +) = P(+ \mid BC)P(BC)$ by the generalized multiplication rule, which is equal to $0.89 * 0.0035 = 0.00312$ (multiplying leaves on a tree). We also have

$$P(+) = P(+ \cap BC) + P(+ \cap !BC) = P(BC)P(+ \mid BC) + P(!BC) \mid P(+ \mid !BC) = 0.0035 \cdot 0.89 + 0.9965 \cdot 0.07 = 0.07288.$$

This is just summing up leaves again. We conclude that $P(BC \mid +) = P(BC \cap +)/P(+) = \frac{0.00312}{0.07288} \approx 0.0428$. Note that we broke down the probabilities $P(BC \cap +) = P(+ \mid BC)P(BC)$ and $P(+) = P(+ \cap !BC) + P(+ \cap BC) = P(+ \mid !BC)P(BC) + P(+ \mid BC)P(BC)$, each into products of conditional and marginal probabilities. Substitute the resulting probability expressions into the numerator and denominator of the original conditional probability to get an application of Bayes' Theorem.

$$P(BC \mid +) = \frac{P(+ \mid BC)P(BC)}{P(+ \mid !BC)P(!BC) + P(+ \mid BC)P(BC)}$$

Bayes' Theorem. *We have*

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

Alternatively, we can write

$$P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_k)P(A_k)},$$

where $A_2, A_3, \cdots, A_k$ represent all other possible outcomes of the first variable by the sum of conditional probabilities.

Basically Bayes' Theorem is how you calculate probabilites for tree diagrams. To apply Bayes' Theorem there are two steps:

(1) Identify the marginal probabilities of each possible outcome of the first varaible: $P(A_1), P(A_2), \cdots, P(A_k)$.
(2) Then identify the probability of the outcome B, conditioned on each possible scenario for the first variable: $P(B \mid A_1), P(B \mid A_2), \cdots, P(B \mid A_k)$.

Applying Bayes' Theorem to a scenario with three outcomes, if we calculate the conditioned probabilities of the first two, we can infer the probability of the third one conditioned on the same statement. This practice of updating beliefs using Bayes' Theorem is the foundation of *Bayesian statistics*.

1.2 Sampling from a small population

Sometimes our sample size is large enough or the population is small enough (sample is 10% of a population) to sample without replacement. Such a notable fraction changes things significantly.

Example 1.2. If there are 15 people in a class and the professor picks without replacement, then if she picks three times your chance of getting picked is $\frac{14}{15} \cdot \frac{13}{14} \cdot \frac{12}{13} = 0.8$. If she picks with replacement (no regard to original selection) then by the multiplication rule for independent processes your probability of not getting picked in three questions is $\frac{14}{15} \cdot \frac{14}{15} \cdot \frac{14}{15} = 0.813$.

Sampling **without replacement** means there is no more independence between observations. In the example, the probability for not being picked for the second question was conditioned on the event that you were not picked for the first question.

For a small sample size this is a drastic effect (30 tickets, 7 prizes). However if we change this to 300 tickets then the results are nearly identical. When the sample size is under 10% observations are nearly independent even when sampling without replacement.

1.3 Random variables

Definition 1.6. A **random variable** is a random process with a numerical outcome. Outcomes of X are labelled with a corresponding lowercase letter x and subscripts.

Definition 1.7. If X takes outcomes $x_1, \cdots, x_k$ with probabilities $P(X = x_1), \cdots, P(X = x_k)$, the **expected value** of X, denoted $E(X)$ or μ, is the sum of each outcome multiplied by its corresponding probability:

$$E(X) = \sum_{i=1} k x_i P(X = x_i).$$

Definition 1.8. If X takes outcomes $x_1, \cdots, x_k$ with probabilities $P(X = x_1), \cdots, P(X = x_k)$ and expected value $\mu = E(X)$, then the **variance** of X, denoted by $\text{Var}(X)$ or the symbol σ^2, is defined as follows:

$$\sigma^2 = \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j)$$

The **standard deviation** is the square root of variance, defined by $\sigma := \sqrt{\sigma^2}$.

Sometimes it's useful to think of processes as modelled by several random variables. Say John commutes to work 5 days a week, and each commute is an average of 18 minutes. Then his expected commute weekly is given by

$$E(W) = E\left(\sum_{i=1}^{5} X_i\right) = \sum_{i=1}^{5} E(X_i) = 5 \cdot 18 = 90.$$

This demonstrates that expectation is linear.

Linearity of Expectation. Let $\sum a_i X_i$ be a linear combination of random variables. Then

$$E\left(\sum a_i x_i\right) = \sum a_i E(X_i).$$

Variance of linear combinations. For $\sum a_i X_i$ a linear combination of random variables, we have

$$\text{Var}\left(\sum a_i X_i\right) = \sum a_i^2 \text{Var}(X_i).$$

todo:normal (triathlon)

1.4 Bernoulli and geometric distribution

Say a health company found 70% of the people they insure yearly stay below their deductible. Each person is called a **trial**, with **probability of success** $p = 0.7$. When an individual trial only has two outcomes (`success` or `failure`), it is called a **Bernoulli random variable**.

Definition 1.9. If X is a random variable that takes value 1 with probability of success p and 0 with probability 1-p, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p, \quad \sigma = \sqrt{p(1-p)}.$$

The geometric distribution describes a waiting time until a success for independent and identically distributed Bernoulli random variables.

Definition 1.10. If the probability of success in one trial is p and the probability of failure is $1-p$, then the probability of finding the first success in the n^{th} trial is given by

$$(1-p)^{n-1}p.$$

Furthermore, we have

$$\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}, \quad \sigma = \sqrt{\frac{1-p}{p^2}}.$$

1.5 Binomial distribution

The binomial distribution describes the number of successes in a fixed number of trials. In general, this is given by

$$\#\text{ of scenarios} \times P(\text{single scenario}).$$

Given k successes and $n-k$ failures, then we have $P(\text{single scenario}) = p^k(1-p)^{n-k}$. The formula for choosing k successes in n trials is $\binom{n}{k}$.

Definition 1.11. Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

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

The mean, variance, and standard deviation of the number of observed successes is are

$$\mu = np, \quad \sigma^2 = np(1-p), \quad \sigma = \sqrt{np(1-p)}.$$

To check whether something is binomial, check the following:

(1) The trials are independent.
(2) The number of trials n is fixed.
(3) Each trial outcome can be classified as a success or a failure.
(4) The probability of a success p is the same for each trial.

2 Introduction

todo:REMOVE THIS FROM GIT We're gonna use R and RMarkdown to do stuff this semester. The goal of this class is to get a foundation in the types of things you do in routine data analytics, like data visualization, data wrangling, etc. We need to be able to manipulate complex data sets and prepare reports, stuff like that. We will not cover inferential statistics, like p-values, confidence intervals, etc. No theoretical justification for machine learning boo.

We will follow the text reasonably closely, it's a good idea to read it before class. Class will consist of half lectures and half worksheets/homework. **Homework is due Sunday 11:59 PM**, and there are **labs closing Tuesday 11:59 PM**. Labs are basically quizzes and homeworks are filling out worksheets. HTML documents are accepted as submissions but try to default to PDFs. Grading is on a point scale: 10 points for a homework, 5 points for a lab, and each of the two projects is worth 150 points. Two lowest scoring homeworks are dropped. There are 465 points total. You need 432 points for an A so try not to lose more than **33 points** total.

Try not to be late but you can request extensions for traveling for athletics, religious, or serious indicidents/emergency. Regrades are possible.

2.1 R stuff

Do you know how to use RMarkdown? Make a markdown file with the extension `.rmd` and knit it in RStudio to compile. Use `:RMarkdown` with the plugins to do it in vim. You can attach things like tweets and Python snippets in rmd files.

3 More R stuff

You can use the terminal/console or a script to interact with R. You can also run R snippets within rmd files. Another way to interact is through RStudio. R can be used as a glorified calculator (how I use it).

See `testing.Rmd` for use cases and basic syntax. Define variables with arrow and you can see them in the "environment" area in RStudio.

4 Probability review

Do the homeworks (homework 0 is a syllabus quiz, homeworks 1 and 2 are probability reviews). We will not review sample spaces, conditional probability, etc; look at the videos on the course website.

Definition 4.1 (Cumulative distribution)**.** For any random variable X, the **cumulative distribution function** (cdf) of X is a function $F_X : \mathbb{R} \to [0,1]$, where $F_X(x) = \mathbb{P}[X \leq x]$ for all $x \in \mathbb{R}$.

todo:valid cdf figure. The cumulative distribution function gives us complete information about the distribution of a random variable. This is a nifty way to encode things in one function/one object. How do we know that the cumulative distribution function is well defined?

Question. What is $\lim_{x\to-\infty} F_X(x)$? This is zero because as we let x approach $-\infty$, the less room our probability has to land. OTOH what is $\lim_{x\to+\infty} F_X(x)$? This is one because as we let x approach $+\infty$, the more room our probability has to land.

Note. The cumulative distribution function is non-decreasing (monotone increasing). Any function with these properties that is right continuous is the cumulative distribution function of some random variable.

Question. What if your cdf is a step function (piecewise flat)? Then your random variable is discrete, and can take countably many values.

Most of the things in the graph of a cdf are useless. The only useful things are where the jumps happen and how big the jumps are. When there is a jump those are the values that the probability can take, and the size is such probability. Taking this data and condensing it gives us the **probability mass function** (pmf). It is usually more convenient to express the distribution of a discrete random variable using this probability (mass) function.

In general, the **support** $\text{supp}(X)$ of a random variable X is (vaguely) the set of all the values it can take. In the discrete case, the support is where the jumps in the cdf happen.

Definition 4.2 (Probability mass function)**.** For these points $x \in \text{supp}(X)$, the pmf is defined as $p_X(x) = \mathbb{P}[X = x]$, which is the size of the jump, or $F_X(x) - F_X(x-)$ (left limit).

What is the simplest random variable one can consider (non-deterministic)? This is the Bernoulli trial, or the coin toss. If we know the exact outcome of a trial then probability is not interesting, we say it's deterministic and the outcomes are degenerate.

Definition 4.3 (Bernoulli distribution)**.** Our first non-trivial example is a Bernoulli trial. There are only two possible outcomes, more precisely, the support of an X with the **Bernoulli distribution** is $\{0, 1\}$. We usually interpret "1" as "success" and "0" as "failure". We denote the probability of success in a single Bernoulli trial by p[1].

Notation. We write $X \sim \text{Bernoulli}(p)$ for

$$X \sim \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

We can also say that $p_X(1) = p, p_X(0) = 1 - p$. Drawing the cdf, we know for sure this is a step function since there are only two points in the support; a line at 0 from $(-\infty, 0)$. Then there is a line at $1 - p = q$ from $[0, 1)$, and finally a line at 1 from $[1, \infty)$.

Definition 4.4 (Binomial distribution)**.** This models the number of successes in a set of *independent* identically distributed (set of probabilities is the same) Bernoulli trials. Denote the probability of success in a single trial as p, and n denote the number of trials. If Y is **binomial**, then we write $Y \sim \text{Binomial}(n, p)$. We have $\text{supp}(Y) = \{0, 1, \cdots, n\}$ (all successes leads to n), and the pmf of Y is given by $p_Y(k) = \binom{n}{k} p^k (1-p)^{n-k}$ (success is p^k, failure is $(1-p)^{n-k}$, choice is $\binom{n}{k}$).

5 Simulations of random variables

True randomness is difficult, computers use different ways to "simulate" randomness enough to pass statistical randomness tests. We can use RNG that comes from physical phenomena (cursor movements, typing, entropy on linux). This stuff is not relevant to the class but one can read more on the links on the website.

5.1 Discrete uniform

Let X be a random variable, where

$$x \sim \begin{cases} x_1 & \text{with probability } \frac{1}{n} \\ x_2 & \text{with probability } \frac{1}{n} \\ \vdots & \text{with probability } \frac{1}{n} \\ x_n & \text{with probability } \frac{1}{n}. \end{cases}$$

[1] Representing *p*roportion of successes.

todo:RMD snippets, consolidate notes Humans are not good at being random number generators.