CME 106: Review Probability theory

Sven Schmit*

April 3, 2015

1 Overview

In the first half of the course, we covered topics from probability theory. The difference between statistics and probability theory is the following: In probability theory, we know everything about the underlying process that generates random variables. We then try to infer characteristics of these random variables. In statistics, however, we do not know the underlying process at all. To the contrary; we observe the outcome of random variables, and then try to infer the underlying process.

There are a few exercises littered throughout this review. They should be straightforward and help solidify your understanding of the material. Putting in some effort to try them will be well worth it, and feel free to reach out to me if you get stuck nonetheless.

2 Probability

While the term probability sounds very intuitive, we need some mathematical rigor to define it properly. In particular, the underlying structure, that we often shove under the rug, is given by the following three

Sample space The sample space $\mathcal{S}$ contains all possible outcomes at the end of our experiment.

Events An event A is a subset of the sample space: $A \subseteq \mathcal{S}$. This can be considered as a set of possible outcomes. We write $\mathcal{F}$ for the set of all possible events.

Probability measure A probability measure $\mathbb{P}$ assigns probability to events, hence it is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$. It denotes the frequency of an event occurring. Any probability measure must satisfy the following three axioms

1. $\mathbb{P}(A) \geq 0$ for all events A in the sample space S,
2. $\mathbb{P}(\mathcal{S}) = 1$,
3. For any two events $A, B \subseteq \mathcal{S}$ that are mutually exclusive, hence $A \cap B = \emptyset$, we have

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

We call the set $(\mathcal{S}, \mathcal{F}, \mathbb{P})$ a probability space. We note that the above are all fixed, there is no randomness involved, yet!

Exercise: Consider throwing two fair coins. Write down the sample space, the event that both coins come up the same side (in terms of elements of the sample space), and the probability of that event.

*Please let me know if you have any comments or spot typos: schmit@stanford.edu.

Exercise: Consider a probability space $(\mathcal{S}, \mathcal{F}, \mathbb{P})$ and two events $A, B \in \mathcal{F}$. Let $C = A \cap B$. Write $\mathbb{P}(A \cup B)$ in terms of $\mathbb{P}(A), \mathbb{P}(B)$, and $\mathbb{P}(C)$. Hint: define mutually exclusive events.

2.1 Independence of events

We say that two events A and B are independent if and only if

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

If two events are independent, the occurrence of one of the events has no impact on the occurrence of the other event. If I throw a die twice, the first and second throw are independent. However, the events that both throws come up heads and that both throws come up tails are not independent.

Also note that two independent events are never mutually exclusive: suppose A and B are mutually exclusive, then once we know event A has happened, we also know B cannot have happened!

3 Counting arrangements

We consider the case where there are finitely many outcomes (think: discrete case), and probability spaces where all elements of $\mathcal{S}$ are equally likely to occur. In this case, we can find the probability of event A by counting the number of elements in A, denoted by $|A|$, and divide that by the total number of events in $\mathcal{S}$, defined by $|\mathcal{S}|$. Hence,

$$\mathbb{P}(A) = \frac{|A|}{|\mathcal{S}|}.$$

Now suppose there are n objects. In how many ways can we arrange these n objects? Well, the first object can go in n places, then there are $n-1$ places left for the second object, etc. Hence, there are $n!$ possible arrangements.

3.1 Permutations

A permutation gives the number of ways we can select r (distinct) objects out of n. Since we are only interested in r objects, this can be calculated as follows: We have n choices for the first object, $n-1$ choices for the second object, up to $n-r+1$ choices for the rth object. Hence, the total number of arrangements is given by

$$n \times (n-1) \times \ldots (n-r+1) = \frac{n!}{(n-r)!} = P(n, r).$$

We can thus also view this as an arrangement of n objects where we are only interested in the ordering of the first r.

An example of the above is given by the following problem: How many ways are there to select a president, a secretary and a member for a student association out of 10 students?

3.2 Combinations

In the above case, we were interested in the arrangement of r (indistinguishable) objects out of n. We revisit the above example problem and now consider the following: How many ways are there to select 3 students to be part of a student association out of 10 students? In this case, we do not care which of the three is the president, the secretary and the member. In general, we note that there

are $r!$ arrangements of r objects, and hence $P(n,r)$ over counts every selection by $r!$ times. Hence, to adjust for this, the number of ways to take r out of n objects without regard to order is

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

3.3 Conclusion

Permutations and combinations are almost the same thing: It's all about what kind of arrangements are we interested in. If we care about the full arrangement, there are $n!$ ways. If we are only interested in the first r items ('select'), then there are $\frac{n!}{(n-r)!}$ ways. If furthermore, we also do not care about the arrangement of the first r items, only whether an item is part of the first or or not, there are $\frac{n!}{r!(n-r)!}$ ways.

4 Random variables

A random variable X is a function that takes elements of the sample space to the real line. More formally, $X : \mathcal{S} \to \mathbb{R}$.

Exercise: Consider the example with two fair coin flips again. Let X denote the number of heads. Write down $X(w)$ for all $w \in \mathcal{S}$.

The defining function for a random variable is the (cumulative) **distribution function**, usually written F_X. F_X is defined by

$$F_X(t) = \mathbb{P}(X \leq t)$$

The distribution function satisfies three properties:

1. F_X is non-decreasing
2. $F_X(-\infty) = 0$ and $F_X(\infty) = 1$
3. F_X is continuous from the right[1]

and it exists for every random variable.

Also note that

$$\mathbb{P}(a \leq X \leq b) = F(b) - F(a).$$

4.1 Discrete random variable

A discrete random variables takes values on a countable set, e.g. $\{1,2,3,4,5,6\}$ to model the outcome of a throw of a dice, or on the positive integers for the number of people entering a shop on a day.

The **probability mass function** gives the probability of a particular outcome. Hence $f_X(x) = \mathbb{P}(X = t)$ For a throw of a dice, this could be

$$f(t) = \begin{cases} 1/6 & \text{for } t \in \{1,2,3,4,5,6\} \\ 0 & \text{otherwise} \end{cases}$$

[1] Not part of the material for the course, so don't worry about this.

4.2 Continuous random variable

A continuous random variable, on the other hand, takes values on an interval, or the entire real line. In this setting, it does not make sense anymore to think about $p(X = t)$, as $p(X = t) = 0$ for all t. Instead, we define the (probability) **density function**, f_X, and define it as:

$$f_X(t) = F'(t) (= \mathbb{P}(X \in dt)).$$

From this it follows that

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(t)dt.$$

A final note: a random variable does not necessarily have to fall in either category, it could also be a mixture of both.

Exercise: Define a random variable that is a mixture of both. Give the distribution function. What can you say about the probability mass function and density function?

4.3 Joint distribution

In case of multiple random variables, we can talk about their **joint distribution**.[2] For example, we can have random variables X and Y, and their joint distribution is then defined as

$$F_{XY}(t, s) = \mathbb{P}(X < t, Y < s).$$

The notion of a joint density also makes sense

$$f_{XY}(t, s) = \frac{\partial^2 F_{XY}}{\partial_t \partial_s}(t, s)$$

which has a discrete analogue

$$\mathbb{P}(X = x, Y = y).$$

The **marginal density** can be found by integrating out (summing over) the other random variable

$$f_X(t) = \int_{-\infty}^{\infty} f_{XY}(t, s)ds.$$

4.4 Independence

Two random variables are independent if their joint density factorizes

$$f_{XY}(t, s) = f_X(t)f_Y(s)$$

Intuitively, the information given by knowing the realization of one of the variables does not help in predicting the other. This notion of independence is the same as for independent events, as defined above.

[2] We treat the continuous case, but the discrete case works the same way. Simply replace the density function with the probability mass function, and integrals by sums.

5 Transformation of variables

Sometimes we are interested in the density function of a function of a random variable. This can be done as follows: If $Y = g(X)$, and g is monotone, then the density of Y can be found by

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|.$$

In the multivariate case, we have to use the Jacobian:

$$f_{X,Y}(x,y) = f_{U,V}(u,v)\left|J\left(\frac{(x,y)}{(u,v)}\right)\right|$$

6 Expectation

The expectation of a random variable gives its 'average' value. The expectation is defined as

$$\mathbb{E}(X) = \sum_{k\in S} k\mathbb{P}(X=k)$$

in the discrete case, where S is the support of X.[3] In the continuous case, the expectation is defined as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} tf(t)dt.$$

We note also that for any function g, we can find its expected or average value:

$$\mathbb{E}(g(X)) = \sum_{k\in S} g(k)\mathbb{P}(X=k) \quad \text{or} \quad \mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(t)f(t)dt.$$

This should not come as a surprise, as a function of a random variable leads to another random variable. Note that we do not have to change variables to find a density in terms of $g(X)$. Also, probability is simply a special case of expectation. For example, $\mathbb{P}(X<c) = \mathbb{E}(\mathbf{1}_{X<1})$ where $\mathbf{1}_C$ is the indicator function:

$$\mathbf{1}_C(x) = \begin{cases} 1 & \text{if } x \in C \\ 0 & \text{otherwise} \end{cases}$$

The expectation of a random variable is *fixed* (not a random variable). Furthermore, the **expectation is linear**, in the sense that

$$\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Please never forget this fact!

Exercise: Prove the linearity of expectation by writing out the integral

However, sometimes it is easy to forget that we cannot do

$$\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$$

unless g is linear/affine. As a simple example: $\mathbb{E}(\frac{1}{X}) \neq \frac{1}{\mathbb{E}(X)}$.

If X and Y are independent random variables, then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

However, the converse it not true, i.e. the above equality does not imply X and Y are independent.

Exercise: Prove this, again by writing out the integral

[3] The support is the set for which the probability is non-zero. In the case of the dice: $S = \{1,2,3,4,5,6\}$.

6.1 Variance

The variance of a distribution gives a notion of the spread of the distribution. A random variable with small variance has most realizations close to the mean, while a random variable with a large variance will see many realizations far away from the mean. The variance of a random variable is defined by

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

The following identity is very useful:

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

Exercise: Prove this

Another commonly used notion of the spread is the **standard deviation**. The standard deviation, σ is simply defined as

$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

6.1.1 Chebyshev's inequality

Chebyshev's inequality gives a bound on the probability that a random variable deviates from the mean:

$$\mathbb{P}(|X - \mathbb{E}(X)| > t) < \sqrt{\frac{\mathrm{Var}(X)}{t^2}} \qquad \text{or} \qquad \mathbb{P}(|X - \mathbb{E}(X)| > k\sigma) < \frac{1}{k^2}.$$

The inequality is extremely useful because it always works, irrespective of the distribution function. However, this also means that the bound can be quite loose: it has to respect the worst case scenario.

6.2 Covariance

The covariance measures whether variables tend to move in the same direction.

$$\mathrm{Cov}[X, Y] = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

If X and Y are independent, then $\mathrm{Cov}[X, Y] = 0$, but the opposite is not necessarily true. When $\mathrm{Cov}[X, Y] = 0$, we also say that X and Y are uncorrelated, as their correlation coefficient, defined by

$$\rho_{X,Y} = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}X\mathrm{Var}Y}}$$

would be zero.

It is not too hard to show that[4]

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y),$$

and thus if X and Y are **independent**, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

[4] We can assume $\mathbb{E}(X) = \mathbb{E}(Y) = 0$ (why?) and then we work out $\mathbb{E}((X + Y)^2) - \mathbb{E}(X + Y)^2$.

6.3 Characteristic function

The characteristic function is defined as

$$\varphi_X(\omega) = \mathbb{E}\left(e^{i\omega X}\right) = \int_{-\infty}^{\infty} e^{i\omega t} f(t) dt.$$

The characteristic function gives another *characterization* of a random variable. In other words, **if two random variables share the same characteristic function, then they also have the same distribution function**! This property can be very useful in proving a random variable has a certain distribution.

Furthermore, the characteristic function can be used to compute the moments of a random variable since

$$-i\mathbb{E}(X) = \varphi_X'(0), \qquad -\mathbb{E}(X^2) = \varphi_X''(0)$$

or more generally

$$(-i)^k \mathbb{E}(X^k) = \varphi_X^{(k)}(0).$$

For characteristic functions, we have the following identity:

$$\varphi_{aX+b}(\omega) = e^{i\omega b}\varphi_X(a\omega).$$

Furthermore, if two random variables X and Y are **independent**, we have

$$\varphi_{X+Y}(\omega) = \varphi_X(\omega)\varphi_Y(\omega),$$

which is can be very useful, so keep this in mind.

7 Conditional probability

Now we turn to conditional probability.

$$\mathbb{P}(A \mid B)$$

gives the probability of event A given that we know event B occurred. Hence, we can restrict our attention to the outcomes in B; we have additional information. Note that $\mathbb{P}(A) = \mathbb{P}(A|S)$.

Note we can express conditional probability as follows

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

which is clear if we draw the sample space.

7.1 Bayes theorem

We can relate $\mathbb{P}(A \mid B)$ to $\mathbb{P}(B \mid A)$ by Bayes' theorem:

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

One often wants to rearrange the above equation, but I feel this one is the easiest to remember.

Exercise: Alice has two coins in her pocket, a fair coin, and a two-headed coin. She picks one at random from her pocket, tosses it and obtains head. What is the probability she flipped the fair coin?[5]

[5] Source: www.statlect.com

7.2 Law of total probability

While not specifically mentioned in class, the law of total probability is worth mentioning. In simple cases, it is very intuitive result, and you probably have used it without realizing it. However, for more complicated cases this is often overlooked

The Law of total probability states that, given mutually exclusive events B_i such that $\bigcup B_i = S$

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A|B_i),$$

or in another form, for two random variables X and Y

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$$

This might seem abstract, so it's useful to consider an example. Recall the craps problem, where we want to find the probability of winning. How did we compute this? First, we conditioned on the value of first card:

$$\mathbb{P}(\text{win}) = \sum_{i=1}^{6} \mathbb{P}(\text{win} \mid \text{first roll} = 1)\mathbb{P}(\text{first roll} = 1)$$

Another example where this is useful is the following: Suppose there are two independent variables X and Y. We are now interested in finding the probability that their sum $Z = X + Y$ is less than some value t. By using the law of total probability we find.

$$\begin{aligned}\mathbb{P}(X + Y < t) &= \int_{-\infty}^{\infty} \mathbb{P}(X + Y < t \mid Y = s) f_Y(s) ds \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X < t - s) f_Y(s) ds \\ &= \int_{-\infty}^{\infty} F_X(t - s) f_Y(s) ds.\end{aligned}$$

Note how this result directly translates to the continuous case.

8 Common Distributions

In this section, we will go over some of the common distributions that are used in this class and beyond.

8.1 Discrete uniform distribution

X is uniform distribution over some finite set S if every outcome of every element in S is equally likely, or

$$p(X = k) = \frac{1}{|S|}$$

for $k \in S$. For example, if X is the points in a roll of a dice, then X is uniform.

8.2 Continuous uniform distribution

The continuous version of the uniform distribution is just as intuitive as the discrete version. We say that X is uniformly distributed over $[a, b]$ if X takes on a value between a and b ($a < b$) and every value is equally likely. The density is given by

$$f_U(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

Also,

$$\mathbb{E}(X) = \frac{a+b}{2} \qquad \text{Var}(X) = \frac{(b-a)^2}{12}$$

8.3 Binomial distribution

The binomial distribution models number of successes given a fixed number of repeated trials. Let X_i for $i = 1 \dots n$ be independent random variables with value 1 (success) with probability p and 0 (failure) otherwise. Then, $Z = \sum_{i=1}^{n} X_i$ has a binomial distribution with parameters n and p. For example, think about the number of heads out of 5 fair coin tosses follows a binomial$(5, 0.5)$ distribution.

We note

$$\mathbb{P}(Z = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

and

$$\mathbb{E}Z = np \qquad \text{Var}Z = np(1-p)$$

which follows directly from the linearity of expectation (why?) and that we can sum the variances under independence.[6] Computing $\mathbb{P}(Z \leq k)$ can be tricky when k and n are large, but k is not close to n (what to do when $k \approx n$?). In this case, it's useful to use a Normal approximation ($\mu = np, \sigma^2 = np(1-p)$). Note that we can only use these approximations reliably when np is sufficiently large (as a rule of thumb, at least 5).

The characteristic function for a binomial random variable is

$$\varphi_X(\omega) = (1 - p + pe^{i\omega})^n.$$

Exercise: Prove this, first for the case $n = 1$, and then for arbitrary n using properties of the characteristic function.

8.4 Poisson distribution

The Poisson distribution models the number of events in a given timeframe where the events arive independently at rate μ. For example, consider the number of visitors that enter a bank on a given day.

Suppose X has a Poisson distribution with parameter μ. Then

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

The expectation and the variance X are equal:

$$\mathbb{E}(X) = \text{Var}(X) = \lambda$$

[6] It's trivial to compute the variance of a single trial.

With $np = \mu$ held fixed, as $n \to \infty$ and $p \to 0$, the binomial distribution converges to a Poisson distribution. The characteristic function is

$$\varphi_X(\omega) = e^{\lambda(e^{i\omega}-1)}$$

8.5 Exponential distribution

The exponential distribution models the time till an event happens. We say that T has an exponential distribution with parameter λ if the density of T is

$$f_T(t) = \lambda e^{-\lambda t}$$

and the CDF is given by

$$F_T(t) = \mathbb{P}(T < t) = 1 - e^{-\lambda t}.$$

The mean and variance are

$$\mathbb{E}(T) = \frac{1}{\lambda} \qquad \mathrm{Var}(T) = \frac{1}{\lambda^2}.$$

It's a good exercise to show that the characteristic function is

$$\varphi_T(\omega) = \frac{\lambda}{\lambda - i\omega}$$

The *inter-arrival time*, that is, the time between ticks, of a Poisson process, has an exponential distribution. Hence, the two distributions are closely related.

Exercise: Show that

$$\mathbb{P}(T > s + t \mid T > s) = \mathbb{P}(T > t)$$

This is called the *memoryless property*, why is that the case?

8.6 Normal distribution

The Normal, or Gaussian, distribution might be the most well known distribution. It is parametrized by μ and σ, which are also the mean and standard deviation. The density function, denoted by a special ϕ, is given by

$$\phi(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(t-\mu)^2}{2\sigma^2}}$$

Unfortunately, there is no closed form expression for the CDF, so do not try to integrate the density function (but do use the fact that if we integrate over $(-\infty, \infty)$, it has to integrate to 1).

However, we can always standardize a Normally distributed random variable X with mean μ and standard deviation σ by considering

$$Z = \frac{X - \mu}{\sigma}.$$

The CDF of this standard normal random variable Z is tabulated. Thus, if X follows a normal distribution with mean μ and standard deviation σ, then we have

$$\mathbb{P}(X < b) = \mathbb{P}\left(\frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = \mathbb{P}\left(Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right).$$

Also note that as n grows, the binomial distribution converges to the normal distribution, just as the Poisson distribution when λ grows.

Furthermore, the characteristic function of the normal distribution is

$$\varphi_X(\omega) = e^{i\omega t + \frac{1}{2}\sigma^2\omega^2}$$

9 Reliability

Finally, we spent quite a bit of time discussing reliability. Apart from using the Poisson and Exponential distributions, we also considered redundancy and multiple components.

In particular, you should know how to compute the probability of failure of a system with components parallel or series configuration. Note it is easiest to compute probability of failure in a parallel system, as there is only one possibility: all components must fail. For a series configuration it's the other way around: it's easiest to compute the probability of success as all components must work.

In the case that r out of n components have to work for a system to work, where each one has some probability p of failing, we can use the binomial distribution.

For reliability computations, practice makes perfect, so here is one last exercise:

Exercise: A system consists of 4 components, and all of them have to work. In the first component, there are 5 units of which 3 have to work. Each fails with probability 0.5. The second component consists of two units in series, that fail with probability 0.25. The third component consists of two parts, of which both must work for the component to work. Both parts have two units of which at leasts one must work, failure probability for each unit is 0.2. The final component is a single unit with failure probability 0.1. Don't forget to make a sketch before starting any computations!