Math Camp - Probability

Probability

A lot of the problems we try to model involve either risk or uncertainty.

Probability gives us a way to model these.

Probability

Formally, we start with three things:

- A set Ω: the sample space - the set of all possible outcomes
- A σ-algebra: $\mathcal{F} \subseteq P(\Omega)$.
- A function $P : \mathcal{F} \to [0, 1]$. which assigns each element of $\mathcal{F}$ a probability.

(A σ-algebra is a set that contains both Ω and $\emptyset$ and is closed under countable unions and complements.)

Probability

Think about rolling a die

- $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- If after I roll the die, I see what number comes up, I want to be able to evaluate the probabilities of each of those outcomes.
- I also want to be able to evaluate the probability of an even number coming up, or either an even number of 5 coming up and so on...
- I also may want to capture that I can't evaluate the probability of a single outcome, but I can figure out how likely it is that an even number comes up.

If the sample space is large, we need this (it's may be impossible to define a probability over everything). It is also convenient for modelling things like dynamics. For most of what we are going to do in the first year, you can just ignore $\mathcal{F}$.

Probability

We want P to behave how we expect:

- $P(\Omega) = 1$, $P(\emptyset) = 0$.
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$, $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ if all sets are pairwise disjoint.

These properties imply (prove on your own)

- $P(A^c) = 1 - P(A)$.
- $P(A \cup B) \leq P(A) + P(B)$
- $P(A \cap B) = P(A) + P(B) - P(A \cup B)$
- $A_1 \subseteq A_2 \subseteq A_3 \ldots$ then $P(\cup_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i)$
- $A_{n+1} \subseteq A_n$ then $P(\cap_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i)$.

Random Variable

We want to describe a variable whose value depends some random outcome.

A Random Variable is a measurable function $X : \Omega \to E$, where E is our set of outcomes. Almost always $E = \mathbb{R}$ for our purposes.

- e.g. the outcome of a bet that pays a dollar if the die comes up on 5 or 6 and loses you a dollar otherwise.

Discrete Random Variable

We say a random variable is discrete if $X(\Omega)$ is countable.

- We call the (closure) of $X(\Omega)$ the support of X.
- Described by a probability mass function (pmf), $f : E \to [0, 1]$, where
$$f(x) = Pr(\{x\})$$
.
- This has a corresponding cumulative distribution function (cdf) $F : E \to [0, 1]$ where
$$F(x) = \sum_{y \leq x} f(y)$$
.

Random Variable

For any random variable with range $E \subseteq \mathbb{R}$, we can define a CDF as $F(x) = P(X \leq x)$.

But, if the support isn't countable then a pmf doesn't exist.

Continuous random variable

We say a random variable is continuous if there exists an $f : E \to \mathbb{R}_+$ such that

$$F(x) = \int_{s \leq x} f(s)\, ds.$$

The function $f(s)$ is called the probability density function.

IF F is differentiable this exists. We can use this to calculate probabilities

$$Pr(X \in A) = \int_A f(x)\, dx.$$

(for the rest of these slides, if the distinction is not important, I'll write things in terms of integrals over the pdf. These can replaced with sums over the pmf in the case of discrete random variables)

Expected value

The expected value is the average value of a random variable:

$$E(x) = \int_{\mathbb{R}} xf(x)\,dx$$

For any $g : E \to \mathbb{R}$, $g(x)$ is a random variable as well, so

$$E(g(X)) = \int_{\mathbb{R}} g(x)f(x)\,dx$$

This is a linear operator, so

$$E(aX + b) = aE(X) + b$$

Similarly for two random variables X, Y

$$E(aX + bY) = aE(X) + bE(Y)$$

Our main goal is to formalize the idea that the expected value actually is the average of a bunch of draws of X.

Expected value

Some more properties worth trying to prove

- If $X = c$ for all ω then $E(X) = c$.
- If $X \geq 0$ for all ω then $E(X) \geq 0$.
- If $X \geq Y$ for all ω then $E(X) \geq E(Y)$.
- If $E(|X|) = 0$ then $X = 0$.
- $E(f(x)) \geq f(E(x))$ if f convex.

Variance

Variance measures how spread out a random variable is around it's mean.

$$Var(X) = E((X - E(X))^2)$$

Some properties

- $Var(X) \geq 0$
- A random variable is constant (a.s.) iff $Var(X) = 0$.
- $Var(aX + b) = a^2 Var(X)$

Some Inequalities

Theorem (Markov's Inequality)

If X is a non-negative random variable and $a > 0$ then

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Corollary (Chebyshev's Inequality)

For any random variable X

$$P(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2}$$

Markov's inequality

Proof:

- Let $1_{\{X \geq a\}}$ be the random varaible that takes the value 1 whenever $X \geq a$ and 0 otherwise.
- Then $a1_{\{X \geq a\}} \leq X$ for all $\omega \in \Omega$ by non-negativity of X.
- So $E(a1_{\{X \geq a\}}) \leq E(X)$.
- Expanding out the left hand side

$$E(a1_{\{X \geq a\}}) = a(1 \cdot P(X \geq a) + 0 \cdot P(X \leq a))$$

- Therefore $P(X \geq a) \leq E(X)/a$

Who Cares

Chebyshev's inequality gives us a quick way to detect how likely departures from the average are.

$$P(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2}$$

For instance, the probability outcome that is 2 standard deviations away from the $E(X)$ is at most $1/4$, the probability of seeing one that's 3 standard deviations away is at most $1/9$ and so on.

Joint Distributions

When we have multiple random variables, the individual random variable's CDF/PDF is not enough to describe them, since it can't capture how they are related.

- Flip a coin, let $X = 1$ if heads, -1 if tails. Flip a coin again and define Y the same way.
- Flip a coin, let $\hat{X} = \hat{Y} = 1$ if heads, and -1 if tails.

All four of these random variables have the same cdf and pmf. But $E(XY) = 0$ and $E(\hat{X}\hat{Y}) = 1$.

Joint Distributions

To resolve this, we define the joint distribution $F(x, y) = P(X \leq x, Y \leq y)$. The corresponding joint pdf solves

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, dv\, du.$$

We can define the corresponding marginal distributions, which describe X ignoring Y as

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

since $F_X(x) = Pr(X \leq x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, v)\, dv\, du$.

Joint Distributions

We say X and Y are independent if

$$F(x, y) = F_X(x)F_Y(y)$$

equivalently if they exist $f(x, y) = f_X(x)f_Y(y)$.

- This means that $E(XY) = E(X)E(Y)$
- The covariance

$$cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

and correlation

$$corr(X, Y) = cov(X, Y)/\sqrt{Var(X)Var(Y)}$$

are 0 if two random variables are independent.

- The converse is not true.

Verify that

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

and

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Joint Distributions

Solving for the distribution of functions of multiple random variables can often be a bit subtle.

Suppose X, Y are independent and uniformly distribution on $[0, 1]$. What is the distribution of $Z = X + Y$

$$\begin{aligned} F_Z(Z) &= Pr(X + Y \leq z) \\ &= \begin{cases} \frac{1}{2}z^2 \text{ if } 0 \leq z \leq 1 \\ 1 - \frac{1}{2}(2 - z)^2 \text{ if } 1 \leq z \leq 2 \end{cases} \end{aligned}$$

Change of Variables

Suppose I want to evaluate

$$\int_{g(a)}^{g(b)} f(x)\,dx$$

f integrable, g differentiable injection.
Applying the fundemental theorem of calculus, we see that this is equal to

$$\int_a^b f(g(s))g'(s)\,ds.$$

Similarly for multiple variables

$$\int_{g(U)} f(x)\,dx = \int_U f(g(s))|det\ Dg(s)|\,ds$$

Change of Variables

Suppose X is distributed according to F with pdf f, what is the distribution of $Y = X^{1/3}$?

$$\begin{aligned} F_y(Y) &= Pr(X \leq Y^3) = F_x(Y^3) \\ &= \int_{-\infty}^{Y^3} f(x)\, dx \\ &= \int_{-\infty}^{Y} f(y^3) 3y^2\, dy. \end{aligned}$$

Change of Variables

Consider X, Y independent. What is the distribution of $X + Y$? Let $U = X + Y$, $V = Y$, then $g(u, v) = (u - v, v)$, which has a Jacobian with determinant 1.

$$\begin{aligned} F_U(U) &= \int_{\{(x,y):x+y\leq U\}} f_X(x) f_Y(y)\, d(x \times y) \\ &= \int_{-\infty}^{U} \int_{-\infty}^{\infty} f_X(u - v) f_Y(v) \cdot 1\, dv du \end{aligned}$$

from change of variables + Fubini's theorem. So the density is

$$f_u(u) = \int_{-\infty}^{\infty} f_X(u - v) f_Y(v)\, dv.$$

Law of Large Numbers

The law of large numbers is the name for a bunch of results that roughly show :

Theorem ("Law of large numbers")

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with finite expected value μ. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.$$

This gives a pretty clear interpretation of the expected value.

Convergence

The sequence $\frac{1}{n}\sum_{i=1}^{n} X_i$ is a sequence of random variable, not numbers. What does it mean for this to converge

Let's step back a bit, and think about sequences of fuctions in $\mathbb{R}$.

Definition

A sequence of functions $f_n : X \to \mathbb{R}^n$, $X \subseteq \mathbb{R}^m$ converges pointwise to f if for all $x \in X$, $f_n(x) \to f(x)$.

Pointwise Convergence

Seems like the natural notion of convergence. But, we can already see it behaves in some counterintutive ways.

Some things to be careful with:

- This does not preserve continuity: e.g. $f_n(x) = \max\{0, 1 - n|x|\}$.
- Limits aren't preserved under integration: $f_n(x) = n1_{\{(0,1/n]\}}$ converges pointwise to 0, but $\int_{0^1} f_n(x)\,dx = 1$.

Some good news, the second example is indeed a bit pathological:

Theorem (Dominated Convergence)

If there exists a $g(x)$ s.t. $g(x) > |f_n(x)|$ for all n and $\int_X |g(x)|\,dx < \infty$, then $\int_X f_n(x)\,dx \to \int_X f(x)\,dx$

Pointwise Convergence

We could fix both these issues by strengthinging our convergence notion.

Definition (Uniform Convergence)

A sequence of functions $f_n : X \to \mathbb{R}^n$, $X \subseteq \mathbb{R}^m$ converges uniformly to f if $\sup_{x \in X} |f_n(x) - f(x)| \to 0$.

This rules out some things that naturally seem like they should converge, i.e. $f_n : [0, 1) \to [0, 1)$, $f_n(x) = x^n$.

- As much as we'd like to preserve continuity, we'd like to go back to random variables.
- With this in mind, even pointwise convergence seems too strong:

$$f_n(x) = \begin{cases} (-1)^n \text{ if } x = 0 \\ 0 \text{ o.w.} \end{cases}$$

clearly doesn't converge pointwise, but if, for instance, X was drawn uniformly from $[0, 1]$, it seems like the random variables $X_n = f_n(X)$ should converge to 0.

Convergence of Random Variables

Random variables are functions, so we could define convergence like we do for any other function.

- Sure-convergence: $X_n \to X$ if for all ω, $X_n(\omega) \to X(\omega)$.
- Uniform-sure convergence, $X_n \to X$ if $\sup_{\omega \in \Omega} |X_n(\omega) - X(\omega)| \to 0$.

This seems like not what we want:

- Requires us to be really careful with the sample space, which we've mostly been ignoring.
- Needs to hold everywhere.

Almost-sure convergence

The first "useful" convergence criteria is Almost-sure convergence, $X_n(\omega) \to X(\omega)$ except on a probability 0 set.

- The natural analogue of pointwise convergence.
- For all intents and purposes X captures long-run behavior of X_n.
- Under mild conditions, if $X_n \xrightarrow{a.s.} X$ then $E(X_n) \to E(X)$.

One could argue this is still too strong. Can constuct sequence of RVs that do not converge pointwise anywhere but seem like they should converge.

Convergence in probability

We could capture a similar idea of differences far along in the sequence never occurring with

$$\lim_{n \to \infty} Pr(|X_n - X| > \epsilon) = 0$$

for all $\epsilon > 0$. This is called Convergence in probability and is weaker than almost sure convergence, and basically does what we want. For large N, X_n behaves like X except on a negligible set.

Convergence

Finally, we could completely abstract away from Ω altogether, and just require that

$$\lim_{n \to \infty} F_n = F(x)$$

This is called convergence in distribution, and is weaker than convergence in probability.

- For bounded continuous function $g(\cdot)$, $E(g(X_n)) \to E(g(X))$.
- This does not imply that the pdfs converge

Convergence

- A.S. Convergence $\Rightarrow$ Convergence in Probability $\Rightarrow$ Convergence in distribution.
- The reverse is generally not true, although convergence in probability and distribution are the same if the limiting RV is a constant.
- All three of these are preserved under continuous functions (continuous mapping theorem).
- All three of these imply $E(g(X_n)) \to E(g(X))$ for any bounded continuous function.

Law of large numbers

Theorem (Weak Law of Large Numbers)

Suppose $X_1, X_2, \ldots$ are iid with finite variance σ^2 and mean μ. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

Law of large numbers

Proof:

- By independence

$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{n\sigma^2}{n^2}$$

and

$$E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \mu$$

- Fix $\epsilon > 0$, by Markov's inequality

$$Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) \leq Var(\frac{1}{n}\sum_{i=1}^{n} X_i)/\epsilon^2 = \sigma^2/(n\epsilon^2).$$

It turns out, this result holds if we replace convergence in probability with almost sure convergence. That is called the strong law of large numbers.

Law of Large Numbers

Let's think about our proof a little more.

- Clearly looking at the limit of $\sum_{i=1}^{n} X_i$ is silly, it's often going to diverge.
- But in many ways we are losing a lot of the properties of the distribution dividing by n, since

$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{n\sigma^2}{n^2}$$

- Let $Y_i = X_i - \mu$. We know from the law of large numbers that $\frac{1}{n}\sum_{i=1}^{n} Y_i$ converges to 0.
- What if we scaled by n^{-k}. Then

$$Var(n^k \sum_{i=1}^{n} Y_i) = n^{1-2k}\sigma^2$$

$k > 1/2$ gives us a version of the weak LLN. Setting $k = 1/2$ seems interesting. The variance doesn't explode or go to 0.

Central Limit Theorem

Theorem (Central Limit Theorem)

Let $X_1, X_2, \ldots$ be a sequence of iid random variables with mean μ and variance $\sigma^2 < \infty$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{dist} Y$$

where Y is $N(0, \sigma^2)$.

(We could similarly look for the scaling factor with the strong LLN breaks, this gives the Law of the iterated logarithm)

Conditional Probability

A natural question to ask is, if an event B happened, what is the probability of A.

For instance, if I know I rolled a die and the number is even, what's the probability it's 2.

- We denote the $P(A|B)$, $P(\cdot|B)$ is a new probability distribution.
- We define for any B s.t. $P(B) > 0$

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

- Some simple arithmetic gives us Bayes Rule

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

and the law of total probability

$$P(A) = P(A|B)P(B) + P(A|B^c)(1 - P(B))$$

Conditional Random Variables

We really want to work with random variables. For discrete random variables this is straightforward:

$$f(x|y) = P(X = x|Y = y) = \frac{Pr(\{X = x\} \cap \{Y = y\})}{Pr(\{Y = y\})} = \frac{f(x, y)}{f_y(y)}$$

For continuous random variables it's a bit less clear, if X, Y are continuous

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}$$

We have another version of Bayes rule

$$f(x|y) = \frac{f(y|x)f_X(x)}{f_Y(y)}$$

Bayes Rule

Bayes rule makes a lot of our calculations easier.

- Suppose I'm flipping a coin. I there's a 50/50 chance it's either fair coin, or a biased coin that comes up heads with probability .75.
- What is $Pr(Biased|Heads)$. $Pr(\text{Biased and Heads})$ seems really hard to work out. On the other hand

$$Pr(Heads|Biased)Pr(Biased) = .75(.5)$$

and the law of total probability gives us

$$Pr(Heads) = .75(.5) + .5(.5)$$

so

$$Pr(Heads|Biased) = \frac{3}{5}$$

Bayes Rule - Example

Suppose we have an unknown parameter $\mu \sim N(0,1)$ and we observe $s = \mu + \epsilon$ where $\epsilon \sim N(0,1)$ and is independent of μ.

What is $f(\mu|s)$? For any $s, \mu \in \mathbb{R}$

$$\begin{aligned} f(s|\mu)f(\mu) &= \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(s-\mu)^2) \cdot \left(\frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(\mu)^2) \right) \\ &= \frac{1}{2\pi} \exp(-\frac{1}{2}s^2) \exp(-(\mu^2 - \mu s)) \\ &= \frac{1}{2\pi} \exp(-\frac{1}{2}s^2 + \frac{1}{4}s^2)) \exp(-(\mu - \frac{1}{2}s)^2) \end{aligned}$$

This means that $f(\mu|s)$ is a bunch of stuff that doesn't depend on μ times $-(\mu - \frac{1}{2}s)^2$. So $\mu|s$ is $N(\frac{s}{2}, \frac{1}{2})$.

Conditional Expectation

We can define conditional expectation

$$E(X|Y=y) = \int_{-\infty}^{\infty} xf(x|y)dx$$

This satisfies the tower rule:

$$E(E(X|Y=y)) = E(X).$$

If X, Y are independent then $E(X|Y=y) = E(X)$.

Conditional expectation

Suppose $N, X_1, X_2, \ldots$ are independent. N has support $\mathbb{N}$, X_i is a real valued random variable. What is

$$E(\sum_{i=1}^{N} X_i).$$

The tower rule makes this pretty easy

$$E(E(\sum_{i=1}^{N} X_i | N = n) = E(NE(X_1)) = E(N)E(X_1)$$