Lecture 1:

We return to foundations.

Definition: Let Ω be a set and F is a collection of subsets of Ω. Then F is a ϭ-algebra on Ω if the following hold:

1. $\Omega \in F$
2. If $A \in F$ then $\Omega \backslash A \in F$
3. If A_n is a countable collection of sets in F (For all natural numbers n $A_n \in F$) then $\cup_n A_n \in F$

Definition: As seen in level 6, a probability measure is a function $P: F \to [0,1]$ with the following properties:

1. $P(\Omega) = 1$
2. If A_n is a disjoint countable collection of sets in F then $P(\cup A_n) = \sum P(A_n)$

We say $P(A)$ is the probability of A. We call the triple $(\Omega, \mathrm{F}, \mathrm{P})$ a probability space.

Definition: The elements of Ω are called outcomes and the elements of F are called events.

It is more useful to talk about probabilities of events than probabilities of outcomes, this is important in the case that Ω is uncountable and the probability of outcomes are 0.

Proposition: Immediately from both of the properties above, $P(A) + P(\Omega \backslash A) = 1$. Also $P(\emptyset) = 0$, also $P(A) \leq P(B)$ if $A \subseteq B$, and $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Example: Suppose all outcomes are equally likely and there are finitely many of them, such as rolling a dice, then we can say $P(A) = \frac{|A|}{|\Omega|}$

Example: Suppose we have n balls labelled 1 to n and they are otherwise indistinguishable. We pick $k \leq n$ balls at random without replacement (Meaning any set of k balls is equally likely). Then we set Ω to be the set of subsets of {1,2,…,n} with exactly k elements. Then the probability of a specific event is exactly $\frac{1}{\binom{n}{k}}$.

Example: The probability of getting a specific shuffle of a well shuffled deck of 52 cards is $\frac{1}{52!}$.

Example: $P(top\ 2\ cards\ are\ aces) = \frac{Number\ of\ ways\ for\ that\ to\ happen}{52!}$

There are 4 choices for the top card then 3 for the next one then 50 for the rest so

$$P(top\ 2\ cards\ are\ aces) = \frac{4 * 3 * 50!}{52!}$$

$$P(top\ 2\ cards\ are\ aces) = \frac{12}{51 * 52} = \frac{1}{221}$$

Example:

Consider a string of n random digits from 0 to 9, then $|\Omega| = 10^n$. The set of n-tuples of digits with none exceeding k has size $(k+1)^n$. So the probability that the largest digit is exactly k is $\frac{(k+1)^n - k^n}{10^n}$.

Example:

Suppose we have n people. What is the probability that at least 2 share a birthday? Lets simplify the model and suppose there are 365 possibilities.

$$|\Omega| = 365^n$$

Let A be n-tuples of {1,2,...,365} with no 2 the same, then the number of subsets of B is given by the formula $365 * 364 * ... * (366 - n) = \frac{365!}{(365-n)!}$. Therefore the probability of the complement of this is exactly

$$1 - \frac{365!}{365^n(365-n)!}$$

By numerical calculations, it turns out that if n=22 the probability is about 0.476 and if n=23 then the probability is about 0.507.

Example:

Let Ω be a finite set with n elements and partition it into k disjoint subsets. Then the number of ways to get that partition from an experiment with n trials where each trial can make something go into one of the k subsets (see level 6 chi squared proof for reason) is $\frac{n!}{n_1!n_2!...n_k!}$

Example:

We want to count the number of strictly increasing functions from {1,2,...,k} to {1,2,...,n} where k is at least n. There are $\binom{n}{k}$ such functions since it all comes down to specifying the range. So the probability of getting such a function if we pick one at random is $\frac{\binom{n}{k}}{n^k}$.

Lecture 2:

Lets count the number of non-decreasing functions f from {1,2,...,k} to {1,2,...,n}. We can find a bijection from this to the set of strictly increasing functions g from {1,2,...,k} to {1,2,...,n+k-1}. We do this by sending f to f(i)+i-1 which is strictly increasing, and this is our bijection, as if the difference is 0 it becomes 1, if it is 1 it becomes 2, etc, and these are uniquely defined by their sequence of differences. So we have $\frac{\binom{n+k-1}{k}}{(n+k-1)^k}$.

Theorem: The ratio between n! and $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ goes to 1 as n goes to infinity

Proof: See level 6 stats, lemma 1 in proof of chi squared tests. In fact we get that the error is a factor of about $e^{\frac{1}{12n}}$ which quickly approaches 1.

Lecture 3:

Proposition: This is a very unsurprising proposition: If A_n is a countable collection in F then $P(\cup A_n) \leq \sum P(A_n)$ even if they are not disjoint.

Proof: Define $B_1 = A_1$ and for n>1 define $B_n = A_n \backslash (A_1 \cup A_2 ... A_{n-1})$. Then these are disjoint and have the same union as $\cup A_n$. So $P(\cup A_n) = \sum P(B_n)$ by countable additivity for disjoint sets. Since B_n is a subset of A_n, $P(B_n) \leq P(A_n)$ so the result follows.

Proposition: (We refer to this as continuity of probability measures). Suppose we have a countable family of sets A_n in F such that each is included in the next. Then the probabilities of A_n are non-decreasing and bounded so they converge. In fact we claim that they converge to $P(\cup A_n)$

Proof: Define B_n as in the previous proposition. Then they are disjoint and $\bigcup_{n=1}^{k} B_n = A_k$. Therefore $P(A_n) = \sum_{k=1}^{n} P(B_k) \to \sum_{k=1}^{\infty} P(B_k) = P(\cup A_n)$.

Proposition: Suppose we have a countable family of sets A_n in F such that each is included in the previous one. Then the probabilities of A_n are decreasing and bounded so they converge. In fact we claim that they converge to $P(\cap A_n)$.

Proof: Apply the above proposition for complements of A_n.

Proposition: Let A,B be in F. Then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Let C be also in F, then $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$. Similarly for n sets.

Proof: This follows from the inclusion-exclusion principle (See numbers and sets).

Alternatively, this holds if there is 1 or 2 sets immediately from the definition, then we can do an easy induction on n and some annoying algebra, ie if H is the union of the first k-1 sets then

$$P(H \cup A_n) = P(H) + P(A_n) - P(H \cap A_n)$$

Then we replace the last term with the terms in the expansion of $P(H)$ but with intersection A_n, and they pick up an additional minus sign.

Corollary:

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

Claim: In the above exclusion-exclusion formulae, the probability is an overestimate if we stop before minus terms and an underestimate if we stop before plus terms.

Proof: Start with $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ and $P(\cup A_n) \leq \sum P(A_n)$ then do careful induction.

Example:

The number of surjections from {1,2,...,n} to {1,2,...,m} is equal to the number of functions minus the union over j ranging from 1 to m of the set of all functions that don't hit j. This is

$$m^n - \sum_{k=1}^{m} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq m} \left| A_{i_1} \cap \ldots A_{i_k} \right| = m^n - \sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} (m-k)^n$$

Lecture 4:

Definition: A derangement is a permutation with no fixed points, that is f(x) is not x for any x.

We want to count the number of derangements of {1,2,...,n}.

Define A_i to be all permutations that fix i. Then we are interested in the intersection of the complement of all the A_i's. This is the complement of the union of all the A_i's. We will use the inclusion exclusion formula

$$P(A) = 1 - P(\cup A_i) = 1 + \sum_{k=1}^{m} (-1)^k \sum_{1 \le i_1 < \cdots < i_k \le m} \left| A_{i_1} \cap \ldots A_{i_k} \right| = 1 + \sum_{k=1}^{m} (-1)^k \binom{n}{k} \frac{(n-k)!}{n!}$$

$$= 1 + \sum_{k=1}^{m} (-1)^k k! = \sum_{k=0}^{m} (-1)^k k!$$

This looks like the taylor series for e^{-1} and in fact as m gets large it converges to e^{-1}.

We will now prove some obvious properties from the definitions of probability spaces.

Definition:

Let A and B be events in a probability space. We say A is independent of B if $P(A \cap B) = P(A)P(B)$.

A countable collection of events A_n is said to be independent if for any k-tuple i of natural numbers, we have that $P\left(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}\right) = P\left(A_{i_1}\right)P\left(A_{i_2}\right) \ldots P\left(A_{i_k}\right)$.

Important note: Even if any pair of events are independent in a collection does not imply the whole collection is independent.

Example: Toss a fair coin twice. Then for any outcome, the probability is $\frac{1}{4}$. Lets say that the events are {0,0}, {0,1}, {1,0} and {1,1}. Lets take A=({0,0},{0,1}), B=({0,0},{1,0}), C=({1,0},{0,1}). Then, for example, we have $P(A)P(B) = \frac{1}{4} = P(A \cap B)$ and similarly we can check this for the other pairs. However, we have that $P(A)P(B)P(C) = \frac{1}{8}$ but $P(A \cap B \cap C) = 0$.

Claim: If A is independent of B then A is independent of the complement of B.

Proof: $P(A \cap B^C) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)\left(1 - P(B)\right) = P(A)P(B^C)$

Now take a probability space and take an event B with P(B)>0. For an event A we define the conditional probability $\left(P(A\ given\ B) \text{ or } P(A|B)\right)$ AS $\frac{P(A \cap B)}{P(B)}$. If A is independent of B then $P(A|B) = P(A)$.

Let A_n be a disjoint sequence of events. Then $P(\cup A_n|B) = \sum P(A_n|B)$. This is because we just need to take intersections with B and divide by P(B) and use normal countable additivity.

Proposition:

Let B_n be a disjoint collection of events and suppose their union is the whole probability space and each has non-zero probability. Then $P(A) = \sum P(B_n)P(A|B_n)$.

Proof:

$$P(A) = P\left(A \cap (\cup B_n)\right) = P\left(\cup (B_n \cap A)\right) = \sum P(B_n \cap A) = \sum P(B_n)P(A|B_n)$$

It now follows that

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

Example:

Suppose there is a rare condition which affects 0.1% of people. Suppose we have a test for it which is positive for 98% of the affected population and 1% of the unaffected population. Pick an individual at random and suppose they tested positive. We want to find the probability they have the disease.

Define A to be the event that the individual has the disease and P to be the event that they test positive. $P(A|P) = \frac{P(P|A)P(A)}{P(P)} = \frac{0.98*0.001}{0.98*0.001+0.01*0.999} \approx 0.0893 = 8.93\%$.

This may be surprising the first time you see it. This basically happens because the false positive rate is much larger than the disease rate.

Lecture 5:

Example: What is the probability that I have two children who are boys if given that one of them is a boy.

The possibilities are as follows:

Older is boy, younger is girl (BG), GG, BB, GB.

There are 3 possibilities that have one being a boy and 1 that has 2 being a boy, so the answer is $\frac{1}{3}$. This is because all possibilities are equally likely.

Example: Now we want to know the probability there are two boys given that the older one is a boy. The probability here is $\frac{1}{2}$.

Example: What is the probability that they are both boys given that one of them is a boy born on a Thursday.

I will write TG = Oldest is boy born on Thursday, youngest is girl

I will write GT = Opposite

I will write TN= Oldest is boy on Thursday, youngest is boy not born on Thursday

Similarly for NT, TT

The probability of each of these possibilities are as follows:

$$P(TT) = \frac{1}{2} * \frac{1}{7} * \frac{1}{2} * \frac{1}{7} = \frac{1}{196}$$

$$P(NT) = P(TN) = \frac{1}{2} * \frac{1}{7} * \frac{1}{2} * \frac{6}{7} = \frac{3}{98}$$

$$P(TG) = P(GT) = \frac{1}{2} * \frac{1}{7} * \frac{1}{2} = \frac{1}{28}$$

Now the probability is

$$\frac{P(TN \cup NT \cup TT)}{P(TN \cup NT \cup TT \cup TG \cup GT)} = \frac{\frac{3}{98} + \frac{3}{98} + \frac{1}{196}}{\frac{1}{28} + \frac{1}{28} + \frac{3}{98} + \frac{3}{98} + \frac{1}{196}} = \frac{13}{27}$$

Ok so this is surprising it does not seem to be relevant which day the boy was born on.

The point is if they're not both born on Thursday you're specifying which one which takes you to the one half world.

Example (Simpson's paradox):

I'm gonna make a table for this

All applicants	Admitted	Rejected	% Admitted
State school	25	25	50%
Independent school	28	22	56%

London schools applicants	Admitted	Rejected	% Admitted
State school	15	22	41%
Independent school	5	8	38%

Cambridge schools applicants	Admitted	Rejected	% Admitted
State school	10	3	77%
Independent school	23	14	62%

The paradox is that in the all applicants table it seems like independent schools have an advantage but when you split the data it seems like state schools have an advantage.

Let A = {An individual is admitted}, B = {They are from london}, C={State}

Now we have

$$P(A|B \cap C) > P(A|B \cap C^C)$$

$$P(A|B^C \cap C) > P(A|B^C \cap C^C)$$

But

$$P(A|C) < P(A|C^C)$$

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

We will rewrite this as

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

$$= P(A|B \cap C)P(B|C) + P(A|B^C \cap C)P(B^C|C)$$

If the first two inequalities above are satisfied then

$$P(A|C) > P(A|B \cap C^C)P(B|C) + P(A|B^C \cap C^C)P(B^C|C)$$

If $P(B|C) = P(B|C^C)$ which is not valid here we get

$$P(A|C) > P(A|B \cap C^C)P(B|C^C) + P(A|B^C \cap C^C)P(B^C|C^C) = P(A|C^C)$$

Which is the reverse of the third inequality.

The intuitive explanation for this is that if you fix in advance the acceptance and rejection rate given the state/independent condition and the london/cambridge situation you only know the acceptance and rejection rate just given the state/independent condition if you know how many applicants you have in each situation – it depends on this number.

Definition: A discrete probability distribution is a probability distribution with countably many outcomes, where F is all subsets of Ω.

We ennumerate the outcomes and write the probability of outcome i as P_i. We know know that for each i, $P_i \geq 0$ and $\sum P_i = 1$.

Definition: The bernoulli distribution is simple. It is any distribution with exactly two outcomes.

We know about the binomial distribution from A level. We can think of it as the sum of n bernoulli distributions. We also know about the multinomial (See level 6 stats chi squared section).

Lecture 6:

We also know from Levels 3-6 about the geometric, negative binomial and poisson distributions.

We would now like to pin down the more precise definition of a random variable.

A random variable can be thought of as a function $X: \Omega \to \mathbb{R}$. Ie, each outcome is a real number. The probability that x takes a specific value is not very useful to talk about in general (in fact this must be 0 at all but possibly countably many points), but you can talk about the probability of x being in an interval. We therefore impose that for all real numbers x we have that $\{X(w) \leq x\}$ is in the ϭ-algebra, which after the rules for a ϭ-algebra implies that any countable combination of intervals is in this, and this is what uniquely characterizes a random variable. This is the probability measure we have used in our level 6 central limit theorem proof.

Let A be in F, then as a reminder we write $1_A(w)$ to be 1 if w is in A and 0 otherwise. This is actually a random variable by definition.

We can define the cumulative distribution function as a non-decreasing function from $\mathbb{R} \to [0,1]$, and in fact such functions are exactly all the real random variables. They don't have to be in $\mathbb{R}$, we could define analagously a random variable in $\mathbb{R}^n$ and in this case we usually impose that all rectangular/cuboid regions are in the ϭ-algebra.

Definition:

We say random variables are independent if for all sets in the ϭ-algebra we have the usual intersection-product independence criterion.

Lecture 7:

Definition: Suppose for now that Ω is either finite or countable. We say that x is non-negative if it never takes a negative value. Then define the expectation of x as $E[x] = \sum_{\omega \in \Omega} X(\omega)P(\omega)$.

We can write the following:

$$E[x] = \sum_{x \in \Omega_X} x \sum_{\omega \in \{X=x\}} P(\omega)$$

Where we split outcomes first by the value the random variable takes and then sum over those.

We end up with

$$E[x] = \sum_{x \in \Omega_x} xP(X = x)$$

Don't worry if you didn't follow this it's obvious what expectation means anyway.

Example: Suppose $X \sim B(n, p)$ ie X is binomial with n trials and probability p.

$$E[x] = \sum_{k=0}^{n} kP(x = k) = \sum_{k=0}^{n} k \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} = n \sum_{k=0}^{n} \frac{(n-1)!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k}$$

$$= n \sum_{k=0}^{n-1} \frac{(n-1)!}{(k)!\,(n-k-1)!} p^{k-1} (1-p)^{n-k-1} = np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-k-1} = np$$

Where I renamed k-1 to k.

Of course we expect it to be np. The last sum is 1 because it is the sum of probabilities of a different binomial distribution which we know should be 1.

Example:

If we have $X \sim Po(\lambda)$ then

$$E[x] = \sum_{k=0}^{\infty} kP(x = k) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} = \lambda \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda$$

By taylor series.

We can define the expectation of a general discrete random variable as follows:

$$E[x] = P(x > 0)E[x \mid x > 0] - P(x < 0)E[-x \mid x < 0]$$

Whenever both of these exist and are finite. We say a random variable is integrable if its expectation exists.

The expectation of an indicator function of a set is the probability of being in that set.

Also $E[g(X)] = \sum g(x)P(X = x)$

I've omitted some technical details and other propositions since they are too boring and obvious – they include stuff like $E[X + Y] = E[X] + E[Y]$.

Example: Suppose $x \geq 0$ and takes integer values, then $E[x] = \sum_{k=1}^{\infty} P(X \geq k)$

Proof:

$X = \sum_{k=1}^{\infty} 1(x \geq k)$ where 1 means the indicator function. Then taking expectations of indicators is taking probabilities. We can swap sums around since all terms are positive.

Lecture 8:

Alternative proof of inclusion exclusion formula:

$$1(A_1 \cup A_2 \cup \ldots \cup A_n) = 1 - \prod_{i=1}^{n}\big(1 - 1(A_i)\big) = 1 - \sum_{i=1}^{n} 1(A_i) + \sum_{i_1<i_2} 1\big(A_{i_1} \cap A_{i_2}\big) + \cdots$$

Using the fact that the expectation of an indicator is the probability and taking the expectation of both sides, the result follows.

Definition: For a natural number r, we call $E[x^r]$ when it exists the r'th moment of x.

Definition (variance): Define $Var(x) = E\left[\big(x - E(x)\big)^2\right] = E[x^2] - E[x]^2$. We call the square root of the variance the standard deviation, this should all be familiar from earlier levels.

We have the following obvious properties directly from the definition:

- Var(x) is non-negative for real random variables
- If Var(x) is 0 then x equals its mean with probability 1
- If c is real then $Var(cx) = c^2 Var(x)$
- If c is real then $Var(c + x) = Var(x)$

Proposition: $Var(x) = \min_{c\in\mathbb{R}} E[(x - c)^2]$ achieved when c is E[x]

Proof: Define $f(c) = E[(x - c)^2] = E[x^2] - 2cE[x] + c^2$.

Now $f'(c) = -2E[x] + 2c$ which is 0 if and only if $E[x] = c$ completing the proof.

Proposition: The variance of a binomial distribution with parameters n and p is np(1-p)

Proof: See level 6

Proposition: The variance of a poisson distribution is equal to its mean

Proof: This is immediate from considering the poisson distribution as the limit of the binomial distribution

Definition: Suppose we have two random variables x and y. Then the covariance of x and y is defined as $E\big[\big(x - E(x)\big)\big(y - E(y)\big)\big]$.

Expanding this gives $E\big[xy - E[x]y - E[y]x + E[x]E[y]\big] = E[xy] - E[x]E[y]$.

Note that $Cov[x, x] = Var[x]$

Properties:

$$Cov(cx, dy) = cdCov(x, y), Cov(x + c, y + d) = Cov(x, y)$$

Proposition: $Var[x + y] = Var[x] + Var[y] + 2Cov[x, y]$

Proof: $Var[x + y] = E[(x + y - E[x] - E[y])^2] = E\left[\big(x - E(x) + Y - E(y)\big)^2\right]$

$= E\left[\big(x - E(x)\big)^2\right] + 2E[(X - E[X])(Y - E[Y])] + E[(Y - E[Y])^2] = Var(x) + 2Cov(x, y) + Var(y)$

Note that if x, y, z are random variables then

$$Cov(x + y, z) = Cov(x, z) + Cov(y, z)$$

In general $Cov(\sum_{i=1}^{n} c_i x_i, d_i x_i) = \sum_{i,j=1}^{n} c_i d_j Cov(x_i y_j)$

In particular, $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var[X_i] + \sum_{i \neq j} Cov(X_i, X_j)$

Note that independence implies pairwise independence for 3 discrete random variables, since

$$P(X_1 = x_1, X_2 = x_2) = \sum_{x_3} P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \sum_{x_3} P(X_1 = x_1)P(X_2 = x_2)P(X_3 = x_3)$$

By indendependence, which adds up to $P(X_1 = x_1)P(X_2 = x_2)$.

This holds for any finite set of discrete random variables with the same proof.

Proposition: Let X and Y be independent and suppose all the expectations below exist, then we have that $E[f(x)g(y)] = E[f(x)]E[g(y)]$

Define Z=(x,y) and $h(z) = f(x)g(y)$ so $E[h(z)] = \sum_{(x,y)} h(x,y)P\big(z = (x,y)\big)$

$$= \sum_{(x,y)} f(x)g(y)p(X = x)p(Y = y)$$

By independence.

We then get that $E\big[X - E[X]\big]E\big[Y - E[Y]\big] = 0$ for independent variables so zero covariance implies independence.

We can now easily see that this is exactly $E[f(x)]E[g(y)]$.

Corrolary: This is how we show more foundationally that variances add for independent variables, compared to the more intuitive slightly less formal treatment I gave in level 6.

Note that zero covariance does not imply independence. We can prove this by counter example. Let x_1, x_2, x_3 be independent coin tosses of a fair coin. Set $y_1 = 2x_1 - 1, y_2 = 2x_2 - 1$. Now define

$z_1 = y_1 x_3, z_2 = y_2 x_3$

Now $E[Y_1] = E[Y_2] = 0$ and by independence $E[Z_1] = E[Y_1]E[X_3] = 0$ Now $E[Z_1 Z_2] = E[Y_1 Y_2 X_3^2]$ which is 0 for the same reason. Therefore $Cov(Z_1, Z_2) = 0$. But we want to show that these are not independent.

To do this, $P(Z_1 = 0, Z_2 = 0) = \frac{1}{2}$ since we need $x_3 = 0$. However, $P(Z_1 = 0)P(Z_2 = 0) = \frac{1}{4}$ hence they are not independent.

Lecture 9:

Proposition (Markov's inequality): If X is non-negative, then $P(X \geq a) \leq \frac{E[x]}{a}$

Proof: See level 6 chi squared proof

Proposition (Chebyshev's inequality): Let X be a random variable with finite expectation, then we have the inequality $P(|x - E[X]| \geq a) \leq \frac{Var[x]}{a^2}$

Proof: See level 6 chi squared proof

Proposition: $E[|XY|] \leq \sqrt{E[X^2]E[Y^2]}$. Assume all expectations here are finite.

Proof: Finally something we didn't do in level 6. Note that $|XY| \leq \frac{1}{2}(x^2 + y^2)$ because we know that $\frac{1}{2}(x^2 \pm 2xy + y^2) = \frac{1}{2}(x \pm y)^2$ and the square of something thus the difference between the two things is non-negative. Assume that x and y are non-negative since we can just prove it for the absolute value of any variable to get it for all real variables.

Now we will do a similar proof to vectors and matrices cauchy schwartz inequality proof.

$$0 \leq E[(x - ty)^2] = E[x^2] - 2tE[xy] + t^2E[y^2]$$

With respect to t the derivative of this is $-2E[xy] + 2tE[y^2]$ which means that the above expression is minimized when $t = \frac{E[xy]}{E[y^2]}$. We don't actually need to know that this is the minimum t, we just need to plug in this value of t to $E[(x - ty)^2]$ and show that it is non-negative. We do the differentiation to derive that this is in some sense the strongest result we can get. Therefore we know that

$$0 \leq E[x^2] - \frac{2E[xy]^2}{E[y^2]} + \frac{E[xy]^2}{E[y^2]}$$

Multiplying through by $E[y^2]$ and simplifying, the result follows.

We have equality if x is equal to ty with probability 1 where $t = \frac{E[xy]}{E[y^2]}$.

Proposition (Jensen's inequality): A function is called convex if it is always the case that for any t in the range 0 to 1 we have $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$. Intuitively this means the graph is curving upwards as it is under the line connecting the points (x, f(x)) and (y, f(y)) between x and y. The inequality says that if X is a random variable and f is a convex function, then $E[f(x)] \geq f(E[x])$.

Proof:

To remember the direction of the inequality just try to apply it to x^2.

We note that if f is a convex function then f is the supremum of all the lines lying below it. We want to show that at each point there is a line that only touches the function at that point and is lower than or equal to the function everywhere else.

Let m be real and suppose $x < m < y$, then set $m = tx + (1-t)y$ for some t in [0,1]. Then we have that $t(m - x) = (1-t)(y - m)$ is equivalent. We have by convexity that $f(m) \leq tf(x) + (1-t)f(y)$. Rearranging gives $t\big(f(m) - f(x)\big) \leq (1-t)\big(f(y) - f(m)\big)$.

Using $t(m - x) = (1-t)(y - m)$ we have $\frac{t(f(m)-f(x))}{t(m-x)} \leq \frac{(1-t)(f(y)-f(m))}{(1-t)(y-m)}$ which simplifies to give the inequality $\frac{(f(m)-f(x))}{(m-x)} \leq \frac{(f(y)-f(m))}{(y-m)}$. Now set $a = \sup_{x<m} \frac{f(m)-f(x)}{m-x}$, then for all x<m<y we have the inequality that $\frac{f(m)-f(x)}{m-x} \leq a \leq \frac{f(y)-f(m)}{(y-m)}$ where the second inequality is by applying the argument to f(-x). We rearrange to get that for any z (in both the case where z is less than or more than m, we use either of the inequalities above), to get $f(z) \geq a(z - m) + f(m)$.

We can now prove Jensen's inequality, as we know that for $m = E[x]$ there exists real numbers a and b determining a line such that $f(x) \geq ax + b$ and $f(m) = am + b$. Taking expectation of both sides of the inequality $f(x) \geq ax + b$ gives $E[f(x)] \geq aE[x] + b = am + b = f(m) = f(E[x])$.

We have an exact equality if f is a function of the form ax+b.

Lecture 10:

Claim: Let f be a convex function, then $\frac{1}{n}\sum_{k=1}^{n} f(x_k) \geq f\left(\frac{\sum_{k=1}^{n} x_k}{n}\right)$

Proof 1: It's obvious just look at the graph

Proof 2: Let X be a random variable with $P(X = x_i) = \frac{1}{n}$. Then $\frac{1}{n}\sum_{k=1}^{n} f(x_k) = E[f(x)]$ and $f\left(\frac{\sum_{k=1}^{n} x_k}{n}\right) = f(E[x])$ so just apply the thing from last lecture.

Example: Take $f(x) = -\log(x)$. This is convex (take the second derivative it is positive). Now if we take $x_1, x_2, \dots, x_n$ real numbers, we get that

$$-\frac{1}{n}\sum_{k=1}^{n} log(x_k) \geq -\log\left(\frac{\sum_{k=1}^{n} x_k}{n}\right)$$

Reversing the signs and logs of both sides we get

Theorem (AM-GM inequality): $(\prod_{k=1}^{n} x_k)^{\frac{1}{n}} \leq \frac{1}{n}\sum_{k=1}^{n} x_k$

Recall that if X is a discrete random variable then the distribution of X is determined by the probability that X is equal to x for all values of x.

Definition: Let $x_1, x_2, \dots, x_n$ be discrete random variables. There joint distribution is defined to be the probability of each of the outcomes, ie $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$.

By summing over all the possibilities that satisfy $X_1 = x_1$ we can recover $P(X_1 = x_1)$. We call this the marginal distribution of X_1. We can also think about the conditional distribution of one variable given certain information about the others.

If we have joint variables x and y then we know $P(X = x) = \sum_y P(X = x|Y = y)P(Y = y)$.

Now we want to understand the distribution of X+Y if we have variables X and Y.

We get $P(X + Y = k) = \sum_y P(X = k - y, Y = y)$. If X and Y are independence we get

$$(X + Y = k) = \sum_y P(X = k - y)P(Y = y)$$

This is called the convolution of the two distributions.

Example: Suppose we have 2 independent poisson random variables. Set $X \sim Po(\lambda), Y \sim Po(\mu)$ and suppose that they are independent. If you have any common sense and understand what the poisson distribution is, it will be immediately obvious to you that $X + Y \sim Po(\lambda + \mu)$. To prove it,

$$P(X + Y = n) = \sum_{k=0}^{n} P(X = n - k)P(Y = k) = \sum_{k=0}^{n} e^{-\lambda}\frac{\lambda^{n-k}}{(n-k)!}e^{-\mu}\frac{\mu^k}{k!}$$

By the binomial theorem, we get

$$= e^{-(\lambda+\mu)}\sum_{k=0}^{n}\frac{\lambda^{n-k}\mu^k\binom{n}{k}}{n!} = e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^n}{n!}$$

Define the conditional expectation of X given the event B as follows:

$$E[x|B] = \frac{E[x * 1(B)]}{P(B)}$$

Where 1 is the indicator function.

Proposition: Let X be a random variable and Ω_n a partition of Ω into disjoint sets and $P(\Omega_n) > 0$ for all n. Then $E[X] = \sum_n E[X|\Omega_n]P(\Omega_n)$

This is intuitively obvious.

Suppose we have two random variables X and Y. We can define the conditional expectation of X given Y=y. We write $E[X|Y = y] = \sum_x xP(X = x|Y = y)$. We can define this as a function of y.

Example: Toss a coin which has probability p to land on heads n times independently. Set X_i to be the indicator that i'th toss is heads. Then set $Y_n = X_1 + X_2 + \cdots + X_n$. We want to calculate the conditional expectation of X_1 given Y_n.

Set $g(y) = E[X_1|Y_n = y]$.

Then this is $\frac{E(x_1 * 1(Y_n = y))}{P(Y_n = y)} = \frac{P(X_1 = 1, Y_n = y)}{P(Y_n = y)} = \frac{P(X_1 = 1)P(X_2 + X_3 + \cdots + X_n = y - 1)}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{p\binom{n-1}{y-1} p^{y-1}(1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{y}{n}$

So now $E[X_1|Y_n = y] = \frac{y}{n}$.

Lecture 11:

Note that everything we are doing above is just for discrete random variables fow now.

We will do some more boring obvious propositions.

Note that $E[E[X|Y]] = E[X]$ as we would expect because

$$E\big[E[X|Y]\big] = E[X|Y] = \sum_y P(Y = y)E[X|Y = y] = \sum_{y,x} P(Y = y)xP(X = x|Y = y)$$

$$= \sum_y xP(X = x) = E[X]$$

Note that if X and Y are independent then $E[X|Y] = E[X]$

This is because

$$E[X|Y] = \sum_y P(Y = y)E[X|Y = y] = \sum_{y,x} P(Y = y)xP(X = x|Y = y) = \sum_y xP(X = x)$$

Ok well this is exactly the same proof but now it is by independence instead of by the fact that we are summing over every possible x.

If Y and Z are independent then $E[E[X|Y]|Z] = E[X]$ because g(y)=E[X|Y] as a function of Y is independent of Z so we have $E[g(y)|Z]$ which by earlier is just $E[g(y)]$ by independence which we know is just $E[X]$.

Also $E[h(Y)X|Y=y]=h(y)E[X|Y=y]$ because when you condition on y, h(y) as a function of y becomes a constant.

From this we know that $E[E[X|Y]|Y]=E[g(y)*1|Y]$ where g(y) is E[X|Y] as a function of y. We then get that this is just g(y), ie $E[E[X|Y]|Y]=E[X|Y]$.

Ok somehow this lecture gets even more trivial. E[X|X=x]=x.

Ok well these may be obvious but they are useful.

Consider a previous example: Toss a coin which has probability p to land on heads n times independently. Set X_i to be the indicator that i'th toss is heads. Then set $Y_n=X_1+X_2+\cdots+X_n$. We want to calculate the conditional expectation of X_1 given Y_n.

Then for each i, $E[X_i|Y_n=y]=E[X_1|Y_n=y]$ for all i.

Now $E[X_1+X_2+\cdots+X_n|Y_n=y]=y$ clearly and then using this we get $\frac{y}{n}$ as desired. Notice this is independent of p.

Now we will talk about random walks.

Definition: A random process or stochastic process is a sequence of random variables X_n.

A random walk is a random process that can be expressed as $X_n=x+Y_1+Y_2+\cdots+Y_n$ where x is a constant and Y_i are independent and identically distributed.

A simple random walk is where $P(Y_i=+1)=p, P(Y_i=-1)=1-p$, and these are the only values y can take. If $p=\frac{1}{2}$ we have a simple symmetric random walk.

Example: Consider the case where this is the fortune of a gambler starting at x where at every time step they gain or lose 1 and they stop if they get to 0. Suppose also that they will stop if they reach a where a>x.

We will write $P_x(A)=P(A|X_0=x)$. Write $h(x)=P_x(X_n\ reaches\ a\ before\ reaching\ 0)$.

Then since we increase by 1 with probability p and decrease by 1 with probability 1-p, we basically reason that the probability equals whatever it is expected to be after the first step, so we have

$$h(x)=ph(x+1)+(1-p)h(x-1), h(\leq 0)=0, h(\geq a)=1$$

Lecture 12:

Lets try to solve this equation.

In the case that $p=\frac{1}{2}$, x is called a simple symmetric random walk. We get that h(x) is the mean of h(x+1) and h(x-1), and then since this is true for all x from 0 to a we get that h(x) must be a linear function of x and so we get $h(x)=\frac{x}{a}$.

If you think about it, this result makes sense since it is consistent with the idea that the expected place you will end up is x.

Now we consider the case where $p\neq\frac{1}{2}$.

We get $h(x) = ph(x+1) + (1-p)h(x-1)$ and we know how to solve recurrence relations like this. I'll walk through the process.

We have $ph(x) - h(x-1) + (1-p)h(x-2) = 0$ and the roots of the corresponding quadratic turn out to be 1 and $\frac{1-p}{p}$. If $p = \frac{1}{2}$ we have the special case of a repeated root.

We know from what the roots are that the general solution is of the form $A + B\left(\frac{1-p}{p}\right)^x$. Given our conditions at 0 and a we get that $h(x) = \frac{\left(\frac{(1-p)}{p}\right)^x - 1}{\left(\frac{(1-p)}{p}\right)^a - 1}$.

We want to know starting from x what is the expected value of the first time that x becomes either 0 or a. We will write this as $E_x[T]$.

Now we will get a recurrence relation. We get that (keeping in mind that after we move 1 step we have to take that into account and add 1 to the time)

$$E_x[T] = p(1 + E_{x+1}[T]) + (1-p)(1 + E_{x-1}[T])$$

We know that $E_0[T] = E_a[T] = 1$.

Setting $k(x) \coloneqq E_x[T]$ we get

$$pk(x) - k(x-1) + (1-p)k(x-2) = -1$$

Conveniently this quadratic has the same roots.

In the case $p = \frac{1}{2}$ then general recurrence relation theory (levels 5-6) implies that we have a particular solution equal to $-x^2$ and a complementary function of the form $Ax + B$. Putting in the initial conditions we get that $k(x) = x(a-x)$.

Otherwise, our complementary function is $A + B\left(\frac{1-p}{p}\right)^x$ and it turns out that $\frac{1}{1-2p}x$ works as a particular solution (we would guess this since we have a constant but we should multiply by x as we have a constant in our complementary function as well). Plugging in our initial conditions we get

$$k(x) = \frac{1}{1-2p}x - \frac{a}{1-2p} * \frac{\left(\frac{(1-p)}{p}\right)^x - 1}{\left(\frac{(1-p)}{p}\right)^a - 1}$$

Recall that a probability generating function for random variables taking positive integers is given by

$$p_x(t) = E[t^x] = \sum_{r=0}^{\infty} P(x = r)t^r$$

The good news is we talked about all the theory of these in levels 5 and 6. Recall that since the sum of the coefficients is 1 we are guranteed to have convergence for $|t| \leq 1$ since we have absolute convergence. Therefore the radius of convergence is at least 1. Therefore these are well defined.

Note that the pgf determines the distribution because the k'th derivative at 0 is related (off by a constant factor of k!) to the probability that x=k, and we know we can differentiate power series inside the radius of convergence.

Recall that the generating function of a binomial distribution with parameters n and p is

$$(pt + 1 - p)^n$$

Theorem: pgf of sums is product of pgf

Proof: See level 6

Ok actually by independence this is easy so I'll just write it.

$$p(t) = E[t^{x_1+x_2+\cdots+x_n}] = E[t^{x_1}]E[t^{x_2}] \dots E[t^{x_n}]$$

By independence.

Example: X is binomial (n,p) and Y is binomial (m,p) and they are independent.

Then the pgf of their product is $(pt + 1 - p)^n(pt + 1 - p)^m = (pt + 1 - p)^{n+m}$ so we easily get the result that we expect which is that X+Y is binomial (n+m,p)

Recall that the pgf of a poisson is $e^{\lambda(t-1)}$, and we can use this to show that the sum of poissons is another poisson much easier.

Proposition: $\lim_{z\to 1^-} p'(z) = E[x]$ when E[x] is finite.

Proof: $p'(z) = \sum r p_r z^{r-1}$

This is increasing and bounded by E[x].

Let ε>0. Then there is N large enuogh such that $\sum_{r=1}^{n} r p_r z^{r-1} \geq E[x] - \varepsilon$ since the infinite sums are the limit of the partial/finite sums and this is how limits are defined.

Now $\lim_{z\to 1^-} p'(z) \geq \lim_{z\to 1^-} \sum_{r=1}^{n} r p_r z^{r-1} \geq E[x] - \varepsilon$

The result follows since ε can be as small as we like.

If E[x] is infinite, then it is the same proof, we say that for each M large enough, $\sum_{r=1}^{n} r p_r z^{r-1} \geq M$ and we do the same limiting argument.

With the same proof, $\lim_{z\to 1^-} p''(z) = E[x(x-1)]$.

Lecture 13:

Let $X_1, X_2, \dots$ be identically distributed independent random variables and let N be an independent random variable that takes integer values. Then take $S_n = X_1 + X_2 + \cdots + X_n$.

Now S_N means we get an outcome ω that determines N and each X then take $\sum_{i=1}^{N(\omega)} X_i(\omega)$.

Lemma: Suppose further that the X's take natural number values. Let $q(z) = E[z^N]$ and $p(z) = E[z^{X_1}]$. Then $r(z) = E[z^{S_N}] = q\big(p(z)\big)$.

Proof: Let $r(z) = E[z^{S_N}] = \sum_{n=0}^{\infty} p(N=n)E[z^{S_n}] = \sum_{n=0}^{\infty} p(N=n)E[z^{X_1}]^n$ by independence of the X_i's and definition of S. We now get $q(E[z^{X_1}]) = q\big(p(z)\big)$.

Now $E[S_n] = q'\big(p(1^-)\big)p'(1^-) = q'(1^-)p'(1^-) = E[N]E[X_1]$ where 1^- means take the limit as we approach 1 from the left. Here we used the fact that $p(1^-) = 1^-$ from basic properties of generating functions.

$E[S_n(S_{n-1})]$ can be found if we differentiate again, we get $q'\big(p(1^-)\big)p''(1^-) + q''\big(p(1^-)\big)p'(1^-)^2$

Using this we could get the variance of S_n.

Now we will talk about branching processes.

Suppose $X_0 = 1$ meaning we start with 1 thing and then they make a random number of offspring equal to X_1 where $P(X_1 = k) = g_k$.

Each thing produces an independent number of offspring with the same distribution as X_1.

Let $Y_{k,n}: k \geq 1, n \geq 0$ be identically distributed independent random variables with the same distribution as X_1. We will consider this the number of offspring the k'th individual of generation n produces. We define

$$X_{n+1} = \begin{cases} Y_{1,n} + \cdots + Y_{x_n,n} : X_n \geq 1 \\ 0 : X_0 = 0 \end{cases}$$

Now this is the distribution of the number of things that are in the n+1'th generation.

Proposition: $E[X_n] = E[X_1]^n$

Proof: This is true by induction using our previous discussion about random numbers of random variables.

Note that because of this we expect the process to go extinct if and only if $E[X_1] < 1$, but we need to discuss what happens if $E[X_1] = 1$ since this is less clear.

Again by previous discussion, if $G(z) = E[z^{X_1}], G_n(z) = E[z^{X_n}]$, then $G_{n+1}(z) = G_n\big(G(z)\big) = G(G(\dots(z)\dots))$ with n+1 G's.

Now we will talk about extinction probability. This is the probability that $X_n = 0$ for some n. Note that this is 1 if $E[X_1] < 1$ since $E[X_n] \to 0$ so $P[X_n \geq 1] \to 0$ by Markov's inequality.

Let q_n be the probability of extinction at the n'th step, then this is an increasing function (since extinction at the n'th step is included in extinction at the (n+1)'th step) that converges to the probability of eventual extinction.

Claim: $q_{n+1} = G(q_n)$ where G is the pgf of X_1.

Proof: Assuming the claim for now (we will prove it), we want to show that the extinction probability is a solution to $q = G(q)$. Since G is continuous, q which is the limit of q_n is also the limit of $G(q_n)$ which is therefore $G(q)$.

Now lets actually prove the claim. We want to relate q_{n+1} to q_n.

Since the pgf at 0 gives the probability our variable is 0, we have

$$q_{n+1} = G_{n+1}(0) = G\big(G_n(0)\big) = G(q_n)$$

So now we just have to solve $q = G(q)$

Lecture 14:

Proposition: If $E[X_n] > 1$ then q<1

Proof: We will shortly prove that in fact we want the smallest root of $q = G(q)$. We know since G is a pgf that $G(1) = 1$ and $G'(1^-) > 1$. So now just draw a picture and it is clear there has to be a root less than 1 since $G(0) > 0$.

Proposition:

1. q is the smallest root of $G(x) = x$
2. If $E[X_n] = 1$ then q is either 1 or 0

Proof:

We can prove we want the smallest root of $q = G(q)$ using another picture. The extinction probability is actually just $G\big(G(\dots(0)\dots)\big)$. Recall that in levels 3 and 4 we discussed cobweb diagrams. Now since G is an increasing function with $G(0) \geq 0$ (since it is a pgf), the cobweb diagram we draw starting at 0 will necessarily converge to the smallest root, just intuitively geometrically.

Now we know that both G and G' are increasing, so the graph of G curves upwards. To have $q = G(q)$ if $G'(1^-) = 1$ have a root any smaller than 1, we see that we must have that $G'(x)$ is identically 1 at all $y < x < 1$ for some y. If this happens then $G''(1^-) = 0$ so there must be 0 variance. Therefore Chebyshev's inequality implies that X_1 must be a random variable which is identically 1 in which case $q = 0$, otherwise if $E[X_1] = 1$ then the extinction probability is 1.

We get another perspective for the result that q is 1 if $E[X_n] < 1$. We have a graph curving upwards with the property that $G(0) > 0, G(1^-) < 1$, and we see this cannot have any roots between 0 and 1.

Now we will start talking about continuous random variables.

In our F (which is the space of events) we include all sets $X(w) \leq x$ as well as the sets that are forced to exist as a result.

We always have a cumulative distribution function which increases from 0 to 1 and approaches 0 and 1 as we go to $-\infty, \infty$ respectively. In fact, these functions are exactly the random variables that we can define on $\mathbb{R}$, almost.

F has to have the property that for every value x, as you approach x from the right, F approaches F(x), we say that F(x) is right continuous. This is because if we had a function like

$$\begin{cases} 0 \; if \; x \leq 67 \\ 1 \; otherwise \end{cases}$$

Then it would essentially be like the random variable always takes the "next real number after 67", which we know is nonsense.

Lecture 15:

Formally the right continuity is because if A_n is some events $X \to x_n$ where x_n approaches x from the right, then from earlier the probability of the intersection is the limit of the probabilities since each A_n is contained in the previous one, which implies F is right continuous.

We say $F(x^-) = \sup_{y<x} F(y)$ which is the probability our variable is strictly less than x.

We say a random variable is a continuous random variable if F is continuous everywhere, or equivalently if it takes every value with probability 0.

We say a random variable is absolutely continuous if F is also differentiable. In this case we call f(x)=F'(x) the probability density function – something which is familiar from previous levels.

Example: The exponential distribution $Exp(\lambda)$ takes positive values and its pdf is given by $\lambda e^{-\lambda x}$.

Here, the distribution of $floor\left(\frac{x}{n}\right)$ as a geometric distribution with probability $e^{-\frac{\lambda}{n}}$. We think of this distribution as a scaled limit of the geometric distribution.

Here $P(X \geq s+t \mid x \geq s) = e^{-t}$ (s,t>0).

Theorem: Let T be a positive random variable which is never 0 or infinity. Then T has the exponential distribution if and only if $P(T \geq s+t \mid T \geq s) = P(T \geq t)$

Proof: If we first assume we have the exponential distribution then by previous discussion we are done. Otherwise, define $g(t) = P(T \geq t)$ and suppose $g(t+s) = g(t)g(s)$ always, then we know that $g(mt) = g(t)^m$ for all integers m, and in particular $g(m) = g(1)^m$. By taking $\lambda := -\log\big(g(1)\big)$ we know now that $g(1) = e^{-\lambda}$. It follows that for all integers and therefore all rational numbers we have that $g(q) = e^{-q\lambda}$ (to see this multiply q by its denominator and consider the functional equation we start with). This would otherwise be a situation where we literally could not conclude anything else (it could be some other exponential for irrational numbers), but luckily we know g is right-continuous and increasing, and therefore we get the result.

Lecture 16:

We define that for expectation of a random variable to exist we need $E[|X|]$ to be finite since otherwise we could run into trouble.

Lemma: $E[x] = \int_{-\infty}^{\infty} P(X \geq x)dx$ for absolutely continuous random variables with integrable densities.

Proof: $E[x] = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty}\int_{0}^{x} 1dy\, f(x)dx$

It is precisely because $E[|x|]$ is finite that we have absolute convergence allowing us to exchange the order of integrals (level 6 technical results), we get

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

Which implies the result.

Example: The expectation of the exponential distribution is $\int_0^{\infty} P(X \geq x)dx = \int_0^{\infty} e^{-\lambda x}dx = \frac{1}{\lambda}$

Recall basic properties of the normal distribution from levels 3,5,6.

Theorem: Let X have density f and let g be a continuously differentiable function with continuously differentiable inverse (so g is strictly increasing or strictly decreasing). Then the random variable g(X) has density given by $f\big(g^{-1}(x)\big)\left|\frac{d}{dx}\big(g^{-1}(x)\big)\right|$

Proof: First suppose that g is increasing. $P(g(x) \leq x) = P\big(X \leq g^{-1}(x)\big) = F\big(g^{-1}(x)\big)$ and the derivative of this is the density and is also the expression we want. Otherwise, if g is decreasing we instead get $P(g(x) \leq x) = P\big(X \geq g^{-1}(x)\big) = 1 - F\big(g^{-1}(x)\big)$ so the result follows.

Example: If X is $N(\mu, \sigma^2)$ then if $Y = Ax + B$ then the inverse of $Ax + B$ is $\frac{x-B}{A}$ so to find the density we can just plug it into the formula and we get the density of Y is

$$\frac{1}{\sqrt{2\pi(a\sigma)^2}} e^{-\left(\frac{y-(a\mu+b)}{2a^2\sigma^2}\right)^2}$$

Lecture 17:

Now consider if X is a random variable in $\mathbb{R}^n$.

We say f is its density if we can write

$$P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f(y_1, \dots, y_n) dy_1 \dots dy_n$$

We say also that $f(x_1, \dots, x_n) = \frac{\partial^N F}{\partial x_1 \dots \partial x_n}(x_1, \dots, x_n)$ assuming that F is something n times continuously differentiable (continuous to ensure we can integrate it).

To define independence for continuous random variables, we say we want $P(X_1 \leq x_1, \dots, X_n \leq x_n)$ to facrorize into $P(X_1 \leq x_1)P(X_2 \leq x_2) \dots P(X_n \leq x_n)$ for all X.

Theorem: Let the random vector X have density f and suppose the components are independent with densities $f_1, f_2, \dots, f_n$ respectively. Then $f(x_1, x_2, \dots, x_n) = f_1(x_1)f_2(x_2) \dots f_n(x_n)$. Also, if f does factorize like this in such a way that all the densities are non-negative and integrate to 1 then the components are independent.

Proof: By independence we know that $P(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1)P(X_2 \leq x_2) \dots P(X_n \leq x_n)$ and then rewrite this as an integral.

Specifically, we know that

$\int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f_1(y_1) \dots f_n(y_n) dy_1 \dots dy_n = \int_{-\infty}^{x_1} f_1(y_1) dy_1 \int_{-\infty}^{x_2} f_2(y_2) dy_2 \dots \int_{-\infty}^{x_n} f_n(y_n) dy_n$

Since that's just how multivariable integrals work.

But by the definition of f we know that $P(X_1 \leq x_1, \dots, X_n \leq x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f(y_1, \dots, y_n) dy_1 \dots dy_n$

Since this holds for every X, we know that the two functions are actually equal.

The converse part of the theorem follows exactly frrom just reversing the logic.

If we know the joint density of a bunch of variables then the density of say, X_1, which we say is the marginal density, is just

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_2 \dots dx_n$$

Because that is what we integrate to get X_1.

Example: Suppose we want $P(X + Y \leq z)$. Then we need to integrate the density over the set of all points such that $X + Y \leq z$. This would be

$$\int_{-\infty}^{\infty} f_x(x) \int_{-\infty}^{z-x} f_y(y) dy\, dx$$

$$= \int_{-\infty}^{\infty} f_x(x) \int_{-\infty}^{z} f_y(y-x)dy\,dx$$

This is equal to

$$\int_{-\infty}^{z} \int_{-\infty}^{\infty} f_x(x) f_y(y-x)dx\,dy$$

By absolute convergence. But we know that

$$P(x+y \leq z) = \int_{-\infty}^{z} Density_{x+y}(y)dy$$

By definition.

Since this is true for all z, we deduce that the density of X+Y is given by

$$f_{x+y}(y) = \int_{-\infty}^{\infty} f_x(x) f_y(y-x)dx$$

We can talk about the conditional density of X given Y=y. This is given by

$$f_{X|Y=y}(x|y) = \frac{f_{x,y}(x,y)}{f_y(y)}$$

Whenever $f_y(y) \neq 0$.

We get (from the definition)

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y=y}(x|y) F_Y(y)dy$$

The conditional expectation of x given y is written as

$$\int_{-\infty}^{\infty} f_{X|Y=y}(x|y)xdx$$

Theorem: Let X be a random variable with values in $D \subseteq \mathbb{R}^n$ with continuous density f (in fact we're fine if f is piecewise continuous). Suppose that

1. D is compact
2. g is a bijection which is continuously differentiable with non-singular derivative in an open region around D

Then the random variable $Y = g(X)$ has density $f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\det\left(\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j=1}^{n}\right)\right|$

Proof:

$P(Y \in B) = P\left(X \in g^{-1}(B)\right) = \int_{g^{-1}(B)} f_X(x)dx$. Now we know the change of variables theorem from the analysis lemmas document, so we can safely write

$$P(Y \in B) = \int_B f_X\left(g^{-1}(y)\right)\left|Det\left(D(g^{-1})(y)\right)\right|dy$$

But by definition of the density of y,

$$P(Y \in B) = \int_B f_Y(y)dy$$

Since this is true for all measurable sets, or all sets in the ϭ-algebra, it follows that the things in the two integrals are equal.

But we know that $\left|Det\left(D(g^{-1})(y)\right)\right| = \det\left(\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j=1}^{n}\right)$.

So the result follows.

Lecture 18:

Let X and Y be independent standard normal random variables. Then consider the random vector (X,Y) and suppose we want to find the distribution of (R,θ). Of course, θ should be a uniform distribution because we know that the normal is rotationally symmetric, but we can use the theorem from last lecture to calculate the density of R and θ.

We have on any compact domain not containing the origin

$$f_{R,\theta}(r,\theta) = f_{X,Y}(r\cos\theta, r\sin\theta)\left|\det\begin{pmatrix} x_r & x_\theta \\ y_r & y_\theta \end{pmatrix}\right| = \frac{1}{2\pi}e^{-\frac{r^2}{2}}\left|\det\begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}\right| = \frac{1}{2\pi}re^{-\frac{r^2}{2}}$$

We know that since this is independent of θ, the marginal density of θ must be constant so θ is uniform.

We know also that the marginal density of R is $\int_{\theta=0}^{2\pi}\frac{1}{2\pi}re^{-\frac{r^2}{2}}d\theta$ because this is the possible range of θ, and this is just $re^{-\frac{r^2}{2}}$ for r positive (0 for r negative since r never takes negative values).

Now let $X_1, X_2, \dots, X_n$ be independent identical random variables with density f. We will use capital F to denote the cumulative distribution function. Now order them from smallest to largest such that $Y_1, Y_2, \dots, Y_n$ are the same as the X's but reordered in this way. Then these Y's are called the order statistics of the sample. Now we want to find the density of $(Y_1, Y_2, \dots, Y_n)$.

We have $P(Y_1 \le x) = 1 - P(Y_1 > x) = 1 - \left(P(X_1 > x)\right)^n$ since the probability the minimum is >x is the same as the probability all X's are >x. This is know $1 - \left(1 - F(x)\right)^n$. To find the density of Y_1 we differentiate this to get $f_{Y_1}(x) = n\left(1 - F(x)\right)^{n-1}f(x)$.

Now to do the opposite,

$P(Y_n \le x)$ requires that all X's are $\le x$ so we get $F(x)^n$ so differentiating gives $nF(x)^{n-1}f(x)$.

Now we will consider $P(Y_1 \le x_1, Y_2 \le x_2, \dots, Y_n \le x_n)$.

This is $n!\,P(X_1 \le x_1, X_2 \le x_2, \dots, X_n \le x_n, X_1 \le X_2 \le \cdots \le X_n)$. This is because we need to take the probability of $P\left(X_1 \le x_1, X_2 \le x_2, \dots, X_n \le x_n,\ X_{\sigma(1)} \le X_{\sigma(2)} \le \cdots \le X_{\sigma(n)}\right)$ and sum over all permutations σ.

We note that by independence, the density is

$$f(Y_1, Y_2, \dots, Y_n) = \begin{cases} n!\,f(x_1)f(x_2)\dots f(x_n)\ if\ x_1 \le x_2 \le \cdots \le x_n \\ 0\ otherwise \end{cases}$$

Clearly, the Y's are not independent.

Now we will talk about order statistics for independent exponential distributions.

Let $X \sim Exp(\lambda), Y \sim Exp(\mu)$ and we want to find the distribution of $Z = \min(X, Y)$.

We have $P(Z \leq z) = 1 - P(Z > z) = 1 - P(X > z)P(Y > z) = 1 - e^{-(\lambda+\mu)z}$.

Therefore $Z \sim Exp(\lambda + \mu)$. It makes sense, if we are waiting for an event to happen that on average will happen λ times per second and another one that will happen μ times per second then if we want either of them to happen it will be $\lambda + \mu$ times per second.

Let $X_1, X_2, \dots, X_n$ all be $\sim Exp(\lambda)$.

Let $Y_1, Y_2, \dots, Y_n$ be the order statistics and set $Z_1 = Y_1, Z_i = Y_i - Y_{i-1}$ for $2 \leq i \leq n$. We will look for the joint density of $(Z_1, Z_2, \dots, Z_n)$. By the same reason as previous discussion we will have

$$Z_i \sim Exp\big((n + 1 - i)\lambda\big)$$

Now $Z = AY$ where A is the matrix $\begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}$. This has determinant 1 (by simple induction if we decompose by the first row). Therefore the density of Y and Z are exactly the same. They are both equal to $n!\left(\lambda e^{-\lambda y_1}\right)\left(\lambda e^{-\lambda y_2}\right) \dots \left(\lambda e^{-\lambda y_n}\right) = n!\,\lambda^n e^{-\lambda(y_1+y_2+\cdots+y_n)}$ conditional on the y's being in the right order.

Definition: The moment generating function of a distribution is defined as $m(\theta) = E\left(e^{\theta X}\right)$ when it exists. This is similar to characteristic functions: In fact $\phi(t) = m(it)$.

When the distribution has a continuous density, $m(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x) dx$.

If we take the k'th derivative of the moment generating function, then we would like to say that we get $E\left(X^k e^{\theta X}\right)$ which is just $E(X^k)$ when evaluated at 0. However, this requires justification of the derivative and expectation interchange. So lets do this.

Lemma: Suppose we have a distribution with and mgf defined on (-a, a). Then $E(|X|^n e^{tX}) < \infty$ for all $t \in (-a, a)$

Proof: It suffices to show it for the cases $X \geq 0$ and $X < 0$.

We first note that because $e^{cy} = \sum \frac{(cy)^m}{m!} \geq \frac{(cy)^k}{k!}$ for $c, y \geq 0$ we can conclude that whenever $c > 0$ and $y \geq 0$ we must have $y^k \leq \frac{k!}{c^k} e^{cy}$.

Now fix $t \in (-a, a)$ and some A with $|t| < A < a$. Then we know that $M(A)$ and $M(-A)$ are both finite.

Now using the inequality with $c = A - t$ and $y = X$ and then multiplying both sides by e^{tX} we get that $X^n e^{tX} \leq \frac{n!}{(A-t)^n} e^{Ax}$. Therefore, for the case of $X \geq 0$, finiteness follows from finiteness of the mgf itself since we just multiply by a constant.

In the case of $X < 0$ it is exactly the same, just applying the previous result to $Y = -X$.

Lemma: $m'(\theta) = E\left[Xe^{\theta X}\right]$

Proof:

Pick $\delta > 0$ such that $|t| + \delta = A$. Define $q_h(x) = \frac{e^{(t+h)x} - e^{tx}}{h}$. For each fixed x, by the definition of the ordinary derivative, $\lim_{h\to 0} q_h(x) = xe^{tx}$.

Now suppose $|h| < \delta$.

By the mean value theorem applied to the function $s \to e^{sx}$ there is some ξ between t and t+h such that $q_h(x) = xe^{\xi x}$. Therefore, $|q_h(x)| = |x|e^{\xi x}$.

If $x \geq 0$ then $\xi \leq t + \delta < A$ so $|q_h(x)| \leq |x|e^{Ax}$ and similarly if $x < 0$ then $|q_h(x)| \leq |x|e^{-Ax}$. Therefore we have $|q_h(x)| \leq |x|(e^{Ax}1_{x\geq 0} + e^{-Ax}1_{x<0})$. By the previous lemma this is always finite.

Now notice that expectation of f(x) is shorthand for integrading f(x)dμ like we did in level 6 stats. Therefore we can apply the dominated convergence theorem:

$\frac{m(t+h)-m(t)}{h} = E[q_h(X)] \to E[Xe^{tX}]$ as $h \to 0$. So the result follows.

Remark: In the above proof, we never actually used the fact that you can differentiate a moment generating function, we actually proved it provided an mgf is finite on an open interval around 0.

We will now define a distribution called the gamma distribution. It is defined by the density (for parameters n and λ)

$$\frac{e^{-\lambda x}\lambda^n x^{n-1}}{(n-1)!}, x > 0$$

We need to check that this is actually a density by integrating it.

Let $I_n = \int_0^\infty \frac{e^{-\lambda x}\lambda^n x^{n-1}}{(n-1)!}dx$. Integrating by parts gives $I_n = \int_0^\infty \frac{(n-1)e^{-\lambda x}\lambda^{n-1}x^{n-2}}{(n-1)!}dx = I_{n-1}$. When we have I_1 it is easy to prove that it integrates to 1, so f is a density.

We will now find the moment generating function of the gamma distribution. It is

$$m_{n,\lambda}(\theta) = \int_0^\infty \frac{e^{-(\lambda-\theta)x}\lambda^n x^{n-1}}{(n-1)!}dx$$

This is just $\frac{\lambda^n}{(\lambda-\theta)^n}$ as dividing by that would make the whole integral integrate to 1 by previous discussion. So we have the mgf for the gamma distribution.

Theorem: Suppose that the moment generating function of a distribution exists on some open interval about 0 which we will call (-a,a), then the distribution is uniquely determined by the moment generating function, similar to characteristic functions.

Proof:

Thank god I attended the complex analysis lectures a year early and that I went through a billion years of effort to prove the central limit theorem in level 6. Because of those two facts, the proof is fairly easy.

Consider the moment generating function as defined on (-a,a) extended to the strip S in the complex plane of complex numbers with real part between -a and a. Since $m(z) = E\left(e^{Re(z)x+iIm(z)x}\right)$, we know that $E\left(\left|e^{Re(z)x+iIm(z)x}\right|\right) = m\left(Re(z)\right)$ by properties of the exponential must also exist. Therefore, since $E\left(\left|e^{Re(z)x+iIm(z)x}\right|\right)$ is finite so is m(z).

We want to show that m defines a holomorphic function on S. So pick $z_0 \in S$. There exists an A such that 0<A<a and for h∈ℂ small enough $Re(z_0 + h) \in (-A, A)$. Consider $\frac{m(z_0+h)-m(z_0)}{h}$, then this is equal to $E\left(\frac{e^{(z_0+h)X}-e^{(z_0X)}}{h}\right)$. As h goes to 0, $\frac{e^{(z_0+h)X}-e^{(z_0X)}}{h}$ goes to Xe^{z_0X} for each fixed X.

By the fundamental theorem of calculus,

$$\frac{e^{(z_0+h)X} - e^{(z_0X)}}{h} = X\int_0^1 e^{(z_0+sh)X}ds$$

Therefore

$$\left|\frac{e^{(z_0+h)X} - e^{(z_0X)}}{h}\right| \leq |X|\int_0^1 \left|e^{(z_0+sh)X}\right|ds = |X|\int_0^1 e^{Re(z_0+sh)X}ds$$

But this is bounded by $|X|(e^{AX}) = |X|\left(e^{AX}1_{\{x\geq 0\}} + e^{AX}1_{\{x<0\}}\right) \leq |X|\left(e^{AX}1_{\{x\geq 0\}} + e^{-AX}1_{\{x<0\}}\right)$ which is integrable by finiteness of m(A) and m(-A). Therefore, we can juse the dominated convergence theorem again, so we have a holomorphic function with derivative $E(Xe^{z_0X})$.

Now by the identity theorem, if the moment generating functions are equal, so are the characteristic functions, and hence so are the distributions (level 6 stats).

Lecture 19:

Claim: If $X_1, X_2, \dots, X_n$ are independent identically distributedrandom variables then $m_{X_1+X_2+\cdots+X_n}(\theta) = m_{X_1}(\theta)m_{X_2}(\theta)\dots m_{X_n}(\theta)$

Proof: This is because $E\left(e^{\theta(x_1+x_2+\cdots+x_n)}\right) = E\left(e^{\theta x_1}\right)E\left(e^{\theta x_2}\right)\dots E\left(e^{\theta x_n}\right)$ by indepdence.

Example: Consider the sum of 2 gamma random variables.

If the parameter is n and m and both λ then the moment generating function is $\left(\frac{\lambda}{\lambda-\theta}\right)^{n+m}$. Since the moment generating function determines the distribution, we get that the sum of a gamma n and a gamma m is a gamma n+m.

A gamma n distribution is the sum of n gamma 1 distributions.

One could generalize this to n non-integers by replacing (n-1)! with $\Gamma(n)$ which is defined to be $\int_0^\infty e^{-x}x^{n-1}dx$ – this is a generalization of the factorial to non-integers.

Definition: The cauchy distribution is a probability distribution with density $\frac{1}{\pi(1+x^2)}$ for all real numbers x. This is a genuine distribution because it integrates to 1, but it does not have finite expectation nor finite variance. We can try to calculate its moment generating function, but the integral will blow up for all values except 0 where it will give 1.

If X has the cauchy distribution then 2X and 3X will have the same mgf but will be different. Therefore the assumption that the mgf converges on an open interval around 0 is necessary.

Example: Let X be $N(\mu, \sigma^2)$. Then the density is given by $\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

$$m(\theta) = E\left(e^{\theta x}\right) = \int_{-\infty}^{\infty} e^{\theta x}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(\theta x - \frac{(x-\mu)^2}{2\sigma^2}\right)dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{x^2}{2\sigma^2} + \frac{x}{\sigma^2}(\mu+\theta\sigma^2) - \frac{\mu^2}{2\sigma^2}\right)dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}\left(x-(\mu+\theta\sigma^2)\right)^2 + \mu\theta + \frac{\theta^2\sigma^2}{2}\right)dx$$

We have some terms not depending on x so

$$m(\theta) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}\left(x-(\mu+\theta\sigma^2)\right)^2\right)dx * \exp\left(\mu\theta + \frac{\theta^2\sigma^2}{2}\right)$$

Note that the first integral is just 1 since it is the integral of the density of some normal distribution. Therefore,

$$m(\theta) = \exp\left(\mu\theta + \frac{\theta^2\sigma^2}{2}\right)$$

Now let X be $N(\mu, \sigma^2)$ and Y be $N(\nu, \tau^2)$. By independence, the mgf of the sum will be

$$\exp\left((\mu+\nu)\theta + \frac{\theta^2(\sigma^2+\tau^2)}{2}\right)$$

This gives an alternative way to see the "Normal + Normal = Normal" result that we saw in level 6.

Let X be a random vector in $\mathbb{R}^n$. Then its mgf is given by $E\left(e^{\theta^T X}\right) = E\left(e^{\theta . X}\right)$.

We define a multidimensional normal to be $A + BZ$ where A is some vector, B is some matrix, and Z is a standard normal. It is easy to see from this definition that the dot product of this with any vector is normal. You can see more about multidimensional normals if you go to the misc results section and look at how the Multivariate CLT is proven. In that document we also prove that these definitions are equivalent. In this course, however, we will use the "every projection is normal" definition, and we won't use the fact that the definitions are equivalent.

Proposition: If X is an n-dimensional normal, A an n*n matrix and B a vector, then Ax+B is another n dimensional normal.

Proof: If u is any vector then $u^T(AX+B) = (A^Tu)^TX + (A^Tu)^Tb$ but because X is a multivariate normal, by definition so is $(A^Tu)^TX$, and $(A^Tu)^Tb$ is just a constant so we're done.

Define the expectation of a random vector by the expectation of each component.

The variance of an n-dimensional normal is $E[(x-\mu)(x-\mu)^T]$. This is the matrix which contains all the pairwise covariances of the components. This is called the covariance matrix, and it is a symmetric matrix.

Theorem: Characteristic function injectivity (ie the fact that it determines the distribution) holds for random vectors, where the multivariate characteristic function is given by $E\left(e^{i\theta^T X}\right)$

Proof:

If you've read the proof of the Cramer Wold theorem in the misc results section, you can skip this proof, since it is an exact copy of the proof of one of the lemmas from there. I just put it here to make the levels self contained.

A random vector is defined by a probability distribution function in $\mathbb{R}^k$, ie sets of k real numbers. We define the cf of a random vector Y to be a function that takes in a vector t and outputs $E\left(e^{i(t.Y)}\right)$. If the space is d dimensional, we define

$$I_\varepsilon := \frac{1}{(2\pi)^d}\int_{R^d}\phi_y(t)e^{-\varepsilon|t|^2}\prod_{j=1}^{d}\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}dt$$

Where that big scary looking thing just means product.

$$=\frac{1}{(2\pi)^d}\int_{R^d}\int_{R^d}e^{i(t.y)}\prod_{j=1}^{d}\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}d\mu e^{-\varepsilon|t|^2}dt$$

Note: Each $\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}$ is bounded for the same reasons as before (ie in level 6 stats), so we have the product of bounded things times a thing which integrates to 1 in the inner dy integral, so that's bounded. Now in total we have a bounded thing times the integral of $e^{-\varepsilon|t|^2}$, which is finite, so we have the conditions to swap the integrals around.

Since $e^{-\varepsilon|t|^2}$ is just the product of $e^{-\varepsilon(t_j{}^2)}$, we can simplify I_ε as follows:

$$=\frac{1}{(2\pi)^d}\int_{R^d}\int_{R^d}e^{i(t.y)}\prod_{j=1}^{d}\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}e^{-\varepsilon|t|^2}dtd\mu$$

$$=\int_{R^d}\int_{R^d}\prod_{j=1}^{d}\frac{1}{2\pi}e^{i(t_jy_j)}\frac{e^{-ia_jt_j}-e^{-ib_jt_j}}{it_j}e^{-\varepsilon(t_j)^2}dtd\mu$$

We know what each term looks like from earlier, so we end up with

$$=\int_{R^d}\int_{R^d}\prod_{j=1}^{d}\frac{1}{2}\left(\operatorname{erf}\left(\frac{y_j-a_j}{2\sqrt{\varepsilon}}\right)-\operatorname{erf}\left(\frac{y_j-b_j}{2\sqrt{\varepsilon}}\right)\right)dtd\mu$$

Which in the limit vanishes exactly when we are outside the (a,b) high dimensional rectangle and goes to 1 when we are inside it, for the same reasons – Each term goes to 1 or 0 and the product is 1 only when all terms go to 1.

Now at last we have, because of dominated convergence again, the same result for random vectors: the characteristic function determines the distribution.

Theorem: MGF injectivity holds for random vectors

Proof:

Suppose the mgfs of X and Y agree on some open neighbourhood of 0.

Let u be a random vector, then the mgf of the 1D random variable u.X with input s is $E\left(e^{s(u.x)}\right)$ which is the moment generating function of X evaluated at su. Similarly, $m_{u.Y}(s)=m_Y(su)$. Since we are in an

open neighbourhood of 0, for each fixed u there exists δ>0 such that su is in U whenever $|s| < \delta$. It therefore follows that $u.X = u.y$ for each u. Therefore, the characteristic functions of u.X and u.Y agree since they have the same distribution. Therefore, if we evaluate the characteristic functions at 1, they will agree. This means that for every u, $E\left(e^{iu.X}\right) = E\left(e^{iu.Y}\right)$. So by the previous lemma, X and Y have the same distribution.

Claim:

Let $(X_1, X_2, \dots, X_n)$ be a random vector, then we can write $m(\theta) = E\left(e^{\theta.X}\right)$ and this is always true, but then we can further write this as $E\left(e^{\theta.X}\right) = \prod_{i=1}^{n} E\left[e^{\theta_i X_i}\right]$ if and only if $X_1, X_2, \dots, X_n$ are independent random variables.

Proof:

If they are independent the claim is trivial. The non-trivial part is showing that $E\left(e^{\theta.X}\right) = \prod_{i=1}^{n} E\left[e^{\theta_i X_i}\right]$ implies independence. We note that there exists some distribution which is $(X_1, X_2, \dots, X_n)$ with these independent, and this distribution will have mgf $E\left(e^{\theta.X}\right) = \prod_{i=1}^{n} E\left[e^{\theta_i X_i}\right]$, so since mgf's uniquely determine the distribution when they are defined on an open interval around 0, the result follows.

Proposition:

Suppose that m is the mgf of a random vector, then the following hold:

i) $\frac{\partial^r m}{\partial \theta_i^r}|_{\theta=0} = E[X_i^r]$

ii) $\frac{\partial^{r+s} m}{\partial \theta_i^r \partial \theta_j^s}|_{\theta=0} = E\left[X_i^r X_j^s\right]$

And the obvious generalization also holds to more than 2 variables.

Proof:

By the 1D theorem, if we take the r'th derivative of $m_{X_i}(t)$ when t=0 we get $E[X_i^r]$. But also, we note that $m_{X_i}(t) = m(te_i)$ so part (i) of the claim follows – differentiating with respect to θ_i is the same as differentiating $m_{X_i}(t)$.

For the general case, pick some cube about the origin such that the mgf is finite and suppose its side length is 2a. Then for every sign vector $\varepsilon = (\pm 1, \pm 1, \dots, \pm 1)$ we have $E[e^{\varepsilon a.X}] < \infty$. Now define $Y \coloneqq e^{a|X_1| + \cdots + a|X_d|}$ where d is the number of variables we are looking at. Then

$$Y = \prod_{k=1}^{d} e^{a|X_k|} \leq \prod_{k=1}^{d} e^{a|X_k|} + e^{-a|X_k|} = \sum_{\varepsilon \in \{\pm 1\}^d} e^{\varepsilon a.X}$$

Therefore by taking expectations $E[Y] < \infty$.

Let $(A_1, A_2, \dots, A_d)$ be a vector of non-negativeintegers such that we want to prove that

$$\frac{\partial^{\sum A_i} m}{\partial \theta_1^{A_1} \partial \theta_2^{A_2} \dots \partial \theta_d^{A_d}}|_{\theta=0} = E\left[\prod X_i^{A_i}\right]$$

By the same inequality as one we used in an earlier proof, for each non-negative real u and non-negative integer n we have $u^n \leq n!\, e^u$. If $u = a|x_k|$ then $|x_k|^{A_k} \leq \frac{A_k!}{a^{A_k}} e^{a|x_k|}$. We can multiply over all such k to get that $\left|\prod x_i^{A_i}\right| = \prod_{k=1}^{d} \frac{A_k!}{a^{A_k}} e^{a\sum|x_i|}$

Therefore there is a constant depending only on A and the distribution such that $\left|\prod x_i^{A_i}\right| \leq CY$, and therefore $E\left[\left|\prod x_i^{A_i}\right|\right] < \infty$.

Fix $\theta = (\theta_1, \theta_2, \dots, \theta_d)$ inside the a-cube. Then $e^{\theta.x} = \prod_{k=1}^{d} e^{\theta_k X_k}$. Now $e^{\theta_k X_k} = \sum_{n_k=0}^{\infty} \frac{(\theta_k X_k)^{nk}}{(n_k)!}$. Write $\theta^A = \prod \theta_i^{A_i}$ and $A! = \prod A_i!$. Then $e^{\theta.x} = \sum_{A\in\mathbb{N}_0^d} \frac{\theta^A X^A}{A!}$.

Now $\sum_{A\in\mathbb{N}_0^d} \left|\frac{\theta^A X^A}{A!}\right| = e^{\sum_{k=1}^{d}|\theta_k||X_k|} \leq e^{a\sum_{k=1}^{d}|X_k|} = Y$ so we have absolute convergence and therefore by the dominated convergence theorem (since an expectation is really an integral with respect to a probability measure) we have $m(\theta) = E\left[e^{\theta.X}\right] = \sum_{A\in\mathbb{N}_0^d} \frac{\theta^A E[X^A]}{A!}$

Now we will try to find the mixed partial derivatives at 0.

Since this is a power series and it is convergent in a neighbourhood around 0, we can differentiate termwise if we just pick some ennumeration of $\mathbb{N}_0^d$. Also it is smooth enough that we have symmetry of mixed partials (as we can differentiate it as many times as we want).

Note that if θ is just an ordinary one-variable function, $D^b(\theta^a) = 0$ at $\theta = 0$ unless a=b, but also $D^a(\theta^a) = a!$. So $D^B f(0) = \frac{E[X^B]}{B!} B! = E[X^B]$ by termwise differentiation, where B is a vector of non-negative integers B_i, and D^B means to partial differentiate the i'th components each with respect to B_i. And now we can see why the result is true.

Lecture 20:

Note that if X is a multivariate normal with mean μ then

$$E[u.X] = E\left[\sum u_i X_i\right] = \sum u_i E[X_i] = u.\mu$$

$$Var[u.X] = Var\left[\sum u_i X_i\right] = \sum_{i,j=1}^{n} u_i Cov(X_i, X_j) u_j = u^T V u$$

Where V is the covariance matrix of X, because to add variances we need to add variances and all the covariances.

It follows that V is positive definite, ie $u^T V u \geq 0$ always, since it is the variance of something. In other words, V is non-negative-definite.

Now lets calculate the mgf of a multivariate normal.

Now $u.X \sim N(u.\mu, u^T V u)$ so $m(u) = m(u.X) = \exp\left(u.\mu + \frac{u^T V u}{2}\right)$ so μ and V determine the distribution.

Let $\mu \in \mathbb{R}^n$ and V be a symmetric non-negative-definite matrix. Then it is possible to construct a multivariate normal $N(\mu, V)$. Here is how to do this:

We can write $X = \mu + \sigma Z$ where Z is $N(0, I)$ but now we want σ to be the square root of a matrix which is dodgey. But luckily, by hypotheses, our matrix is diagonalizable and has non-negative eigenvalues.

Therefore, if our matrix is $U^T DU$ then we can write $\sigma = U^T D^{\frac{1}{2}} U$ where we just square root each element of the diagonal matrix D.

Now we want to verify that $\mu + \sigma Z \sim N(\mu, \sigma^2)$ in the multivariate setting. But this is easy. To verify that the variance is correct note that $E[(X-\mu)(X-\mu)^T] = E[\sigma Z Z^T \sigma^T] = E[\sigma\sigma^T] = \sigma^2$. And it is a normal because it is the linear transformation of a known normal.

Now we will try to compute the density of $N(\mu, \sigma^2)$.

In the 1 dimensional case the density is $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Case 1: None of the eigenvalues of V are 0

In this case V is positive definite so all eigenvalues are >0. We can write $Z = \sigma^{-1}(X-\mu)$ and so by a previous theorem about how to change variables,

$$f_X(X) = f_Z\left(\sigma^{-1}(X-\mu)\right)|Det(\sigma^{-1})| = f_Z(z)|Det(\sigma^{-1})| = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{|z|^2}{2}\right)|Det(\sigma^{-1})|$$

$$f_X(X) = \frac{1}{\left(2\pi^n Det(V)\right)^{\frac{1}{2}}} \exp\left(-\frac{(X-\mu)^T V^{-1}(X-\mu)}{2}\right)$$

Since $Det(V) = \det(\sigma^2)$ and $|z|^2 = Z^T Z = (X-\mu)^T(\sigma^{-1})^T(\sigma^{-1})(X-\mu) = (X-\mu)^T V^{-1}(X-\mu)$

Case 2: V is not invertible

Here there is no density its like we have a normal on some lower dimensional space. However, if you look at the normal as only being in that space, then the density is

$$\frac{1}{\left(2\pi^n Det(U)\right)^{\frac{1}{2}}} \exp\left(-\frac{(X-\mu)^T U^{-1}(X-\mu)}{2}\right)$$

Where U is a diagonal matrix with all the non-zero eigenvalues and we call its size n*n.

Lecture 21:

Let X be a multivariate normal with mean μ and variance V.

Claim: If $X_1, \ldots, X_n$ are independent then V is diagonal

Proof: Because independent implies zero covariance

Claim: If V is a diagonal matrix then a normal with covariance matrix V has independent components

Proof: Earlier we proved multivariate normals are uniquely determined by their covariance matrix. Since there is some normal with independent components with covariance matrix V, the result follows.

Claim: Let σ_1, σ_2 be >0 and let $p \in [-1,1]$. Then the matrix $\begin{pmatrix} \sigma_1^2 & p\sigma_1\sigma_2 \\ p\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ is non-megative definite.

Proof: If we expand out U^TVU we get (by checking the cases p>0, p<0), that

$$U^TVU = (1-p)(\sigma_1^2u_1^2 + \sigma_2^2u_2^2) + (\sigma_1u_1 + \sigma_2u_2)^2 = (1+p)(\sigma_1^2u_1^2 + \sigma_2^2u_2^2) - (\sigma_1u_1 - \sigma_2u_2)^2$$

Is positive.

Now note that $\begin{pmatrix} \sigma_1^2 & p\sigma_1\sigma_2 \\ p\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ is always the covariance matrix of a 2*2 normal because p can be thought of as the pmcc of the two components, which is always between -1 and 1.

Now suppose we have a 2D gaussean vector (which is just a multivariate normal) and p is not -1 or 1 so its covariance matrix has non-zero determinant.

Now we will consider $E[X_2|X_1]$. We can write $X_2 = X_2 - aX_1 + aX_1$, where there exists some value of a such that $X_2 - aX_1$ is independent of X_1.

Consider the vector $\begin{pmatrix} X_1 \\ X_2 - aX_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -a & 1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. Then this is a multivariate normal. We just need to find a such that $Cov(X_1, X_2 - aX_1) = 0$. This is $Cov(X_1, X_2) - aVar(X_1)$, since we can decompose the covariance by the second component. So take $a = \frac{Cov(X_1,X_2)}{Var(X_1)}$.

Now by independence, $E[X_2|X_1] = E[X_2 - aX_1|X_1] + aX_1 = E[X_2 - aX_1] + aX_1 = \mu_2 - a\mu_1 + aX_1$.

Definition: As in level 6, we define convergence in distribution to mean pointwise convergence of the cumulative distribution function at all points where the cdf is continuous.

We say convergence in probability means: A sequence of random variables $X_n \to X$ in probability if for every $\varepsilon \to 0$, $P(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$

This makes sense if X is a constant, however, this can still work with X a random variable if X_n depends on X. For example, if X is any random variable and $X_n = X + \frac{1}{n}$ then clearly the definition is satisfied.

Theorem (Weak law of large numbers): If X_n is a sequence of identical random variables with finite expectation and finite variance then $\frac{\sum X_n}{n}$ converges in probability to $E[X_n]$. Write $S_n = \sum X_n$.

Lecture 22:

Proof:

We want to show that $P\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \to 0$ for all ε. By chebyshevs inequality, this is

$$P(|S_n - n\mu| > n\varepsilon) \leq \frac{E[|S_n - n\mu|^2]}{n^2\varepsilon^2} = \frac{Var(S_n)}{n^2\varepsilon^2} = \frac{nVar(X_n)}{n^2\varepsilon^2} \to 0$$

Definition: If X_n is a sequence of random variables and X is another random variable we say $X_n \to X$ almost surely if $P\left(\lim_{n\to\infty} X_n = X\right) = 1$. To say $\lim_{n\to\infty} X_n = X$ what that means is for all ε>0 there exists a natural number N such that for all n>N, $|X_n - X| < \varepsilon$.

Claim: If $X_n \to 0$ almost surely then $X_n \to 0$ in probability

Proof:

$$P(|X_n| \leq \varepsilon) \geq P\left(\bigcap_{m=n}^{\infty} \{|X_m| \leq \varepsilon\}\right)$$

Let $A_n = \bigcap_{m=n}^{\infty}\{|X_m| \leq \varepsilon\}$, then $A_n \subseteq A_{n+1}$ so $P(A_n) \to P(\cup A_n) = P(\cup \bigcap_{m=n}^{\infty}\{|X_m| \leq \varepsilon\})$ so therefore $\lim_{n\to\infty} P(|X_n| \leq \varepsilon) \geq P(\forall \varepsilon > 0, |X_m| \leq \varepsilon \; for\; all\; m\; sufficiently\; large) = 1$ because we are assuming almost sure convergence.

Theorem (Strong law of large numbers):

If X_n is a sequence of identical random variables with finite expectation and finite $E[x^4]$ then $\frac{\sum X_n}{n}$ converges almost surely to $E[X_n]$. Write $S_n = \sum X_n$.

Proof:

Set $Y_i = X_i - \mu$. Then $E[Y_i] = 0$ and $E[Y_i^4] < \infty$ as it is a sum of finite multiples of $E[X^a]$ with $a \leq 4$.

We will show that with probability 1, $\sum \left(\frac{S_n}{n}\right)^4 < \infty$ with probability 1 so that the terms go to 0, so we will be done.

$$E\left[\sum \left(\frac{S_n}{n}\right)^4\right] =_? \sum_{n=1}^{\infty} \frac{1}{n^4} E[S_n^4] = \sum_{n=1}^{\infty} \frac{1}{n^4} E[(Y_1 + Y_2 + \cdots + Y_n)^4]$$

We will justify swapping the sum and expectation later.

$$E[(Y_1 + Y_2 + \cdots + Y_n)^4] = E\left[\sum_{i=1}^{n} Y_i^4 + O(n^4)\; terms\; like\; Y_i^2 Y_j^2, Y_i Y_j^3, Y_i^2 Y_j Y_k, Y_i Y_j Y_k Y_l\right]$$

The expectation of $Y_i Y_j^3$ factorizes by independence and $E[Y_i]$ is 0 so that gives 0, and similarly for two of the other terms. And there are only $O(n^2)$ terms like $Y_i^2 Y_j^2$.

By independence this is the same as $(E[Y_1^2])^2 < \infty$. Therefore

$$E[S_n^4] = nE[Y_1^4] + O(n^2)E[Y_1^2]^2$$

So

$$\sum E\left[\left(\frac{S_n}{n}\right)^4\right] = \sum \frac{Finite * O(n^2)}{n^4} = \sum O(n^{-2}) < \infty$$

Note that $E[x^a]$ converging implies expectation of lower powers converges because we can split it into the part of the distribution between -1 and 1 and the other part, and each has finite $E[x^a]$, the first part has finite $E\left[x^{anything}\right]$ and in the second part we are taking something strictly smaller. Hopefully this makes sense, it's just a technical detail.

Now, for doing the sum swap thing from earlier, we just proved the sum is dominated by something so we can apply the dominated convergence theorem.

The central limit theorem helps us analyze what the fluctuation around the mean will look like in these cases.

Theorem (Central limit theorem): Let X_i be random variables that are independent and identically distributied and let their sum be S_n and suppose $E[X_1] = \mu, Var[X_1] = \sigma^2 < \infty$. Then $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ converges in distribution to $N[0,1]$.

Proof: I've never felt so satisfied to say "See level 6" in my life.

Lecture 23:

Claim: Convergence in probability implies convergence in distribution

Proof:

Let x be a continuity point of F_X. We want to show that $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ given only that we have convergence in probability. Let ε>0. Then

$$\{X \le x - \varepsilon\} \subseteq \{X_n \le x\} \cup \{|X_n - X| > \varepsilon\}$$

$$\{X_n \le x\} \subseteq \{X \le x + \varepsilon\} \cup \{|X_n - X| > \varepsilon\}$$

Now we will take probabilities of both sides.

$$P\{X \le x - \varepsilon\} \le P\{X_n \le x\} + P\{|X_n - X| > \varepsilon\}$$

But note that $P\{|X_n - X| > \varepsilon\} \to 0$ by convergence in probability so

$$F_x(x - \varepsilon) \le \text{liminf}_{n\to\infty} P(X_n \le X) = \text{liminf}_{n\to\infty} F_{X_n}(x)$$

By the same argument, $\text{limsup}_{n\to\infty} F_{X_n}(x) \le F_X(x + \varepsilon)$. So

$$F_x(x - \varepsilon) \le \text{liminf}_{n\to\infty} F_{X_n}(x) \le \text{limsup}_{n\to\infty} F_{X_n}(x) \le F_X(x + \varepsilon)$$

Now by continuity of F at x, we get that

$$\lim_{\varepsilon\to 0} F_x(x - \varepsilon) = \lim_{\varepsilon\to 0} F_x(x + \varepsilon) = \lim_{\varepsilon\to 0} F_x(x)$$

Lecture 24:

Suppose we have a thing where a bunch of people vote yes with probability p and no with probability 1-p independently.

Now sample N individuals (where N is large) and record their votes. Let X_i be 1 if the i'th individual voted yes and 0 otherwise. Set $S_N = X_1 + X_2 + \cdots + X_N$, ie the number of yes votes. Now lets denote $\frac{S_N}{N}$ by $\hat{p}_N$ which converges to p almost surely by the strong law of large numbers. And $S_N \sim B(N, p)$. And by CLT, S_N can be approximated by $N\big(Np, Np(1-p)\big)$.

We will investigate how large N needs to be to ensure that p has an error of $\pm 4\%$ with $> 99\%$ probability.

Write $S_N \approx Np + \sqrt{Np(1-p)}N(0,1)$. We will find N such that

$$P(|\hat{p}_n - p| \ge 0.04) \le 0.01$$

Now $\frac{S_N}{N} \approx p + \sqrt{\frac{p(1-p)}{N}}N(0,1)$. Therefore $|\hat{p}_n - p| \approx \sqrt{\frac{p(1-p)}{N}}|Z|$ where Z is a standard normal. By symmetry of the normal distribution, $P(|Z| \geq z) = 2\big(1 - P(Z \leq z)\big)$. Now if $z \approx 2.58$, by numerical data, $P(|Z| \geq z) = 0.01$. So we want

$$\sqrt{\frac{p(1-p)}{N}}2.58 \leq 0.04$$

Note that $\sqrt{p(1-p)} \leq \frac{1}{2}$ always. We get that $N \geq 1040$ is sufficient. Note that the CLT gives an approximation so it is not clear if this is actually sufficient, but it's probably good enough.

Definition: If a random variable's cumulative distirbution function does not have an inverse, define its generalized inverse as $G(u) = \inf\{x \in \mathbb{R}: u \leq f(x)\}$. So if our cumulative is something simple like a Bernoulli distribution with parameter ½, then G will be 0 if u is between 0 and ½ and 1 if u is between ½ and 1. So we see that this is a sensible definition: We will make this more precise.

Proposition: $G(u) \leq x$ if and only if $u \leq f(x)$

Proof: If $u \leq f(x)$ then it is by definition that $G(u) \leq x$. Now suppose that $G(u) \leq x$. Since $G(u)$ is the infimum of $x \in \mathbb{R}: u \leq f(x)$, there exists a decreasing sequence $x_n \to G(u)$ such that $u \leq F(x_n)$. Since F is right continuous, $F(x_n) \to F\big(G(u)\big)$. Therefore, since u is a lower bound for the sequence $F(x_n)$, $u \leq F\big(G(u)\big)$. Therefore, is $G(u) \leq x$ then $u \leq F\big(G(u)\big) \leq F(x)$ because F is increasing.

This allows us to simulate any random variable just given how to simulate a Uniform (0,1) distribution, by just taking a U(0,1) and applying to the result the generalized inverse of the distribution we want.

Since it is difficult to find an algorithm to find the generalized inverse of the normal (although nowadays we have good ways to approximate it numerically), we will show a trick to simulate a normal distribution using only a uniform distribution.

One idea if to generate a bunch of U(0,1)'s and use the CLT to approximate it. But here is a nicer idea.

Let X and Y be standard normals and consider (X,Y) as a 2D random variable. Recall that

$R = \sqrt{X^2 + Y^2}$, $\theta = U(0,2\pi)$, and that R has density $re^{-\frac{r^2}{2}}$. This is something which we know how to integrate – it has a cumulative distribution function $\begin{cases} 0: r < 0 \\ 1 - e^{\frac{r^2}{2}}: otherwise \end{cases}$.

Now we take the generalized inverse and get $R = \sqrt{-2\log(V)}$, and it is much easier to calculate log. This allows us to get R, and it is easy to get θ as we just need to multiply the result of U(0,1) by 2π. We can now take $R * cos(\theta)$ then we are done.

We can generalize uniform random variables to say that for a measurable set A in $\mathbb{R}^d$, we can come up with a random variable with density $f(x) = \frac{1(x \in A)}{|A|}$.

We can get a uniform on $[0,1]^d$ by taking a uniform in each coordinate. We can generate a random variable with density any subset of this by just generating these until we get one in the set we want.