Definitions and Theorems

Probability Theory, Fall 2022

by:

Padmini Mukkamala

Budapest University of Technology and Economics

Last updated: December 4, 2022

Contents

Sample questions

Sample theory questions (from Fall 2021)

1. State the following definition/theorem.

   (a) When are random variables $X_1, X_2, ..., X_n$ said to be (jointly) independent? $(n > 0)$

   (b) State the linear regression line of Y in terms of X, giving the coefficients in terms of covariance, standard deviation and expected value of X and Y.

   Solution: Random variables $X_1, X_2, ..., X_n$ are said to be (jointly) independent if for every $x_1, x_2, ..., x_n \in \mathbb{R}$, the events $\{X_1 \leq x_1\}, ..., \{X_n \leq x_n\}$ are independent.

   If $Var(X), Var(Y)$ and $Cov(X, Y)$ are finite, and $Var(X) \neq 0$, then the linear regression line of Y in terms of X is defined as $\beta X + \alpha$, where,

   $$\beta = \frac{Cov(X,Y)}{Var(X)}, \alpha = E(Y) - \beta E(X)$$

   .

2. State the following definition/theorem.

   (a) What is the correlation coefficient of random variables X and Y in terms of covariance and standard deviations of X and Y, and under what conditions is it defined?

   (b) What conditions must a Riemann integrable function $f : \mathbb{R} \to \mathbb{R}$ statisfy so that there is a random variable X such that f is its probability density function?

   Solution: If $Cov(X, Y), Var(X), Var(Y)$ are finite and $\sigma_X \neq 0$ and $\sigma_Y \neq 0$, then the correlation $\rho(X, Y)$ is denifed as,

   $$\rho(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

   .

   For f to be a density function, f must be non-negative and

   $$\int_{-\infty}^{\infty} f(x)dx = 1$$

   .

3. State the following definition/theorem.

   (a) Define the expected value of a simple random variable.

   (b) Under what conditions can we express the expected value of the product of two random variables X and Y in terms of $E(X)$ and $E(Y)$? What is the relation under those conditions?

   Solution: The expected value for a simple random variable X is given by,

   $$E(X) = \sum_{k \in Range(X)} k \cdot P(X = k)$$

   If X and Y are independent and if $E(XY), E(X)$ and $E(Y)$ exist, then,

   $$E(XY) = E(X)E(Y)$$

   .

4. State the following definition/theorem.

   (a) Let (X, Y) be a continuous random variable vector. What is the condition density function of Y given X?

   (b) Let X be a simple (discrete) random variable and $g : \mathbb{R} \to \mathbb{R}$ a function, such that $E(g(X))$ exists. State $E(g(X))$ using the distribution of X.

Solution:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,u)du}$$

for those values $x, y \in \mathbb{R}$, where $f_X(x) \neq 0$. It is defined as $f_{Y|X}(y|x) = 0$ if $f_X(x) = 0$.

$$E(g(X)) = \sum_{k \in Range(X)} g(k) \cdot P(X = k)$$

Lecture 1

Definition of Probabilistic measure

De Morgan's Laws for two events: $\overline{A \cup B} = \overline{A} \cap \overline{B}$ and $\overline{A \cap B} = \overline{A} \cup \overline{B}$.
De Morgan's Laws for many events: $\overline{\bigcup_{i=1}^{\infty} A_i} = \bigcap_{i=1}^{\infty} \overline{A_i}$ and $\overline{\cap_{i=1}^{\infty} A_i} = \cup_{i=1}^{\infty} \overline{A_i}$.

Mutually exclusive: Two events A and B are said to be mutually exclusive if $A \cap B = \emptyset$.

In both the definitions that follow, the word 'proper' signifies a sigma algebra, or in other words, some conditions the collection of subsets must satisfy. This, however, is not in syllabus, and we only state here that when Ω is an infinite set and if we are not careful about the collection of subsets we consider, then weird paradoxes might arise when we define the probability measure.

Probabity measure: Given a sample space Ω and a 'proper' collection $\mathcal{F}$ of subsets (events) of Ω, a measure $P : \mathcal{F} \to [0, 1]$ is said to be a probability measure if,

- $P(\Omega) = 1$
- (sigma additivity) For any countable collection of mutually exclusive events $A_1, A_2, ... \in \mathcal{F}$, $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Note: Sigma additivity is defined for a collection of mutually exclusive events, that is, for any $i \neq j$, $A_i \cap A_j = \emptyset$.

Probability space: The triple $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ a 'proper' collection of events, and P a probability measure on $\mathcal{F}$, is said to be a probability space.

Some consequences of the definition of Probability measure:

- $P(\emptyset) = 0$
- $P(\overline{A}) = 1 - P(A)$
- If $A \subseteq B$, then $P(A) \leq P(B)$
- $P(A \cap B) + P(A \cap \overline{B}) = P(A)$

Inclusion-Exclusion or Poincare's Formula.
For two events: $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$.
For three events: $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_2 \cap A_3) - P(A_3 \cap A_1) + P(A_1 \cap A_2 \cap A_3)$.
For many events:

$$P(\cup_{i=1}^{n} A_i) = \sum_{i} P(A_i) - \sum_{i_1<i_2} P(A_{i_1} \cap A_{i_2}) + ... + \sum_{i_1<i_2<...<i_r} (-1)^{r+1} P(A_{i_1} \cap ... \cap A_{i_r}) + ...$$

$$... + (-1)^{n+1} P(A_1 \cap A_2 \cap ... \cap A_n)$$

Boole's Inequality: For any collection $A_1, ...A_n$ of events and a probability measure P,

$$P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$$

Limit properties.
Property 1: Given a sequence $A_1 \subseteq A_2 \subseteq ... \subseteq A_n \subseteq ...$ of increasing events and a probability measure P, then,

$$P(\cup_i A_i) = \lim_{n\to\infty} P(A_n)$$

Property 2: Given a sequence $A_1 \supseteq A_2 \supseteq ... \supseteq A_n \supseteq ...$ of decreasing events and a probability measure P, then,

$$P(\cap_i A_i) = \lim_{n\to\infty} P(A_n)$$

Lecture 2

Conditional Probability

Conditional Probability: Given two events A, B and a probability measure P. If $P(B) > 0$, that is the probability of the event B is non-zero, then the conditional probability of the event A given that B is true is defined as $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Independence: Given two events A, B and a probability measure P, we say that the events are independent if $P(A \cap B) = P(A)P(B)$.
Note: if the probabilities of the events are non-zero, then for independent events, $P(A|B) = P(A)$, and $P(B|A) = P(B)$, but we don't take this as the definition of independence because the conditional probabilities are not always defined.

Lemma: If two events A, B are independent, then A and $\overline{B}$ are also independent.

Multiplication Rule (two events): For any two events A_1, A_2, not necessarily independent, if the conditional probability $P(A_2|A_1)$ exists, then: $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1)$.
Note: it is good to think of the events A_1 and A_2 occuring chronologically in that order and this rule is useful when the conditional probabilities are defined and much more straightforward to analyse than the probabilities of the intersection of the events.

Multiplication Rule (many events): For any events $A_1, A_2, ..., A_n$, not necessarily independent, if for all $1 < i \leq n$ the conditional probability $P(A_i|A_{i-1} \cap ... \cap A_1)$ exists, then:

$$P(\cap_{i=1}^{n} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)...P(A_n|A_1 \cap ... \cap A_{n-1})$$

Note: as before it is good to think of the events $A_1, ..., A_n$ as occuring chronologically in that order.

Pairwise Independence: Events $A_1, A_2, ..., A_n$ are said to be pairwise independent if $\forall i \neq j$, the events A_i and A_j are independent, that is, $P(A_i \cap A_i) = P(A_i)P(A_j)$.

Total Independence: Events $A_1, A_2, ..., A_n$ are said to be totally independent if for every $I = \{i_1, i_2, ..., i_k\} \subset \{1, 2, ..., n\}$, $P(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_k}) = P(A_{i_1})P(A_{i_2})...P(A_{i_k})$.

Partition: A partition of a sample space Ω is a collection of mutually exclusive events whose union is Ω. That is, events $A_1, A_2, ..., A_n$ are said to be a partition of Ω if $\forall i \neq j$, $A_i \cap A_j = \emptyset$ and $\cup_{i=1}^{n} A_i = \Omega$.

Law of Total Probability: Given a partition $A_1, A_2, ..., A_n$ of a sample space Ω such that $P(A_i) > 0$, $\forall i$, and another event B, then, $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$.
Note: The right hand side in the equation above can be rewritten, using the definition of conditional probability as, $\sum_{i=1}^{n} P(B \cap A_i)$.

Bayes Theorem: Given a partition $A_1, A_2, ..., A_n$ of a sample space Ω such that $P(A_i) > 0$, $\forall i$, and another event B such that $P(B) > 0$, then, for a given k,

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Note: Again, this can be simplified using the definition of conditional probability as, $\frac{P(B \cap A_k)}{\sum_{i=1}^{n} P(B \cap A_i)}$, which using the law of total probability is $\frac{P(B \cap A_k)}{P(B)}$.

Lecture 3

Discrete Random Variables

Random Variable: Any numerical function $X : \Omega \to \mathbb{R}$ is called a Random variable. We can further classify random variables based on the range of X:

- If the range of X is finite, then it is called a simple random variable. Example: For a single coin toss, let $X(Heads) = 1$ and $X(Tails) = 0$ is a simple random variable. The outcome of a dice roll is a simple random variable.
- If the range of X is discrete (countable), then X is called a discrete random variable. Simple random variables are necessarily discrete. Consider the experiment of tossing a fair coin until it lands on Heads. Let the number of tosses be the random variable X. Then X is a discrete (but not a simple) random variable.
- If the range of X is conitnuous, then X is a continuous random variable. For example, consider a unit circle and let X be the random variable denoting the distance of a randomly chosen point from the center. Then X is a continuous random variable.

Indicator Random Variable: Given a sample space Ω and an event $A \subseteq \Omega$, the indicator random variable of the event A is defined as $1_A : \Omega \to \{0, 1\}$ such that

$$X(\omega) = \begin{cases} 1, \omega \in A \\ 0, \omega \notin A \end{cases}$$

If $P(A) = p$, then this is also denoted by $1(p)$.

Bernoulli Random Variable: If the sample space Ω has only two outcomes, which we will term as Success and Failure, then we define the Bernoulli random variable as the indicator of success.

$$X(\omega) = \begin{cases} 1, \omega \text{ is a Success} \\ 0, \omega \text{ is a Failure} \end{cases}$$

Binomial Random Variable: A Binomial random variable $Bin(n, p)$ is used to denote the number of successes in n independent, identical Bernoulli trials, each with a probability p of success. So, if $X \sim Bin(n, p)$, by which we mean X is a random variable with $Bin(n, p)$ distribution, then, $Range(X) = \{0, 1, ...n\}$.

Probability Mass Function (pmf): Given a probability space $\{\Omega, \mathcal{F}, P\}$, and a discrete random variable $X : \Omega \to \mathbb{R}$, a function $p_X : Range(X) \to [0, 1]$ is called a probability mass function of X, if for any $x \in Range(X)$, $p_X(x) = P(X = x)$ and $\sum_x p_X(x) = 1$.
Note: We assume here that all sets $A_x = \{\omega | X(\omega) = x\}$ defined in the $Range(X)$ belong to $\mathcal{F}$ and we think of $P(X = x) = P(A_x)$.

Cumulative Distribution Function (cdf): Given a random variable X and its probability mass function p_X, we define the cumulative distribution function $F_X(a)$ as follows:

$$F_X(a) = P(X \leq a) = \sum_{x \leq a} p_X(x)$$

This is also called a step function for discrete random variable because of its shape.
Note: $F_X(a)$ is also defined in a lot of literature as $P(X < a)$. All the discussion that follows can be carried out with this definition also, just with minor adjustments in the proofs.

- $F_X : \mathbb{R} \to [0, 1]$
- F_X is a monotine increasing function, i.e. $\forall a \leq b, F(a) \leq F(b)$
- F_X is right continuous, i.e. $\lim_{x \to a^+} F_X(x) = F_X(a)$.

Expected Value: The expected value of a discrete random variable X, denoted $E(X)$ or μ_X, is defined as

$$E(X) = \mu_X = \sum_{x \in \text{Range}(X)} xP(X = x) = \sum_{x \in \text{Range}(X)} x p_X(x)$$

Note: The expected value is always defined for simple random variables. For general discrete random variables, it is defined if the sum is absolutely convergent, i.e. $\sum_x |x| p_X(x) = L$ for some limit L.

Properties of Expected Value:

- Translation: $E(X+b) = E(X) + b$
- Scaling: $E(aX) = aE(X)$
- LOTUS (Law of the Unconscious Statistician):

$$E(g(X)) = \sum_{x \in \text{Range}(X)} g(x)p_X(x)$$

Variance: The Variance of a random variable X, denoted $Var(X)$, is defined as $E((X-\mu_X)^2)$ which can be further simplified to $E(X^2) - (E(X))^2$.
Note: As for the expected value, variance is defined if the series its computation is absolutely convergent.

Standard Deviation: For any random variable X, the standard deviation $\sigma_X = \sqrt{Var(X)}$. It is used to find the 'spread' of the random variable in the same units as the variable.

Moments: The i^{th} moment of a random variable X is defined as $E((X^i))$.

Moment Generating Function: For any random variable X, its moment generating function, denoted by $M_X(t)$ is defined as $M_X(t) = E(e^{Xt})$.

Linearity of Expectation: For any two (not necessarily independent) random variables X and Y, $E(X+Y) = E(X) + E(Y)$.

Product of RVs: For any two independent random variables X and Y, $E(XY) = E(X)E(Y)$.
Note: Here independence is crucial, the result may not be true if X and Y are not independent.

Geometric Distribution: A random variable $X \sim Geo(p)$ is said to have geometric distribution with parameter p, if $Range(X) = \{1, 2, ...\}$, the natural numbers, and $p_X(i) = (1-p)^{i-1}p, i \geq 1$.
Note: It is good to think of Geometric distribution as a random variable counting the number of times a Bernoulli experiment (with probability of success p) is repeated until it results in a success.

Lecture 4

Moment Generating Function - On Oct 15th!

Moment Generating Function: For any random variable X, its moment generating function, denoted by $M_X(t)$ is defined as $M_X(t) = E(e^{Xt})$.

Properties of MGF:

- $M_X(0) = 1$
- $M_X^{(i)}(0) = E(X^i)$, that is, the i^{th} derivative of the MGF evaluated at 0 gives the i^{th} moment.
- Positivity: $M_X(t) \geq 0, \forall t \in \mathbb{R}$
- Translation: $M_{X+b}(t) = e^{bt} M_X(t)$
- Scaling: $M_{aX}(t) = M_X(at)$
- Sum: For any two independent random variables X and Y, $M_{X+Y}(t) = M_X(t) M_Y(t)$
- The MGF determines the distribution of the random variable, so two random variables with the same MGFs must have the same distribution. Mathematically, if $M_X(t) = M_Y(t), \forall t \in \mathbb{R}$, then, $F_X(a) = F_Y(a), \forall a \in \mathbb{R}$.
- Limits of MGF: For a sequence of random variables X_n and another random variable X, if $M_{X_n}(t) \to M_X(t)$, then, $f_{X_n} \to f_X$. Note: this property has deliberately been phrased a little vaguely because we haven't really discussed what notion of 'convergence' of functions we are using.

Central Limit Theorem (Simplified version): Let $X_1, X_2, X_3, ...$ be a sequence of independent, identically distributed random variables, with $E(X_i) = 0$ and $Var(X_i) = E(X_i^2) = 1$, $\forall i$. Further let $Z_n = \frac{X_1 + X_2 + ... + X_n}{\sqrt{n}}$. Then,

$$\lim_{n \to \infty} M_{Z_n}(t) \to e^{\frac{1}{2} t^2}$$

Central Limit Theorem: Let $X_1, X_2, X_3, ...$ be a sequence of independent, identically distributed random variables. Let $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, $\forall i$. Further let $Z_n = \frac{X_1 + X_2 + ... + X_n - n\mu}{\sigma\sqrt{n}}$. Then,

$$\lim_{n \to \infty} M_{Z_n}(t) \to e^{\frac{1}{2} t^2}$$

Lecture 5

Continuous Random Variables

Continuous Random variable: X is said to be a continuous random variable, if there exists a Riemann integrable function $f_X : \mathbb{R} \to [0, \infty)$ such that for any $a \in \mathbb{R}$, $F_X(a) = P(X \leq a) = \int_{-\infty}^{a} f_X(x) dx$. In particular, $\int_{-\infty}^{\infty} f_X(x) dx = 1$.
f_X is said to be the **probability density function** of the random variable X.

Properties of the Cumulative distribution function:

- F_X is a monotone increasing function, i.e., $F_X(a) \leq F_X(b)$, $\forall a \leq b$.
- It is continuous from right, i.e., $\lim_{x \to a^+} F_X(x) = F(a)$.
- Limits: $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.

For any continuous random variable X, the probability of interval $(a, b]$, or, $P(a < x \leq b) = \int_a^b f_X(x)dx$. Note: this is also the probability of the intervals $[a, b), [a, b]$ and (a, b).

Uniform Distribution: $X \sim U(a, b)$ is said to have uniform distribution if its pdf is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

Expected value: Given a continuous random variable X with probability density function $f_X(x)$, if $\int_{-\infty}^{\infty} |x| f_X(x)dx = L$ for some real number L (that is, the integral is absolutely convergent), then the expected value of X is given by,

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

Note: We have the same properties for expected value (scaling, translation, Lotus, linearity) as in the discrete case.

Transformation: Given a random variable X, $Y = g(X)$ is called a transformation of X. It is called a linear transformation if $g(X) = aX + b$ for some constants a, b.

Steps for solving problems involving transforms, given $Y = g(X)$:

- Find the $range(Y)$.
- $F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$, here we use g^{-1} (but carefully!) to write $F_Y(y)$ in terms of $F_X()$.
- differentiate F_Y to get f_Y.

Lecture 6

Poisson, Exponential, Normal distributions

The Poisson distribution is used when we have a very large collection of independent small probability events and we are interested in the number of occurances in a fixed time interval. For example, number of cars that will have a flat tyre on a certain day (there are millions of cars, and each has a very minor probability of having a flat tyre, and each of these are independent events). We notice that for two disjoint intervals of time of same length, since all events are independent, the average number of occurances in both intervals must be the same. We call this average λ.

Poisson distribution: A random variable $X \sim Pois(\lambda)$ has Poisson distribution with parameter λ (lambda), if its probability mass function is given by,

$$p_X(i) = P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}, i = 0, 1, 2, ...$$

Here, λ denotes the average number of occurances in a fixed time interval. Note: in $P(X = i)$, we are computing the probability of i occurrances in the **same** time interval. If the duration of the time interval is changed, then λ should be changed accordingly.

Poisson approximation of Binomial distribution: Let $X \sim Bin(n,p)$, where n is large and the parameter p is small, so that $\lambda = np$ is moderate. Then,
$P(X=i) = \binom{n}{i}p^i(1-p)^{n-i} = \frac{\binom{n}{i}}{n^i}\lambda^i(1-\frac{\lambda}{n})^{n-i}$
If $np \to \lambda$ as $n \to \infty$, then the above tends to $e^{-\lambda}\frac{\lambda^i}{i!}$.
So, for large n and a small p, we can approximate $Bin(n,p)$ with $Pois(np)$.

The probability distribution of the **time** between two consecutive events in a Poisson process (many small probability independent events with a constant average rate of occurance) has exponential distribution.

Exponential distribution: $X \sim Exp(\lambda)$ is said to have exponential distribution if it is a continuous random variable with the following probability density function,

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

for some $\lambda > 0$. Here λ denotes the average number of occurances in unit time.

The cumulative distribution function for the Exponential distribution is given by
$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$.

Memoryless property: A random variable X (continuous or discrete) is said to have memoryless property if the following is true:

$$P(X > t+s | X > s) = P(X > t)$$

Theorem: The distribution of a discrete (continuous) RV is memoryless if and only if it is Geometric (Exponential).
Proof: Let us assume X is discrete and has memoryless property. Then, $P(X > t+s|X > s) = P(X > t)$ $\implies \frac{P(X>t+s)}{P(X>s)} = P(X > t) \implies P(X > t+s) = P(X > s)P(X > t)$.
Let us set $P(X = 1) = p$, then $P(X > 1) = (1-p)$. Using the equation above, $P(X > i) = (1-p)^i$ and so $P(X = i) = P(X > i-1) - P(X > i) = (1-p)^{i-1}p$.
Similar proof can be used to show that any continuous distribution that has the memoryless property is necessarily Exponential (upto a translation).

Normal distribution: A continuous random variable X is said to have Normal distribution $N(\mu, \sigma^2)$, if its probability density function is defined as follows:

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

Where $\mu \in \mathcal{R}$ and $\sigma \in [0, \infty)$ or $\mathcal{R}^+$.

Standard Normal Distribution: When the parameters $\mu = 0$ and $\sigma^2 = 1$, then the distribution is called the Standard Normal Distribution and denoted by $N(0,1)$. We usually use the letter Z for a random variable with standard normal distribution. So,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$\Phi(a)$: The CDF of the standard Normal distribution is denoted by $\Phi(a)$ (pronounced as Fi in Five). So, for $Z \sim N(0,1)$, $\Phi(a) = P(Z \leq a) = \int_{-\infty}^{a} \frac{1}{2\pi} e^{-\frac{1}{2}z^2} dz$. This integral doesn't have a nice solution in elementary functions but can be approximated.
The density function $f_Z(z)$, or the bell curve, is symmetric about 0. This gives us the very useful property that $P(Z \leq -a) = P(Z \geq a)$, $\forall a$.

Below is the table for $\Phi(a)$:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Standardization of the general normal distribution: $X \sim N(\mu, \sigma^2)$, then the transform $\frac{X-\mu}{\sigma}$ has standard normal distribution. So to find $P(X \leq a)$, we use the fact that this is equal to $P(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}) = \Phi(\frac{a-\mu}{\sigma})$.

Variance and Standard Deviation: For $Z \sim N(0,1)$, the standard normal variable,

$$E(Z) = 0, \text{ and } Var(Z) = \sigma_Z = 1$$

For $X \sim N(\mu, \sigma^2)$, a variable with general normal distribution,

$$E(X) = \mu, Var(X) = \sigma^2 \text{ and } \sigma_X = \sigma$$

Lecture 7

Joint distributions (discrete case), Independence, Covariance

Joint Probability Mass Function: Given two discrete random variables X and Y, the **joint probability mass function** is a function, $p_{X,Y} : \mathcal{R}^2 \to [0, 1]$, such that,

$$\sum_x \sum_y p_{X,Y}(x, y) = 1$$

Here we think of $p_{X,Y}(x, y)$ as the probability that $X = x$ and $Y = y$.
Where its obvious, the subscript X, Y is dropped and it is written as $p(x, y)$.

Marginal Probability Mass Functions: The probability mass functions of X, $p_X(x)$ and of Y, $p_Y(y)$ are called the marginal probability mass functions.

The marginal probability mass functions can be derived from the joint mass function as follows:

$$p_X(x) = \sum_y p_{X,Y}(x, y)$$

$$p_Y(y) = \sum_x p_{X,Y}(x, y)$$

The expected value of any function $g(X, Y)$ is given by

$$\sum_x \sum_y g(x, y) p_{X,Y}(x, y)$$

Independence: Two random variables X and Y are said to be independent if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$, $\forall x, y \in \mathcal{R}$.

Let X and Y be two independent random variables. Then,

$$E(XY) = E(X)E(Y)$$

.

Covariance: Let X, Y be two random variables. Then, the covariance of X and Y, denoted by $Cov(X, Y)$ is defined as $E((X - \mu_X)(Y - \mu_Y))$. Because of Linearity of expectation, this is $E(XY) - \mu_X \mu_Y$.

If two random variables X, Y are independent, then $E(XY) = \mu_X \mu_Y$, so $Cov(X, Y) = 0$.
Note: If X, Y are two random variables such that $Cov(X, Y) = 0$, it does not imply that X, Y are independent!

Properties of Covariance

- (Commutative) $Cov(X,Y) = Cov(Y,X)$.
- If X, Y are independent, $Cov(X,Y) = 0$. (Note: The converse is not true! $Cov(X,Y) = 0$ does not imply that X, Y are independent.
- $Cov(aX+b, Y) = aCov(X,Y)$
- $Cov(X+Y, Z) = Cov(X,Z) + Cov(Y,Z)$

Recall that we defined the Variance of a random variable X as $Var(X) = E((X-\mu_X)^2)$. This we can see is $E((X-\mu_X)(X-\mu_X)) = Cov(X,X)$.

Properties of Variance

- $Var(X) \geq 0$. Also, if $Var(X) = 0$, then it is 'almost surely' a constant.
- $Var(aX+b) = a^2 Var(X)$. Then, $\sigma_{aX+b} = |a|\sigma_X$.
- $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$. **In particular, if X, Y are independent, then $Var(X+Y) = Var(X) + Var(Y)$.**
- $Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X,Y)$.

Correlation coefficient : Given two random variables, X, Y, the correlation coefficient, denoted by $\rho(X,Y)$ (pronounced 'Row' X, Y), is defined as $\frac{Cov(X,Y)}{\sigma_X \sigma_Y}$.

Theorem: For any two random variables X and Y, $-1 \leq \rho(X,Y) \leq 1$.

All the discussion of multiple random variables is much more succintly expressed as vectors. Again consider X_1, X_2 as two random variables with joint pmf p_{X_1,X_2}. We could instead think of them as a random variable vector, $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. The joint pmf will be the same, only denoted by the vector $p_{\boldsymbol{X}}$.
Expected value of the random variable vector can be obtained by taking the expected value of each of its components, $\boldsymbol{\mu_X} = E(\boldsymbol{X}) = E\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix}$.

Covariance Matrix: The Covariance Matrix for $\boldsymbol{X}$ is denoted by $\boldsymbol{\Sigma}$ and is defined as follows: $\boldsymbol{\Sigma} = Cov(\boldsymbol{X}, \boldsymbol{X}) = E((\boldsymbol{X} - \boldsymbol{\mu_X})(\boldsymbol{X} - \boldsymbol{\mu_X})^T)$
$= E\begin{pmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) \\ (X_1-\mu_1)(X_2-\mu_2) & (X_2-\mu_2)^2 \end{pmatrix} = \begin{pmatrix} Var(X_1) & Cov(X_1,X_2) \\ Cov(X_1,X_2) & Var(X_2) \end{pmatrix}$

Properties of the Covariance Matrix: Let $\boldsymbol{X}$ be a random variable vector and $\boldsymbol{A}$ a matrix of constants. Then $\boldsymbol{\Sigma_A} = Cov(\boldsymbol{AX}, \boldsymbol{AX}) = \boldsymbol{A}Cov(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{A}^T = \boldsymbol{A\Sigma A}^T$.
Where $\boldsymbol{\Sigma}$ is the Covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{\Sigma_A}$ is the Covariance matrix of $\boldsymbol{AX}$.

Lecture 8

Joint continuous distributions, Convolutions

Joint Probability Density Function: Random variables X and Y, are said to be jointly continuous, if there exists a non-negative Riemann integrable function $f_{X,Y}(x,y) : \mathcal{R}^2 \to \mathcal{R}$, such that

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)dxdy$$

Here, $f_{X,Y}(x,y)$ is said to be the joint probability density function, and $F_{X,Y}(x,y)$ is the joint cumulative distribution function.
Note that, in particular this would imply that,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$$

It follows that given the joint CDF $F_{X,Y}(x,y)$,

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

Marginal probability density functions: Given two random variables X, Y and their joint probability density function $f_{X,Y}(x,y)$, the marginal probability density functions are as follows:

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

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Many jointly continuous variables: Random variables $X_1, X_2, ...X_n$ are said to be jointly continuous, if there exists a non-negative Riemann integrable function $f_{X_1,X_2,...X_n}(x_1,x_2,...,x_n) : \mathcal{R}^n \to \mathcal{R}$, such that

$$F_{X_1,X_2,...X_n}(a_1,a_2,...,a_n) = \int_{-\infty}^{a_1} ... \int_{-\infty}^{a_n} f_{X_1,X_2,..X_n}(x_1,x_2,...,x_n)dx_1dx_2...dx_n$$

and in particular, this would imply that,

$$\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1,X_2,..X_n}(x_1,x_2,...,x_3)dx_1dx_2...dx_n = 1$$

Independence: Two random variables X and Y are said to be independent if $F_{X,Y}(x,y) = F_X(x)F_Y(y), \quad \forall x, y \in \mathbb{R}$.
Taking partial derivative of the above with respect to x and y, we can see that,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \qquad \forall x, y \in \mathcal{R}$$

Expected Value: Let X, Y be two random variables. Then, for any function $g(X,Y)$, the expected value is defined as

$$E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy$$

provided that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|g(x,y)|f_{X,Y}(x,y)dxdy$ is defined.

Covariance, Variance and their properties are the same as in the previous lecture and will not be repeated here again.

Convolution: Let X, Y be two **independent** random variables. Then, the **convolution** of X and Y is the random variable $Z = X + Y$.

Convolution for discrete random variables: Let X and Y be independent discrete random variables, and let $Z = X + Y$. Then,

$$p_Z(z) = \sum_y p_X(z-y)p_Y(y)$$

Convolution for continuous random variables: Let X and Y be independent continuous random variables, and let $Z = X + Y$. Then,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy$$

Convolution using Moment Generating Functions: Let X, Y be independent random variables. Then the moment generating function of the convolution is given by $M_X(t)M_Y(t)$.

Lecture 9

Normal distribution, Bivariate Normal Distribution

Normal Distribution: A continuous random variable X is said to have Normal distribution $N(\mu, \sigma^2)$, if its probability density function is defined as follows:

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

Where $\mu \in \mathcal{R}$ is the mean and $\sigma \in [0, \infty)$ or $\mathcal{R}^+$ is the standard deviation of X .

Standard Normal Distribution: When the parameters $\mu = 0$ and $\sigma^2 = 1$, then the normal distribution is called the **Standard Normal Distribution** and denoted by $N(0, 1)$. We usually use the letter Z for a random variable with standard normal distribution. So,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Transformation of a Normal Distribution: The linear transform of a random variable with normal distribution will also have normal distribution. More formally, let $X \sim N(\mu, \sigma^2)$. Further, let $Y = aX+b$. Then $Y \sim N(a\mu + b, a^2\sigma^2)$.
Note: This gives us the potential to transform any normal distribution to the standard normal distribution. This process is called Standardization. If $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0, 1)$. And so every normal distribution is just a linear transform of the standard normal distribution, that is, it is obtained by scaling and translating the standard normal distribution.

Moments of the Standard Normal Distribution: Let $Z \sim N(0,1)$. Then, $E(Z^i) = (i-1)E(Z^{(i-2)})$.

Moment Generating Function: The Moment Generating Function of the Standard Normal Distribution $Z \sim N(0,1)$ is given by $M_Z(t) = e^{\frac{1}{2}t^2}$.
Using Linearity and Scaling properties of MGFs, we can see that the MGF of the general normal distribution is $= e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

Radially symmetric distributions: Suppose X, Y are two **independent** random variables, where we also know that the joint density function only depends on the distance from the origin. Then $(X,Y)^T$ must have the standard normal distribution (upto a constant factor). That is $X, Y \sim N(0, \sigma^2)$ for some $\sigma > 0$.

Standard Bivariate Normal distribution: Denoted by $\boldsymbol{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$, the standard bivariate normal random variable consists of two **independent** random variables with standard normal distribution. So, the joint probability density function is given by,

$$f_{Z_1,Z_2}(z_1, z_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(z_1^2 + z_2^2)}$$

$$E(\boldsymbol{Z}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The covariance matrix is the identity matrix. Recall,

$$\boldsymbol{\Sigma} = Cov(\boldsymbol{Z}) = \begin{pmatrix} Var(Z_1) & Cov(Z_1, Z_2) \\ Cov(Z_1, Z_2) & Var(Z_2) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Convolution of Normals: Given two independent random variables $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, any linear combination of them also has normal distribution. In particular, for any two non-zero constants $c_1, c_2 \in \mathcal{R}$, $c_1 X_1 + c_2 X_2 \sim N(c_1\mu_1 + c_2\mu_2, c_1^2\sigma_1^2 + c_2^2\sigma_2^2)$.
From this we can conclude that for any linear transform matrix $\boldsymbol{A}$, the two components of the transform $\boldsymbol{AZ}$ where $\boldsymbol{Z}$ is the random variable vector with standard bivariate normal distribution, also have normal distribution. We use this to define the general bivariate normal distribution vector.

General Bivariate Normal distribution: A random variable vector $\boldsymbol{X}$ has a bivariate normal distribution if $\exists \boldsymbol{A} \in \mathcal{R}^{2\times 2}$ and a $\boldsymbol{\mu} \in \mathcal{R}^2$ such that $\boldsymbol{X} = \boldsymbol{AZ} + \boldsymbol{\mu}$, where $\boldsymbol{Z}$ is the standard bivariate normal distribution random variable vector.

We can notice the following properties of $\boldsymbol{X}$:

$$E(\boldsymbol{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \boldsymbol{\mu}$$

$$\boldsymbol{\Sigma_X} = Cov(\boldsymbol{X}) = \boldsymbol{AA}^T$$

Joint Probability Density Function: The joint probability density functon of $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by, $f_{\boldsymbol{X}}(x_1, x_2) = \frac{1}{(\sqrt{2\pi})^2|\Sigma|} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T(\boldsymbol{\Sigma})^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}$.

It is useful to note that for any matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, let the determinant $D = ad - bc$ be non-zero. Then, the inverse of the matrix is given by $\frac{1}{D}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

Properties of the bivariate normal distribution: Let X_1, X_2 be the components of a bivariate normal distribution. Then the following are true:

- Linear combination $c_1X_1 + c_2X_2$, for non-zero c_1, c_2, has normal distribution.
- If X_1, X_2 are uncorrelated, then they are independent. (Counter example to this in general case is let X be standard normal and $Y = WX$ where W takes values 1 and -1 with probability $\frac{1}{2}$.
- Regression $E(X_2|X_1)$ is the linear regression.

Lecture 10

Inequalities, Central Limit Theorem

Markov Inequality: Given a **positive** random variable X ($P(X < 0) = 0$), the following is true for any real number $a > 0$:

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

Chebyshev's Inequality: Given **any** random variable X with mean μ and standard deviation σ, the following is true for any real number $a > 0$:

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

Another way of writing it using the **standardization** of X:

$$P(\left|\frac{X - \mu}{\sigma}\right| \geq a) \leq \frac{1}{a^2}$$

In this second inequality, $a = 2, 3, 4$ gives the probability of X being within 2, 3, 4 standard deviations of the mean is at least $75\%, 89\%$ and 93.75% respectively.

Chernoff's bounds: Given **any** random variable X, the following is true for any real number a (not necessarily positive):

$$P(X \geq a) \leq \frac{M_X(t)}{e^{ta}}, \forall t > 0$$

Note that we optimize the parameter t to get the best bound.

For any random variable X, $E(|X - c|)$ is minimized when c is the median, that is, $\int_c^\infty f_X(x)dx = \frac{1}{2}$.

Steiner Equality: For any random variable X, $E((X-c)^2)$ is minimized when $c = \mu$, where $\mu = E(X)$.

Independent Identical Distributions: A sequence of random variables $X_1, X_2, X_3, ..., X_n$ are said to be **i.i.d**, or "Independent Identically Distributed" if they all have the same probability distribution and if they are totally independent. (In some theorems assumption of pairwise independence is sufficient).

Weak Law of Large Numbers: Given a sequence $X_1, X_2, ...,$ of i.i.d random variables (pairwise independence is sufficient) with mean μ and standard deviation σ, we define a new sequence of averages, $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}$. Then the following is true for all $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\overline{X_n} - \mu| > \epsilon) = 0$$

Strong Law of Large Numbers: Given a sequence $X_1, X_2, ...,$ of i.i.d random variables with mean μ, we define a new sequence of averages, $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}$. Then the following is true for all $\epsilon > 0$:

$$P(\lim_{n\to\infty} \overline{X_n} \to \mu) = 1$$

or in other words, $\overline{X_n}$ almost surely (with probability 1) converges to the mean μ.

Central Limit Theorem: Given a sequence $X_1, X_2, ...,$ of i.i.d random variables with mean μ and standard deviation σ, we define a new sequence of **standardized random variables**, $Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$. Then the limiting random variable $Z = \lim_{n\to\infty} Z_n$ has the standard normal distribution. Or,

$$Z_n \to N(0,1)$$

Note: For a $n > 20$ it is standard practice to approximate the distribution of Z_n with the standard normal distribution $N(0,1)$.

De-Moivre Laplace Theorem: Given a sequence $X_1, X_2, ...,$ of i.i.d random variables with distribution $1(p)$. We define a new sequence of **standardized random variables**, $Z_n = \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}$. Then the limiting random variable $Z = \lim_{n\to\infty} Z_n$ has the standard normal distribution.

$$Z_n \to N(0,1)$$

Lecture 11

Simple Linear Regression, Conditional distributions

Method of Least squares for Simple Linear Regression: The linear transform of X given by $\beta X + \alpha$, where $\alpha, \beta \in \mathbb{R}$, which minimize the error in estimation of Y, represented by $E((Y-(\beta X+\alpha))^2)$ (method of least squares), is given by:

$$\beta = \frac{Cov(X,Y)}{\sigma_X^2}$$

$$\alpha = E(Y) - \frac{Cov(X,Y)}{\sigma_X^2} E(X)$$

and the error of such an approximation is,

$$= \sigma_Y^2(1 - (\rho(X,Y))^2)$$

where $\rho(X,Y)$ is the correlation coefficient.

Joint Conditional Mass Function: For two random variables X, Y, the joint conditional probability mass function of X conditioned on a specific value of $Y = y$ where $p_Y(y) \neq 0$, is given by,

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

We can think of $p_{X|Y}(\cdot|y)$ as an updated probability mass function for the variable X. In particular,

$$\sum_x p_{X|Y}(x|y) = 1$$

Some properties of the joint conditional probability mass function:

- $\sum_x p_{X|Y}(x|y) = 1$, or $p_{X|Y}$ is a probability mass function for X.
- $p_{X|Y}(x|y) = p_X(x)$ when X and Y are independent.
- $F_{X|Y}(a|y) = \sum_{x \leq a} p_{X|Y}(x|y)$, since $p_{X|Y}$ is a pmf, we have to sum it as before to get the joint conditional cumulative distribution function.

Joint Conditional Probability Density Function: For two continuous random variables X, Y with joint pdf given by $f_{X,Y}(x,y)$, for the values of $Y = y$ where the density function $f_Y(y) \neq 0$, there the joint conditional probability density function is given by,

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

Joint Conditional Cumulative Distribution Function:

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

Lecture 12

Regression, Law of Total Expectation, Law of Total Probability

Conditional Expectation: The conditional expected value of a discrete random variable Y, conditioned on $X = c$ is as follows:

$$E(Y|X = c) = \sum_y y p_{Y|X}(y|c)$$

while if Y is continuous, then its given by,

$$E(Y|X = c) = \int_{-\infty}^{\infty} y f_{Y|X}(y|c)dy$$

Regression: The function $E(Y|X=x)$, also written as $E(Y|X)$, is called the **Regression** of the variable Y in terms of X. It is the function of x which minimizes the least squares error of approximating Y with a function of X, or, $E((Y-g(X))^2)$ is minimized if $g(X)=E(Y|X=x)$.
Here Y is called the 'dependent' variable, while X is called the 'independent' variable.

Properties of Regression:

- For any function $h(X)$, $E(h(X)Y|X)=h(X)E(Y|X)$. In particular, $E(h(X)|X)=h(X)$.
- $E(aY+b|X)=aE(Y|X)+b$.
- $E((Y-g(X))^2)$ is **minimum** for $g(X)=E(Y|X)$.

Law of Total Expectation: For any two random variables X,Y, we have $E(E(Y|X))=E(Y)$.
Proof sketch for simple random variables (range is finite).
$E(E(Y|X))=\sum_x E(Y|X=x)p_X(x)=\sum_x\big(\sum_y yp_{Y|X}(y|x)\big)p_X(x)=\sum_x\big(\sum_y y\frac{p_{X,Y}(x,y)}{p_X(x)}\big)p_X(x)$
For simple random variables X and Y, it is easy to see that we can rearrange the above terms to get,
$=\sum_x\sum_y yp_{X,Y}(x,y)=E(Y)$
Note: the above proof, with more care, can be modified to work for infinite and continuous random variables.

Note: It is important to note that in $E(E(Y|X))$, the outer expected value is taken in terms of the random variable X, while the inner one in terms of Y where Y here has the conditional distribution (either $p_{Y|X}$ or $f_{Y|X}$).

Proof of Property 3: Let $g(X)$ be a function that minimizes $E((Y-g(X))^2)$.
Using Law of Total expectation, we can write this as,

$$E(E((Y-g(X))^2|X=x))$$

where the outer expectation is over the variable X and the inner over the random variable Y conditioned on $X=x$. We recall Steiner's inequality that, $E((X-c)^2)$ is minimized when $c=E(X)$. Then the above expected value is minimized when,

$$g(X)=E(Y|X=x)$$

So regression function, $E(Y|X=x)$ can be thought of as the best approximation of Y as a function of X (best in terms of least squared errors).

Law of Total Probability: For any event A,

$$P(A)=\int_{-\infty}^{\infty}P(A|X=x)f_X(x)dx$$

Lecture 13

Interval estimates and examples from The Probabilitic Method

Interval Estimates when Sigma is known: Consider samples $X_1, X_2, ..., X_n$ taken independently randomly from some data set. If we knew the standard deviation of X_i is σ for all i, then we can estimate the mean of the data set as follows:
Let θ be the expected value of a population parameter that we want to estimate. Let a sample of n items have mean $\hat{\theta}$. Suppose we know from previous data that the standard deviation of the population is σ. Let the confidence level we require of our interval estimate be $1-\alpha$, where $0 < \alpha < 1$. Further let c be a constant such that $\Phi(c) = 1 - \frac{\alpha}{2}$, where Φ is the CDF of $N(0,1)$. Then, the confidence interval estimate for θ with $1-\alpha$ confidence is given by

$$[\hat{\theta} - c\frac{\sigma}{\sqrt{n}}, \hat{\theta} + c\frac{\sigma}{\sqrt{n}}]$$

We will now look at some examples where probability theory can be used to derive some very beautiful results in mathematics.
Note: What we will use many times is if $E(X) = c$ and X is not a constant random variable, then there must be an event where X takes value more than c and there must be an event where X takes a value less than c.
First a warm up. This is a theorem we did in ITC2, but we didn't use probability there.

Theorem: Every graph $G(V, E)$ contains a bipartite subgraph with at least $|E|/2$ edges.
Proof: Pick partition A by picking vertices randomly and independently with probability $\frac{1}{2}$. We let remaining unpicked vertices be B and keep only the edges between A and B to get a bipartite graph. Let X_i be the indicator that the edge e_i gets selected in our bipartite graph. Then $P(X_i = 1) = 1/2$. Then, $E(X_i) = \frac{1}{2}$, and if we let $X = \sum X_i$, then $E(X) = \frac{|E(G)|}{2}$. Then there must be a choice of A, B with at least as many edges between them.

Our first major theorem uses Boole's inequality.

The Ramsey number $R(k, l)$ is defined as the minimum n such that any coloring of a complete graph K_n with two colors red and blue will either contain a monochromatic red K_k or a monochromatic blue K_l.

Theorem (P.Erdos, 1947): If $\binom{n}{k} \cdot 2^{1-\binom{k}{2}} < 1$, then $R(k, k) > n$. Then, $R(k, k) > \lfloor 2^{k/2} \rfloor$ for all $k \geq 3$.
Proof: Take a random coloring of the edges of K_n where an edge gets red or blue color with probability $1/2$.

This second theorem uses Linearity of Expectation.

A set $S \subseteq V$ in a graph $G(V, E)$ is said to be a dominating set, if every $v \in V \setminus S$ has a neighbor in S.

Theorem: Let $G(V, E)$ be a graph with minimum degree δ. Then, G has a dominating set of size at most $n(1 + ln(1 + \delta))/(1 + \delta)$.
Proof: We make a subset A of V by selecting every vertex independently with probability p. Then for any vertex v, the probability that it is not in A and that none of its neighbors are also in A (that is, v is not dominated by A) is at most $(1-p)^{\delta+1}$. We let all such vertices (not dominated by A) be a subset B and we notice that $A \cup B$ will be a dominating set. Now we let $X = |A|$ and $Y = |B|$. By linearity of expectation, $E(X + Y) = E(X) + E(Y) \leq np + n(1-p)^{\delta+1}$. Then there must be a way of picking a dominating set of size at most this quantity. If we optimize on the value of p, we will get the above bound.

Now a result from Combinatorial number theory. But first a definition.

A set A is said to be sum-free if $A + A \cap A = \phi$, or, there are no two elements $a_1, a_2, a_3 \in A$, all not necessarily distinct, such that $a_1 + a_2 = a_3$.

Theorem (P.Erdos, 1965): Every set $B = \{b_1, b_2, ..., b_n\}$ of n non-zero integers contains a sum-free subset A of size $> \frac{n}{3}$.
Proof: Let $p = 3k + 2$ be a prime such that $p > 2\text{max}_i|b_i|$. Our idea is to map B into $\mathbb{Z}_p$ and we take p to be large enough so that the b_i's all map to distinct elements. (Note: in the rest of the discussion, assume we are taking mod p when we do additions or multiplications). Let $C = \{k+1, k+2, ..., 2k+1\}$ and we notice that C is a sum-free subset of $\mathbb{Z}_p$ containing more that $\frac{1}{3}$ of the elements of $\mathbb{Z}_p$. If we pick X uniformly from $\mathbb{Z}_p \setminus \{0\}$ and consider the distribution of Xb_i for some fixed b_i, then we note that Xb_i will range over all values of $\mathbb{Z}_p \setminus \{0\}$ with equal probability. So $P(Xb_i \in C) > \frac{1}{3}$. Let X_i be the indicator of $Xb_i \in C$. We notice that if $A \subset of B$ such that $XA \subset C$, then A is necessarily sum-free, because if $a_1 + a_2 = a_3$, then $a_1x + a_2x = a_3x(\text{mod p})$. Let $Y = \sum_i X_i$. Then by linearity of expectation, $E(Y) > \frac{n}{3}$, so there is a choice of X such that more that $\frac{n}{3}$ elements of B map to C, and these elements give us the required sum-free subset.

Theorem: Let $v_1, v_2, ..., v_n \in \mathbb{R}^n$, where $|v_i| = 1$, $\forall i$. Then there exist $\epsilon_1, \epsilon_2, ..., \epsilon_n = \pm 1$ weights, so that

$$|\epsilon_1 v_1 + \epsilon_2 v_2 + ... + \epsilon_n v_n| \leq \sqrt{n}$$

, and also there exist $\epsilon_1, \epsilon_2, ..., \epsilon_n = \pm 1$ weights, so that

$$|\epsilon_1 v_1 + \epsilon_2 v_2 + ... + \epsilon_n v_n| \geq \sqrt{n}$$

,
Proof: Pick ϵ_i's uniformly as ± 1 with probability $\frac{1}{2}$.

Let $\mathcal{F}$ be a family of sets. It is said to be intersecting if for all $A, B \in \mathcal{F}$, $A \cap B \neq \phi$.

Theorem (Erdos-Ko-Rado, proof due to Katona 1972): Let $\mathcal{F}$ be an intersecting family of k-element subsets of $\{0, 1, ..., n-1\}$, where $n \geq 2k$. Then $|\mathcal{F}| \leq \binom{n-1}{k-1}$.
Proof: We uniformly pick a permutation σ of $\{0, 1, ..., n-1\}$ and independently and uniformly pick a number $i \in \{0, 1, ..., n-1\}$. We let A_i denote the set of elements $\{\sigma(i), \sigma(i+1), ..., \sigma(i+k-1)\}$ where we think of the numbers modulo n. We are interested in $P(A_i \in \mathcal{F}$. We will evalute this in two different ways.
Firstly we will use law of total probability. $P(A_i \in \mathcal{F}) = \sum_\sigma P(A_i \in \mathcal{F}|\sigma)P(\sigma)$, where the summation runs over all permutations and they do form a partition of the sample space. But we notice that for a fixed σ, we are looking at how many consecutive k-element subsets can belong to $\mathcal{F}$. But since $\mathcal{F}$ is intersecting, at most k of these can. We also note that since we chose i uniformly from the set $\{0, 1, ..., n-1\}$, $P(A_i \in \mathcal{F}|\sigma) \leq \frac{k}{n}, \forall \sigma$. So, since σ is uniformly picked from all permutations, we get that $P(A_i \in \mathcal{F}) \leq \frac{k}{n}$.
The second way we try to estimate the same probability is by noticing that by uniformly picking σ and uniformly independently picking i, we have created a uniform probability space over all possible k subsets of $\{0, 1, ..., n-1\}$. So, $P(A_i \in \mathcal{F}) = \frac{|\mathcal{F}|}{\binom{n}{k}}$.
Putting these together we have $\frac{|\mathcal{F}|}{\binom{n}{k}} \leq \frac{k}{n}$, solving which we get the result.

Wish you Good Luck in all future endeavors!