Final Review

Alexander Van Roijen

December 3, 2018

Please complete the following:

1. Lecture 1. Basics of Probability

   (a) **Definition (Sample space and event).** The sample space S of an experiment is the set of all possible outcomes of the experiment. An event A is a subset of the sample space S , and we say that A occurred if the actual outcome is in A.

   (b) **Definition (General definition of probability).** A probability space consists of a sample space S and a probability function P(.) which takes an event A $\subset$ S as input and returns P(A), a real number between 0 and 1, as output. The probability function must satisfy the following axioms: $P(\emptyset) = 0\, P(S) = 1$ and for a union of disjoint events, we get $P(A_1 \cup ...A_n) = P(A_1) + ...P(A_n)$

   (c) **Theorem. Properties of probability.** A probability function has the following properties, for any events A and B.

      i. $P(A^c) = 1 - P(A)$

      ii. if $A \subset BP(A) \leq P(B)$

      iii. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ which can be extended
      $P(A \cup B \cup C) = P(A)+P(B)+P(C)-P(A \cap B)-P(A \cap C)-P(B \cap C)-P(A \cap B \cap C)$

   (d) **Definition (Conditional probability).** If A and B are events with $P(B) > 0$, then the conditional probability of A given B, denoted by $P(A|B)$, is defined as: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ Further note that all probabilities are in fact conditional. We like to think of P(A) as our prior beliefs of an event, and P(A—B) as our posterior, or what we think it is given something is already known.

   (e) **Theorem.** For any events A1,..An with positive probabilities,
   $P(A1, ...An) = P(A1)P(A2|A1)P(A3|A1, A2)...P(An|A1, ...A_{n-1})$

   (f) **Theorem (Bayes' rule).** $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

   (g) **Theorem (Law of total probability (LOTP)).** Let A1,...An be a partition of the sample space S with $P(Ai) > 0$ for all i. Then
   $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$
   When we condition on an event E, we update our beliefs to be consistent with this

knowledge, effectively putting ourselves in a universe where we know that E occurred. Within our new universe, however, the laws of probability operate just as before. Conditional probability satisfies all the properties of probability!

(h) **Theorem (Bayes' rule with extra conditioning).** Provided that $P(A \cap E) > 0$ and $P(B \cap E) > 0$, we have $P(A|B, E) = \frac{P(B|A,E)P(A|E)}{P(B|E)}$

(i) T**heorem (Law of total probability (LOTP) with extra conditioning).** Let A1... An be a partition of the sample space S with $P(Ai|E) > 0$ for all i. Then $P(B|E) = \sum_{i=1}^{n} P(B|A_i, E)P(A_i|E)$

(j) **Definition (Independence of two events).** Events A and B are independent if $P(A \cap B) = P(A)P(B)$. If $P(A) > 0$ and $P(B) > 0$, then this is equivalent with $P(A|B) = P(A)$, and also equivalent with $P(B|A) = P(B)$. Independence is a symmetric relation.

(k) **Proposition.** If A and B are independent, then A^c and B are independent, A^c and B^c are independent, and A and B^c are independent.

(l) Definition (Independence of three events). Events A, B and C are said to be independent if all of the following relations hold:
$P(A \cap B) = P(A)P(B)$;
$P(A \cap C) = P(A)P(C)$;
$P(B \cap C) = P(B)P(C)$;
$P(A \cap B \cap C) = P(A)P(B)P(C)$

(m) **Definition (Conditional independence).** Events A and B are said to be conditionally independent given E if $P(A \cap B|E) = P(A|E)P(B|E)$.

(n) Problems shown: Monty Hall, and Positive test of conditionitis and bayes rule

2. (a) **Definition (Random variable).** Given an experiment with sample space S , a random variable is a function from the sample space S to the real numbers R. It is common, but not required, to denote random variables by capital letters. P(X=x) = P(X=X(s))

(b) Discrete PMFs are non negative, and sum to one over their support.

(c) **Definition (Bernoulli distribution).** An random variable X is said to have a Bernoulli distribution with parameter p if P(X = 1) = p and P(X = 0) = 1 ∼ p, where $0 < p < 1$. We write this as X ∼ Bern(p).

(d) **Theorem.** Let X ∼ $Bin(n, p)$, and $q = 1 - p$ (often taken to denote the failure of a Bernoulli trial). Then $n - X \sim Bin(n, q)$.

(e) **Theorem (Hypergeometric PMF).** Consider an urn with w white balls and b blacks balls. We draw n balls out of the urn at random without replacement such that all the $\binom{w+b}{n}$ samples are equally likely. Let X be the number of white balls in the sample. Then X is said to have the Hypergeometric distribution with parameters w, b and n: $X \sim HGeom(w, b, n)$. Then the PMF of X is
$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{b+w}{n}}$

(f) **Theorem.** If $X \sim HGeom(w, b, n)$ and $Y \sim HGeom(n, w + b - n, w)$, then X and Y have the same distribution.

(g) **Theorem.** If X ∼ Bin(n,p), Y ∼ Bin(m,p), and X is independent of Y, then the conditional distribution of $X|X + Y = r \sim$ hgeom(n,m,r).

(h) **Theorem (Binomial as a limiting case of the Hypergeometric).** If X∼ HGeom(w, b, n) and N = w+b approaches infinity such that p = w/(w + b) remains fixed, then the PMF of X converges to the Bin(n, p) PMF.

(i) **Theorem (PMF of g(X))**. Let X be a discrete random variable and $gR- > R$. Then the support of g(X) is the set of all y such that g(x) = y for at least one x in the support of X, and the PMF of g(X) is $P(g(X) = Y) = \sum_{x:g(x)=y} P(X = x)$

(j) **Definition (Independence of two random variables).** Random variables X and Y are said to be independent if $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ for all $x, y \in R$. In the discrete case, this is equivalent to the condition P(X = x, Y = y) = P(X = x)P(Y = y); for all x in the support of X and all y in the support of Y.

3. chapter 3: Expectation

(a) **Proposition (Monotonicity of expectation).** Let X and Y be random variables such that X ≥ Y with probability 1. Then E(X)≥E(Y), with equality holding if and only if X = Y with probability 1.

(b) A hypergeometric can be considered a sum of bernoulli random variables, but with their probabilities conditioned on the previous iteration. However, when calculating hte expectation, each of the bernoullis are equally like to be picked first, so the expecation is n * p where p = w/w+b

(c) **Theorem. Properties of Indicator Random Variables**

1. $I_A^k = I_A$
2. $I_{A^c} = 1 - I_A$
3. $I_{A \cap B} = I_A I_B$
4. $I_{A \cup B} - I_A + I_B - I_{A \cap B}$

(d) Inclusion Exclusion

$P(A_1 \cup ...A_n) = \sum_i P(A_i) - \sum_{j>i} P(A_i \cap A_j) + ...(-1)^n P(A_1...A_n)$

(e) Negative Binomial: Number of failures till a fixed number of successes. $=> NBin(r,p) P(X = x) = \binom{x+r-1}{r-1} p^r q^x$

(f) **Poisson** Note $\sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!} = e^x$ Useful for determining expectations and such. its a kernel

(g) **Theorem. Sum of independent poisson is poisson** if $X \sim Pois(\lambda_1)$ & $Y \sim Pois(\lambda_2)$ $X + Y \sim Pois(\lambda_1 + \lambda_2)$

(h) **Theorem. Poisson given a sum of Poissons is Binomial** $X \sim Pois(\lambda_1)$ & $Y \sim Pois(\lambda_2)$ $PoisX = k|X + Y = n \sim Binom(n, \lambda_1/(\lambda_1 + \lambda_2))$

4. Chapter 4: Continuous

(a) **Universality of the Uniform**. This is useful as it can simulate a RV with any discrete Distribution from the Unif.

ex

$F(x) = \frac{e^x}{1+e^x}$ $F^{-1}(u) = log(\frac{u}{1-u})$ ~logistic with $U \sim Unif(0,1)$

(b) **Normal distribution**

Note that the normal is symmetric, meaning tail areas are symmetric, andi n general $\phi(z) = \phi(-z)$

(c) **exponential distribution** This is the continuous version of the geometric distribution. $f(x) = \lambda e^{-\lambda x} x > 0$

(d) **Generalize exponential** $Y = \frac{X}{\lambda} \sim Expo(\lambda)$ with $X \sim Expo(1)$

(e) $E[X]$ with $X \sim Expo(\lambda) = \frac{1}{\lambda}$

(f) $Var[X]$ with $X \sim Expo(\lambda) = \frac{1}{\lambda^2}$

(g) Exponential RV is the only memoryless Random Variable => $P(X \geq s+t|X \geq s) = P(X \geq t)$

(h) '**Poisson Process** A poisson process is the process of arrivels with a rate λ if

1.The number of arrivals that occur in an interval of length t is a $Pois(\lambda t)$ random variable.

2.The number of arrivals that occur in disjoint time intervals are independent of each other.

(i) doisjoint intervals of the poisson process are all exponential distributions, but their sums are not exponential, they are gamma

(j) see notes for further elaboration on this section and bayesian view point of statistics

5. Chapter 5: Moments

(a) **Definition (Median).** We say that c is a median of a random variable X if $P(X \leq c) \geq .5 and P(X \geq c) \geq .5$

(b) **Definition (Mode).** For a discrete random variable X, we say that c is a mode of X if it maximizes the PMF: $P(X = c) \geq P(X = x)$ for all x. For a continuous random variable X with PDF f , we say that c is the mode if it maximizes the PDF: $f(c) \geq f(x)$ for all x.

Note we can have multiple modes and medians but not means!

(c) **Theorem.** Let X be a random variable with mean μ, and let m be the median of X. The value of c that minimizes the mean squared error $E(X-c)^2$ is c = μ.

A value of c that minimizes the mean absolute error $E|X-c|$ is c = m.

(d) **Definition (Symmetry of a random variable).** We say that a random variable X has a symmetric distribution about μ if $X - \mu$ has the same distribution as $\mu - X$. The number μ must be the mean E(X) if it exists, and must also be a median of the distribution of X.

(e) **Proposition (Symmetry in terms of the PDF).** Let X be a continuous random variable with PDF f . Then X is symmetric about μ if and only if $f(x) = f(2\mu - x)$ for all x.

(f) **Definition (Kinds of moments).** Let X be a random variable with mean μ and variance σ^2. For any positive integer n, the nth moment of X is $E(X^n)$, the nth central moment is $E[(X - \mu)^n]$, and the nth standardized moment is$E((\frac{X-\mu}{\sigma^2})^n)$ The mean is the first moment and the variance is the second central moment

(g) **Definition (Skewness).** The skewness of a random variable X with mean μ and variance σ^2 is the third standardized moment of X
Skew(X) = $E[(\frac{X-\mu}{\sigma^2})^3]$

(h) **Proposition (Odd central moments of a symmetric distribution).** Let X be symmetric about its mean μ. Then for any odd number m, the mth central moment $E(X - \mu)^m$ is 0 if it exists.

(i) **kurtosis** This represents how diffent a distributiobns pdf is , large kurtosis means sharp peak in the middle, heavy tails, and low shoulders.

(j) **Definition (Sample moments).** Let X1, X2, ... Xn be i.i.d. random variables. The kth sample moment is the random variable
$M_k = \frac{1}{n}\sum_{j=1}^{n} X^k$
and $E[M_k] = E[\frac{1}{n}\sum_{j=1}^{n} X^k] = E[X_1^k]$ Hence it is an unbiased estimator of the true kth moment

(k) $Var(\bar{X}_n) = \frac{\sigma^2}{n}$

(l) **Sample variance** = $S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$

(m) **Example: Bernoulli MGF** For $X \sim Bern(p)$, the MGF is $M(t) = E(e^{tX}) = e^{t0}P(X = 0) + e^tP(X = 1) = pet + q$ where q = 1 - p. Since M(t) is finite for any t $\in$ R, the MGF of a Bernoulli random variable is defined on the entire real line.

(n) **Example: Geometric MGF** For $X \sim Geom(p)$, the MGF is $M(t) = E(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}P(X = k) = \sum_{k=0}^{\infty} e^{tk}q^kp = \frac{p}{1-qe^t}$ for $q^{et} < 1$ or, equivalently, for $(-\infty, log(1/q))$,

(o) **Theorem (Moments via derivatives of MGFs).** Given the MGF M(t) of a random variable X, we can get the nth moment of X by evaluating its nth derivative of the MGF at 0 $E(X^n) = M^{(n')}(0)$

(p) **Theorem (MGF determines the distribution).** The MGF of a random variable determines its distribution: if two random variables have the same MGF, they must have the same distribution.

(q) **Theorem (MGF of a sum of independent random variables).** If X and Y are independent, then the MGF of X + Y is the product of the individual MGFs: $M_{X+Y}(t) = M_X(t)M_Y(t)$

Using this result, we can obtain the MGFs of the Binomial and Negative Binomial, which are sums of independent Bernoulli and Geometric i.i.d. random variables.

Example: Binomial MGF The MGF of a Bern(p) random variable is$pe^t + q$, so the MGF of a Bin(n, p) random variable is $M(t) = (pe^t + q)^n$

6. Chapter 6 : Joint Distributions

(a) **Using 2d lotus to get Expected Value**

$X, Y \sim N(0,1)$ find $E[|X-Y|] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|x-y|\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-y^2/2}dxdy$

we know $X - Y \sim N(0,2) => X - Y = \sqrt{2}Z$ with $Z \sim N(0,1) => E[|X-Y|] = \sqrt{2}E[|Z|] = \sqrt{2}\int_{-\infty}^{\infty}|z|e^{-z^2/2}dz$

$= 2\sqrt{2}\int_0^{\infty} ze^{-z^2/2}dz = \frac{2}{\sqrt{\pi}}$

(b) **Properties of Covariance**// 4. Cov(aX,Y) = aCov(X,Y) for any constant a in R

5. Cov(X + Y,Z) = Cov(X,Z) + Cov(Y,Z).

6. Cov(X + Y,Z + W) = Cov(X,Z) + Cov(X,W) + Cov(Y,Z) + Cov(Y,W).

7. Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y).

8. For any X1 X2 ... Xn $Var(X1 + X2 + ..Xn) = Var(X1) + ...Var(Xn) + 2\sum_{i<j} Cov(X_i, X_j)$

(c) **Theorem (Multinomial joint PMF).** If $X \sim Multk(n,p)$, then the joint PMF of the random vector X is $P(X_1 = n_1...X_k = n_k) = \frac{n!}{n_1!...n_k!}p_1^{n_1}...p_k^{n_k}$ where n1 + n2 + ... + nk = n.

(d) **Theorem (Multinomial margins).** If $X \sim Multk(n,p)$, then $Xj \sim Bin(n, p_j)$.

(e) **Theorem (Multinomial lumping).** If$X \sim Multk(n,p)$, then for any distinct i and j,$Xi + Xj \sim Bin(n, pi + pj)$. The random vector of counts obtained from merging

categories i and j is still Multinomial. For example, merging categories 1 and 2 gives $(X1 + X2, X3, ...Xk) \sim Mult_{k1}(n, (p1 + p2, p3, ..., pk))$

(f) **Theorem (Multinomial conditioning)**. If$X \sim Multk(n, p)$, then $(X2, .., Xk)|X1 = n1 \sim Mult_{k-1}(n - n1, (p2', ...pk'))$; where$p'_j = \frac{p_j}{p_2+...p_k}$

(g) **Definition (Multivariate Normal distribution).** A random vector X = (X1,...,Xk) is said to have a Multivariate Normal (MVN) distribution if **every** linear combination of the Xj has a Normal distribution. That is, we require t1X1 + t2X2 + ... + tkXk to have a Normal distribution for any t1, t2,..., tk in R. For k = 2 this distribution is called the Bivariate Normal (BVN). The marginal distribution of any component Xj of the random vector X is Normal which can be seen by taking tj = 1 and tj' = 0 for j, ! = j.

(h) Consider Z andW to be i.i.d. N(0, 1). Then (Z,W) is Bivariate Normal since the sum of independent Normals is also Normal. Furthermore, (Z + 2W, 3Z + 5W) is also Bivariate Normal since any linear combination of its components can be expressed as a linear combination of Z and W: t1(Z + 2W) + t2(3Z + 5W) = (t1 + 3t2)Z + (2t1 + 5t2)W.

(i) Note that subsets of Multivariate normals are also multivariate normal. Further, combinations of multivariate normals which are **independent of one another** they are also multivariate normals jointly.

(j) these are described with a mean vector and variance covariance matrix

(k) **Definition (Joint Moment Generating Function (MGF))**. The joint MGF of a random vector X = (X1,...,Xk) is the function which takes a vector of constants t' = (t1,...,tk) and returns

$M(t) = E[e^{t'X}] = E[e^{t_1X_1+...+t_kX_k}]$

With $= e^{t_jE[X_j]+\frac{1}{2}t_j^2Var(X_j)}$

$=> E[e^{t'X}] = E[e^{t_1X_1+...+t_kX_k}] = e^{t_1E[X_1]+...t_kE[X_k]+\frac{1}{2}Var(t_1X_1+...t_kX_k)}$

(l) **Theorem.** Within a Multivariate Normal random vector, uncorrelated implies independence. That is, if X = (X1,X2) is Multivariate Normal, where X1 and X2 are subvectors, and every component of X1 is uncorrelated with every component of X2,

then X1 and X2 are independent. In particular, if (X, Y) is Bivariate Normal with Corr(X, Y) = 0, then X and Y are independent.

7. Transformations

(a) **Theorem (Change of variables).** Let X = (X1,...,Xn) be a continuous random vector with joint PDF $f_X(x)$, and let Y = g(X) where g is an invertible function from $R^n to R^n$. Let y = g(x) which implies $x = g^{-1}(y)$. We consider the Jacobian matrix which is the matrix of all the partial derivatives $\frac{dx_i}{dy_j}$ that are assumed to exist and be continuous:
$\frac{dX}{dY} = \begin{bmatrix} \frac{dx_1}{dy_1} & \frac{dx_1}{dy_2} & \cdots \frac{dx_1}{dy_n} \\ \frac{dx_2}{dy_1} & \frac{dx_2}{dy_2} & \cdots \\ \frac{dx_n}{dy_1} & \cdots & \frac{dx_n}{dy_n} \end{bmatrix}$
and joint pdf $f_Y(y) = f_X(X)|\frac{dX}{dY}|$ and $|\frac{dX}{dY}| = |\frac{dY}{dX}|^{-1}$

(b) **Example: Bivariate Normal joint PDF** We find the PDF of the vector (Z,W) which follows a Bivariate Normal distribution with N(0, 1) marginals and $Corr(Z, W) = \rho \in (-1, 1)$. We construct (Z,W) by transforming another vector (X, Y) with $X, Y \sim N(0, 1)$, and X independent of Y as follows: $Z = X$ $W = \rho X + \sqrt{1-\rho^2}Y$; The inverse transformation that maps (Z,W) into (X, Y) is $X = Z$ $Y = -\frac{\rho}{\sqrt{1-\rho^2}}Z + \frac{1}{\sqrt{1-\rho^2}}W$
The Jacobian matrix is $\frac{d(x,y)}{d(z,w)} = \begin{bmatrix} 1 & 0 \\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{bmatrix}$
$=> f_{Z,W} = f_(x,y)(x,y) * |\frac{d(x,y)}{d(z,w)}| = f_(x,y)(x,y) * \frac{1}{\sqrt{1-\rho^2}}$

(c) **Theorem (Convolution sums and integrals).** If X and Y are independent discrete random variables, then the PMF of their sum T = X + Y is $P(T = t) = \sum_x P(Y = t - x)P(X = x) = \sum_y P(X = t - y)P(Y = y)$
and
$f_T(t) = \int_{-\infty}^{\infty} f_y(t - x)f_x(x)dx = \int_{-\infty}^{\infty} f_x(t - y)f_y(y)dx$

(d) **Example: Uniform convolution** Let X, Y be i.i.d. Unif(0; 1). Find the distribution of T = X + Y. Solution: The PDF of X and of Y is $g(x) = 1 if x \in (0, 1) and g(x) =$ 0, otherwise. The convolution formula gives:
$\int_{-\infty}^{\infty} f_y(t - x)f(x)dx = \int_{-\infty}^{\infty} g(t - x)g(x)dx$
we can see that we must have

$0 < t - x < 1 \& 0 < x < 1 \Rightarrow x < t < 1 + x \& 0 < x < 1 \Rightarrow 0 < x < t < 1 + x < 2$

Depending on the value of t, $t < 1$ or $t \geq 1$ x falls in either $0 < x < t$ or $t - 1 < x < 1$

$$\Rightarrow f_T(t) = \begin{cases} \int_0^t dx = t & 0 < t \leq 1 \\ \int_{t-1}^1 dx = t & 1 < t \leq 2 \end{cases}$$

(e) **Beta Distribution** A random variable X is said to have the Beta distribution with parameters $a > 0$and $b > 0$ if its PDF is

$\frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}$ $0 < x < 1$

(f) $\Gamma(a) = (a-1)!$

(g) **Definition (Gamma distribution).** A random variable Y is said to have the Gamma distribution with parameters a and λ, where $a > 0$ and $\lambda > 0$, if its PDF is

$f(y) = \frac{1}{\Gamma(a)}(\lambda y)^a e^{-\lambda y} \frac{1}{y}$

(h) The general $Gamma(a, \lambda)$ distribution can be constructed from the $X \sim Gamma(a, 1)$ by a scale transformation.

Consider the random variable $Y = \frac{X}{\lambda}$ for some $\lambda > 0$. By the change of variables formula with $x = \lambda y$ and $\frac{dx}{dy} = \lambda$, the PDF of Y is

$f_y(y) = f_x(x)|\frac{dx}{dy}| = \frac{1}{\Gamma(a)}(x)^a e^{-x} \frac{1}{x} * \lambda$

substituting for $x = \lambda y$ and cancelling we get

$\frac{1}{\Gamma(a)}(\lambda y)^a e^{-\lambda y} \frac{1}{y}$

8. Conditional Expectation:

(a) **Definition (Conditional expectation given an event).** Let A be an event with positive probability, P(A) ¿ 0.

Discrete: $E[X|A] = \sum_{x \in X} x * P(X = x|A)$

Continuous $E[X|A] = \int_{-\infty} x f_x(x|A)dx$ with $f_x(x|A) = \frac{d}{dx}P(X \leq x|A) = \frac{d}{dx}F(x|A)$

Using Bayes Rule $f(x|A) = \frac{P(A|x=x)f_x(x)}{P(A)}$

(b) **Theorem(Law of total expectation).** Let A1,A2,...,An be a partition of a sample space,with P(Ai) ¿ 0 for all i = 1,2,...,n. Let Y be a random variable on this sample space. Then $E(Y) = \sum_{i=1}^n E(Y|Ai)P(Ai)$.

The law of total probability is a specific case of this law using indicator random variables and the fundadmental bridge.

(c) Conditional Expectations are functions of the conditioned variable.

$E[Y|X] = g(X)$. thus we can ask for $E[g(x)]Var(g(X))$ ...

(d) **Theorem(Taking out what is known).** For any function h(·), we have E(h(X)Y — X) = h(X)E(Y — X).

Note that the above equality means that the random variable g1(X) = E(h(X)Y — X) is equal with the random variable g2(X) = h(X)E(Y — X).

(e) Linearity : $E[X_1 + X_2..X_n|Y] = E[X_1|Y] + E[X_2|Y]..E[X_n|Y]$

(f) Adams law: $E[E[X|Y]] = E[X]$

(g) Adams Law w/Extra conditioning : $E[E[Y|X, Z]|Z] = E[Y|Z]$

(h) **Theorem(Projection interpretation).** For any function $h(\Delta)$,the random variable $Y - E[Y|X]$ is uncorrelated with $h(X)$. Since $E(Y - E(Y|X)) = E(Y) - E(E(Y|X)) = E(Y) - E(Y) = 0$, this is equivalent with $E((Y - E(Y|X))h(X)) = 0$. Main point being that $E[Y|X]$ minimizes the function $|Y - E[Y|X]|$

(i) **Definition(Conditional variance).** The conditional variance of Y given X is $Var(Y|X) = E[[Y - E(Y|X)]^2|X]$

This is equivalent to $Var(Y|X) = E(Y^2|X) - (E(Y|X))^2$.

(j) **Theorem(Eve'slaw: connecting conditional variance to unconditional variance).** For any random variables X and Y, we have $Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$. This relation is known as the law of total variance, or as the variance decomposition formula.

Aka there is variance within a group, and there is a variance between groups

Happy holidays!