CSE 312: Foundations of Computing II
Advanced Topics Session #3

Joint Distributions, Random Vectors, and the Multinomial Distribution

Lecturer: Alex Tsun
Date: April 25, 2017

0 Introduction

Being able to do analysis with a single random variable is great, but we often need to consider multiple! In particular, machine learning and artificial intelligence often require us to represent joint distributions over features, and we will learn some of the tools needed to do more advanced probabilistic analysis.

1 The Negative Binomial Distribution

Consider the following problem: we flip a coin with $P(head) = p$ independently until we get our first success. Let X be the number of flips up to and including the first head. We know that $X \sim Geo(p)$ and therefore $E[X] = \frac{1}{p}$. Recall the pmf for the geometric rv is

$$p_X(k) = P(X = k) = (1-p)^{k-1}p, \;\; k = 1,2, \ldots$$

This is because the first $k-1$ flips must've been tails and the last flip must have been heads.

Now let Y be the number of flips up to and including the r^{th} head. What is the probability mass function for Y, $p_Y(k)$?

$$p_Y(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \;\; k = r, r+1, \ldots$$

This is because the last flip must be a head. But out of the first $k-1$ flips, we choose $r-1$ positions for the other heads, and we have r heads total and $k-r$ tails total. We say that Y has a **negative binomial distribution with parameters r and p**, and write that $\boldsymbol{Y \sim NegBin(r,p)}$**.**

Exercise: What is $E[Y]$, where $Y \sim NegBin(r,p)$? (Hint: Use Linearity of Expectation).
Solution: Let $X_1, \ldots, X_r \sim Geo(p)$ be **independent and identically distributed (iid)** random variables. Then, $Y = X_1 + \cdots + X_r = \sum_{i=1}^{r} X_i$ (Why?).

$$E[Y] = E\left[\sum_{i=1}^{r} X_i\right] = \sum_{i=1}^{r} E[X_i] = \sum_{i=1}^{r} \frac{1}{p} = \frac{r}{p}$$

2 Jointly Distributed Random Variables

Let X and Y be random variables. We define the **joint distribution (joint probability mass function) of X and Y** as $p_{X,Y}(x,y) = P(X = x, Y = y)$. Let $\Omega_{X,Y}$ denote the **joint range/support of X and Y**, that is, $\Omega_{X,Y} = \{(x,y)|p_{X,Y}(x,y) > 0\}$. Because these are probabilities, we have the natural requirement that

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

We will learn more through examples. Suppose we roll a fair four-sided die independently twice, and let X be the value on the first roll, and Y be the value on the second roll.

Exercise: Specify $\Omega_{X,Y}$ and find $p_{X,Y}(x,y)$.
Solution: Let $[n] = \{1,2,\dots,n\}$. $\Omega_{X,Y} = \{(x,y)|x \in [4], y \in [4]\}$. Since these are fair dice and rolls are independent,

$$p_{X,Y}(x,y) = P(X = x, Y = y) = \frac{1}{16}, \quad (x,y) \in \Omega_{X,Y}$$

Now let's say we only cared about the distribution of X. What is $p_X(t) = P(X = t)$? Independence allows us to ignore Y, and we know the die is fair, so

$$p_X(t) = \frac{1}{4}, \quad t \in [4]$$

Symmetrically, $p_Y(t) = \frac{1}{4}, t \in [4]$. Notice that, because X and Y are "independent",

$$p_{X,Y}(x,y) = \frac{1}{16} = \frac{1}{4} \cdot \frac{1}{4} = p_X(x)p_Y(y)$$

We say that two random variables X and Y are **independent** if and only if $\Omega_{X,Y} = \Omega_X \times \Omega_Y$ and

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \quad \forall (x,y) \in \Omega_{X,Y}$$

and we write that $\boldsymbol{X \perp Y}$. Recall for two sets A, B, their Cartesian product is $A \times B = \{(a,b)|a \in A, b \in B\}$. Here $\Omega_{X,Y} = \{(x,y)|x \in [4], y \in [4]\} = \{x|x \in [4]\} \times \{y|y \in [4]\} = \Omega_X \times \Omega_Y$.

Now suppose $S = \max\{X,Y\}$ and $T = \min\{X,Y\}$. Find $p_{S,T}(s,t)$. Are S and T independent?

No they are not. This is because $S \geq T$ always. We can also see that the joint range $\Omega_{S,T} = \{(s,t)|s \in [4], t \in [4], s \geq t\} \neq [4] \times [4] = \Omega_S \times \Omega_T$, so we can immediately claim that they are dependent.

$p_{S,T}(s,t)$

$s\backslash t$	1	2	3	4
1	1/16	0	0	0
2	2/16	1/16	0	0
3	2/16	2/16	1/16	0
4	2/16	2/16	2/16	1/16

Now let's say we only care about the distribution of the maximum, or $p_S(s)$. How would we go about it?

For example, let's consider finding $p_S(2) = P(S = 2)$. It is very natural to say, "sum up the entire row where $S = 2$". And this is completely correct! We also call a distribution with fewer variables (in our case, just 1) than the joint distribution a **marginal distribution**. We define it as

$$\boldsymbol{p_S(s) = P(S = s) = \sum_t p_{S,T}(s,t)}$$

So here,

$$p_S(s) = \begin{cases} 1/16, & s = 1 \\ 3/16, & s = 2 \\ 5/16, & s = 3 \\ 7/16, & s = 4 \end{cases}$$

Notice that

$$\sum_s p_S(s) = 1$$

as it should!

What if we wanted to know the value of the maximum, conditioned on the value of the minimum? Or the random variable $S|T$? We define the **conditional distribution of S given T** as $\boldsymbol{p_{S|T}(s|t) = P(S = s|T = t) = \frac{P(S=s,T=t)}{P(T=t)} = \frac{p_{S,T}(s,t)}{p_T(t)}}$. This is just the definition of conditional probability! Again, naturally, we have

$$\sum_s p_{S|T}(s|t) = 1 \ \ \forall t$$

Exercise: Find $p_{S|T}(s|3) = P(S = s|T = 3)$. What is $\Omega_{S|T}$?
Solution: If $T = 3$, then the minimum value was 3, so $S|T \in \{3,4\}$ only!

$$p_T(3) = \sum_s p_{S,T}(s,3) = \frac{1}{16} + \frac{2}{16} = \frac{3}{16}$$

$$p_{S|T}(s|3) = \frac{p_{S,T}(s,3)}{p_T(3)}$$

$$p_{S|T}(1|3) = p_{S|T}(2|3) = 0$$
$$p_{S|T}(3|3) = \frac{p_{S,T}(3,3)}{p_T(3)} = \frac{1/16}{3/16} = 1/3$$
$$p_{S|T}(4|3) = \frac{p_{S,T}(4,3)}{p_T(3)} = \frac{2/16}{3/16} = 2/3$$

Now let's talk about expectation and variance. Recall that for a single variable X,

$$E[X] = \sum_x x p_X(x)$$
$$E[g(X)] = \sum_x g(x) p_X(x)$$

We define expectation for a function of two random variables similarly!

$$E[g(X,Y)] = \sum_x \sum_y g(x,y) p_{X,Y}(x,y)$$

Exercise: Prove that $E[X+Y] = E[X] + E[Y]$, for any two random variables X and Y (even if not independent)!
Solution:

$$E[X+Y] = \sum_x \sum_y (x+y) p_{X,Y}(x,y) = \sum_x \sum_y x p_{X,Y}(x,y) + \sum_y \sum_x y p_{X,Y}(x,y)$$
$$= \sum_x x \sum_y p_{X,Y}(x,y) + \sum_y y \sum_x p_{X,Y}(x,y) = \sum_x x p_X(x) + \sum_y y p_Y(y) \ [def\ of\ marginal]$$
$$= E[X] + E[Y]$$

Exercise: Prove that, if $X \perp Y$, then $E[XY] = E[X]E[Y]$.
Solution:

$$E[XY] = \sum_x \sum_y xy p_{X,Y}(x,y) = \sum_x \sum_y xy p_X(x) p_Y(y) \ [independence]$$
$$= \sum_x x p_X(x) \sum_y y p_Y(y) = \left(\sum_x x p_X(x)\right)\left(\sum_y y p_Y(y)\right) = E[X]E[Y]$$

Note that the **converse is not true**! If $E[XY] = E[X]E[Y]$, it does not necessarily imply $X \perp Y$.

Recall the variance of a random variable X was defined as $E[(X-\mu_X)^2] = E[X^2] - E^2[X]$, where $\mu_X = E[X]$. We define the **covariance of X and Y** as $\boldsymbol{Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]}$. This is

very similar to why we defined variance: it describes how far on average X varies about its mean while Y varies about its mean. If $Cov(X,Y) > 0$, we say that X and Y are **positively correlated** (that is, very roughly, if X is higher, then Y is typically higher, and vice versa). If $Cov(X,Y) < 0$, we say that X and Y are **negatively correlated**. Notice that $\boldsymbol{Cov(X,X) = E[(X-\mu_X)^2] = Var(X)}$.

Exercise: Show from the definition of covariance that $Cov(X,Y) = E[XY] - E[X]E[Y]$.
Solution:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y]$$
$$= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X\mu_Y = E[XY] - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y$$
$$= E[XY] - \mu_X\mu_Y = E[XY] - E[X]E[Y]$$

Exercise: If $Cov(X,Y) = 0$, should we say that X and Y are independent?
Solution: No!! If $Cov(X,Y) = 0$, then $E[XY] = E[X]E[Y]$, and we cannot claim independence (stated earlier).

We define the **correlation of X and Y as $\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$**. This measure is always between -1 and 1, and you may have seen this before in linear regression! Correlation is a standardized measure of covariance – it is just normalized by the variances of X and Y. When does $\rho_{X,Y} = \pm 1$? This is the case when $Y = aX + b$ for any $a \neq 0, b \in \mathbb{R}$! That means, as X changes, we can exactly predict Y because Y is some linear function of X. We will show this for $a = 1, b = 0$. Suppose $X = Y$. $\rho_{X,Y} = \frac{Cov(X,X)}{Var(X)} = \frac{Var(X)}{Var(X)} = 1$.

Properties of Covariance:

- $Cov(aX+b, cX+d) = acCov(X,Y)$
- $Cov(X+Y, W+Z) = Cov(X,W) + Cov(Y,W) + Cov(X,Z) + Cov(Y,Z)$
- $Cov(X,Y) = Cov(Y,X)$

Recall in class we showed that, if $X \perp Y$, then $Var(X+Y) = Var(X) + Var(Y)$.

Exercise: Show that, in general, $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$. Notice that this doesn't contradict what we said earlier: if $X \perp Y$, $Cov(X,Y) = 0$ and so $Var(X+Y) = Var(X) + Var(Y)$.
Solution:

$$Var(X+Y) = E[(X+Y)^2] - E^2[X+Y] = E[X^2 + 2XY + Y^2] - (\mu_X + \mu_Y)^2$$
$$= E[X^2] + 2E[XY] + Y^2 - \mu_X^2 - 2\mu_X\mu_Y - \mu_Y^2$$
$$= (E[X^2] - \mu_X^2) + (E[Y^2] - \mu_Y^2) + 2(E[XY] - \mu_X\mu_Y)$$
$$= Var(X) + Var(Y) + 2Cov(X,Y)$$

Now we continue with the example. Recall X and Y were independent rolls of a fair four-sided die, and $S = \max\{X,Y\}$ and $T = \min\{X,Y\}$.

Exercise: Find $E[ST]$.
Solution: We use the clever observation that $ST = XY$ (why?), then exploit that $X \perp Y$.

$$E[ST] = E[XY] = E[X]E[Y] = \frac{5}{2} \cdot \frac{5}{2} = \frac{25}{4}$$

Note that $E[ST]$ may not equal $E[S]E[T]$ since S and T are not independent. However, $ST = XY$ because S is either X or Y, and T is the other (since one is the min and one is the max). Then since X and Y are independent, we can write $E[XY] = E[X]E[Y]$.

Exercise: Do you expect $Cov(S,T) > 0$ or $Cov(S,T) < 0$?
Solution: I expect $Cov(S,T) > 0$. This is because higher values of S correspond to higher values of T. If $T = 3$ for example, $S \geq 3$ because $S \geq T$ always.

3 Random Vectors

Let $X_1, \dots, X_n$ be arbitrary real-valued random variables, and stack them into a vector

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

We call $\boldsymbol{X}$ an **n-dimensional random vector (rvtr)**. We will see why these are useful later. What can we do with these? We define the expectation of a random vector just as you would hope: coordinate-wise.

$$E[\boldsymbol{X}] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix}$$

Define the **covariance matrix (or variance-covariance matrix)** of a rvtr $\boldsymbol{X} \in \mathbb{R}^n$ with $E[\boldsymbol{X}] = \boldsymbol{\mu}$ as the matrix $Var(\boldsymbol{X}) = \Sigma$ whose entries $\Sigma_{ij} = Cov(X_i, X_j)$.

$$\Sigma = Var(\boldsymbol{X}) = Cov(\boldsymbol{X}) = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T] = E[\boldsymbol{X}\boldsymbol{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$$
$$= \begin{bmatrix} Cov(X_1,X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Cov(X_2,X_2) & \dots & Cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \dots & Cov(X_n,X_n) \end{bmatrix}$$
$$= \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \dots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Var(X_2) & \dots & Cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \dots & Var(X_n) \end{bmatrix}$$

where the second equality is because $Cov(X_i, X_i) = Var(X_i)$. Notice that the covariance matrix is symmetric ($\Sigma_{ij} = \Sigma_{ji}$), and contains variances along the diagonal. For those with more linear algebra background, covariance matrices are also positive semi-definite. That is, $\forall \boldsymbol{v} \in \mathbb{R}^n, \boldsymbol{v}^T \Sigma \boldsymbol{v} \geq 0$.

Proof:

First, $Cov(X_i, X_j) = Cov(X_j, X_i)$, so Σ is symmetric. Second, let $\boldsymbol{v} \in \mathbb{R}^n$ be arbitrary. Recall $\Sigma = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\boldsymbol{T}}]$, and define the random variable (not random vector) $Y \equiv \boldsymbol{v}^{\boldsymbol{T}}(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{v} \cdot (\boldsymbol{X} - \boldsymbol{\mu})$. Then $\boldsymbol{v}^{\boldsymbol{T}}\Sigma\boldsymbol{v} = \boldsymbol{v}^{\boldsymbol{T}}E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\boldsymbol{T}}]\boldsymbol{v} = E[\boldsymbol{v}^{\boldsymbol{T}}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{v}] = E[(\boldsymbol{v}^{\boldsymbol{T}}(\boldsymbol{X} - \boldsymbol{\mu}))((\boldsymbol{X} - \boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{v})] = E[(\boldsymbol{v} \cdot (\boldsymbol{X} - \boldsymbol{\mu}))((\boldsymbol{X} - \boldsymbol{\mu}) \cdot \boldsymbol{v})] = E[Y^2] \geq 0$.

Q.E.D.

Properties of expectation and variance still hold for rvtrs. Let $\boldsymbol{X}$ be an n-dimensional rvtr, $A \in \mathbb{R}^{n \times n}$ be a constant matrix, $\boldsymbol{c} \in \mathbb{R}^n$ be a constant vector, and suppose $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{c}$. Then,

$$E[\boldsymbol{Y}] = AE[\boldsymbol{X}] + \boldsymbol{c}$$
$$Var(\boldsymbol{Y}) = A\,Var(\boldsymbol{X})\,A^T$$

Just as an FYI, for two rvtrs $\boldsymbol{X}, \boldsymbol{Y}$ in $\mathbb{R}^n$ with $\boldsymbol{\mu_X} = E[\boldsymbol{X}]$ and $\boldsymbol{\mu_Y} = E[\boldsymbol{Y}]$, their **cross-covariance matrix** is given as $Cov(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - \boldsymbol{\mu_X})(\boldsymbol{Y} - \boldsymbol{\mu_Y})^T]$.

4 The Multinomial Distribution

Recall the binomial distribution: if we flip a coin with $P(head) = p$ independently n times, and let X be the number of heads, then we say X has the **binomial distribution with parameters $\boldsymbol{n}$ and $\boldsymbol{p}$**, and we write $\boldsymbol{X \sim Bin(n, p)}$.

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \;\; k = 0,1, \dots, n$$

This is because a particular sequence with exactly k heads out of n flips has probability $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ ways to choose which flips were heads and the rest must be tails. As an exercise, verify that $E[X] = np$ and $Var(X) = np(1-p)$ (Hint: use linearity of expectation/variance by writing X as the sum of iid $Ber(p)$.)

A generalization of the binomial model is when there are r different outcomes in a sequence of n independent trials, with $P(outcome\ i) = p_i$ for $1 \leq i \leq r$, and $p_1 + \cdots + p_r = 1$. Let $(X_1, \dots, X_r)$ be the random vector such that X_i is the number of times we observed outcome i in n independent trials, where clearly $X_1 + \cdots + X_r = n$. We write that $\boldsymbol{(X_1, \dots, X_r) \sim Mult_r(n, p_1, \dots, p_r)}$. Find the joint probability mass function for the **multinomial random vector**, $p_{X_1,\dots,X_r}(k_1, \dots, k_r)$.

$$p_{X_1,\dots,X_r}(k_1, \dots, k_r) = \frac{n!}{k_1! \dots k_r!} p_1^{k_1} \dots p_r^{k_r} = \binom{n}{k_1, \dots, k_r} \prod_{i=1}^{r} p_i^{k_i}, k_1 + \cdots + k_r = n$$

The derivation is almost the same as that of the binomial. Think about how to define the "negative multinomial distribution!

Find the covariance matrix Σ for $\boldsymbol{X}$, when $\boldsymbol{X} \sim Mult_r(n, \boldsymbol{p})$, where $\boldsymbol{X}$ is r-dimensional and $\boldsymbol{p} \in \mathbb{R}^r$ is a probability vector (a vector whose entries are nonnegative and sum to 1). If you think about it, the distribution of a single rv X_i is $Bin(n, p_i)$ (why?). So $\Sigma_{ii} = Var(X_i) = np_i(1 - p_i)$. Now we need to find $\Sigma_{ij} = Cov(X_i, X_j)$ for $i \neq j$. Before we compute this, should we expect a positive or negative covariance? We should expect a negative covariance because if one increases, the other cannot be as high (since they all sum to n). Calculating this covariance is too hard for us right now, but it turns out that $Cov(X_i, X_j) = -np_ip_j$. So this fully specifies Σ.

5 Conclusion

These are all very important topics! Please make sure you understand everything written above – everything here is crucial for success in future courses in machine learning, artificial intelligence, natural language processing, computer vision, and randomized algorithms.

6 Additional Exercises

1. Let X and Y be random variables with ranges Ω_X and Ω_Y, respectively. Write an expression for $P(X = Y)$. (It may be a sum, or product).

$$P(X = Y) = \sum_{t \in \Omega_X \cap \Omega_Y} p_{X,Y}(t, t)$$

2. Suppose we are measuring particle emissions, and the number of particles emitted follows a Poisson distribution with parameter λ, $X \sim Poi(\lambda)$. Suppose our device to measure emissions is not always entirely accurate – sometimes we fail to observe particles that actually are emitted. So for each particle actually emitted, say we have probability p close to 1 of actually recording it, independently of other particles. Let Y be the number of particles we actually measure. What is $p_Y(y)$?

$$p_Y(y) = P(Y = y) = \sum_{x=y}^{\infty} P(Y = y | X = x) P(X = x)$$

$$= \sum_{x=y}^{\infty} \binom{x}{y} p^y (1-p)^{x-y} e^{-\lambda} \frac{\lambda^x}{x!}$$

$$= e^{-\lambda} p^y \sum_{x=y}^{\infty} \frac{x!}{y!\,(x-y)!} (1-p)^{x-y} \frac{\lambda^x}{x!}$$

$$= \frac{e^{-\lambda} p^y}{y!} \sum_{x=y}^{\infty} \frac{\lambda^x}{(x-y)!} (1-p)^{x-y}$$

$$k = x - y$$

$$\frac{e^{-\lambda} p^y}{y!} \sum_{k=0}^{\infty} \frac{\lambda^{k+y}}{k!} (1-p)^k$$

$$= \frac{e^{-\lambda}(\lambda p)^y}{y!} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}(1-p)^k \quad [Taylor\ series\ for\ e^{-\lambda(1-p)}]$$

$$= \frac{e^{-\lambda}(\lambda p)^y}{y!} e^{-\lambda(1-p)} = \frac{e^{-p\lambda}(\lambda p)^y}{y!}$$

So $Y \sim Poi(p\lambda)$.

3. Suppose $X_1, \dots, X_n$ are iid random variables with common mean μ and common variance σ^2. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (the sample mean). Find $E[\bar{X}]$ and $Var(\bar{X})$.

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n\mu = \mu$$

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

4. Suppose X, Y, Z are discrete random variables with the following joint probability mass function

$x = 0$

$y\backslash z$	5	6
3	1/4	1/16
4	1/16	1/8

$x = 1$

$y\backslash z$	5	6
3	1/16	5/16
4	1/16	1/16

There are two possible values of each random variable, and as an example, $p_{X,Y,Z}(1,3,6) = 5/16$.

a) Find the marginal joint probability mass function, $p_{Y,Z}(y,z)$, and specify the values of this function for all $y, z \in \mathbb{R}$.

$$p_{Y,Z}(y,z) = \begin{cases} 5/16, & (y,z) = (3,5) \\ 6/16, & (y,z) = (3,6) \\ 2/16, & (y,z) = (4,5) \\ 3/16, & (y,z) = (4,6) \end{cases}$$

b) Find $E[(Y-1)(Z-3)]$.

$$E[(Y-1)(Z-3)] = \sum_{(y,z)\in\Omega_{Y,Z}} (y-1)(z-3) \cdot p_{Y,Z}(y,z)$$

$$= 2 \cdot 2 \cdot \frac{5}{16} + 2 \cdot 3 \cdot \frac{6}{16} + 3 \cdot 2 \cdot \frac{2}{16} + 3 \cdot 3 \cdot \frac{3}{16} = \frac{20 + 36 + 12 + 27}{16} = \frac{95}{16}$$

c) Find the marginal probability mass function, $p_X(x)$ and specify the values of this function for all $x \in \mathbb{R}$.

$$p_X(x) = \begin{cases} 1/2, & x = 0 \\ 1/2, & x = 1 \end{cases}$$

d) Identify X as one of the named distributions, and give $E[X]$ and $Var(X)$.

$X \sim Ber\left(\frac{1}{2}\right)$ so $E[X] = \frac{1}{2}$, $Var(X) = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}$. Alternatively, $X \sim Unif(0,1)$ so $E[X] = \frac{1}{2}$ and $Var(X) = \frac{(1-0)(1-0+2)}{12} = \frac{1}{4}$.

e) Find $E[16(Y-1)(Z-3) + 2X^7 + 2]$.

Notice $X^7 \equiv X$, so $E[X^7] = E[X] = \frac{1}{2}$.

$$E[16(Y-1)(Z-3) + 2X^7 + 2] = 16E[(Y-1)(Z-3)] + 2E[X^7] + 2 = 95 + 1 + 2 = 98$$

5. Suppose W, X, Y, Z are arbitrary random variables. Write an expression for $p_{W,Y}(w,y)$ and then $p_{Z|W,Y}(z|w,y)$, only in terms of $p_{W,X,Y,Z}$ and summations.

$$p_{W,Y}(w,y) = \sum_x \sum_z p_{W,X,Y,Z}(w,x,y,z)$$

$$p_{Z|W,Y}(z|w,y) = \frac{p_{W,Y,Z}(w,y,z)}{p_{W,Y}(w,y)} = \frac{\sum_x p_{W,X,Y,Z}(w,x,y,z)}{\sum_x \sum_z p_{W,X,Y,Z}(w,x,y,z)}$$

6. A generalization of the hypergeometric model is when there are r different colors of balls in a bag, having K_i balls of each color, $1 \le i \le r$. Let $N = K_1 + \cdots + K_r$, the total number of balls in the bag, and suppose we draw n without replacement. Let $(X_1, \ldots, X_r)$ be the random vector (vector of random variables) such that X_i is the number of balls of color i we drew, where clearly $X_1 + \cdots + X_r = n$. We write that $(\boldsymbol{X_1, \ldots, X_r}) \sim \boldsymbol{MVHG_r(N, n, K_1, \ldots, K_r)}$. Find the joint probability mass function for the **multivariate hypergeometric random vector**, $p_{X_1,\ldots,X_r}(k_1, \ldots, k_r)$.

$$p_{X_1,\ldots,X_r}(k_1, \ldots, k_r) = \frac{\binom{K_1}{k_1} \cdots \binom{K_r}{k_r}}{\binom{N}{n}} = \frac{\prod_{i=1}^r \binom{K_i}{k_i}}{\binom{N}{n}}, \quad k_1 + \cdots + k_r = n$$

Derivation identical to that of the hypergeometric.