ASSIGNMENT 1 - PROBABILITY, STATISTICAL MODELS & TESTS
INF2190 - FALL 2022
AUDREY MEDAINO-TARDIF

Note this assignment is to be done individually. Make a copy of this doc and use it as a template for the assignment. Share your google doc with me: tegan.maharaj@utoronto.ca. Don't share it with anyone else. I may ask you for your source R code at any time during the semester. Submit a PDF of your doc via Quercus.

PART I: PROBABILITY
Tutorial: https://michael-franke.github.io/IDA-2019/tutorials/Tutorial_5.zip

1. Probability Vocab: elementary outcomes, events, probability, distributions

1.1 M&Ms

A. A probability distribution is a **function** that assigns a **probability** to each possible **event**

B. The distribution of colours in the bag is given by **its elementary outcome**

C. Randomly drawing M&Ms and checking the colour is an example of a random **process.**

D. What is the sample space of this process? **QMnM = {brown, blue, orange, red}**

E. Give an example of an event for this process: **A = {blue}**

F. What is the probability of picking a brown M&M? **A = {brown} P= 3/10 = 0.3**

1.2 CARDS

A. What is Ω and what does it mean when I say that P(Ω)=1 Why is the statement true? #
The set of all possible results (elementary outcomes). It is true because as above with the m&ms if any subset A of Ω is called an event, then P is equally a subset event of Ω and in this case we are considering a random process of picking one card out of the deck. Therefore, the event will always result in 1, as we are picking 1 card out of the deck.

B. Name a few events that could happen as a result of our random process.

P(Ω) = {heart}

P(Ω) = {seven}

C. What is the probability of picking a black card? **P(black) = 28/56 = 0.50 (50%)**

D. What is the probability of picking a Queen? **P(queen) = 4/56 = 0.07 (7%)**

E. What is the probability of picking a spade or a red King? **P(spade, red king) = 16/56 = 0.29 (29%)**

2. Probability Distributions from Samples

A. Why do we sometimes need to approximate probability distributions?

We need to approximate probability distribution, as real world problems don't usually come in easy visual and quantifiable fractions, as was the case with the MnMs and the deck of cards. Therefore, we need to approximate the distribution for the data to be more easily calculated and dealt with.

B. When we approximate a distribution, we define them as a **function** that returns a **representative sample of the distribution.**

C. Insert plot and R code of the density of a distribution with mean 3 and standard deviation 2.

```
y_dist <- dnorm(x, mean = 3, sd = 2)

qplot(x, y_dist, geom = "line")
```

D. Is the above plot a standard normal distribution?

I believe it is a standard normal deviation, as it is a bell shape with 95% of its data being found within 2 standard deviations of mean. In this case, we have a standard deviation of 2.

E. Does the density of this function change as you increase the sample size

No.

F. Insert plots and R code for 2 different sample sizes here:

```
sample_size <- 2000

x <- seq(-5, 5, length = sample_size)

y_dist <- dnorm(x, mean = 0, sd = 1)

qplot(x, y_dist, geom = "line")


sample_size <- 4500

x <- seq(-5, 5, length = sample_size)

y_dist <- dnorm(x, mean = 0, sd = 1)

qplot(x, y_dist, geom = "line")


sample_size <- 25
```

```
x <- seq(-5, 5, length = sample_size)

y_dist <- dnorm(x, mean = 0, sd = 1)

qplot(x, y_dist, geom = "line")
```

G. Give an example of real-world data that seems to be well-modeled by a normal distribution (search online; insert a plot of the real data that you find, and give 1-2 lines explanation of why you think it is well-modeled by a normal distribution.)

The average temperature possibilities in Canada, rounded by 0.5

```
x <- seq(-40, 40, by = 0.5)

print(mean(x))

y <- dnorm(x, mean = 3.5, sd = 1.0)

plot(x,y, main = "Normal Distribution", col = "blue")
```

H. Plot a binomial distribution with values different from in the tutorial. Insert plots and R code here:

```
x_axis <- 1:25

binom_dist_02 <- as_tibble(dbinom(x_axis, size = length(x_axis), prob = 0.2))

binom_dist_01 <- as_tibble(dbinom(x_axis, size = length(x_axis), prob = 0.1))

binom_dist_08 <- as_tibble(dbinom(x_axis, size = length(x_axis), prob = 0.8))


ggplot(mapping = aes(x = x_axis, y = value)) +
  geom_line(data = binom_dist_02, color = "red") +
  geom_line(data = binom_dist_01, color = "blue") +
  geom_line(data = binom_dist_08, color = "green")
```

I. Give an example of real-world data that seems to be well-modeled by a binomial distribution (search online; insert a plot of the real data that you find, and give 1-2 lines explanation of why you think it is well-modeled by a binomial distribution.)

An example would be the success rate of a basketballers free throws into the basket. Eacg time he shoots he has the same probability of success, independent of his previous shot.

```
succtrial<- 0:20

plot(succtrial, dbinom(succtrial, size=20, prob=.3),type='h')
```

J. Plot a colour-coded Poisson distribution with values different from in the tutorial. Insert plots and R code here:Sa

```
x_axis <- 1:75

pois_dist_1 <- as_tibble(dpois(x_axis, lambda = 2))

pois_dist_4 <- as_tibble(dpois(x_axis, lambda = 6))

pois_dist_10 <- as_tibble(dpois(x_axis, lambda = 25))


ggplot(mapping = aes(x = x_axis, y = value)) +
  geom_line(data = pois_dist_1, color = "red") +
  geom_line(data = pois_dist_4, color = "blue") +
  geom_line(data = pois_dist_10, color = "green")
```

K. Give an example of real-world data that seems to be well-modeled by a poisson distribution (search online; insert a plot of the real data that you find, and give 1-2 lines explanation of why you think it is well-modeled by a poisson distribution.)

Firstly, the poisson distribution is a discrete distribution meaning that it is a countable number of values and it calculates the rate, or the number of events or counts that occur within a fixed period of time. Therefore, one of the real world examples I found was the number of website visitors per hour:

```
hours <- 1:25

plot(hours, dpois(hours, lambda=5),
    type='h',
    main='Number of Website Visitors per Hour',
    ylab='Probability',
```

xlab ='# of visitors',

lwd=3)

It is well-modeled by a poisson distribution as it correctly demonstrates the rate and probability of visitors, which is exactly what the poisson model seeks to represent.

3. Joint, Marginal and Conditional Probabilities

3.1 Linda

"Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."

A. Rank the following statements in order of their probability (reorder them here:)

a. 1. Linda is active in the feminist movement
b. 2. Linda is a bank teller
c. 3. Linda is a bank teller and is active in the feminist movement

B. What is a Conjunction Fallacy? (describe in words)

Conjunction fallacy is an argument put forth by Tversky and Kahneman where a scenario like the #3 seems to be more probable because it seems to better represent Linda as a whole rather than simply one aspect of her.

C. C. Using mathematical notation for the probability of C, D, and their joint probability, show their relationship in terms of magnitude.. Use = (or \eq) for equals, >/= (or \geq) for greater-than-or-equal, </= (or \leq) for less-than-or-equal

I don't understand the question.

P(C) ≥ P(C^D)≤ P(D)

D. What is a marginal probability? (describe in words)

Marginal probability is the calculation or chance for which an event will happen, the probability of an event happening notwithstanding other conditions, variables or other outcomes.

E. What is the marginal probability of Linda being active in the feminist movement? (Include both the equation and the final answer)

P(feminist) = P (0.9) = 0.9

F. What is the joint probability of Linda being both a bank teller and active in the feminist movement? (Include both the equation and the final answer)

P(bank ^ feminist) = P(bank) x P(feminist) = 0.2 x 0.9 = 0.18

G. What is a conditional probability? (describe in words)

The outcome of this event relies on a condition, that is on the result of another event. The probability of me stopping for gas on my way to work is conditional to the probability of my car needing gas.

H. What is a base rate? (you may need to search online for this).

The base rate is also known as prior probabilities or the likelihood. It is probability based in the absence of other information.

https://link.springer.com/referenceworkentry/10.1007/978-0-387-79061-9_289

https://en.wikipedia.org/wiki/Base_rate

I. Why does it matter to account for base rates?

It matters for transparency and unbiased (or as unbiased as possible!). For example, if a class has a rate of absence of 25%, it is a static number which simply states that on a student list, you have 75% probability of landing on a name of a student that is present in class. However, often not accounting for this base rate, causes "base rate neglect" or fallacy causing someone to use this information to draw other conclusions, such as if a student has a 4.0 GPA, it can wrongly be assumed that the student is present 75% of the time in class.

https://link.springer.com/referenceworkentry/10.1007/978-0-387-79061-9_289

3.2 Cards again

If we roll a 3 or a 4, we'll draw a diamond.

If we roll a 5, we'll draw a club.

If we roll a 6 we'll draw a spade.

A. What are the marginal probabilities?

P(diamonds) = 2/6 = 0.33

P(club) = 1/6 = 0.167

P(spade)= 1/6 = 0.167

B. What is the conditional probability P(face | hearts)?

P (face|heart) = P(A and B) JOINT / P(B) MARGINAL

3/56 | 14/56

P(face|hearts) = 0.21

C. What is the conditional probability P(face | spades)? Why is this equal to P(face | hearts)?

Same thing, as the amount of faces versus no face cards is the same for each colour. Therefore, the results are the same.

i.e Number of faces in hearts = 3, no faces = 11

Number of faces in spades = 3, no faces = 11

4. Bayes' Rule

A. Write Bayes' Rule, using X to represent the data and Z to represent the parameters of our model.

P(Z | X) = P(X | Z) P(Z)/P(X)

B. Give mathematical expressions for the following terms, using the formula and notation above:
   a. Prior over parameters *P (Z)*
   b. Posterior over parameters *P(Z|X)*
   c. Likelihood of data *P(X|Z)*
   d. Marginal likelihood of data *P(X)*

C. Describe the Bayesian view of probability:

It is a way to quantify assumptions that we may have based on an observation for which we cannot have an absolute certain cause, but rather we assume or have a belief that Cause 1 or Cause 2 may have an effect on the observation. Thus, we measure the degree of

belief that the cause may be affecting the observation. Bayesian view of probability necessarily requires a prior and can have variable interpretations depending on who is observing.

D. Describe the Frequentist view of probability:

Frequentists base themselves on the likelihood of something happening, on the frequency of an event and do not require a prior. The result is therefore based on the proportions of a random process occurring in an infinite number of times.

E. Would a Frequentist estimation of the probability of an event change as we get more data?

Yes, as it is based on the law of large numbers, which states that the more observations are collected the more the particular outcome converges with the probability of the outcome. The example given in our notes works well: is it more surprising to flip 3 x the head if we flip the coin 10 x, 100x or 1000x… obviously 3x in the 1000 flips would be quite surprising.

F. When we use Bayes' rule for inference, with the accumulation of each data point the posterior distribution becomes the *prior* distribution for the next step

5. Random Variables: Expectation & Variance

A. What is a random variable?

A random variable's value depends on a random process.

B. Is every random variable characterised by a probability distribution?

Yes

C. For a random variable X, if X is discrete, what is its distribution called?

Probability Mass Function

D. For a random variable X, if X is continuous, what is its distribution called?

Probability Density Function

E. Can a continuous probability distribution integrate to a value >1 over some interval?

No

F. Can discrete probabilities in a distribution sum to more than 1?

No

G. Write the formula for the expectation of a discrete random variable:

$\mathbb{E}_X = \sum_{i} x_i f_X(x_i)$

H. Describe in words what an expectation computes:

An expectation is the average value of a random variable.

I. X is a discrete random variable such that it takes values from the finite set S = {3,4,5,6,7}. If X were *equally likely* to take any of the 5 values, and you were to sample X 10 times, what would be the average (mean) of these values of X? Include R code and final answer.

```
x <- c(3,4,5,6,7)

px <- rep(0.2, 5)


X <- tibble(x, px) %>% print


samples <- sample(x, prob=px,
         10,
         replace = T) %>% print %>%  mean %>% print
```

Mean = 4.7

J. As the number of samples increases to 10000, the average of these 10000 realizations of X gets closer and closer to the mean. If X were biased (more likely) towards taking the value 4 (80% probability, with 5% for the other values) and if now you were to sample 10000 realizations of X_biased, what would your average value be close to? Include R code and final answer.

```
x <- c(3,4,5,6,7)

px_biased <- c(0.05,0.05,0.05,0.05,0.8)


X_biased <- tibble(x, px_biased) %>% print

samples <- sample(x, prob=px_biased,
         10000,
         replace = T) %>%  mean %>% print
```

Mean = 6.48

K. What is variance (in statistics)? Describe in words:

The variance in random variables is the measure we use to calculate the dispersion od the data points in a data set.

L. What is the variance of the unbiased X from (I)?

Variance = 1.97

M. What is the variance of X_biased?

Variance = 1.29

6. Cumulative distributions

A. What is a Cumulative distribution? Describe in words:

"It is a function that gives the probability that a random variable is less than or equal to the independent variable of the function"
https://www.merriam-webster.com/dictionary/cumulative%20distribution%20function

B. Plot the cumulative distribution for the sum of numbers on rolls of three dice (include labelled plot and R code):

```
diceroll_sum %>%
ggplot(mapping = aes(x = sum, y = cumsum(probability)))+
  geom_col(fill="#0072B2") +
  scale_x_continuous(name = "sum of the three dice' outcomes", breaks =
c(3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18))+
  ylab("probability")+
  ggtitle("Cumulative Distribution of the Sum of Numbers on Roll of Three Dice ")
```