DANA 4800 HW2 Answer Keys

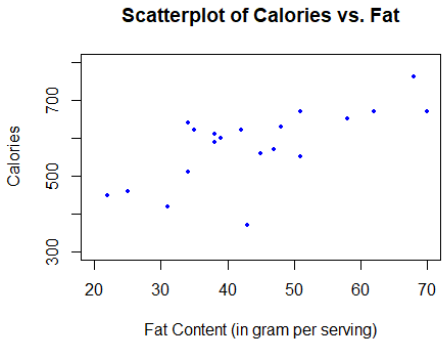
Note: The first 3 questions are from Q10-12 in HW1.

1. A dietician wanted to find out how the total fat content (**Fat**; measured in grams per serving) is dependent on the amount of calories (**Calories**; measured in calories) among chicken burgers made from different fast food chains in Canada. A random sample of 20 chicken burgers was collected from different fast food chains (one burger per fast food chain) and the information was recorded. The data set is from "DANA4800_HW2_Q1_Data.xlsx" on BrightSpace.

   a) Provide a description of the subjects of interest. **[1 mark]**
   Subjects: chicken burgers

   b) Identify the role (or use) of the two variables. **[1 mark]**
   Dependent variable: Fat Content or simply Fat
   Independent variable: Calories

   c) Use the *plot()* function to produce a Scatterplot, based on the roles you defined in the above part. **[2 marks]**
   Note: Make sure you label your graph and axes appropriately.



[2]

   d) Provide a description of the above scatterplot. **[2 marks]**
   Note: Please make sure the title and axes are properly labeled.
   Direction: the relation between calories and fat content has a positive relation.
   Strength: Hard to tell from the plot, but it seems to show a strong relation.
   Outliers: There is an outlier on the top right - 760 calories with 68 grams of fat. (Other answers are acceptable too.)
   Form: It seems the relation has a linear form.

   e) Use the *cov()* function to find the Variance-Covariance matrix. Please keep one decimal place only and identify which number is the covariance and which numbers are the variances of what. **[1+1 marks]**

	Calories	Fat
Calories	9451.6	872.6
Fat	872.6	173.3

Var(Calories) = 9451.6
Var(Fat) = 173.3
Covariance between Calories and Fat = 872.6

   f) Use the cor() function to find the Correlation Coefficient. Please keep four decimal places and identify the value of the correlation coefficient. **[1 mark]**

	Calories	Fat
Calories	1.0000	0.6818
Fat	0.6818	1.0000

Correlation coefficient = 0.6818

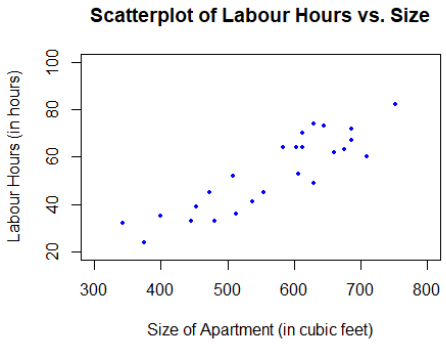
g) Provide the description the above correlation coefficient (or most appropriate statistic in this situation). **[2 marks]**
The correlation coefficient between Calories and Fat among 20 randomly selected chicken burgers. [2]

2. Trying to accurately allocate labour hours in a moving job, the manager of a moving company would like to develop a method of predicting the labour hours (**Labour**; measured in hours) based on the size of the high-rise apartment (**Size**; measured in cubic feet). A random sample of 25 high-rise apartment moves was randomly selected in downtown Vancouver in the previous calendar year. The data set is in "DANA4800_HW2_Q2_Data.xlsx" on BrightSpace.

a) Provide a description of the subjects of interest. **[1 mark]**
Subjects: high-rise apartment moves in downtown Vancouver
Note: the keyword is "moves", not just the "apartment"

b) Identify the role (or use) of the two variables. **[1 mark]**
Dependent variable: Labour Hours or simply Labour
Independent variable: Apartment Size or simply Size

c) Use the *plot()* function to produce a Scatterplot, based on the roles you defined in the above part. **[2 marks]**
Note: Make sure you label your graph and axes appropriately.



[2]

d) Provide a description of the above scatter diagram. **[2 marks]**
Direction: the relation between Labour Hour and Apartment Size has a positive relation.
Strength: the scatterplot shows a strong relation between Labour Hour and Apartment Size.
Outliers: There is potentially one outlier on the upper right corner - 750 cubic feet and takes about 80 hours to complete. Otherwise, it is not too bad.
Form: The relation has a linear form.

e) Use the cor() function to find the Correlation Coefficient. Please keep four decimal places and identify the value of the correlation coefficient. **[1 mark]**

`        Size Labour`
`Size   1.0000 0.8857`
`Labour 0.8857 1.0000`
	Correlation Coefficient = 0.8857

f) Upon seeing the above correlation coefficient, an assistant reported it to the manager and said the following. Identify two major flaws of the statements. Briefly justify your answers. **[2+2 marks]**

"Because the correlation coefficient 0.8857 cubic feet per hour is close to one, the reason of working long work hour is because of the high-rise apartment size only. "

1) Correlation coefficient does not have any units. It should be reported as "0.8857" only.
2) Association does not imply causation. In this case, the labour hours would also include how high the apartment is (i.e. the higher it is, the longer the elevator wait and hence longer labour hour).

g) Provide the description the most appropriate statistic used in this situation (about apartment moving). **[2 marks]**

The correlation coefficient between Labour and Size among 25 randomly selected high-rise apartment moves in downtown Vancouver. [2]

3. The PopularKids data set was about opinions of a group of primary school students, who were stratified by their origin (rural, suburban and urban). More information about the data set can be in the following link: https://www.openml.org/search?type=data&sort=runs&id=1100&status=active
*Note: In this question, let us only use **Gender** (boy and girl) as the row variable and **Goal** (Grades, Popular, and Sports) as the column variable. Every subsequent mentioning of "row" and "column" refer to this definition.*
The data set is in "DANA4800_HW2_Q3_Data.xlsx" on BrightSpace.

a) Use the *table()* function to produce a Two-Way Table (or Contingency Table) with frequency. **[1 mark]**

```
      Goal
Gender Grades Popular Sports
  boy     117      50     60
  girl    130      91     30
```

b) Use the frequency table from part (a), calculate and enter expected frequencies in the following table. Please keep one decimal place in all entries. **[2 marks]**

		Goals			
		Grade	**Popular**	**Sports**	
Gender	**Boy**	**117 (117.3)**	**50 (67.0)**	**60 (42.7)**	**227**
	Girl	**130 (129.7)**	**91 (74.0)**	**30 (47.3)**	**251**
		247	**141**	**90**	**478**

c) Manually calculate the χ^2-statistic. **[3 marks]**
Calculation of χ^2 is not shown here, but its value is 21.46.

Please use the above two-way table with frequency to answer the following 3 questions.
Hint: You are expected to do this manually. But you could also use the margin.table() function to find the marginal totals first. There is an argument called MARGIN with three options. Please look up the R documentation for details.

d) Find the percentage of students who are boys and their main goal is being popular. **[1 mark]**

Ans = 50/478 or 0.1046 or 10.46%

e) Find the percentage of boys whose main goal is being popular. **[1 mark]**
Ans = 50/227 or 0.2203 or 22.03%

f) Among the students whose main goal is being popular, find the percentage of them who are boys. **[1 mark]**
Ans = 50/141 or 0.3546 or 34.46%

Note: There are the same 3 MARGIN options in the proportion() function. Please look up the R documentation for details.

g) Use the *proportions()* function to produce a two-way table with Table Percentages. Please keep only two decimal places. **[1 mark]**

```
       Goal
Gender Grades Popular Sports
  boy    0.24    0.10   0.13
  girl   0.27    0.19   0.06
```

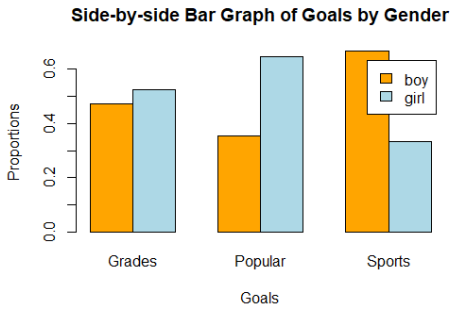
h) Use the *proportions()* function to produce a two-way table with Row Percentages. Please keep only two decimal places. **[1 mark]**

```
       Goal
Gender Grades Popular Sports
  boy    0.52    0.22   0.26
  girl   0.52    0.36   0.12
```

i) Use the *proportions()* function to produce a two-way table with Column Percentages. Please keep only two decimal places. **[1 mark]**

```
       Goal
Gender Grades Popular Sports
  boy    0.47    0.35   0.67
  girl   0.53    0.65   0.33
```

j) Use the *barplot()* function to produce a Side-by-Side Bar Graph, with the variable Goals on the x-axis, column percentages on the y-axis, and including a legend. The title of the graph should say "Side-by-side Bar Graph of Goals by Gender". **[2 marks]**



[2]

k) Provide a description of the above graph. **[1 mark]**
The three groups of bars do not look alike. (We could make a conclude that the two categorical variables are not independent to each other.)

4. In a local high school, 25% of all Grade 11 students play basketball and 20% of all Grade 11 students play volleyball. It is also estimated that 5% of all Grade 11 students play both sports. Use a two-way table, or otherwise any other method, to find the following probabilities.
   Subjects: Grade 11 students
   # of variables: 2
   Variable 1 desc. & type: whether a Grade 11 student plays basketball (class: Yes, No); categorical variable
   Variable 2 desc. & type: whether a Grade 11 student plays volleyball (class: Yes, No); categorical variable
   a) When a Grade 11 student is randomly selected, what is the probability that one is playing either basketball or volleyball? Provide an interpretation of the answer. **[2+2 marks]**
   Define event A as Grade 11 students playing basketball and event B as Grade 11 students playing volleyball. Given: P(A) = 0.25, P(B) = 0.20 and P(A and B) = 0.05.

		Play Basketball		
		Yes (A)	***No (A^c)***	
Play Volleyball	***Yes (B)***	***5***	***15***	***20***
	No (B^c)	***20***	***60***	***80***
		25	***75***	***100***

   Both the two-way table method and formula method are presented in this answer key.
   Note: Work of how to get this two-way table is not shown here. Please ask if you are not sure how to do it.

   Two-way table method: (5+15+20)/100 = 40/100
   Formula method: P(A or B) = P(A) + P(B) – P(A and B) = 0.25 + 0.20 – 0.05 = 0.40 [2]
   Interpretation: When the experiment of "selecting a random Grade 11 student" is repeated infinite number of times, 45% of them play either basketball or volleyball. [2]

   b) When a Grade 11 student is randomly selected, what is the probability that one is playing neither basketball nor volleyball? **[1 mark]**
   Two-way table method: 60/100
   Formula method: P(neither A nor B) = 1 – P(A or B) = 1 – 0.40 = 0.60 [1]

   c) When a Grade 11 student who is playing volleyball is randomly selected, what is the probability that one is also playing basketball? **[2 marks]**
   Two-way table method: 5/20 = 0.25
   Formula method: $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{0.05}{0.20} = 0.25$ **[2]**

   d) When a Grade 11 student who is playing basketball is randomly selected, what is the probability that one is also playing volleyball? **[2 marks]**
   Two-way table method: 5/25 = 0.20
   Formula method: $P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{0.05}{0.25} = 0.20$ **[2]**

e) Define event A as Grade 11 students playing basketball and event B as Grade 11 students playing volleyball. Are the two events A and B mutually exclusive? Briefly justify your answer using some probability calculations. **[0+2 marks]**
1) If the two events are mutually exclusive, then P(A and B) = 0
2) P(A and B) = 0.05 (from the two-way table)
3) Since P(A and B) is not zero, the two events A and B are not mutually exclusive. Heuristically, you can see that there are 5% of students playing both sports.

f) Define event A as Grade 11 students playing basketball and event B as Grade 11 students playing volleyball. Are the two events A and B independent? Briefly justify your answer using some probability calculations. **[0+2 marks]**
1) If the two events are independent, then P(A and B) = P(A)*P(B).
2) LHS = P(A and B) = 0.05
RHS = P(A)*P(B) = 0.25*0.20 = 0.05
3) Since LHS is the same as RHS, the two events A and B are independent to each other.

5. Suppose that 85% of shoppers buy coffee at this local café, 65% of shoppers buy baked goods (like muffins and cookies etc.) at this local café, and 90% of shoppers buy either coffee or baked goods.
a) Construct a two-way table of the situation above. **[2 marks]**

		Buy Coffee		
		Yes (A)	**No (A^c)**	
Buy Baked Goods	**Yes (B)**	**60**	**5**	**65**
	No (B^c)	**25**	**10**	**35**
		85	**15**	**100**

Please try to use the two-way table method to do the following 3 parts.

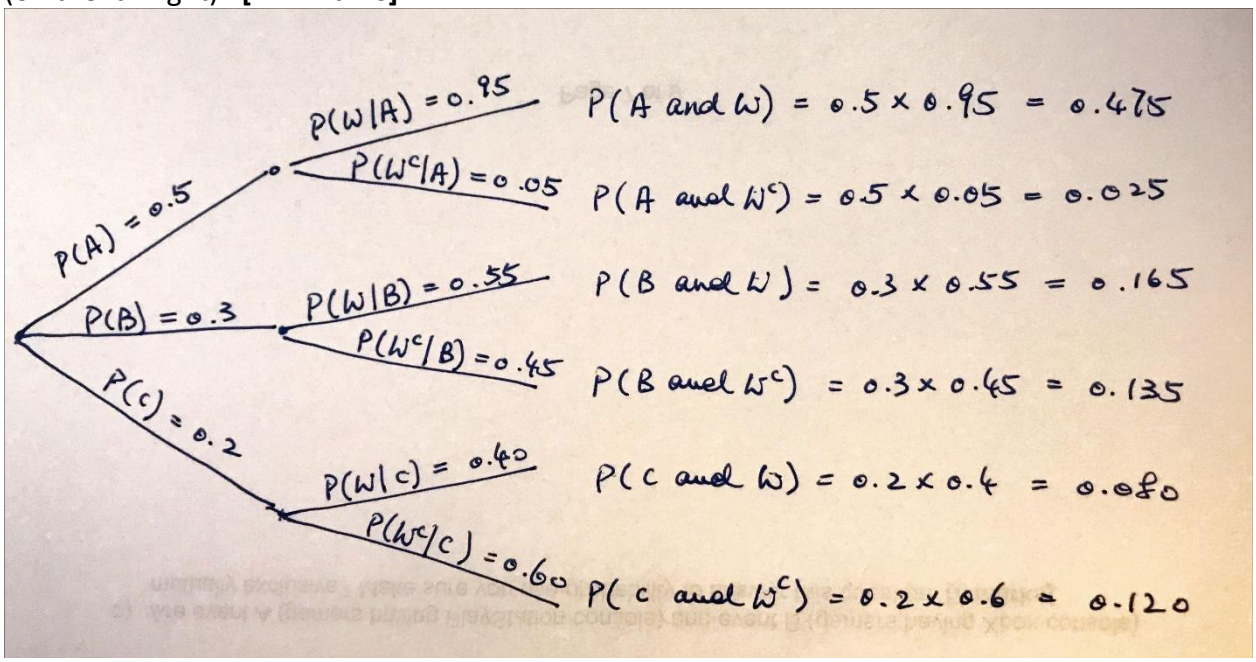
b) When a shopper visiting this local café is randomly selected, what is the probability that one is buying coffee and baked goods? **[1 mark]**
Answer = 60/100 [1]

c) Given a shopper visiting this local café is buying baked goods, what is the probability that one is also buying coffee? **[1 mark]**
Answer = 60/65 = 0.923 [1]

d) Given a shopper visiting this local café is buying coffee, what is the probability that one is also buying baked goods? **[1 mark]**
Answer = 60/85 = 0.706 [1]

6. Suppose that a city has three airports (Airport A, Airport B, and Airport C). Suppose Airport A handles 50% of the air airline traffic, Airport B and Airport C handle 30% and 20% of the air traffic respectively. Suppose the detection rates of weapon (W = weapon) at the three airports are 0.95, 0.55 and 0.40 respectively.
   a) Define all necessary events, assign a symbol (or letter) to each one and identify all six probabilities. **[3 marks]**
   P(A) = 0.5 = probability of all air traffic handled by Airport A,
   P(B) = 0.3 = probability of all air traffic handled by Airport B,
   P(C) = 0.2 = probability of all air traffic handled by Airport C,
   P(W|A) = 0.95 = probability of detecting weapon given that it happens in Airport A,
   P(W|B) = 0.55 = probability of detecting weapon given that it happens in Airport B,
   P(W|C) = 0.40 = probability of detecting weapon given that it happens in Airport C.

   b) Draw a tree diagram showing the above situation. Please also compute all 6 joint probabilities (on the far right). **[2+2 marks]**



   c) If a passenger is randomly selected at one of the three airports and is found to be carrying a weapon through the boarding gate, what is the probability that the passenger is using Airport A? **[2 marks]** Note: Please try to use the Bayes' Theorem to do this.
   Using Bayes' Theorem:

$$P(A \mid W) = \frac{P(W \mid A) \times P(A)}{P(W \mid A) \times P(A) + P(W \mid B) \times P(B) + P(W \mid C) \times P(C)}$$

$$\frac{0.95 \times 0.5}{0.95 \times 0.5 + 0.55 * 0.3 + 0.40 * 0.2} = \frac{0.475}{0.475 + 0.165 + 0.080} = \frac{0.475}{0.720} = \mathbf{0.6597}$$

d) Repeat part (c) with Airport C? **[2 marks]** Note: Please try to use two-way table to do this.
Using the two-way table method:

		Airport			
		A	**B**	**C**	
Weapon Detection	**Yes (W)**	**475**	**165**	**80**	**720**
	No (W^c)	**25**	**135**	**120**	**280**
		500	**300**	**200**	**1000**

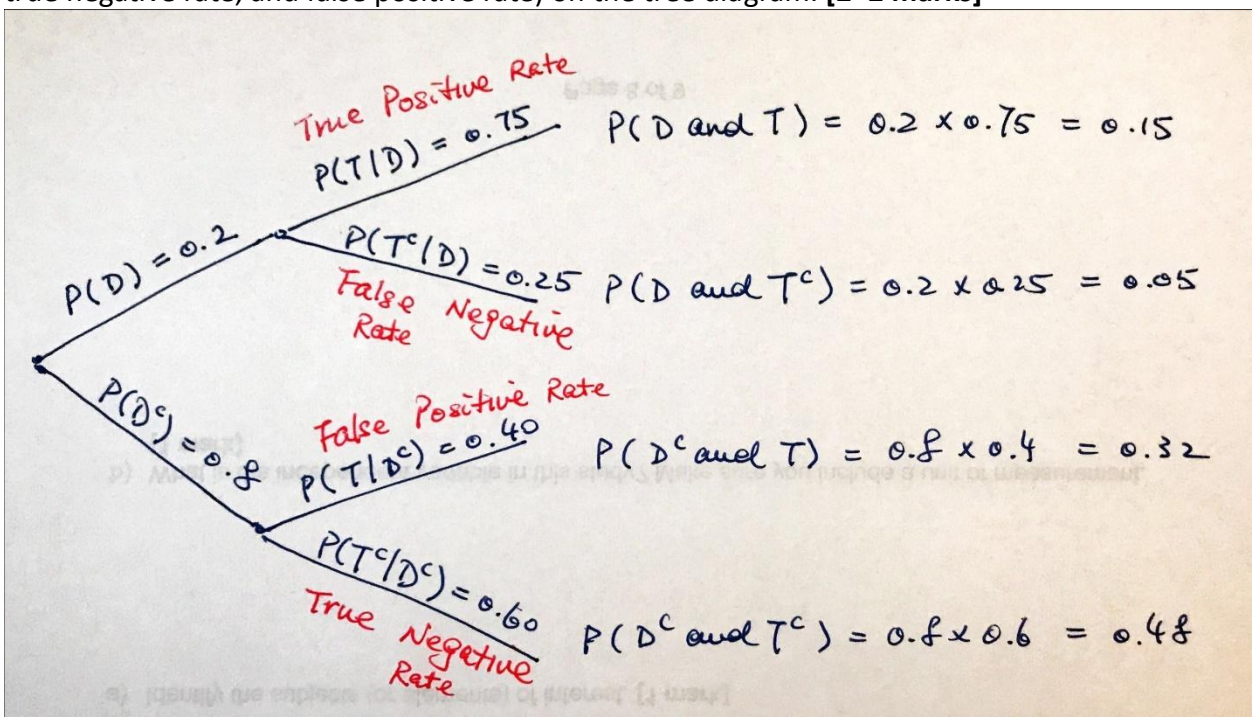
P(C|W) = 80/720 = 0.1111

e) Note that P(A) < P(A|W). (Note also that P(C) > P(C|W).) What implication does it have? Try to explain this to the Director of Aviation at Transport Canada, who has no clue what Bayes' Theorem is about. **[2+2 marks]**
P(A) = 0.50 vs. P(A|W) = 0.66
Without considering the weapon detection, i.e. P(A), it only considers the volume of traffic Airport A handles.
Given the high usage of Airport A in P(A) = 0.50 and high weapon detection rate in Airport A in P(A|W) = 0.95, it gives us a better understanding (or better knowledge) about Airport A in P(A|W).
Same (but rather opposite) argument for Airport C.

7. Suppose that a certain disease is present in 20% of the population, and that there is a screening test designed to detect if this disease is present. If the screening test is applied to those who have this disease, 75% of the time it will give a positive result. Another way of saying it is that the "**true positive rate**" is 75%. Also, if the screening test is applied to those who do not have this disease, 60% of the time it will give a negative result. It is also called "**true negative rate**".
   a) Define all necessary events, assign a symbol (or letter) to each one and identify all three probabilities. **[3 marks]**
   P(D) = 0.20 = probability of population who have this disease
   P(T|D) = 0.75 = probability of getting a positive result given someone has the disease = true positive
   $P(T^c|D^c)$ = 0.60 = probability of getting a negative result given someone does not have the disease = true negative

   b) Google the terms "**false negative rate**" and "**false positive rate**". (1) Provide a definition using the context, (2) identify them using the event and symbol defined above, and (3) find their values in this question. **[2+2 marks]**
   Note: There might be a few conflicting courses. So, please try not to rely on one webpage only.
   False negative rate = probability of getting a negative result given that the result is wrong (or someone indeed has the disease)
   -> False negative rate = $P(T^c|D)$ = 1 – P(T|D) = 1 – 0.75
   False positive rate = probability of getting a positive result given that the result is wrong (or someone does not have the disease)
   -> False positive rate = $P(T|D^c)$ = 1 – $P(T^c|D^c)$ = 1 – 0.60 = 0.40

c) Draw a tree diagram showing the above situation. Please also compute all 4 joint probabilities (on the far right) and identify clearly the four new terms (true positive rate, false negative rate, true negative rate, and false positive rate) on the tree diagram. **[2+2 marks]**



d) When a person is tested positive, what is the probability that they have the disease? **[2 marks]**
Again, I will use the Bayes' Theorem formula here and two-way table method in the next part.

$$\boldsymbol{P(D \mid T) = \frac{P(T \mid D) \times P(D)}{P(T \mid D) \times P(D) + P(T \mid D^c) \times P(D^c)} = \frac{0.75 \times 0.2}{0.75 \times 0.2 + 0.4 \times 0.8} = 0.3191}$$

e) When a person is tested negative, what is the probability that they do not have the disease? **[2 marks]**
Using the two-way table method:

		Test Result		
		Positive (T)	**Negative (T^c)**	
Has Disease?	**Yes (D)**	**15**	**5**	**20**
	No (D^c)	**32**	**48**	**80**
		47	**53**	**100**

$$\boldsymbol{P(D^c \mid T^c) = \frac{48}{53} = 0.9057}$$

f) In a 2x2 two-way table (a two-way table with two rows and two columns that contain frequencies or counts only), the "**odds ratio**" is defined as "the product of the number of true positives and the number of true negatives divided by the product of the number of false negatives and the number of false positives". Use this definition to find the odds ratio in this question. **[2 marks]**
Note: "True positive rate" is a probability, but "number of true positives" is a count, which you can find from the two-way table.

Using the two-way table:

		Test Result		
		Positive (T)	**Negative (T^c)**	
Has Disease?	**Yes (D)**	**15 (true positive)**	**5 (false negative)**	**20**
	No (D^c)	**32 (false positive)**	**48 (true negative)**	**80**
		47	**53**	**100**

-> The odds ratio is $\frac{15\times48}{32\times5} = 4.5$.

g) Note that the above odds ratio is greater than 1. Please look up the internet to find the meaning when the odds ratio bigger than 1 (and when the odds ratio is less than 1). What is the implication of "odds ratio greater than 1" here, in terms of the "power" or "ability" of the screening test to identify the disease. **[2 marks]**
General definition says that when the odds ratio is greater than 1, the "exposure" is a risk factor. Let's apply in this context here. It basically means that the screening test is about 4.5 times better to get to the right decision than getting to the wrong decision. In other words, this screening test is pretty good.

h) What does it mean when the odds ratio is equal to one? Come up with a "hypothesis" or "postulation" about the relationship between "odds ratio being one" and "the independence relation between two categorical variables". Then, create a two-way table that has an odds ratio of 1, and shows that the two binary categorical variables are independent to each other. **[2+2 marks]**
Note: I am sure you can find thousands of websites about this, but I would challenge you to figure this part yourself without the help of internet. Believe in your own ability. :)
Using a similar example but with different numbers. To create a two-way table with an odds ratio of 1, you just need to make sure A/B is the same as C/D in the below table.

		Test Result		
		Positive (T)	**Negative (T^c)**	
Has Disease?	**Yes (D)**	**A**	**B**	
	No (D^c)	**C**	**D**	

As an illustration, I make the ratio 3:7.

		Test Result		
		Positive (T)	**Negative (T^c)**	
Has Disease?	**Yes (D)**	**6**	**14**	**20**
	No (D^c)	**24**	**56**	**80**
		30	**70**	**100**

Odds ratio = (6*56)/(14*24) = 1.
To show independence, we just need to show one identity $P(T) = P(T|D) = P(T|D^c) = 0.30$