MET CS 555 - Data Analysis and Visualization Quiz - 6

1. A study of stroke patient who survived 6 months after the stroke found that 6/45 men and 22/63 women lived in an institution (e.g., nursing home or assisted living facility). Is there evidence of a difference in risk of living in an institution after stroke for men versus women? What is the risk difference (using men as the reference group)?

   A. -26%
   B. -22%
   C. -16%
   D. 16%
   E. 22%
   F. 26%

   Answer (E)

   Description.

   The risk difference is just subtracting these two risks from each other. We want to use men as the reference group so that we do $22/63 - 6/45 = 0.215873 = 22\%$

   In R code you can write.

```
> p1 <- 22/63
> p2 <- 6/45
> p1-p2
[1] 0.215873
```

---

2. A study of stroke patient who survived 6 months after the stroke found that 6/45 men and 22/63 women lived in an institution (e.g., nursing home or assisted living facility). Is there evidence of a difference in risk of living in an institution after stroke for men versus women? What is the odds ratio (using men as the reference group)?

   A. 0.15
   B. 0.29
   C. 0.54
   D. 1.86
   E. 2.61
   F. 3.49
   G. 6.50

   Answer (F)

   Description. It is about the odds ratio calculation.

$$\frac{\frac{p1}{(1-p1)}}{\frac{p2}{(1-p2)}}$$

   In R code you can write.

```
> p1 <- 22/63
> p2 <- 6/45
> (p1/(1-p1)) /(p2/(1-p2))
[1] 3.487805
```

3. A study of stroke patient who survived 6 months after the stroke found that 6/45 men and 22/63 women lived in an institution (e.g., nursing home or assisted living facility). Is there evidence of a difference in risk of living in an institution after stroke for men versus women? What is the risk ratio (using men as the reference group)?

    A. 0.22

    B. 0.29

    C. 0.38

    D. 2.61

    E. 3.49

Answer (D)

Description. The risk ratio is just the ratio of two risks.

In R code you can write.

```
> p1 <- 22/63
> p2 <- 6/45
> p1 /p2
[1] 2.619048
```

---

4. A study of stroke patient who survived 6 months after the stroke found that 6/45 men and 22/63 women lived in an institution (e.g., nursing home or assisted living facility).

    What is the z-statistic for testing the null hypothesis of $H_0 : p_1 = p_2$?

    A. 0.22

    B. 0.49

    C. 1.12

    D. 2.52

    E. 2.74

Answer (D)

Description. It is asking about the two-sample z-statistic

$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

We need to calculate our pooled proportion $\hat{p}$

And in R we can calculate.

```
> p1 <- 22/63
> p2 <- 6/45
> phat <- (22+6)/(45+63)
> (p1-p2)/sqrt(phat * (1-phat)* (1/63 + 1/45))
[1] 2.523845
```

---

5. A study of stroke patient who survived 6 months after the stroke found that 6/45 men and 22/63 women lived in an institution (e.g., nursing home or assisted living facility).

    What is the 95% confidence interval for the risk difference (with men as the reference group)?

    A. -37.0 to -6.2

B. -34.9 to -13.3

C. -23.0 to -20.2

D. 6.2 to 37.0

E. 13.3 to 34.9

F. 20.2 to 23.0

Answer (D)

Description.

We want to calculate this formula

$(\hat{p}_1 - \hat{p}_2) \pm z \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

And in R we can calculate.

```
> (p1 - p2) - 1.96 * sqrt (  (p1 * (1-p1)/63) + (p2 * (1-p2)/45) )
[1] 0.06185114
> (p1 - p2) + 1.96 * sqrt (  (p1 * (1-p1)/63) + (p2 * (1-p2)/45) )
[1] 0.3698949
```

---

6. A study of stroke patient who survived 6 months after the stroke found that 6/45 men and 22/63 women lived in an institution (e.g., nursing home or assisted living facility). If the point estimates remained the same, what effect would increasing the sample size have on the confidence interval?

   A. It would get smaller.

   B. It would get larger.

   C. Since the point estimates are remaining the same, it would have no effect on the confidence intervals.

   Answer (A)

   Description.

   We know that as the sample size increases the variability decreases.

   If you look at this formula $(\hat{p}_1 - \hat{p}_2) \pm z \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

   You can see if the sample size increases only the two parameter $n1$ and $n2$ increase, so that the make the values $\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}$ and $\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$ smaller. As a result the confidence interval range would be smaller.

---

7. We have 30-day follow-up data on 350 stroke patients and want to investigate whether the risk of recurrent stroke and/or death depends on the type of stroke (cerebral embolism or not). The results of the simple logistic regression of the dummy variable for cerebral embolism (1 = yes, 0 = no) are shown below. Use the output to calculate the odds ratio for recurrent stroke and/or death for those who had a cerebral embolism versus those who did not?

Parameter	Estimate	Standard Error	p-value
β_0	-2.80	0.51	<0.001
β_1	1.87	0.65	0.0040

   A. 0.06081

   B. 0.154124

   C. 6.488296

   D. 16.44465

Answer (C)

Description. To get the odds ratio here we just need to calculate e to the power of $beta_1$ estimate which is $e^{\hat{\beta}_1}$

In R we can do

```
> exp(1.87)
[1] 6.488296
```

---

8. We have 30-day follow-up data on 350 stroke patients and want to investigate whether the risk of recurrent stroke and/or death depends on the type of stroke (cerebral embolism or not) and age. The results of the multiple logistic regression of the dummy variable for cerebral embolism (1 = yes, 0 = no) and age are shown below. What is the risk of recurrent stroke/death for a patient without a cerebral embolism who is 60 years of age?

Parameter	Estimate	Standard Error	p-value
β_0	-15.32	9.50	0.1069
$\beta_{\text{cerebral embolism}}$	2.07	0.68	0.0024
β_{age}	0.18	0.13	0.1840

A. 0.0108

B. 0.0109

C. 0.086294

D. 0.079439

Answer (A)

Description. To calculate the risk in multiple logistic regression we used the following formula

$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\epsilon}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\epsilon}} = \frac{e^L}{1+e^L}$

We just need to plugin our values.

$\beta_0 = -15.32$

$\beta_{cerebral\ embolism} = 2.02$

$\beta_{age} = 0.18$ and $x_age = 60$

In R we can do

```
> exp(-15.32+2.02*0+0.18*60) / (1 + exp(-15.32+2.02*0+0.18*60))
[1] 0.01077173
```

---

9. We have 30-day follow-up data on 350 stroke patients and want to investigate whether the risk of recurrent stroke and/or death depends on the type of stroke (cerebral embolism or not) and age. The results of the multiple logistic regression of the dummy variable for cerebral embolism (1 = yes, 0 = no) and age are shown below. Calculate the odds ratio comparing the odds of recurrent stroke/death for a patient who is 65 versus 64.

Parameter	Estimate	Standard Error	p-value
β_0	-2.80	0.51	<0.001
β_1	1.87	0.65	0.0040

A. 0.180

B. 0.835

C. 1.197

D. 6.050

Answer (C)

Description.

We need to calculate the odds ratio for age which is the exponentiation of β_{age} (0.18) .

In R we can do

```
> exp(0.18)
[1] 1.197217
```

---

10. We have 30-day follow-up data on 350 stroke patients and want to investigate whether the risk of recurrent stroke and/or death depends on the type of stroke (cerebral embolism or not) and age. The results of the multiple logistic regression of the dummy variable for cerebral embolism (1 = yes, 0 = no) and age are shown below. Calculate the odds ratio comparing the odds of recurrent stroke/death for a patient who is 65 versus 55.

Parameter	Estimate	Standard Error	p-value
β_0	-2.80	0.51	<0.001
β_1	1.87	0.65	0.0040

A. 0.004

B. 0.022

C. 0.026

D. 0.180

E. 1.197

F. 6.050

Answer (F)

Description. We need to calculate the odds ratio for increase of 10 units in age variable. exp(0.18) is the odds ratio for one unit increase, we need to multiply the beta estimate for the 10 units and the exponentiate it.

```
> exp(0.18*10)
[1] 6.049647
```

---

11. We have 30-day follow-up data on 350 stroke patients and want to investigate whether the risk of recurrent stroke and/or death depends on the type of stroke (cerebral embolism or not) and age. The results of the multiple logistic regression of the dummy variable for cerebral embolism (1 = yes, 0 = no) and age are shown below. Calculate the 95% confidence interval for the odds ratio comparing the odds of recurrent stroke/death for patients with a cerebral embolism versus those without.

Parameter	Estimate	Standard Error	p-value
β_0	-15.32	9.50	0.1069
$\beta_{\text{cerebral embolism}}$	2.07	0.68	0.0024
β_{age}	0.18	0.13	0.1840

A. 0.32 to 4.54

B. 0.93 to 1.54

C. 1.37 to 45.68

D. 2.09 to 30.05

E. 2.59 to 24.25

Answer (D)

Description.

$$e^{\left(\hat{\beta}_1 \pm z_{\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_1}\right)} = e^{(2.07 \pm 1.960 \cdot 0.68} \tag{1}$$

$$= (2.09, 30.05) \tag{2}$$

In R we can do

```
> exp(2.07 - 1.96*0.68)
> exp(2.07 - 1.96*0.68)
[1] 2.090075
> exp(2.07 + 1.96*0.68)
[1] 30.04812
```

---

12. A multiple logistic regression was run predicting risk of lung cancer from age, gender, and smoking history. The odds ratio for males versus females was 2.05.

    What is the correct interpretation of this result?

    A. The odds of lung cancer are approximately 2 times higher for females versus males.

    B. The odds of lung cancer are approximately 2 times higher for females versus males, after adjusting for age and smoking history.

    C. The risk of lung cancer are approximately 2 times higher for females versus males.

    D. The risk of lung cancer are approximately 2 times higher for females versus males, after adjusting for age and smoking history.

    E. The odds of lung cancer are approximately 2 times higher for males versus females.

    F. The odds of lung cancer are approximately 2 times higher for males versus females, after adjusting for age and smoking history.

    G. The risk of lung cancer are approximately 2 times higher for males versus females.

    H. The risk of lung cancer are approximately 2 times higher for males versus females, after adjusting for age and smoking history.

    Answer (F)

    Description. Here we select statements that are about odds ratio and including the statement of *"after adjusting for age and smoking history"*. Because the odds ratio is for males versus females was 2.05, we can say that the *"The odds of lung cancer are approximately 2 times higher for males versus females"*

---

13. A multiple logistic regression was run predicting risk of lung cancer from age, gender, and smoking history. The beta estimate for age was significant at the 0.05 level.

    Which of the following are possible for the 95% confidence interval for the odds ratio for a 10 unit increase in age? Select all that apply.

A. -1.5 to 5.0

B. 0.5 to 5.0

C. 0.5 to 0.75

D. 0.75 to 5.0

E. 1.5 to 5.0

F. 5.0 to 7.5

Answer (C), (E) and (F)

Description.

This question is a bit tricky.

Since β_age is significant, which means that we are 95% confident that β_age does not equal to 0.

It in turn means that odds ratio, e^{β_age}, does not equal to 1.

So the 95% confidence interval won't cover 1.

So all the intervals that do not cover 1 are possible answers.

---

14. A multiple logistic regression was run predicting risk of lung cancer from age, gender, and smoking history. The regression indicated that the effect for smoking history (yes versus no) was significant and indicated a higher risk of lung cancer for those with smoking history.

    Which of the following are possible values for the odd ratio **comparing smokers to non-smokers**? Select all that apply.

    A. -2

    B. -0.15

    C. 0.15

    D. 0.50

    E. 1.25

    F. 1.5

    G. 2.5

    Answer (E), (F) and (G)

    Description. The risks for the non-smokers are lower. We should look for odds ratios that are higher than 1.0 (**comparing smokers to non-smokers**), and not negative (odds ratio is never negative).

---

15. A multiple logistic regression was run predicting risk of lung cancer from age, gender, and smoking history. The regression indicated that the effect for smoking history (yes versus no) was significant and indicated a higher risk of lung cancer for those with smoking history.

    Which of the following are possible values for the beta estimate comparing smokers to non-smokers? Select all that apply.

    A. -2

    B. -0.15

    C. 0.15

    D. 0.50

    E. 1.25

    F. 1.5

G. 2.5

Answer (C), (D) (E), (F) and (G)

Description. If we have positive beta estimates we will have odds ratios grater than one.

Here we just want to have all of the positive answers.

---

16. A study was conducted to determine key predictors of chromosomal fetal abnormalities. Using the multiple logistic regression model and a cut off selected by the investigator, 78 fetuses were predicted by the model of having an abnormality. However, only 14 of the 78 that were predicted to have the abnormality actually did. Of the 122 fetuses that the model predicted did not have the abnormality, 6 of them actually did. Construct a 2 by 2 table of these results to help you calculate the **sensitivity** of the model using this cutoff.

    A. 5%

    B. 18%

    C. 20%

    D. 30%

    E. 64%

    F. 70%

    G. 82%

    Answer (F)

    Description.

    We can calculate the following table (you can consider this as a table in excel) for the actual and predicted values based on the givens from the question.

		Actual		
		N	Y	total
Predicted	N	116	**6**	122
	Y	64	**14**	78

    Sensitivity is the proportion of true events that were classified correctly. In our table we can see that we have 6 +14 = 20 as the total number, and then ratio of events that we classified correctly is 14/20=0.70 or 70 percent.

---

17. A study was conducted to determine key predictors of chromosomal fetal abnormalities. Using the multiple logistic regression model and a cut off selected by the investigator, 78 fetuses were predicted by the model of having an abnormality. However, only 14 of the 78 that were predicted to have the abnormality actually did. Of the 122 fetuses that the model predicted did not have the abnormality, 6 of them actually did.

    Construct a 2 by 2 table of these results to help you calculate the **specificity** of the model using this cutoff.

    A. 5%

    B. 18%

    C. 20%

    D. 30%

    E. 64%

    F. 70%

G. 82%

Answer (E)

Description.

The **specificity** is the proportion of true non-events that were classified correctly.

Same data as above. We calculate the sum to be 116 + 64 = 180 and the ratio of true non-event is 116/180 = 0.6444

Or 64 percent.
