Print

Quiz, Lesson 5: Categorical Data Analysis

Your Score: 100%

Congratulations! Your score of 100% indicates that you've mastered the topics in this lesson. If you'd like, you can review the feedback for each question.

1. In a PROC FREQ step, which statement or set of statements creates a frequency table for **Country**, a frequency table for **Size**, and a crosstabulation table for **Country** by **Size**?

   a. `tables Country, Size, Country*Size;`
   b. `tables Country*Size;`
   c. `tables Country | Size;`
   d. `tables Country Size Country*Size;`

   Your answer: d
   Correct answer: d

   You use the TABLES statement in PROC FREQ to create frequency and crosstabulation tables. In the TABLES statement, you separate table requests with a space. In a table request for a crosstabulation table, you specify an asterisk between the variable names.

   Review: Crosstabulation Tables

---

2. This table shows frequency statistics for the variables **country** and **size** in a data set that contains data about people and the cars they drive. What evidence in the table indicates a possible association?

Frequency
Percent
Row Pct
Col Pct

Table of country by size

country(country)	size(size) Large	Medium	Small	Total
American	36	53	26	115
	11.88	17.49	8.58	37.95
	31.30	46.09	22.61	
	85.71	42.74	18.98	
European	4	17	19	40
	1.32	5.61	6.27	13.20
	10.00	42.50	47.50	
	9.52	13.71	13.87	
Japanese	2	54	92	148
	0.66	17.82	30.36	48.84
	1.35	36.49	62.16	
	4.76	43.55	67.15	
Total	42	124	137	303
	13.86	40.92	45.21	100.00

   a. The frequency statistics indicate that the values of each variable are equally distributed across levels.
   b. The row percentages indicate that the distribution of **size** changes when the value of **country** changes.
   c. The column percentages indicate that most of the cars of each size are manufactured in Japan.

Your answer: b
Correct answer: b

To see a possible association, you look at the row percentages. A higher percentage of American-made cars are large as opposed to small. The opposite is true for European cars and especially for Japanese cars.

Review: Association between Categorical Variables, Crosstabulation Tables

---

3. Suppose you are testing for an association between student ratings of teachers and student grades. The **Rating** variable has the values 1 (for poor), 2 (for fair), 3 (for good) and 4 (for excellent). The **Grade** variable has the values A, B, C, D, and F. Which of the following TABLES statements in PROC FREQ produces the appropriate chi-square statistics and measure of strength for these variables?

   a. `tables Rating*Grade / chisq measures;`
   b. `tables Rating*Grade / chisq;`
   c. `tables Rating*Grade / mhchisq;`
   d. `tables Rating*Grade / mhchisq clodds=pl;`

Your answer: a
Correct answer: a

Both variables are ordinal and have logically-ordered values, so the Mantel-Haenszel test (for ordinal association) is a stronger test than the Pearson chi-square test (for general association) in this situation. The CHISQ option produces both the Pearson and Mantel-Haenszel statistics. The MEASURES option produces the Spearman correlation statistic, which measures the strength of an ordinal association. MHCHISQ is not a valid option, and the CLODDS= option is not a valid option in PROC FREQ.

Review: The Mantel-Haenszel Chi-Square Test, The Spearman Correlation Statistic, Performing a Mantel-Haenszel Chi-Square Test of Ordinal Association

---

4. Suppose you are analyzing the relationship between hot dog ingredients and taste. Which of the following statistics provides evidence of a relatively strong association between the variables **Type** (which has the values *Beef*, *Meat*, and *Poultry*) and **Taste** (which has the values *Bad* and *Good*)?

   a. A Cramer's V statistic that is close to 1
   b. An odds ratio that is greater than 1
   c. A Spearman correlation statistic that is close to 1

Your answer: a
Correct answer: a

Cramer's V statistic is the only appropriate statistic to use in this example. When Cramer's V is close to 1, there is a relatively strong general association between two categorical variables. You cannot use an odds ratio because the predictor **Type** is not binary. You cannot use the Spearman correlation statistic because the predictor **Type** is not ordinal.

Review: Cramer's V Statistic, Odds Ratios, The Spearman Correlation Statistic

---

5. Which statement about binary logistic regression is false?

   a. Binary logistic regression uses predictor variables to estimate the probability of a specific outcome.
   b. To model the relationship between a predictor variable and the probability of an outcome,

you must use a nonlinear function.

c. The mean of the response in binary logistic regression is a probability, which is between 0 and 1.

d. The response variable can have more than two levels as long as one of the levels is coded as 0.

Your answer: d
Correct answer: d

In binary logistic regression, the response variable can only have two levels.

Review: Modeling a Binary Response

6. Suppose you want to investigate the relationship between the gender of elementary school students and their focus in school. The variable **Gender** indicates the gender of each student as *Boy* or *Girl*. The variable **Focus** identifies each student's main focus in school as *Grades* or *Sports*. Which of the following MODEL statements correctly completes this PROC LOGISTIC step for your analysis?

```
proc logistic data=school.students;
   class Gender;

   _____________________________________
run;
```

a. `model Focus(event='Sports*Grades')=Gender;`
b. `model Focus(event='Sports')=Gender;`
c. `model Focus(ref='Sports')=Gender;`
d. `model Focus*Gender(ref='Sports');`

Your answer: b
Correct answer: b

In the MODEL statement, the response variable name is followed by the EVENT= option in parentheses (which specifies the event category—the level of the response variable that you're interested in), an equal sign, and the predictor variable name.

Review: The LOGISTIC Procedure

7. Which statement about the backward elimination method is **false**?

a. Backward elimination is a method of selecting variables for a logistic regression model.
b. Backward elimination removes effects and interactions one at a time.
c. All main effects and interactions that remain in the final model must be significant.
d. To obtain a more parsimonious model, you specify a smaller significance level.

Your answer: c
Correct answer: c

Backward elimination results in a final model that can contain one or more main effects and (if specified) interactions. Any interactions in the final model must be significant. Main effects that are involved in interactions must appear in the final model, whether or not they are significant.

Review: The Backward Elimination Method of Variable Selection

---

8. Suppose you want to fit a multiple logistic regression model to determine which of two

rehabilitation programs is more effective. The categorical response variable **Relapsed** (*Yes* or *No*) indicates whether study participants stayed clean after one year. The categorical predictor variables are **Program** (*1* or *2*) and **Gender** (*Male* or *Female*). **Age** is a continuous predictor variable.

Assume that you want to use reference cell coding with the default reference levels. Which of the following CLASS statements correctly completes the PROC LOGISTIC step for this analysis?

```
proc logistic data=programs.rehabilitation;
   _____________________________________
   model Relapsed (event='Yes') = Program | Gender | Age @2;
run;
```

a.
```
class Program(param=ref ref='2')
      Gender(param=ref ref='Male');
```

b.
```
class Program(param=ref ref='2')
      Gender (param=ref ref='Male')
      Age (param=ref units=1);
```

c.
```
class Program(param=ref ref='1')
      Gender(param=ref ref='Female');
```

Your answer: a
Correct answer: a

The CLASS statement lists all the categorical predictor variables. For each categorical predictor, you use the PARAM= option to specify reference cell coding (REF or REFERENCE) instead of the default parameterization method, effect coding. The default reference level is the level with the highest ranked value when the levels are sorted in ascending alphanumeric order.

Review: Specifying a Parameterization Method in the CLASS Statement, Reference Cell Coding

---

✔

9. Suppose you want to fit a multiple logistic regression model to determine how the method of administering a drug affects patients' response to the drug. The binary variable **Response** has the values *0* and *1*. There are three predictors: **Amount** identifies the dosage amount in *mg*, **Frequency** has the values *Daily* and *Weekly*, and **Meal** has the values *Yes* and *No*.

   You want to calculate three odds ratios:

   •an odds ratio for **Amount** at 20 mg intervals
   •an odds ratio for **Frequency** against the reference level (*Daily*) as compared to all levels of **Meal**
   •an odds ratio for **Meal** against the reference level (*Yes*) as compared to all levels of **Frequency**

   Which of the following blocks of code below correctly completes the following PROC LOGISTIC program?

```
proc logistic data=newdrug;
   class Frequency (param=ref ref='Daily')
         Meal (param=ref ref='Yes');
   model Response (event='1') = Frequency | Meal | Amount @2;
   ____________________________________________________
run;
```

   a.
```
oddsratio Amount (units=20);
oddsratio Frequency / diff=ref at (Meal=all);
oddsratio Meal / diff=ref at (Frequency=all);
```

   b.
```
units Amount=20;
oddsratio Amount;
oddsratio Frequency / diff=all at (Meal='Yes');
oddsratio Meal / diff=all at (Frequency='Daily');
```

   c.
```
units Amount=20;
```

```
oddsratio Amount;
oddsratio Frequency / diff=ref at (Meal=all);
oddsratio Meal / diff=ref at (Frequency=all);
```

d.
```
oddsratio Amount (units=20);
oddsratio Frequency / diff=all at (Meal='Yes');
oddsratio Meal / diff=all at (Frequency='Daily');
```

Your answer: c
Correct answer: c

You must specify the intervals of **Amount** in the UNITS statement, not in the ODDSRATIO statement. To calculate odds ratios for the two categorical variables as described, each of the two ODDSRATIO statements must set DIFF= to REF against all levels of the interacting variable.

Review: The ODDSRATIO Statement, The UNITS Statement

10. According to the goodness-of-fit statistics shown below, which multiple logistic regression model would be the best to use?

Statistic	Model 1	Model 2	Model 3
AIC	501.5	520.4	501.5
SC	501.5	520.4	501.5
c	0.675	0.675	0.655

a. Model 1
b. Model 2
c. Model 3

Your answer: a
Correct answer: a

Models 1 and 3 are better than Model 2 because they have lower values of AIC and SC. Model 1 also has the highest values of the c statistic so it is the best of the three models.

Review: Comparing the Binary and Multiple Logistic Regression Models, Fitting a Binary Logistic Regression Model

---

Close
