The dataset "binary" contains data on the admittance status of 400 undergraduates to a graduate program at a select postgraduate institution. In addition to the admittance status, data on the ranking of the undergraduate school of the applicant is provided (1 to 4 with 1 = best). Also provided is the applicants GRE score and their GPA.

1) **Using a contingency table analysis, determine whether there is an overall association between admittance status (outcome) and the rank of the undergraduate school (exposure). Which test should you use in this case? Explain.**

```
data admit; set sasdata.binary; run;

proc freq data=admit;
      tables rank*admit / chisq norow nocol nopercent; run;
```

The FREQ Procedure

Frequency

Table of RANK by ADMIT

RANK	ADMIT 0	ADMIT 1	Total
1	28	33	61
2	97	54	151
3	93	28	121
4	55	12	67
Total	273	127	400

Statistics for Table of RANK by ADMIT

Statistic	DF	Value	Prob
Chi-Square	3	25.2421	<.0001
Likelihood Ratio Chi-Square	3	25.0098	<.0001
Mantel-Haenszel Chi-Square	1	23.4662	<.0001
Phi Coefficient		0.2512	
Contingency Coefficient		0.2436	
Cramer's V		0.2512	

Sample Size = 400

The Pearson chi-square statistic X^2 exceeds the 99% critical value of 13.277 allowing us to reject the null hypothesis of there being no association between RANK and ADMIT at a high level of confidence.

However, in this case, both the rows and columns are ordinally measured. In such cases, a more powerful test is the **Mantel-Haenzel** correlation statistic, M^2. It is more powerful because it requires only 1 dof, not the 3 in this case. So, if using X^2 implies the H_0 cannot be rejected, using M^2 may indicate there is enough evidence to reject the H_0 of no association.

Using SAS, M^2 is obtained by using the CMH option to the TABLES statement:

```
proc freq data=admit;
       tables rank*admit / chisq norow nocol nopercent cmh;
run;
```

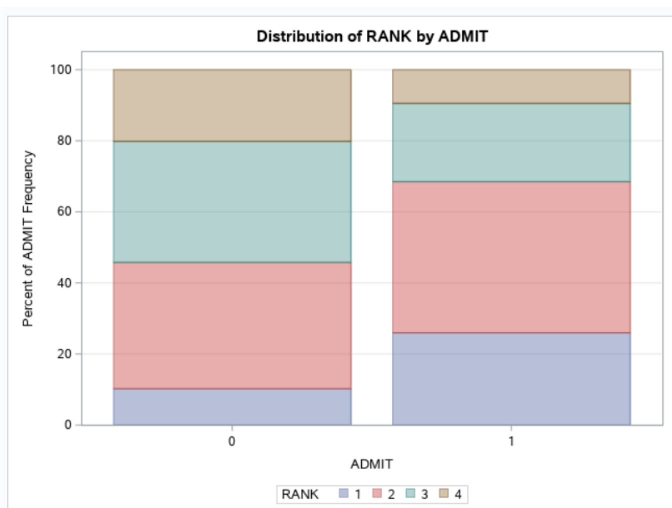
Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	23.4662	<.0001
2	Row Mean Scores Differ	3	25.1790	<.0001
3	General Association	3	25.1790	<.0001

M^2 is given by the "Nonzero Correlation" which also indicates that the H_0 of no association can be rejected. So, in this case, both X^2 and M^2 yield the same answer…but it may not always be the case.

The eagle-eyed reader will see that the Mantel-Haenszel Chi-Square statistic produced by the CHISQ option matches the Nonzero Correlation CMH statistic produced by the CMH option. These should be the same when both the row and column variables are ordinally measured.

Finally, plotting the frequency data as a bar chart reveals obvious differences across the ADMIT categories, revealing a "trend":

```
proc freq data=admit;
       tables rank*admit / chisq norow nocol nopercent cmh
       plots=freqplot(groupby=column twoway=stacked scale=grouppct);
run;
```



2) Calculate how much greater (?) are the odds of admittance having attended an undergraduate school with a ranking of "1" vs a ranking of "4". Do the same for a rank of 2 vs 4 and 3 vs 4.

Based on the contingency table produced in #1, we can calculate the odds and odds ratios as follows:

Variable	Odds	Odds Ratio Relative to RANK = 4
RANK 1	=33/28 = 1.179	=1.179/0.218 = **5.408**
RANK 2	=54/97 = 0.557	= 0.557/0.218 = **2.555**
RANK 3	=28/93 = 0.301	= 0.301/0.218 = **1.381**
RANK 4	=12/55 = 0.218	---------

So, the odds of admittance having graduated from a school with a #1 ranking are **5.4** times greater than having graduated from a school ranked #4. Likewise, the odds of admittance having graduated from a school with a #2 and #3 ranking are **2.6** and **1.4** times greater than having graduated from a school ranked #4.

And, yes, the odds are all "greater" since the odds ratios are all > 1.

3) Using PROC LOGISTIC, estimate the model admit = f(rank). Specifically, use the following:

```
proc logistic data=admit descending;
      class rank / param=ref;
      model admit = rank;
run;
```

The relevant output is shown below:

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
RANK	3	23.7795	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.5224	0.3186	22.8318	<.0001
RANK	1	1	1.6867	0.4093	16.9820	<.0001
RANK	2	1	0.9367	0.3610	6.7315	0.0095
RANK	3	1	0.3220	0.3847	0.7008	0.4025

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
RANK 1 vs 4	5.402	2.422	12.049
RANK 2 vs 4	2.552	1.257	5.177
RANK 3 vs 4	1.380	0.649	2.933

a. What does the DESCENDING option do?

This option orders the dependent variable from 1 to 0 and makes interpretation of the coefficient signs easier. Specifically, β > 0 means the variable is positively related to being admitted; β < 0 means the variable is negatively related.

b. **As a group, is the variable RANK statistically significant? Explain.**

The CLASS statement tells SAS to include n-1 categories of the CLASS variable in the model (in this case RANK 1,2 and 3). So, we can see the statistical association for each category and admittance. To see if the variable RANK overall (across all categories) is associated with admittance, we look at the table Type 3 Analysis of Effects. Here we see that RANK is indeed statistically associated with admittance at a high level of confidence (i.e. we can reject the null hypothesis of no association).

c. **Interpret the coefficient on RANK 2.**

The β on RANK 2 is 0.9367. Because we used a CLASS statement with the `param=ref` option, we are setting the reference category (by default = RANK = 4). Taking exp(0.9367) yields an OR of 2.552 which is the OR relative to RANK 4 as was saw above.

For more on the param=ref statement: https://stats.idre.ucla.edu/sas/faq/in-proc-logistic-why-arent-the-coefficients-consistent-with-the-odds-ratios/

d. **Compare the odds ratios you calculated in #2 with those produced by PROC LOGISTIC. Any differences?**

Except for rounding, the OR are the same. And they should be since we set the reference category.

e. **How precise is the estimated OR for RANK 1 vs. 4 compared to RANK 2 vs. 4 and RANK 3 vs. 4? Explain.**

The 95% CI on RANK 1 vs. 4 is much wider than that for RANK 2 and RANK 3. Although the point estimate is 4.718, the lower bound is 2.080 and the upper bound is 10.701, a range of 8.621. The CI for RANK 2, for example, is much narrower (3.757) suggesting a more precise point estimate for RANK 2 compared with that for RANK 1.

4) **Add the control variables GRE and GPA to your model:**

a. **What is your new model?**

ADMIT = f(RANK (1,2,3,4), GRE, GPA)

b. **Rerun PROC LOGISTIC adding these two control variables**

```
proc logistic data=admit descending;
```

```
        class rank / param=ref;
        model admit = gre rank gpa;
        oddsratio gre / cl=wald; /* yields cool CI charts */
run;
```

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
GRE	1	4.2842	0.0385
RANK	3	20.8949	0.0001
GPA	1	5.8714	0.0154

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.5414	1.1381	23.7081	<.0001
GRE		1	0.00226	0.00109	4.2842	0.0385
RANK	1	1	1.5514	0.4178	13.7870	0.0002
RANK	2	1	0.8760	0.3667	5.7056	0.0169
RANK	3	1	0.2112	0.3929	0.2891	0.5908
GPA		1	0.8040	0.3318	5.8714	0.0154

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
GRE	1.002	1.000	1.004
RANK 1 vs 4	4.718	2.080	10.701
RANK 2 vs 4	2.401	1.170	4.927
RANK 3 vs 4	1.235	0.572	2.668
GPA	2.235	1.166	4.282

c. Are GRE and GPA statistically related to admittance status? Explain.

Based on the Wald chi-square statistics, both GRE and GPA are statistically related to admittance at the 95% confidence level. That is, we can reject the null hypothesis that there is no association with admittance (i.e., that $\beta = 0$).

d. Compare the odds ratios for RANK with those produced by PROC LOGISTIC in #2. Any differences? If so, why?

We can see that by adding control variables, the ORs have been reduced with the RANK 1 OR relative to RANK 4 being reduced the most (13%). Adding control variables accounts for other variables that can affect admittance. To some extent by not including these, RANK was being attributed to having more of an effect on admittance than it really has. These ORs are called “adjusted” ORs.

Variable	No Control	With Control
RANK 1	5.402	4.718
RANK 2	2.552	2.401
RANK 3	1.380	1.235
RANK 4	-----	-----

e. Interpret the coefficient on GPA.

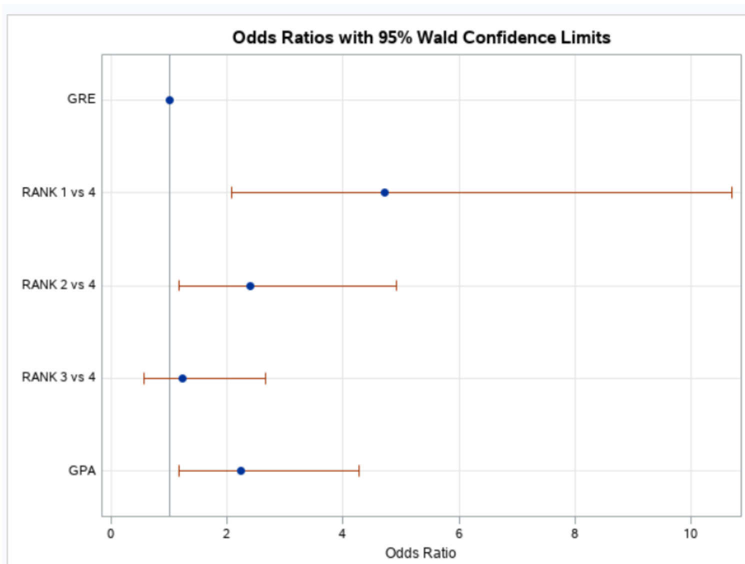
The β on GPA is 0.8040. Thus, for every 1-point increase in GPA, the log odds of admittance increase by 0.8040. Taking (exp(0.8040)-1)*100 leads us to conclude that every 1-point increase in GPA is associated with a 123% increase in the odds of admittance.

f. What is more important, GRE or GPA? Explain.

The β on GPA is 0.8040 and β on GRE is 0.0023. Both β's differ statistically from 0 at the 95% level of confidence. But **practically**, GRE has very little association with admittance. We can see this by looking at the estimated odds ratio, which is calculated relative to the average GRE score. The OR is essentially 1. The ODDSRATIO statement produces this cool CI chart which supports this conclusion:

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
GRE	1.002	1.000	1.004
RANK 1 vs 4	4.718	2.080	10.701
RANK 2 vs 4	2.401	1.170	4.927
RANK 3 vs 4	1.235	0.572	2.668
GPA	2.235	1.166	4.282



5) Using the model estimated in #4 above, explain what each of the following tell us about how well the model fits the data:

a. Log likelihood/AIC

Answers the question, “Is this model better than some other model?”. There is no H_0. Smaller values of -2LogL or the AIC indicate a better fitting model.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	501.977	470.517
SC	505.968	494.466
-2 Log L	499.977	458.517

b. Likelihood Ratio

Answers the question, “Is this model better than no model?”. That is, better than one with just an intercept. The H_0 is that all β on the current model = 0. In this case we can reject this H_0.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	41.4590	5	<.0001
Score	40.1603	5	<.0001
Wald	36.1390	5	<.0001

c. Deviance (use AGGREGATE SCALE=NONE option in MODEL statement)

Answers the question, “Is there a better model than this one?” Specifically, a saturated model (one with interactions). The H_0 is that the coefficients on the interaction terms are jointly are = 0. Here the evidence is mixed. The deviance says we can reject the $H_{0,}$ but the Pearson deviance does not. We should investigate interactions and see what we find.

However, with a continuous variable, we have many groupings or profiles.[1] In such cases the deviance test is questionable (the deviance test is more applicable when there is a limited number of profiles, as would occur if all variables were categorical), and the HL should be used.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	446.3806	385	1.1594	0.0167
Pearson	388.0596	385	1.0079	0.4467

Number of unique profiles: 391

[1] A model is fully saturated when there is 1 estimated parameter for each grouping or profile of the data. With a continuous variable, such a model is not possible. In this example we have 391 profiles.

d. Hosmer-Lemeshow (HL) test (use LACKFIT option in MODEL statement)

Answers the same question as the deviance test, "Is there a better model than this one?" That is, one with interactions and non-linearities. The H_0 is that the current model fits the data well. Here we see that the HL test indicates that the H_0 cannot be rejected suggesting that a more complex model is not warranted.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
11.0854	8	0.1969

Note that we can always estimate a model using interactions. Here is one using all possible 2 and 3-way interactions. None are statistically significant at the 95% confidence level.

```
proc logistic data=admit descending;
      class rank / param=ref;
      model admit = gre gpa rank gpa*gre gpa*rank gre*rank
      gre*gpa*rank;
run;
```

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-22.0616	18.6273	1.4027	0.2363
GRE		1	0.0292	0.0313	0.8714	0.3506
GPA		1	5.5758	5.3341	1.0927	0.2959
RANK	1	1	-13.9383	24.5875	0.3214	0.5708
RANK	2	1	20.0163	20.7147	0.9337	0.3339
RANK	3	1	0.6710	22.8830	0.0009	0.9766
GRE*GPA		1	-0.00775	0.00894	0.7522	0.3858
GPA*RANK	1	1	4.6669	7.1203	0.4296	0.5122
GPA*RANK	2	1	-5.4385	5.9902	0.8243	0.3639
GPA*RANK	3	1	-0.3121	6.5547	0.0023	0.9620
GRE*RANK	1	1	0.0231	0.0401	0.3311	0.5650
GRE*RANK	2	1	-0.0312	0.0346	0.8137	0.3670
GRE*RANK	3	1	0.00298	0.0385	0.0060	0.9382
GRE*GPA*RANK	1	1	-0.00689	0.0115	0.3566	0.5504
GRE*GPA*RANK	2	1	0.00884	0.00997	0.7858	0.3754
GRE*GPA*RANK	3	1	-0.00051	0.0110	0.0021	0.9631