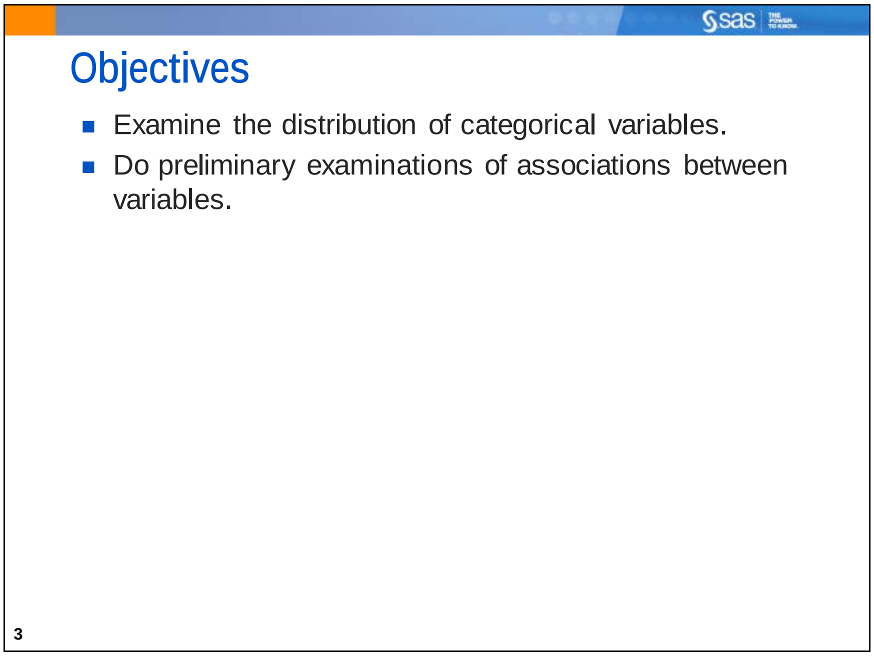
Chapter 5   Categorical Data Analysis

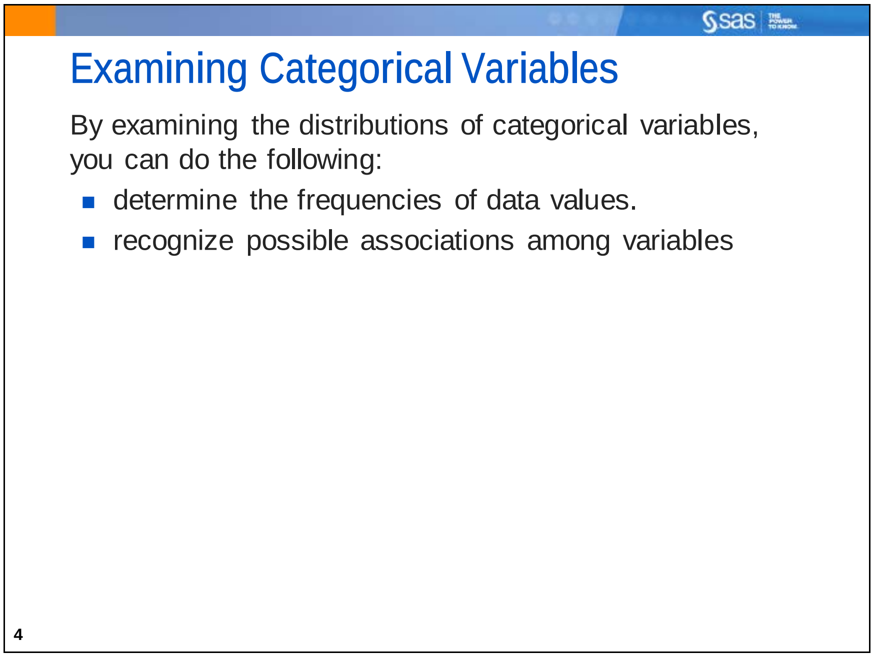
5.1 Describing Categorical Data



Objectives

- Examine the distribution of categorical variables.
- Do preliminary examinations of associations between variables.

3

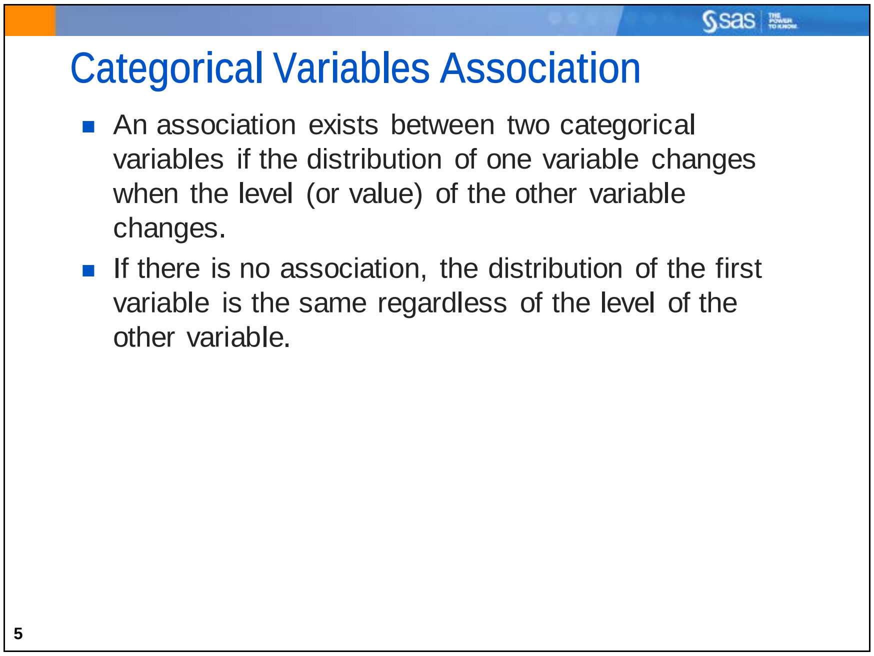


Examining Categorical Variables

By examining the distributions of categorical variables, you can do the following:

- determine the frequencies of data values.
- recognize possible associations among variables

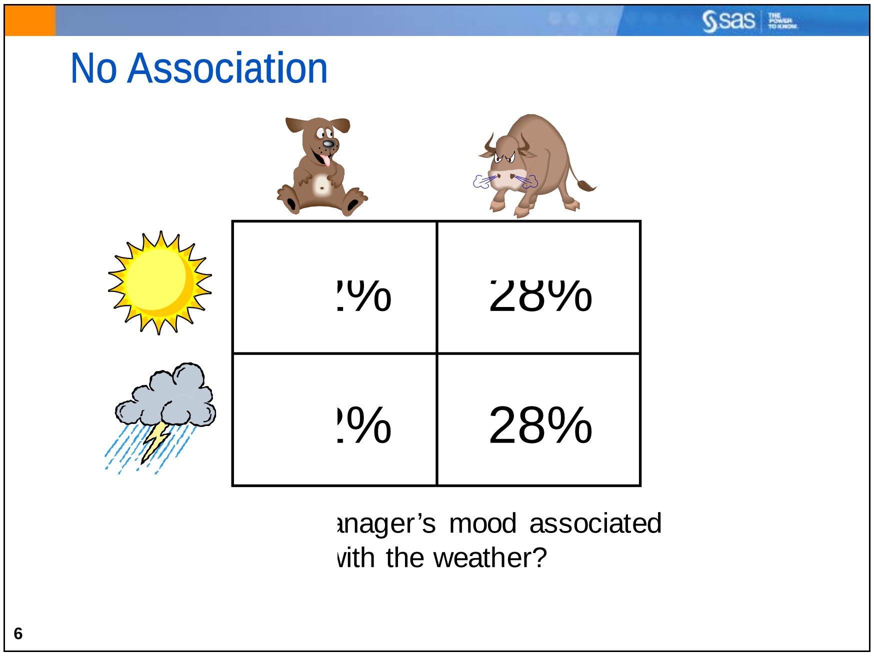
4



Categorical Variables Association

- An association exists between two categorical variables if the distribution of one variable changes when the level (or value) of the other variable changes.
- If there is no association, the distribution of the first variable is the same regardless of the level of the other variable.

5



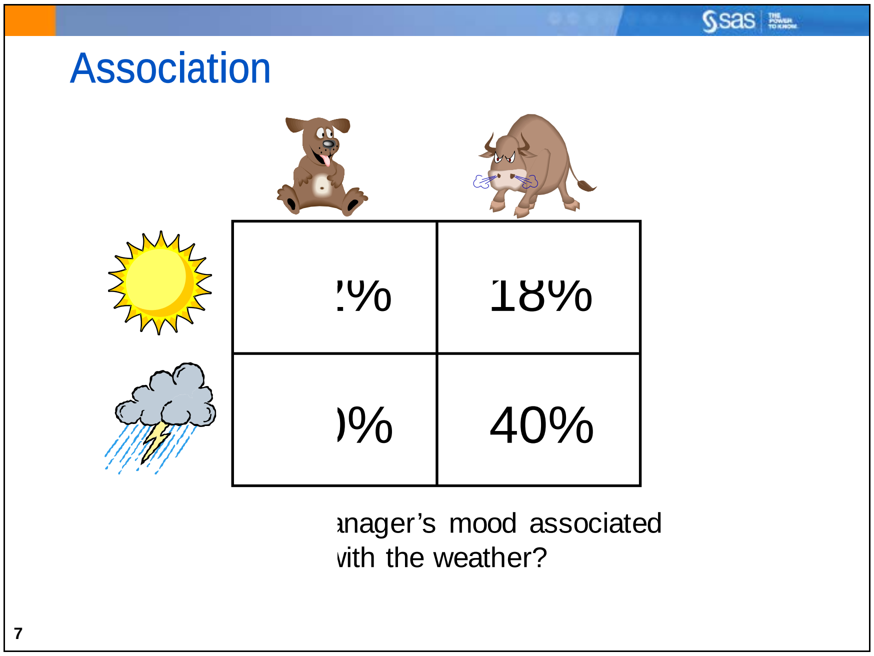
No Association

2%	28%
2%	28%

anager's mood associated
vith the weather?

6

There appears to be no association between your manager's mood and the weather here because the row percentages are the ***same*** in each column.

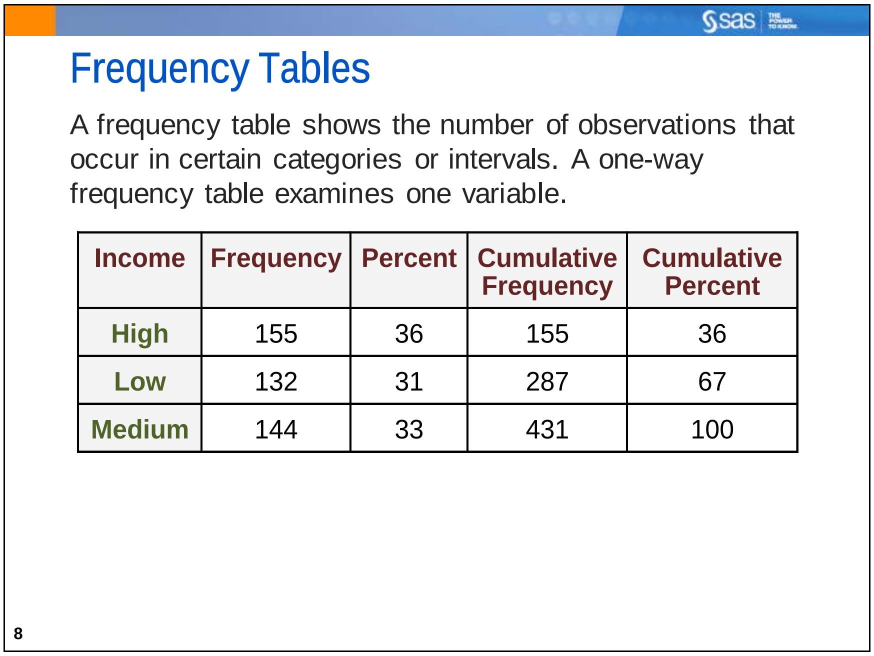


Association

2%	18%
0%	40%

anager's mood associated
vith the weather?

There appears to be an association here because the row percentages are ***different*** in each column.



Frequency Tables

A frequency table shows the number of observations that occur in certain categories or intervals. A one-way frequency table examines one variable.

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	155	36	155	36
Low	132	31	287	67
Medium	144	33	431	100

8

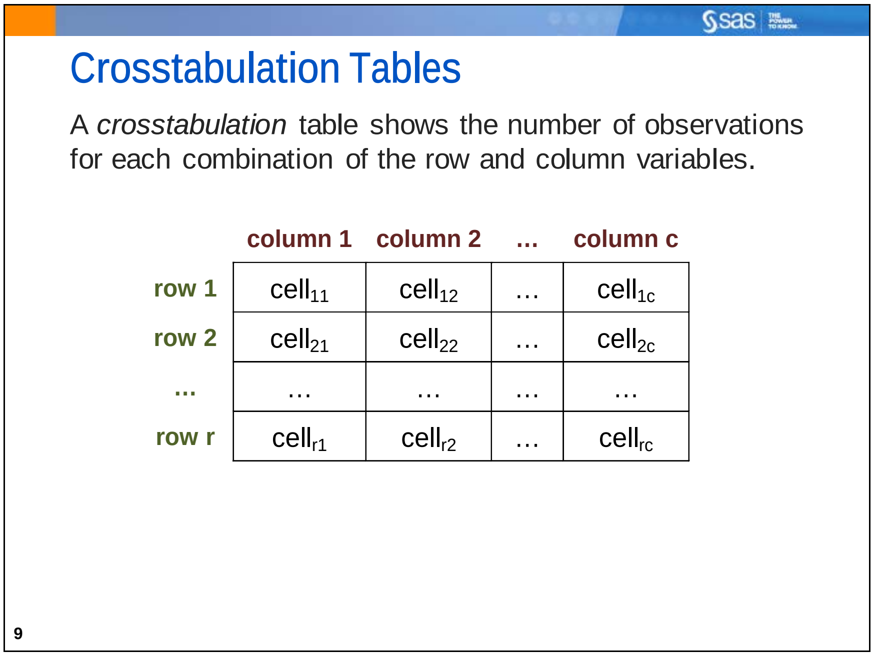
Typically, there are four types of frequency measures included in a frequency table:

Frequency is the number of times the value appears in the data set.

Percent represents the percentage of the data that has this value.

Cumulative Frequency accumulates the frequency of each of the values by adding the second frequency to the first, and so on.

Cumulative Percent accumulates the percentage by adding the second percentage to the first, and so on.



Crosstabulation Tables

A *crosstabulation* table shows the number of observations for each combination of the row and column variables.

	column 1	column 2	…	column c
row 1	cell$_{11}$	cell$_{12}$	…	cell$_{1c}$
row 2	cell$_{21}$	cell$_{22}$	…	cell$_{2c}$
…	…	…	…	…
row r	cell$_{r1}$	cell$_{r2}$	…	cell$_{rc}$

9

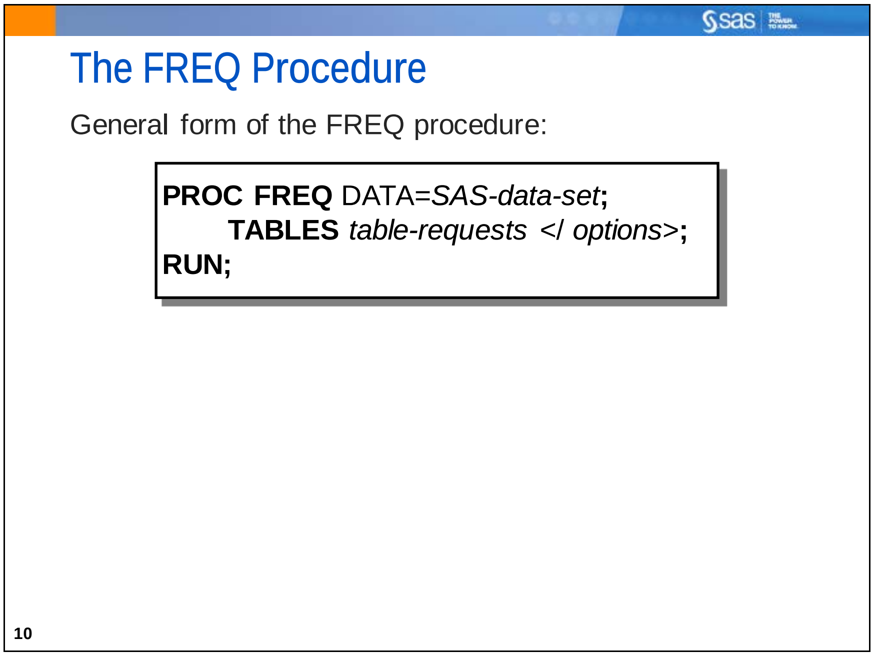
By default, a crosstabulation table has four measures in each cell:

Frequency Number of observations falling into a category formed by the row variable value and the column variable value

Percent Number of observations in each cell as a percentage of the total number of observations

Row Pct Number of observations in each cell as a percentage of the total number of observations in that row

Col Pct Number of observations in each cell as a percentage of the total number of observations in that column



The FREQ Procedure

General form of the FREQ procedure:

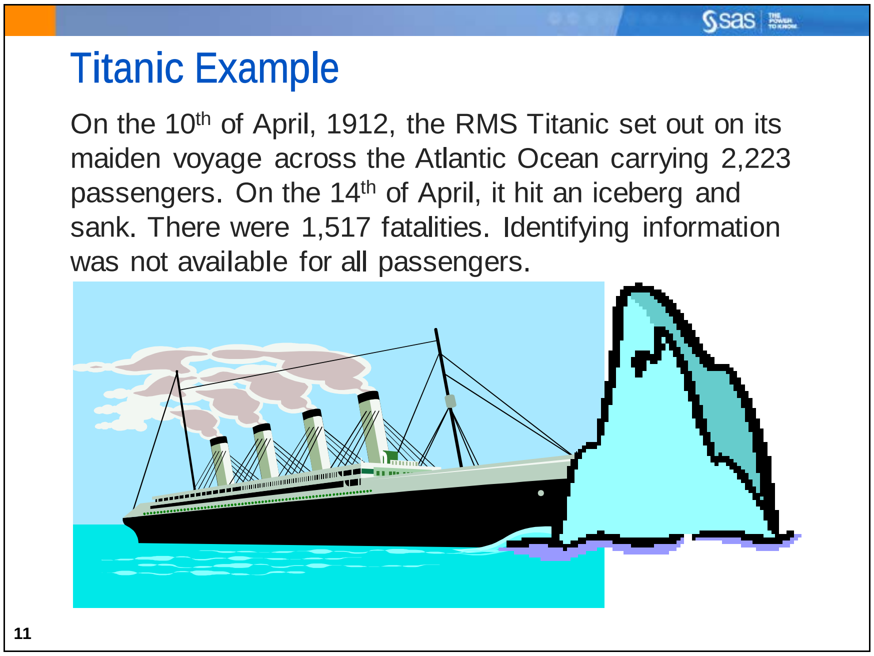
```
PROC FREQ DATA=SAS-data-set;
      TABLES table-requests </ options>;
RUN;
```

10

Selected FREQ procedure statement:

TABLES requests tables and specifies options for producing tests. The general form of a table request is *variable1*variable2**…, where any number of these requests can be made in a single TABLES statement. For two-way crosstabulation tables, the first variable represents the rows and the second variable represents the columns.

PROC FREQ can generate large volumes of output as the number of variables or the number of variable levels (or both) increases.



Titanic Example

On the 10th of April, 1912, the RMS Titanic set out on its maiden voyage across the Atlantic Ocean carrying 2,223 passengers. On the 14th of April, it hit an iceberg and sank. There were 1,517 fatalities. Identifying information was not available for all passengers.

11

Example: The data are stored in the **sasuser.Titanic** data set.

These are the variables in the data set:

Survival survival status (1=**Survived**, 0=**Died**)

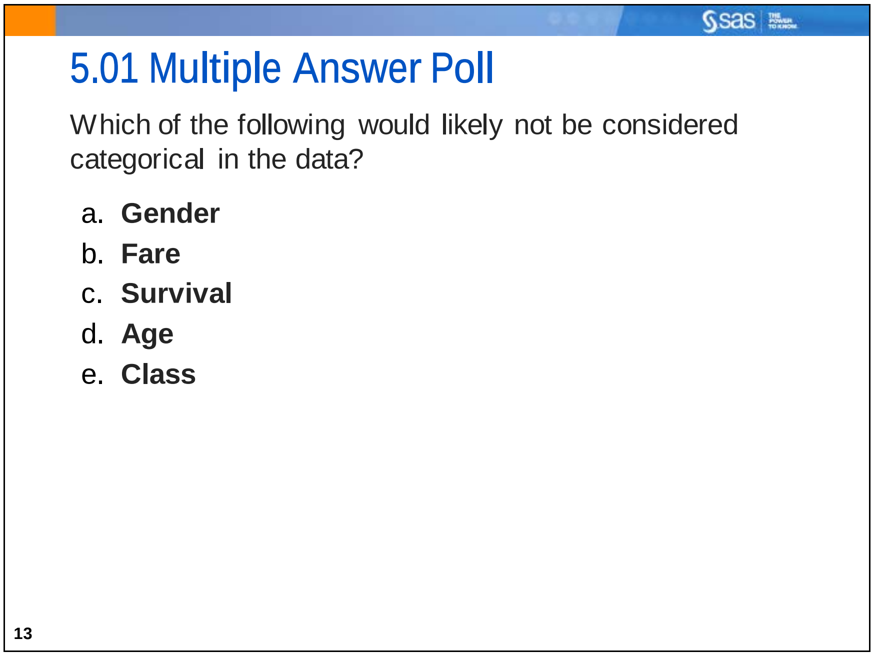
Age age of passenger in years

Gender gender of passenger (**male**, **female**)

Class ticket class (**1**, **2**, **3**)

Fare ticket fare (This variable is misleading because it is shown as the cumulative total for a purchase for each person in a party.)

This is a publically available data set.



5.01 Multiple Answer Poll

Which of the following would likely not be considered categorical in the data?

a. **Gender**
b. **Fare**
c. **Survival**
d. **Age**
e. **Class**

13

Examining Distributions

Example: Invoke PROC FREQ and create one-way frequency tables for the variables **Gender**, **Class**, and **Survival** and create two-way frequency tables for the variables **Survival** by **Gender**, and **Survival** by **Class**. For the continuous variable, **Age**, create histograms for each level of **Survival**. Use a CLASS statement in PROC UNIVARIATE.

Use the FORMAT procedure to format the values of **Survival**.

```
/*st105d01.sas*/
title;
proc format;
   value survfmt 1="Survived"
                 0="Died"
                 ;
run;

proc freq data=sasuser.Titanic;
   tables Survived Gender Class
          Gender*Survived Class*Survived /
          plots(only)=freqplot(scale=percent);
   format Survived survfmt.;
run;

proc univariate data=sasuser.Titanic noprint;
   class Survived;
   var Age;
   histogram Age;
   inset mean std median min max / format=5.2 position=ne;
   format Survived survfmt.;
run;
```

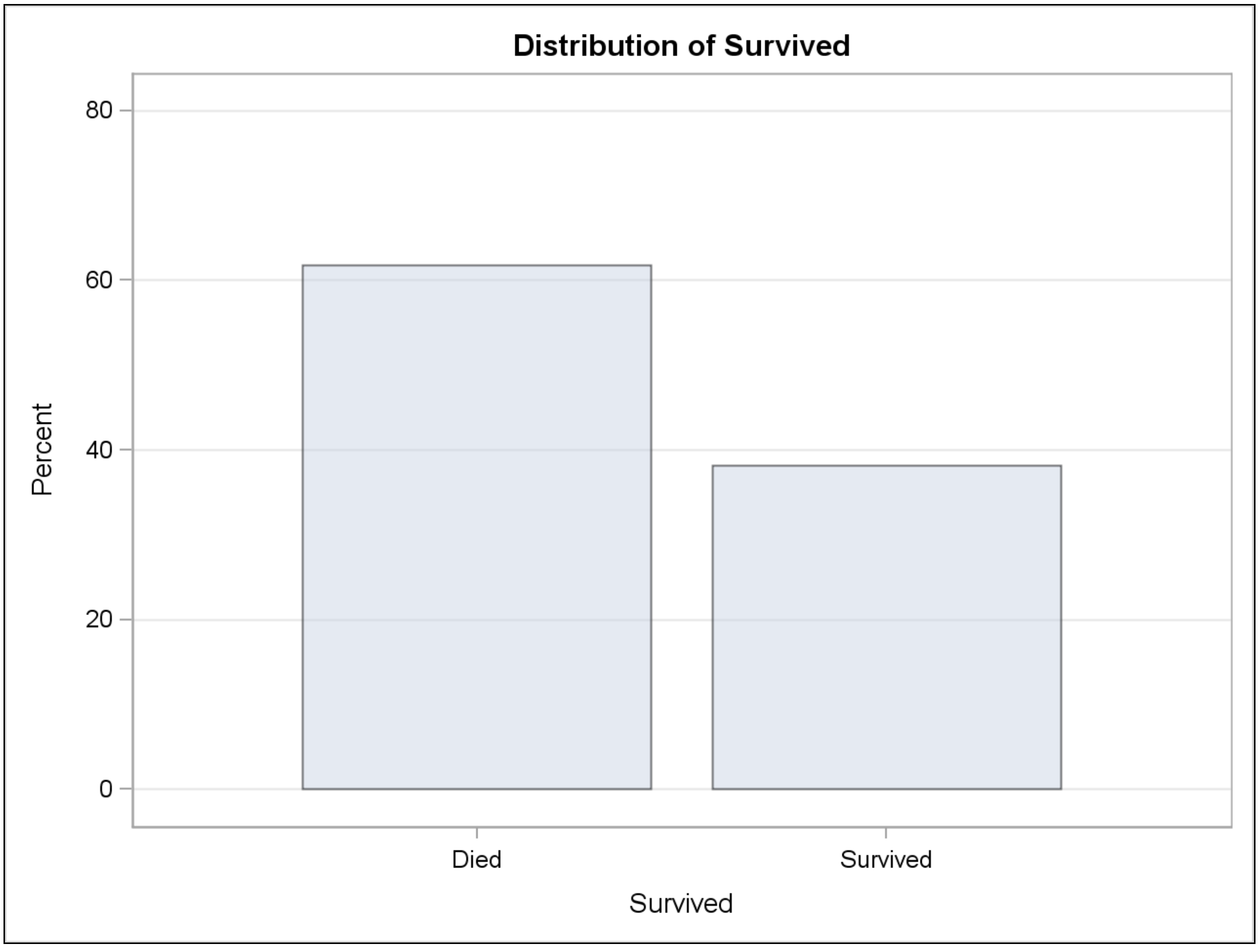
Selected TABLES statement PLOTS option and suboptions:

FREQPLOT(<*suboptions*>) requests a frequency plot. Frequency plots are available for frequency and crosstabulation tables. For multiway tables, PROC FREQ provides a two-way frequency plot for each stratum.

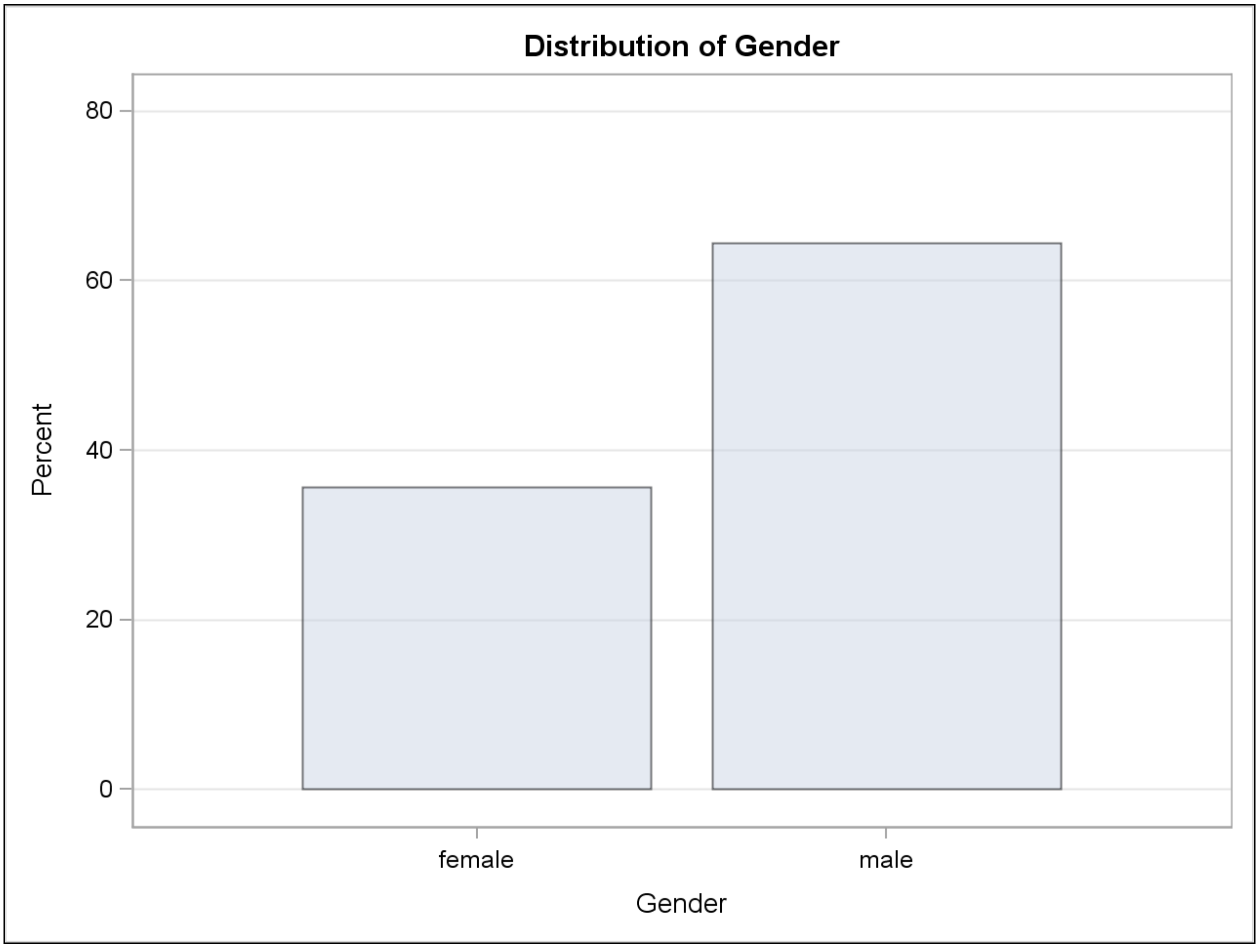
(SCALE=) specifies the scale of the frequencies to display. The default is SCALE=FREQ, which displays unscaled frequencies. SCALE=PERCENT displays percentages (relative frequencies).

PROC FREQ Output

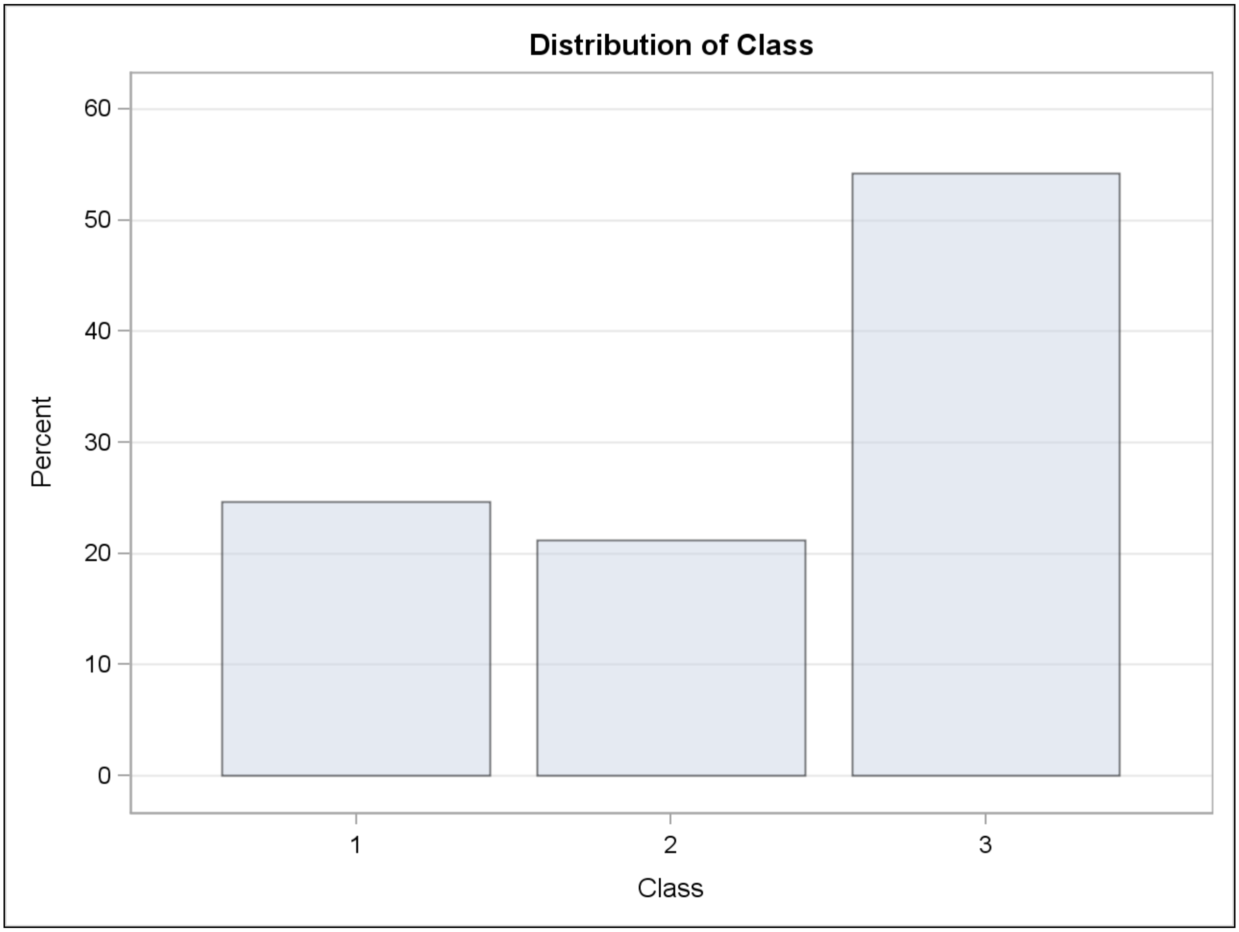
Survived	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Died	809	61.80	809	61.80
Survived	500	38.20	1309	100.00



Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
female	466	35.60	466	35.60
male	843	64.40	1309	100.00



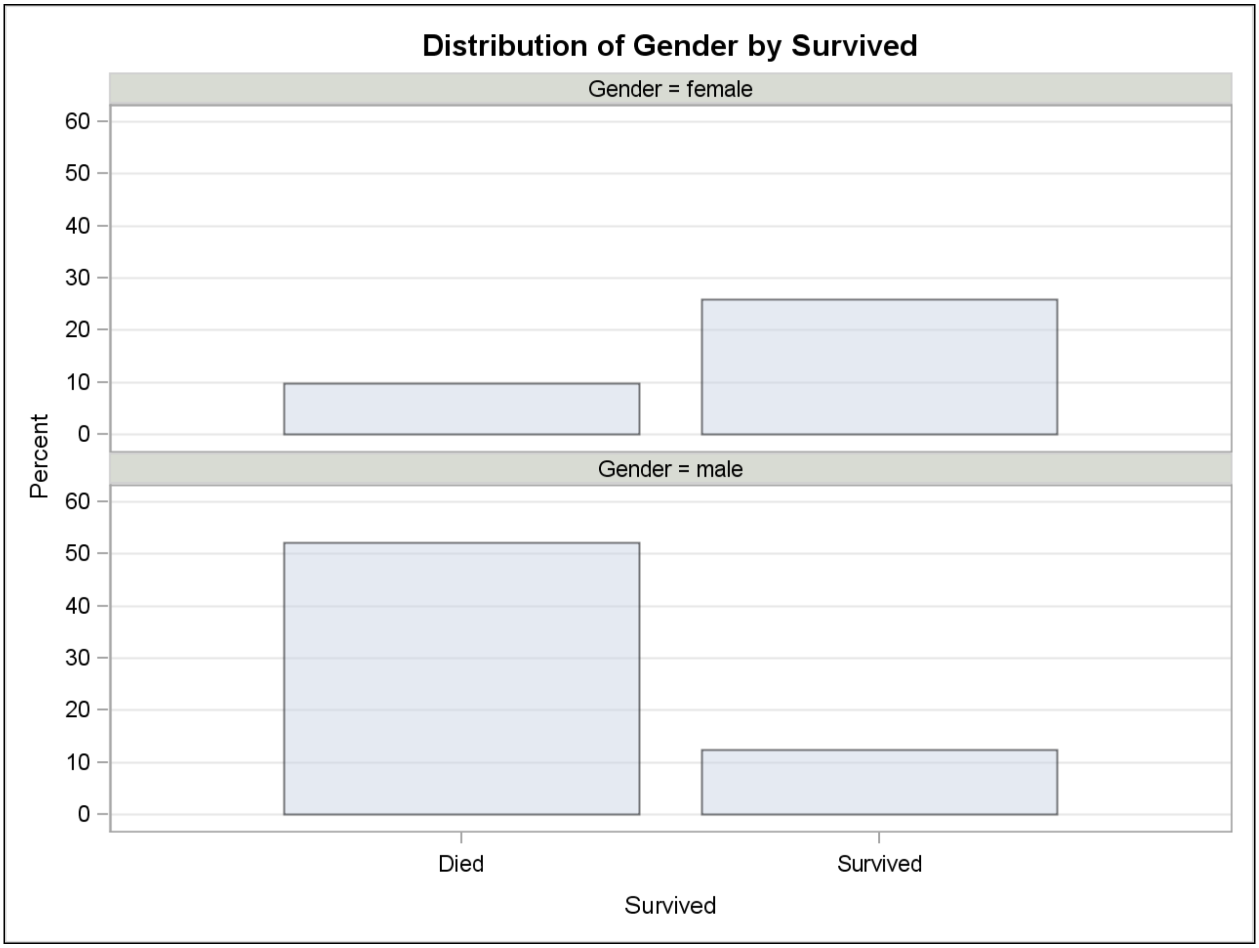
Class	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	323	24.68	323	24.68
2	277	21.16	600	45.84
3	709	54.16	1309	100.00



There seem to be no unusual data values that could be due to coding errors for any of the categorical variables.

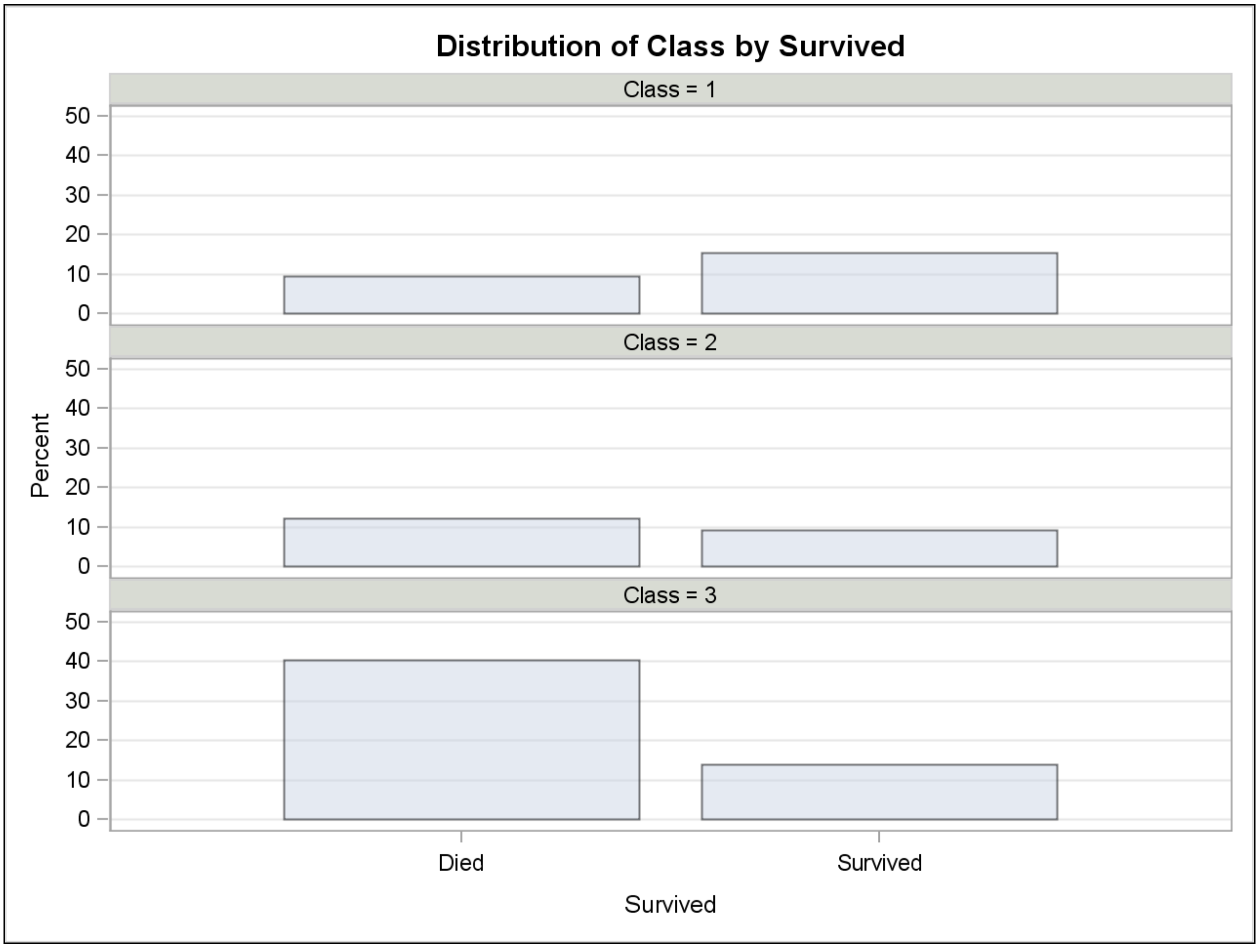
The requested two-way frequency tables follow. You can get a preliminary idea whether there are associations between the outcome variable, **Survival**, and the predictor variables, **Gender** and **Class**, by examining the distribution of **Survival** at each value of the predictors.

Table of Gender by Survived			
Gender	**Survived**		
Frequency **Percent** **Row Pct** **Col Pct**	**Died**	**Survived**	**Total**
female	127 9.70 27.25 15.70	339 25.90 72.75 67.80	466 35.60
male	682 52.10 80.90 84.30	161 12.30 19.10 32.20	843 64.40
Total	809 61.80	500 38.20	1309 100.00



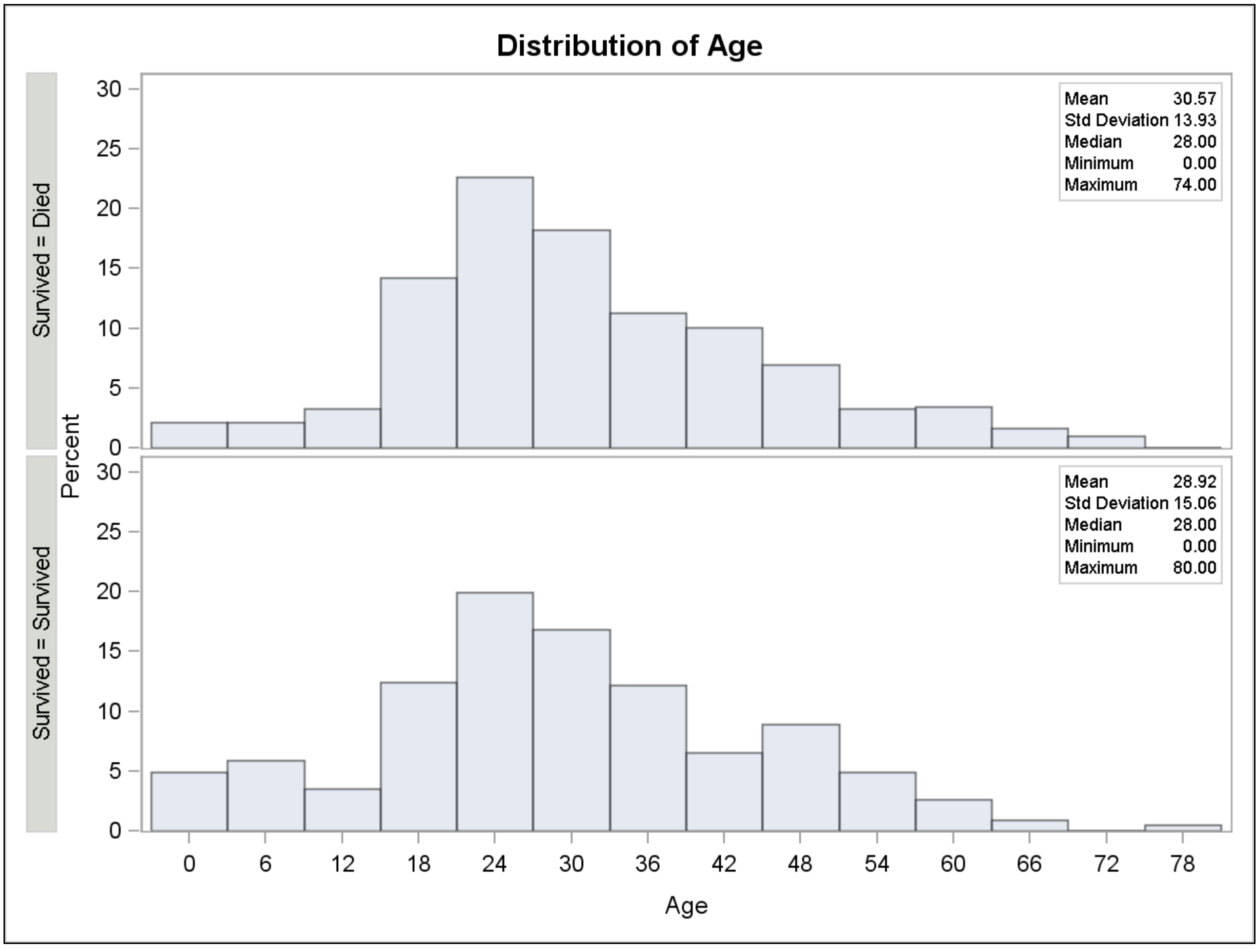
By examining the row percentages, you see that **Survival** is associated with **Gender**.

Table of Class by Survived			
Class	**Survived**		
Frequency Percent Row Pct Col Pct	**Died**	**Survived**	**Total**
1	123 9.40 38.08 15.20	200 15.28 61.92 40.00	323 24.68
2	158 12.07 57.04 19.53	119 9.09 42.96 23.80	277 21.16
3	528 40.34 74.47 65.27	181 13.83 25.53 36.20	709 54.16
Total	809 61.80	500 38.20	1309 100.00



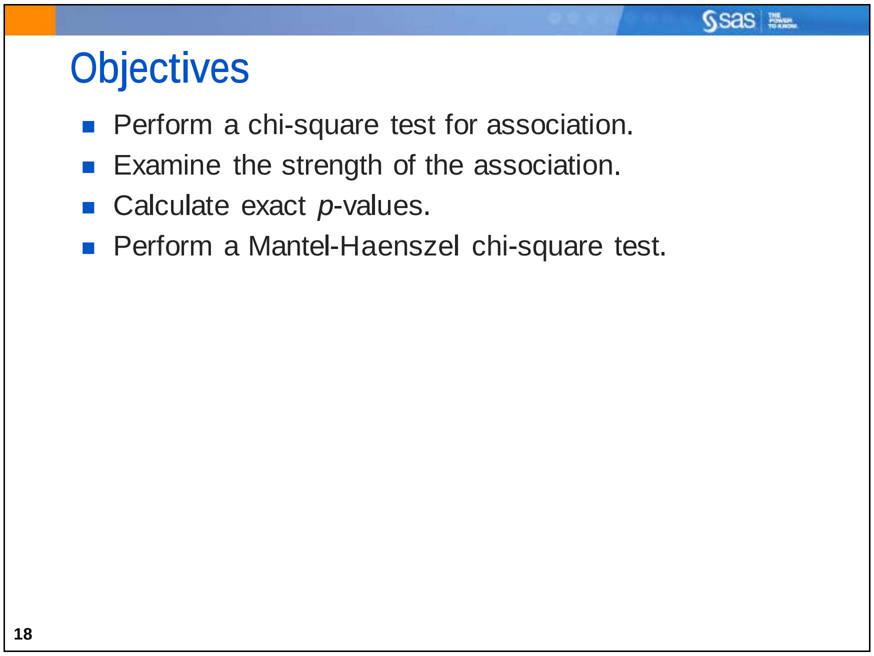
There also seems to be an association between **Survival** and **Class**, with a far greater chance of surviving in higher classes.

The plot below shows the distribution of the continuous variable, **Age**, by survival status.



The distribution of **Age** appears to have no obvious outliers or strange shape for either group.

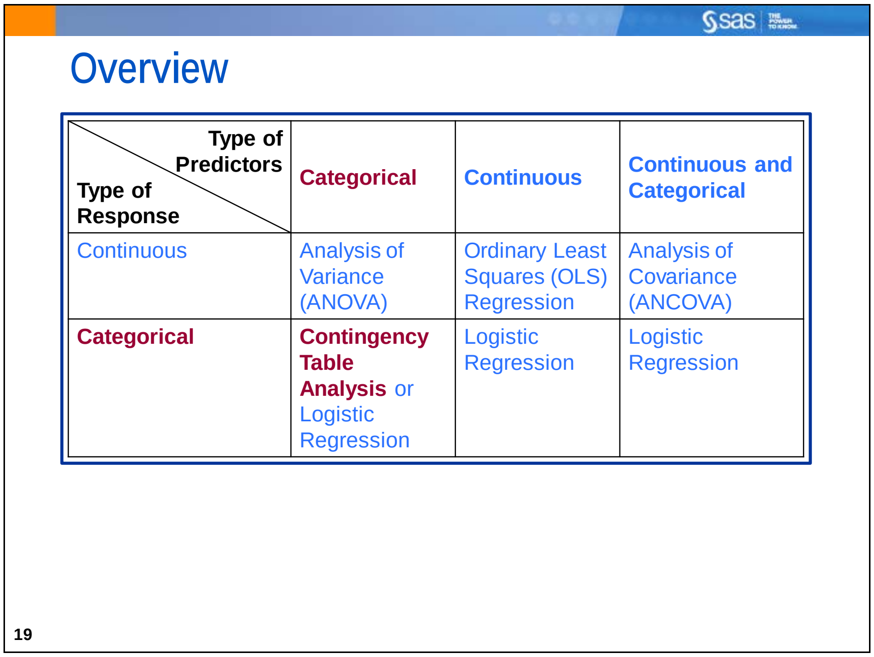
5.2 Tests of Association



Objectives

- Perform a chi-square test for association.
- Examine the strength of the association.
- Calculate exact *p*-values.
- Perform a Mantel-Haenszel chi-square test.

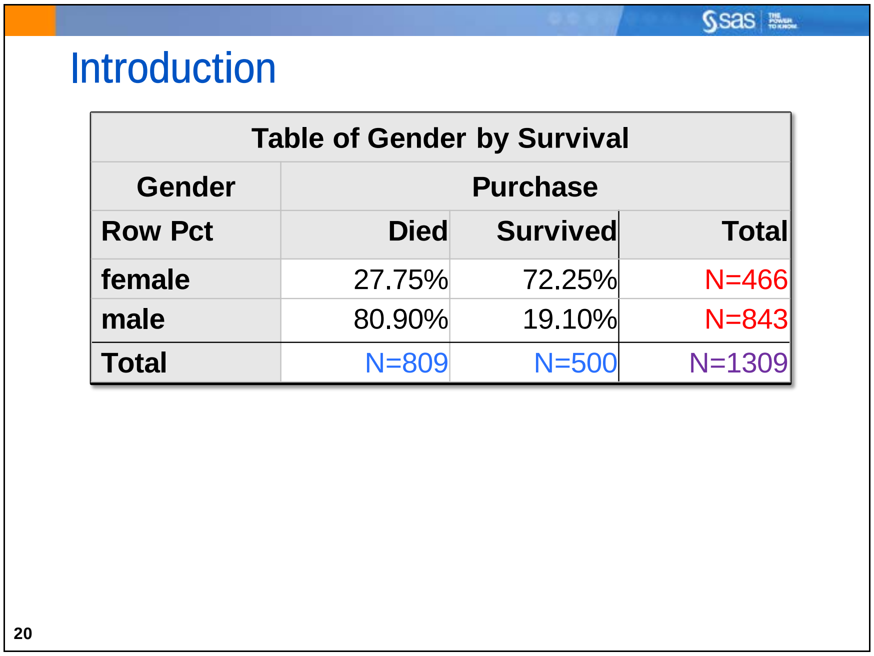
18



Overview

Type of Predictors / **Type of Response**	**Categorical**	**Continuous**	**Continuous and Categorical**
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	**Contingency Table Analysis** or Logistic Regression	Logistic Regression	Logistic Regression

19

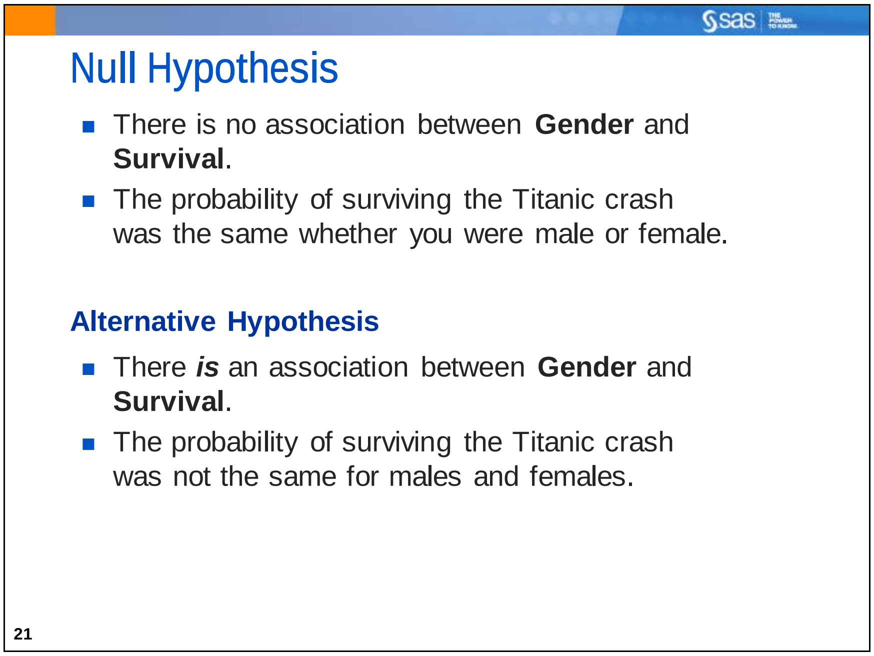


Introduction

Table of Gender by Survival			
Gender	**Purchase**		
Row Pct	**Died**	**Survived**	**Total**
female	27.75%	72.25%	N=466
male	80.90%	19.10%	N=843
Total	N=809	N=500	N=1309

20

There appears to be an association between **Gender** and **Survival** because the row probabilities are different in each column. To test for this association, you assess whether the difference between the probabilities of females surviving (72.25%) and males surviving (19.10%) is greater than would be expected by chance.



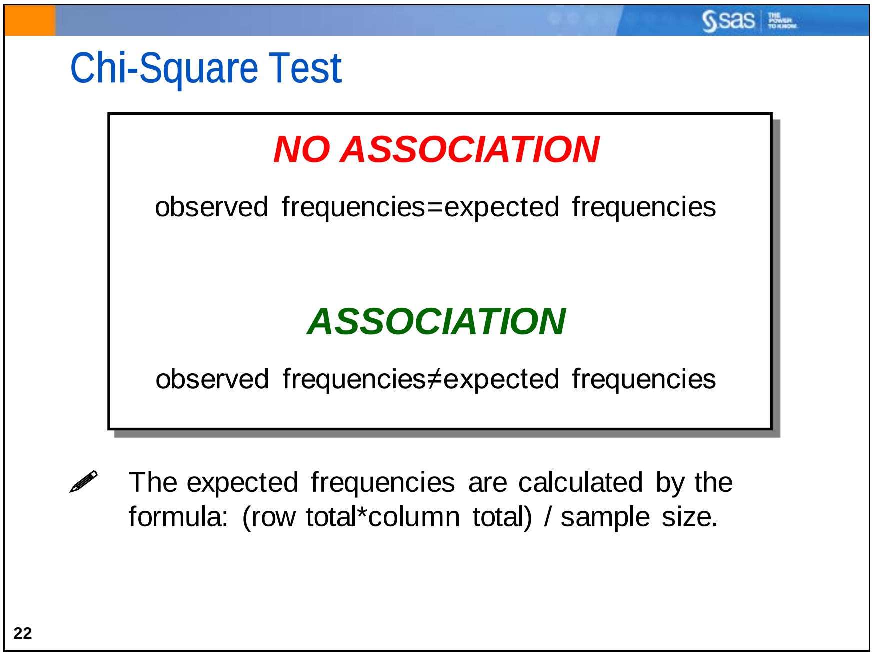
Null Hypothesis

- There is no association between **Gender** and **Survival**.
- The probability of surviving the Titanic crash was the same whether you were male or female.

Alternative Hypothesis

- There ***is*** an association between **Gender** and **Survival**.
- The probability of surviving the Titanic crash was not the same for males and females.

21



Chi-Square Test

NO ASSOCIATION

observed frequencies=expected frequencies

ASSOCIATION

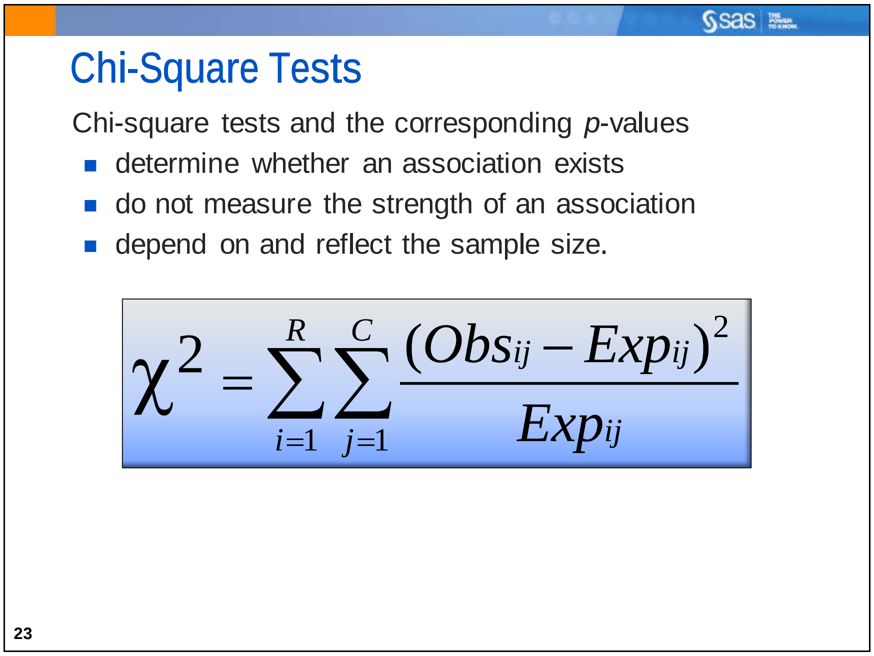
observed frequencies≠expected frequencies

✎ The expected frequencies are calculated by the formula: (row total*column total) / sample size.

22

A commonly used test that examines whether there is an association between two categorical variables is the Pearson chi-square test. The chi-square test measures the difference between the observed cell frequencies and the cell frequencies that are expected if there is no association between the variables. If you have a significant chi-square statistic, there is strong evidence that an association exists between your variables.

✎ Under the null hypothesis of no association between Row and Column variable, the “expected” percentage in any R*C cell will be equal to the percent in that cell’s row (R / T) times the percent in the cell’s column (C / T). The expected count is then only that expected percentage times the total sample size. The expected count=(R/T)*(C/T)*T=(R*C)/T.



Chi-Square Tests

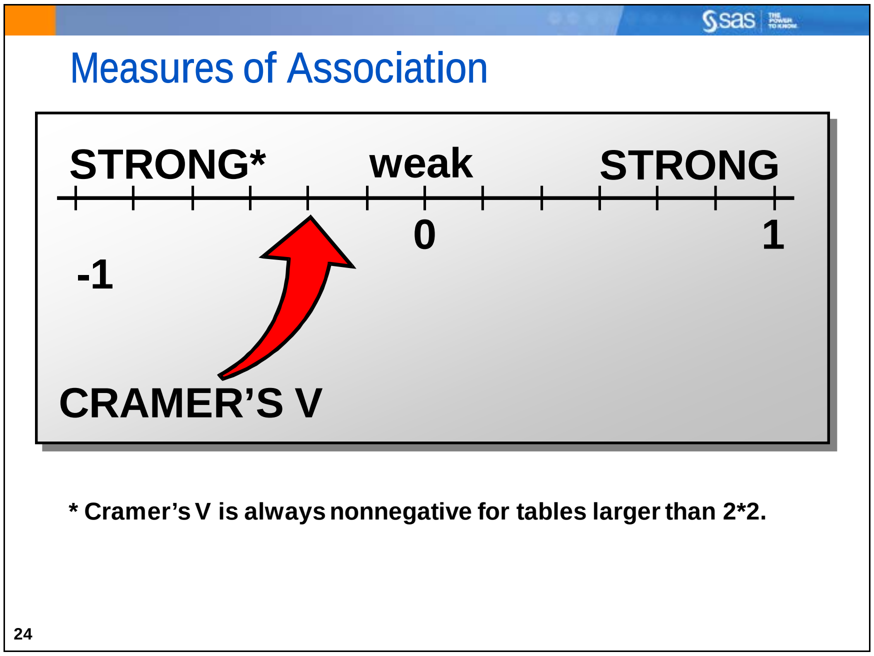
Chi-square tests and the corresponding *p*-values

- determine whether an association exists
- do not measure the strength of an association
- depend on and reflect the sample size.

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

23

The *p*-value for the chi-square test only indicates how confident you can be that the null hypothesis of no association is false. It does not tell you the magnitude of an association. The value of the chi-square statistic also does not tell you the magnitude of the association. If you double the size of your sample by duplicating each observation, you double the value of the chi-square statistic, even though the strength of the association does not change.



Measures of Association

STRONG* weak STRONG

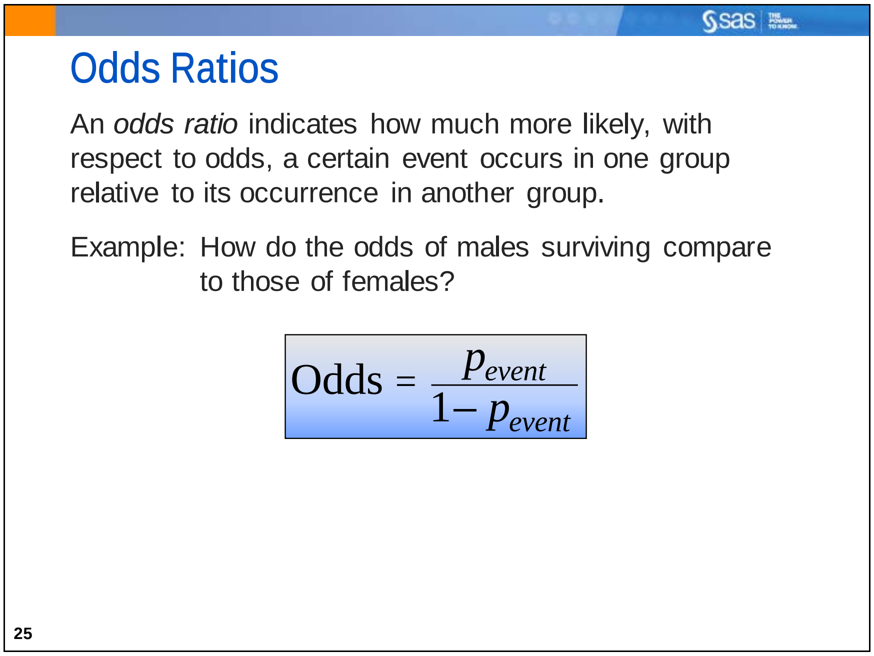
-1 0 1

CRAMER'S V

* Cramer's V is always nonnegative for tables larger than 2*2.

24

One measure of the strength of the association between two nominal variables is Cramer's V statistic. It has a range of –1 to 1 for 2-by-2 tables and 0 to 1 for larger tables. Values farther from 0 indicate stronger association. Cramer's V statistic is derived from the Pearson chi-square statistic.



Odds Ratios

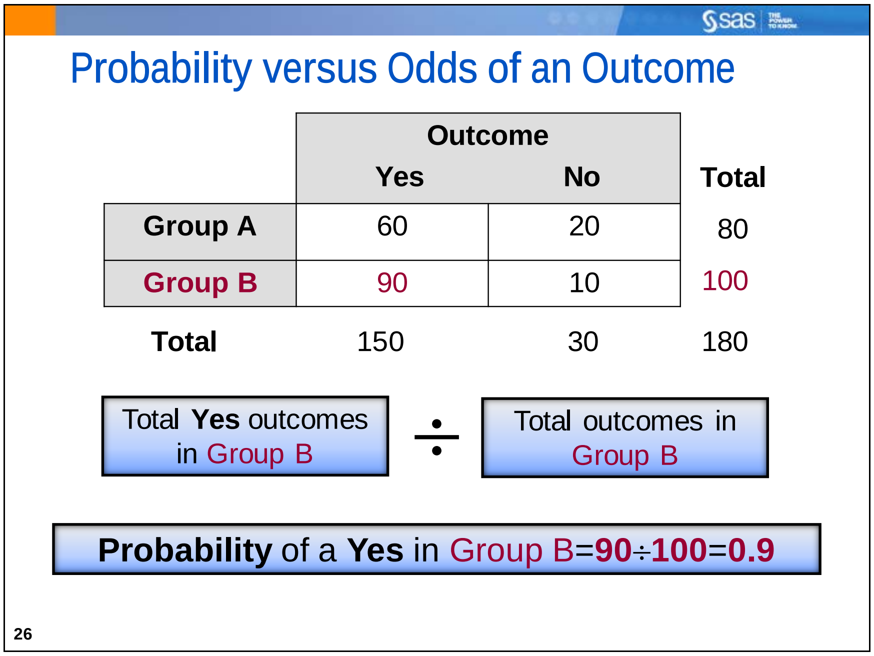
An *odds ratio* indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.

Example: How do the odds of males surviving compare to those of females?

$$\text{Odds} = \frac{p_{event}}{1 - p_{event}}$$

25

The odds ratio can be used as a measure of the strength of association for 2 * 2 tables. Do not mistake odds for probability. Odds are calculated from probabilities as shown in the next slides.



Probability versus Odds of an Outcome

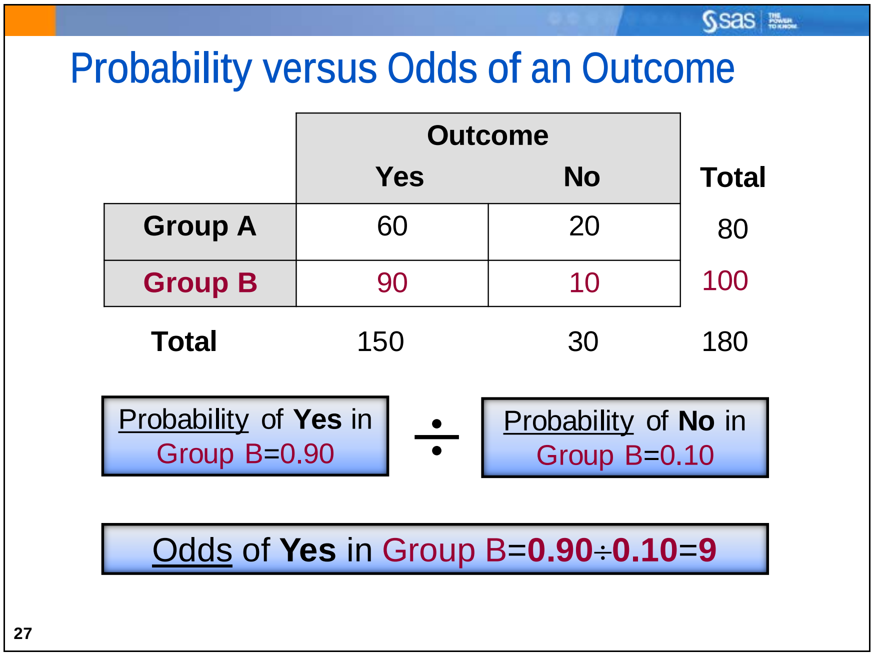
	Outcome		
	Yes	**No**	**Total**
Group A	60	20	80
Group B	90	10	100
Total	150	30	180

Total **Yes** outcomes in Group B ÷ Total outcomes in Group B

Probability of a **Yes** in Group B=**90**÷**100**=**0.9**

26

There is a 90% probability of having the outcome in group B. What is the probability of having the outcome in group A?



Probability versus Odds of an Outcome

	Outcome Yes	 No	 Total
Group A	60	20	80
Group B	90	10	100
Total	150	30	180

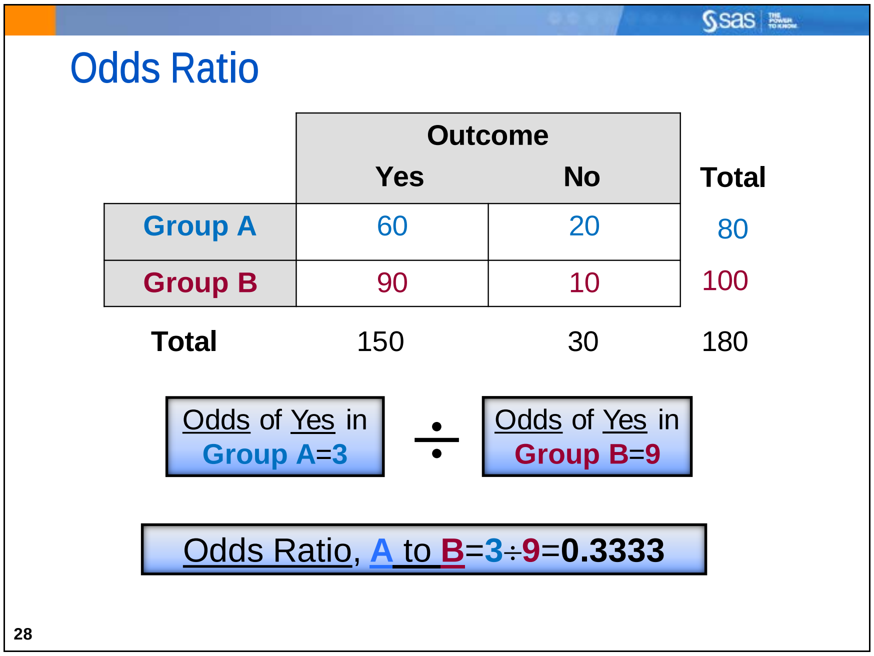
Probability of **Yes** in Group B=0.90 ÷ Probability of **No** in Group B=0.10

Odds of **Yes** in Group B=**0.90÷0.10=9**

27

The odds of an outcome are the ratio of the expected probability that the outcome will occur to the expected probability that the outcome will ***not*** occur. The odds for group B are 9, which indicate that you expect nine times as many occurrences as non-occurrences in group B.

What are the odds of having the outcome in group A?



Odds Ratio

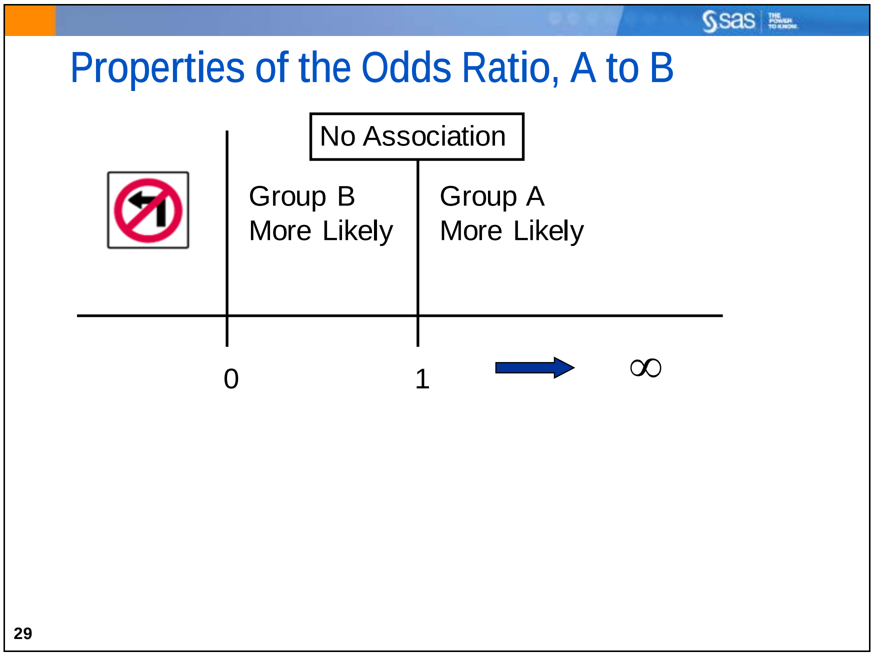
	Outcome Yes	 No	 Total
Group A	60	20	80
Group B	90	10	100
Total	150	30	180

Odds of Yes in **Group A=3** ÷ Odds of Yes in **Group B=9**

Odds Ratio, A to B=3÷9=**0.3333**

28

The odds ratio of group A to group B equals 1/3, or 0.3333, which indicates that the odds of getting the outcome in group A are one third those in group B. If you were interested in the odds ratio of group B to group A, you would simply take the inverse of 1/3 to arrive at 3.



Properties of the Odds Ratio, A to B

No Association

Group B
More Likely

Group A
More Likely

0

1

∞

29

The odds ratio shows the strength of the association between the predictor variable and the outcome variable. If the odds ratio is 1, then there is no association between the predictor variable and the outcome. If the odds ratio is greater than 1, then group A, the numerator group, is more likely to have the outcome. If the odds ratio is less than 1, then group B, the denominator group, is more likely to have the outcome.

Chi-Square Test

Example: Use the FREQ procedure to test for an association between the variables **Gender** and **Survived**. Also generate the expected cell frequencies and the cell's contribution to the total chi-square statistic.

```
/*st105d02.sas*/
ods graphics off;
proc freq data=sasuser.Titanic;
   tables (Gender Class)*Survived
         / chisq expected cellchi2 nocol nopercent
           relrisk;
   format Survived survfmt.;
   title1 'Associations with Survival';
run;
ods graphics on;
```

Selected TABLES statement options:

CHISQ produces the chi-square test of association and the measures of association based on the chi-square statistic.

EXPECTED prints the expected cell frequencies under the hypothesis of no association.

CELLCHI2 prints each cell's contribution to the total chi-square statistic.

NOCOL suppresses printing the column percentages.

NOPERCENT suppresses printing the cell percentages.

RELRISK prints a table with risk ratios (probability ratios) and odds ratios.

The frequency table is shown below.

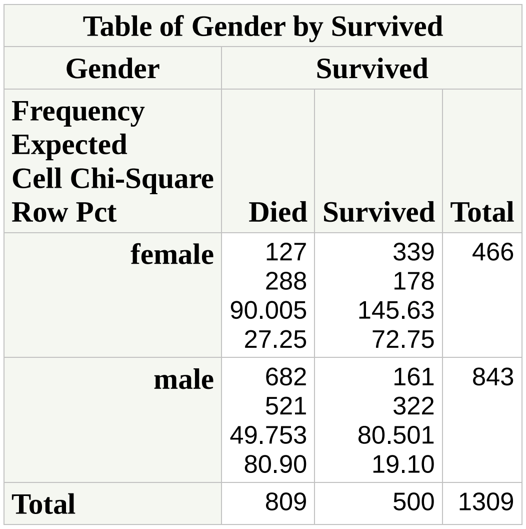


Table of Gender by Survived			
Gender	**Survived**		
Frequency **Expected** **Cell Chi-Square** **Row Pct**	**Died**	**Survived**	**Total**
female	127 288 90.005 27.25	339 178 145.63 72.75	466
male	682 521 49.753 80.90	161 322 80.501 19.10	843
Total	809	500	1309

It appears that the cell for **Survived**=`1` (`Survived`) and **Gender**=`female` contributes the most to the chi-square statistic. The Cell Chi-Square value is 145.63.

The cell chi-square is calculated using the formula (observed frequency–expected frequency)^2/expected frequency.

The overall chi-square statistic is calculated by adding the cell chi-square values over all rows and columns: $\Sigma\Sigma$((observed$_{rc}$–expected$_{rc}$)^2/expected$_{rc}$).

Below is the table that shows the chi-square test and Cramer's V.

Statistic	DF	Value	Prob
Chi-Square	1	365.8869	<.0001
Likelihood Ratio Chi-Square	1	372.9213	<.0001
Continuity Adj. Chi-Square	1	363.6179	<.0001
Mantel-Haenszel Chi-Square	1	365.6074	<.0001
Phi Coefficient		-0.5287	
Contingency Coefficient		0.4674	
Cramer's V		-0.5287	

Because the *p*-value for the chi-square statistic is <.0001, which is below .05, you reject the null hypothesis at the 0.05 level and conclude that there is evidence of an association between **Gender** and **Survived**. Cramer's V of -0.5287 indicates that the association detected with the chi-square test is relatively strong.

Fisher's Exact Test	
Cell (1,1) Frequency (F)	127
Left-sided Pr <= F	7.351E-83
Right-sided Pr >= F	1.0000
Table Probability (P)	6.705E-83
Two-sided Pr <= P	7.918E-83

Exact tests are often useful where asymptotic distributional assumptions are not met. The usual guidelines for the asymptotic chi-square test are generally 20-25 total observations for a 2 × 2 table, with 80% of the table cells having counts greater than 5. Fisher's Exact Test is provided by PROC FREQ when tests of association are requested for 2*2 tables. Otherwise, the exact test must be requested using an EXACT statement. The two-sided *p*-value of 7.918 * 10^{-83} is exceedingly small and statistically significant.

Estimates of the Relative Risk (Row1/Row2)			
Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.0884	0.0677	0.1155
Cohort (Col1 Risk)	0.3369	0.2894	0.3921
Cohort (Col2 Risk)	3.8090	3.2797	4.4239

The Relative Risk table shows another measure of strength of association.

The odds ratio is shown in the first row of the table, along with the 95% confidence limits. The odds ratio can be interpreted as the odds of a top row (**`female`**, in this case) value to be in the left column (**`Died`**), compared with the same odds in the bottom row (**`male`**). The value of 0.0884 says that a female has about 9% of the odds of dying, compared with a male. This is equivalent to saying that a male has about 9% of the odds of surviving, compared with a female.

Cohort estimates for each column are interpreted as probability ratios, rather than odds ratios. You get a choice of assessing probabilities of the left column (Col1) or the right column (Col2). For example, the Col1 risk shows the ratio of the probabilities of females to males being in the left column (27.25/80.90=0.3369).

If is often easier to report odds ratios by first transforming the decimal value to a percent difference value. The formula for doing that is (OR-1) * 100. In the example, you have (0.0884-1)*100=-91.16%. In other words, males have 91.16 percent lower odds of surviving compared with females.

The 95% odds ratio confidence interval goes from 0.0677 to 0.1155. That interval does not include 1. This confirms the statistically significant (at alpha=0.05) result of the Pearson chi-square test of association. A confidence interval that included the value 1 (equality of odds) would be a non-significant result.

Table of Class by Survived

Class	Survived		
Frequency Expected Cell Chi-Square Row Pct	**Died**	**Survived**	**Total**
1	123 199.62 29.411 38.08	200 123.38 47.587 61.92	323
2	158 171.19 1.0169 57.04	119 105.81 1.6453 42.96	277
3	528 438.18 18.411 74.47	181 270.82 29.788 25.53	709
Total	809	500	1309

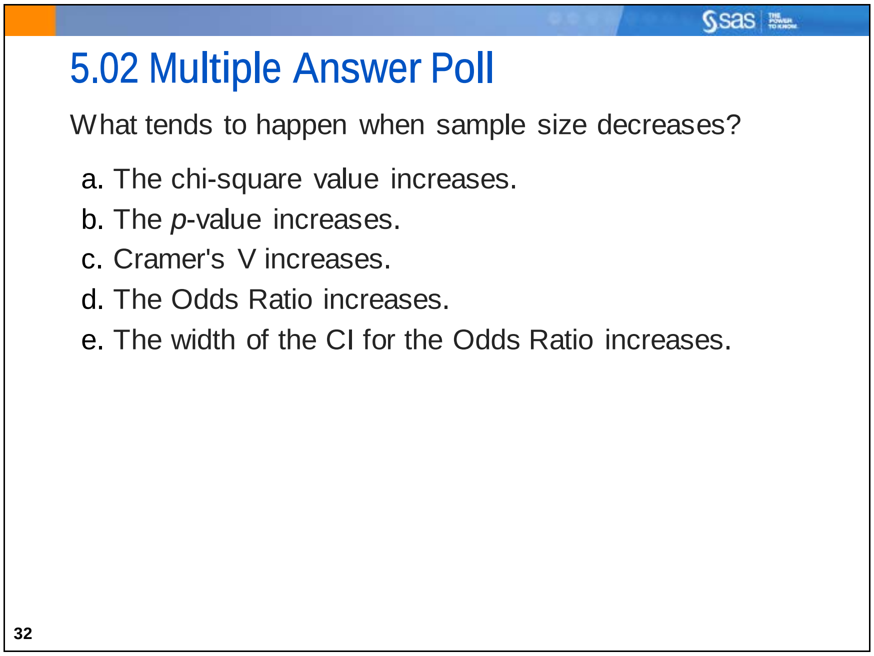
Statistics for Table of Class by Survived

Statistic	DF	Value	Prob
Chi-Square	2	127.8592	<.0001
Likelihood Ratio Chi-Square	2	127.7655	<.0001
Mantel-Haenszel Chi-Square	1	127.7093	<.0001
Phi Coefficient		0.3125	
Contingency Coefficient		0.2983	
Cramer's V		0.3125	

Sample Size=1309

There also seems to be an association between **Class** and **Survival** (Chi-Square(2 df)=127.8592, p<.0001). Cramer’s V for that association is 0.3125.

Mantel-Haenszel chi-square is a test of an ordinal association between **Class** and **Survival**.

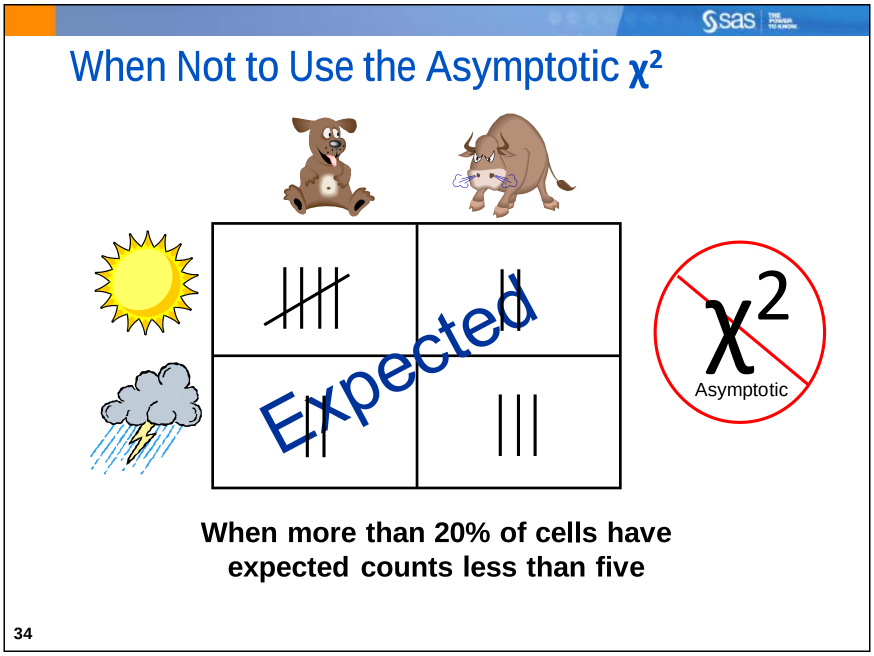


5.02 Multiple Answer Poll

What tends to happen when sample size decreases?

a. The chi-square value increases.
b. The *p*-value increases.
c. Cramer's V increases.
d. The Odds Ratio increases.
e. The width of the CI for the Odds Ratio increases.

32

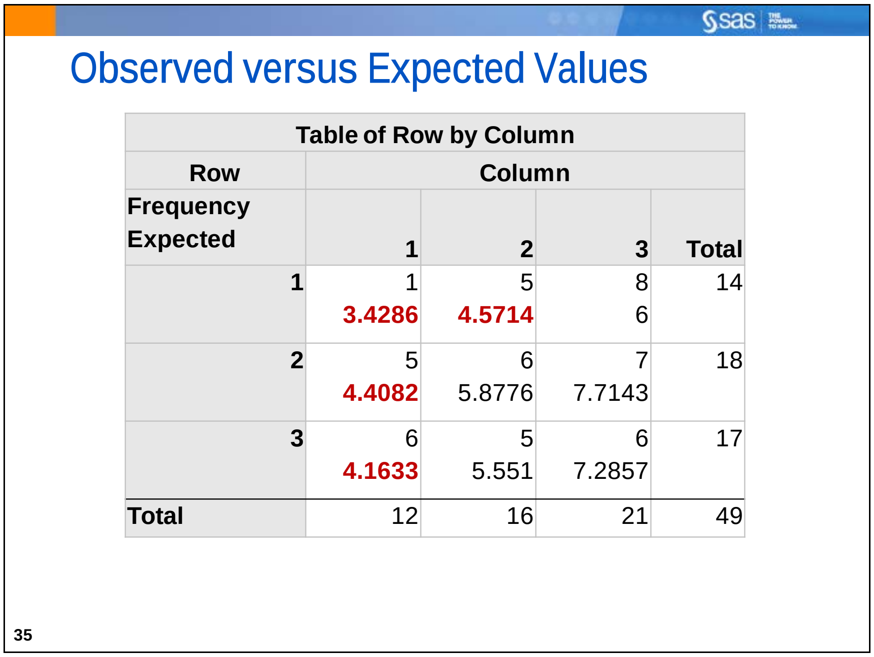


When Not to Use the Asymptotic χ^2

When more than 20% of cells have expected counts less than five

34

There are times when the chi-square test might not be appropriate. In fact, when more than 20% of the cells have expected cell frequencies of less than 5, the chi-square test might not be valid. This is because the *p*-values are based on the assumption that the test statistic follows a particular distribution when the sample size is sufficiently large. Therefore, when the sample sizes are small, the asymptotic (large sample) *p*-values might not be valid.



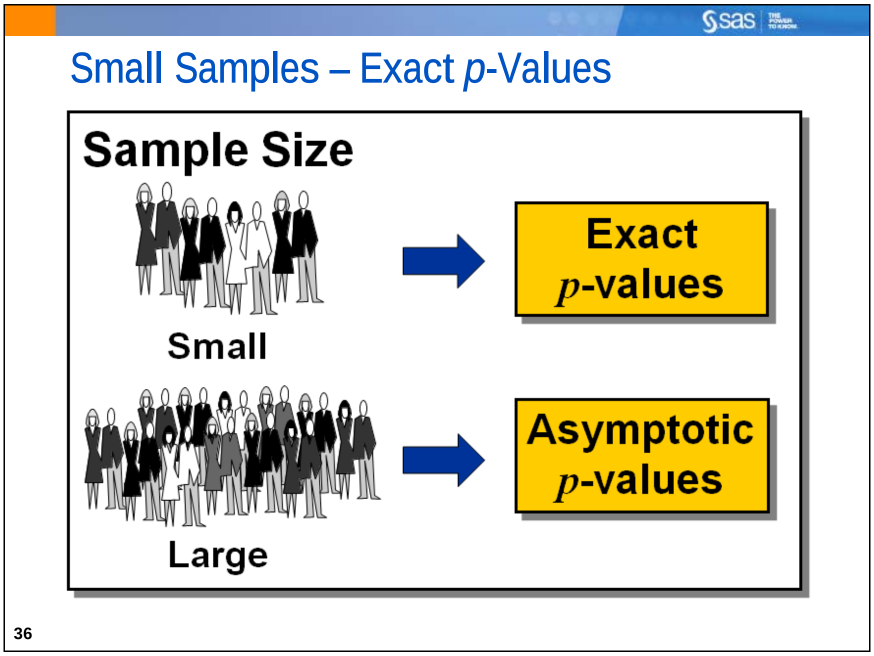
Observed versus Expected Values

Table of Row by Column

Row	Column			
Frequency Expected	1	2	3	Total
1	1 **3.4286**	5 **4.5714**	8 6	14
2	5 **4.4082**	6 5.8776	7 7.7143	18
3	6 **4.1633**	5 5.551	6 7.2857	17
Total	12	16	21	49

35

The criterion for the chi-square test is based on the expected values, not the observed values. In the slide above, 1 out of 9, or 11% of the cells, has an ***observed*** count less than 5. However, 4 out of 9, or 44%, of the cells have ***expected*** counts less than 5. Therefore, the chi-square test might not be valid.



Small Samples – Exact *p*-Values

Sample Size

Small

Large

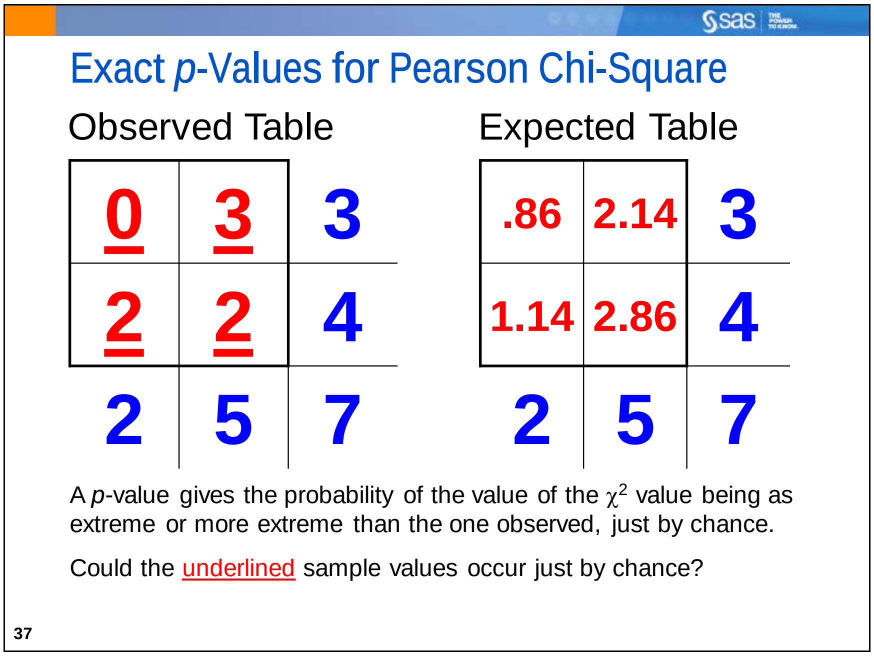
Exact *p*-values

Asymptotic *p*-values

36

The EXACT statement provides exact *p*-values for many tests in the FREQ procedure. Exact *p*-values are useful when the sample size is small. In this case, the asymptotic *p*-values might not be useful.

However, large data sets (in terms of sample size, number of rows, and number of columns) can require a prohibitive amount of time and memory for computing exact *p*-values. For large data sets, consider whether exact *p*-values are needed or whether asymptotic *p*-values might be quite close to the exact *p*-values.



Exact *p*-Values for Pearson Chi-Square

Observed Table

0	3	3
2	2	4
2	5	7

Expected Table

.86	2.14	3
1.14	2.86	4
2	5	7

A *p*-value gives the probability of the value of the χ^2 value being as extreme or more extreme than the one observed, just by chance.

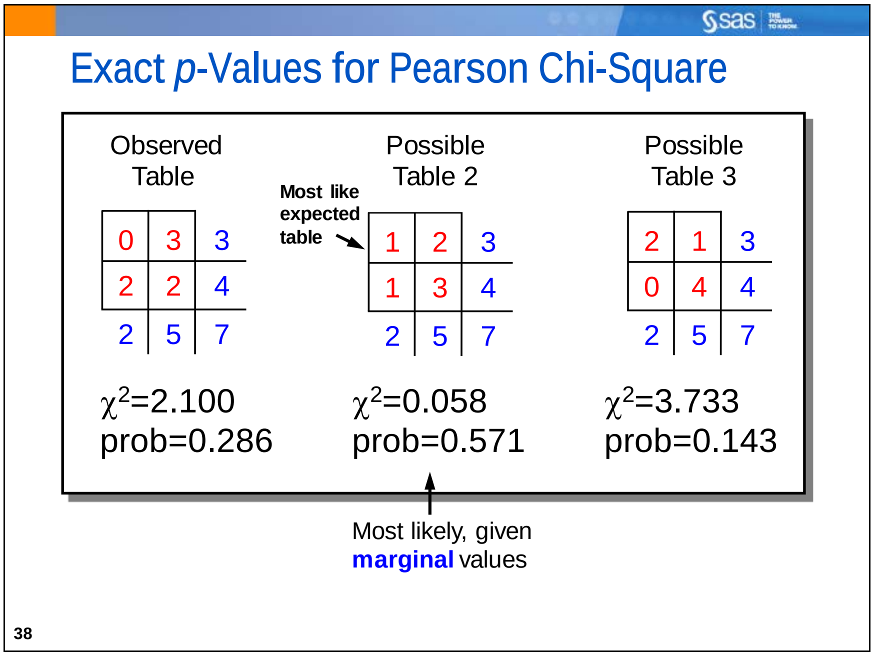
Could the underlined sample values occur just by chance?

37

Consider the table at left above. With such a small sample size, the asymptotic *p*-values would not be valid, because the accuracy of those *p*-values depends on large enough expected values in all cells.

Exact *p*-values reflect the probability of observing a table with at least as much evidence of an association as the one actually observed, given there is no association between the variables.

✎ Recall that expected count within each cell is calculated by expected count=(R*C)/T.



Exact *p*-Values for Pearson Chi-Square

Observed Table

0	3	3
2	2	4
2	5	7

χ^2=2.100
prob=0.286

Possible Table 2

Most like expected table

1	2	3
1	3	4
2	5	7

χ^2=0.058
prob=0.571

Possible Table 3

2	1	3
0	4	4
2	5	7

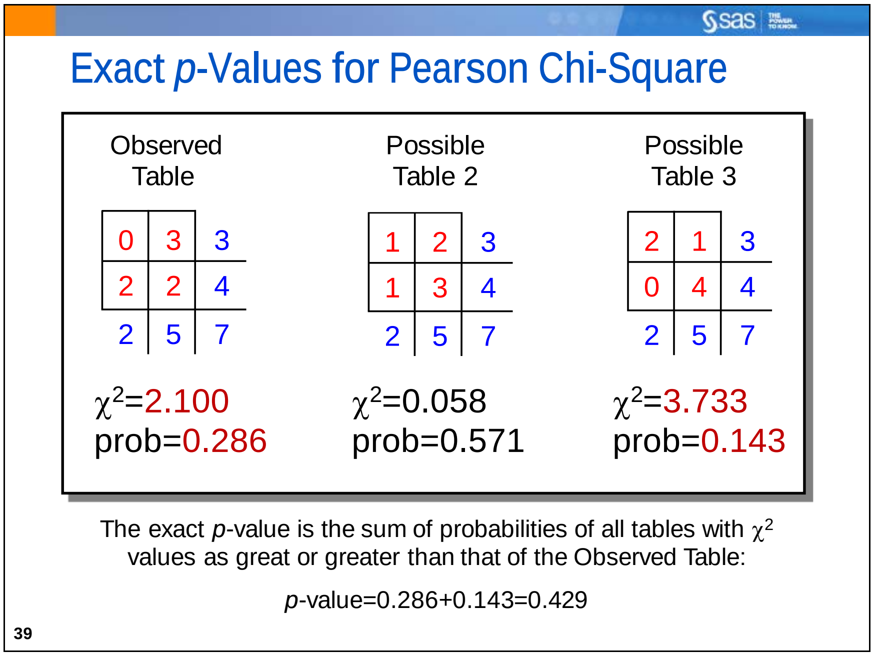
χ^2=3.733
prob=0.143

Most likely, given **marginal** values

38

A key assumption behind the computation of exact *p*-values is that the column totals and row totals are fixed. There are only three possible tables, including the observed table, given the fixed marginal totals.

Possible Table 2 is most like the Expected Table of the previous slide. So, the probability (0.571) that its cell values would occur in a table, given these row and column total values, is greatest of any possible table that could occur by chance.



Exact *p*-Values for Pearson Chi-Square

Observed Table

0	3	3
2	2	4
2	5	7

χ^2=2.100
prob=0.286

Possible Table 2

1	2	3
1	3	4
2	5	7

χ^2=0.058
prob=0.571

Possible Table 3

2	1	3
0	4	4
2	5	7

χ^2=3.733
prob=0.143

The exact *p*-value is the sum of probabilities of all tables with χ^2 values as great or greater than that of the Observed Table:

p-value=0.286+0.143=0.429

39

To compute an exact *p*-value for this example, examine the chi-square value for each table and the probability that the table should occur by chance if the null hypothesis of no association were true. (The probabilities add up to 1.)

Remember the definition of a *p*-value. It is the probability, if the null hypothesis is true, that you would obtain a sample statistic **as great as or greater than** the one you observed just by chance.

In this example, this means the probability of obtaining a table with a χ^2 value as great as or greater than the 2.100 for the Observed Table. The probability associated with every table with a χ^2 value of 2.100 or higher would be summed to compute the two-sided exact *p*-value.

The exact *p*-value would be 0.286 (Observed Table)+0.143 (Possible Table 3)=0.429. This means you have a 42.9% chance of obtaining a table with at least as much of an association as the observed table simply by random chance.

Fisher's Exact *p*-Values for the Pearson Chi-Square Test

Example: Invoke PROC FREQ and produce exact *p*-values for the Pearson chi-square test. Use the **sasuser.exact** data set, which has the data from the previous example.

```
/*st105d03.sas*/
ods graphics off;
proc freq data=sasuser.exact;
   tables A*B / chisq expected cellchi2 nocol nopercent;
   title "Exact P-Values";
run;
ods graphics on;
```

The frequency table is shown below.

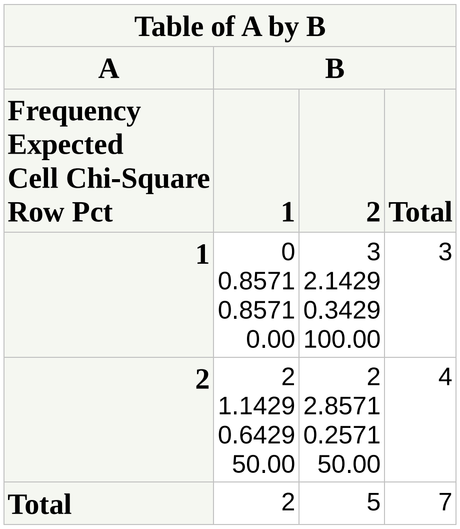
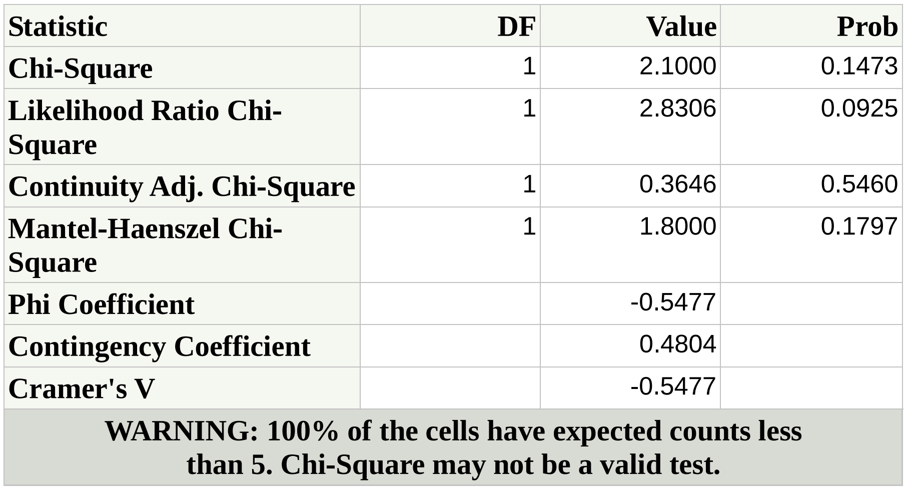
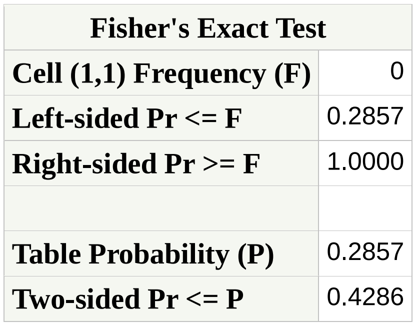


Table of A by B			
A	**B**		
Frequency **Expected** **Cell Chi-Square** **Row Pct**	**1**	**2**	**Total**
1	0 0.8571 0.8571 0.00	3 2.1429 0.3429 100.00	3
2	2 1.1429 0.6429 50.00	2 2.8571 0.2571 50.00	4
Total	2	5	7



Statistic	DF	Value	Prob
Chi-Square	1	2.1000	0.1473
Likelihood Ratio Chi-Square	1	2.8306	0.0925
Continuity Adj. Chi-Square	1	0.3646	0.5460
Mantel-Haenszel Chi-Square	1	1.8000	0.1797
Phi Coefficient		-0.5477	
Contingency Coefficient		0.4804	
Cramer's V		-0.5477	
WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

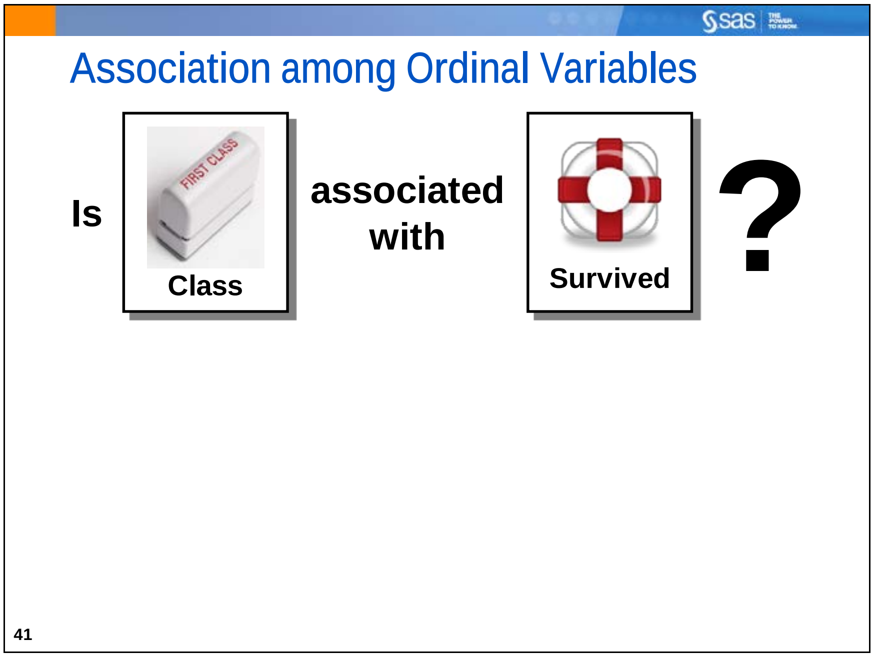
The warning tells you that you should not trust the reported *p*-value in this table.



Fisher's Exact Test	
Cell (1,1) Frequency (F)	0
Left-sided Pr <= F	0.2857
Right-sided Pr >= F	1.0000
Table Probability (P)	0.2857
Two-sided Pr <= P	0.4286

The Two-sided Pr <= P value is the one you will report. Notice the difference between the exact *p*-value (0.4286) and the asymptotic *p*-value (0.1473) in the Pearson chi-square test table. The exact *p*-values are larger. Exact tests tend to be more conservative than asymptotic tests.

For tables larger than 2*2, an EXACT statement must be submitted to obtain exact *p*-values. For large tables, this can take a long time and use a great deal of computational resources.

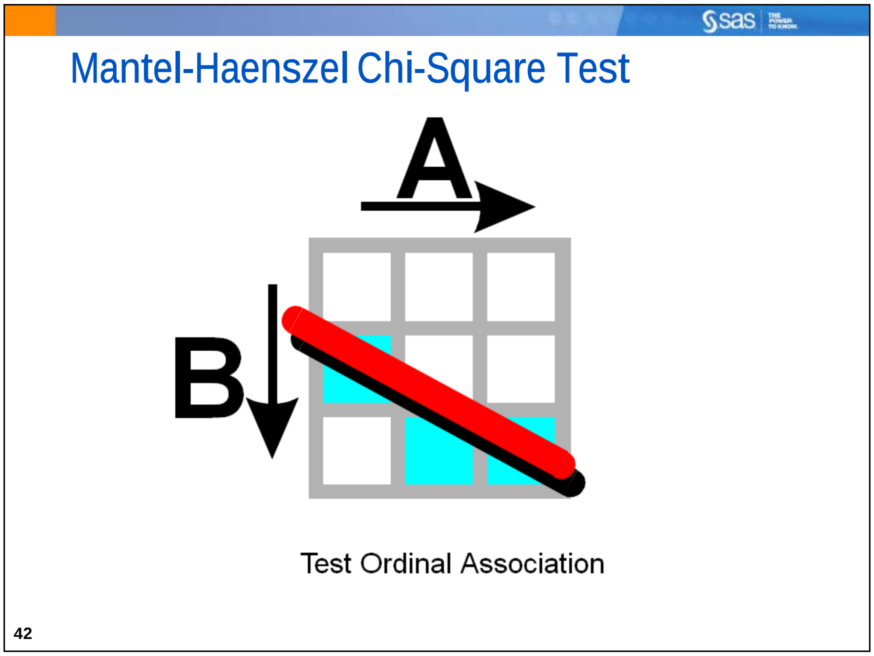


Association among Ordinal Variables

Is Class associated with Survived ?

41

You already saw that **Survived** and **Class** have a significant general association. Another question that you can ask is whether **Survived** and **Class** have a significant ordinal association. The appropriate test for ordinal associations is the Mantel-Haenszel chi-square test.



Mantel-Haenszel Chi-Square Test

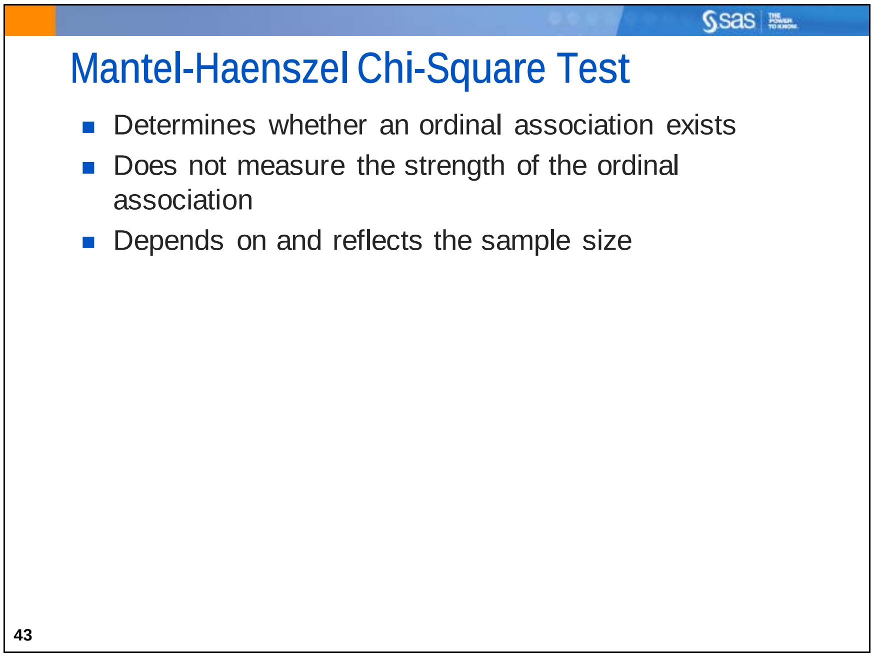
Test Ordinal Association

42

The Mantel-Haenszel chi-square test is particularly sensitive to ordinal associations. An *ordinal association* implies that as one variable increases, the other variable tends to increase, or decrease. For the test results to be meaningful when there are variables with more than two levels, the levels must be in a logical order.

Null hypothesis: There is no ordinal association between the row and column variables.

Alternative hypothesis: There is an ordinal association between the row and column variables.



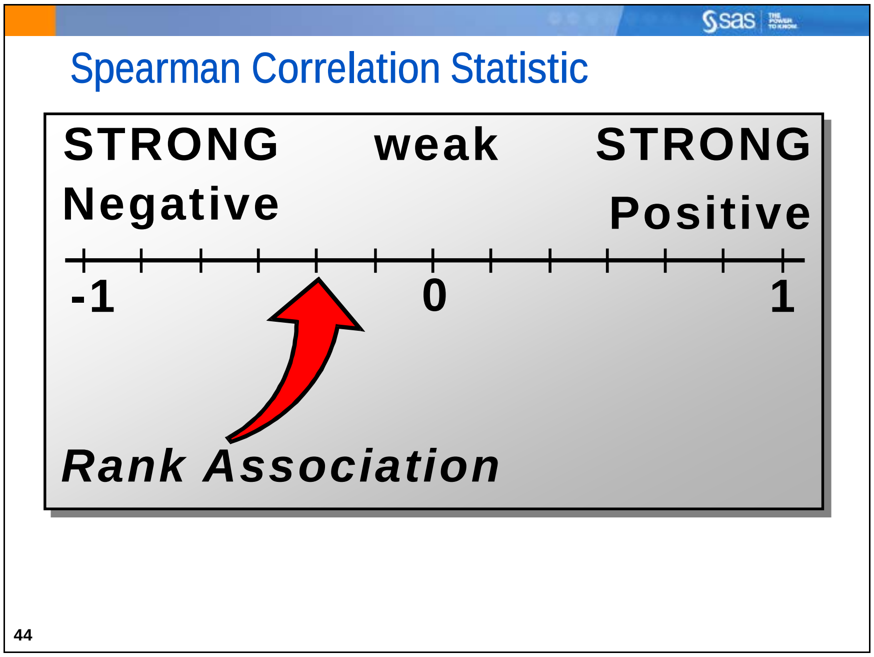
Mantel-Haenszel Chi-Square Test

- Determines whether an ordinal association exists
- Does not measure the strength of the ordinal association
- Depends on and reflects the sample size

43

The Mantel-Haenszel chi-square statistic is more powerful than the general association chi-square statistic for detecting an ordinal association. The reasons are that

- all of the Mantel-Haenszel statistic’s power is concentrated toward that objective
- the power of the general association statistic is dispersed over a greater number of alternatives.



Spearman Correlation Statistic

STRONG Negative

weak

STRONG Positive

-1

0

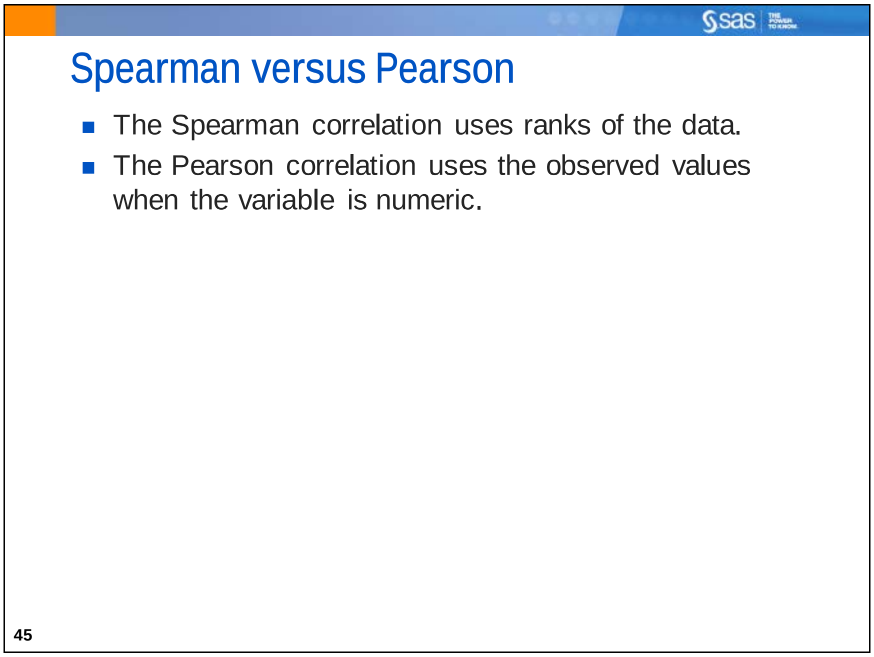
1

Rank Association

44

To measure the strength of the ordinal association, you can use the Spearman correlation statistic. This statistic

- has a range between –1 and 1
- has values close to 1 if there is a relatively high degree of positive correlation
- has values close to –1 if there is a relatively high degree of negative correlation
- is appropriate only if both variables are ordinal scaled and the values are in a logical order.



Spearman versus Pearson

- The Spearman correlation uses ranks of the data.
- The Pearson correlation uses the observed values when the variable is numeric.

45

The Spearman statistic can be interpreted as the Pearson correlation between the ranks on variable X and the ranks on variable Y.

For character values, SAS assigns, by default, a 1 to column 1, a 2 to column 2, and so on. You can change the default with the SCORES= option in the TABLES statement.

Detecting Ordinal Associations

Example: Use PROC FREQ to test whether an ordinal association exists between **Survived** and **Class**.

```
/*st105d04.sas*/
ods graphics off;
proc freq data=sasuser.Titanic;
   tables Class*Survived / chisq measures cl;
   format Survived survfmt.;
   title1 'Ordinal Association between CLASS and SURVIVAL?';
run;
ods graphics on;
```

Selected TABLES statement options:

CHISQ produces the Pearson chi-square, the likelihood-ratio chi-square, and the Mantel-Haenszel chi-square. It also produces measures of association based on chi-square such as the phi coefficient, the contingency coefficient, and Cramer’s V.

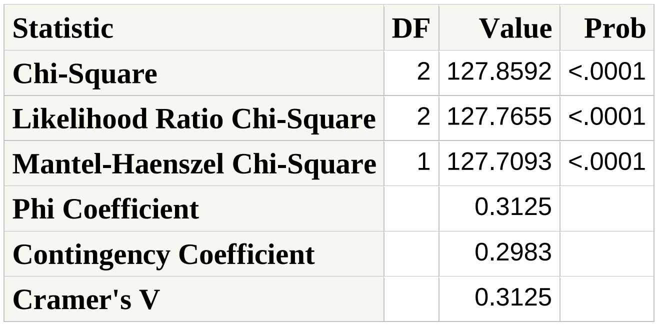
MEASURES produces the Spearman correlation statistic along with other measures of association.

CL produces confidence bounds for the MEASURES statistics.

The crosstabulation is shown below.

Table of Class by Survived			
Class	**Survived**		
Frequency Percent Row Pct Col Pct	**Died**	**Survived**	**Total**
1	123 9.40 38.08 15.20	200 15.28 61.92 40.00	323 24.68
2	158 12.07 57.04 19.53	119 9.09 42.96 23.80	277 21.16
3	528 40.34 74.47 65.27	181 13.83 25.53 36.20	709 54.16
Total	809 61.80	500 38.20	1309 100.00

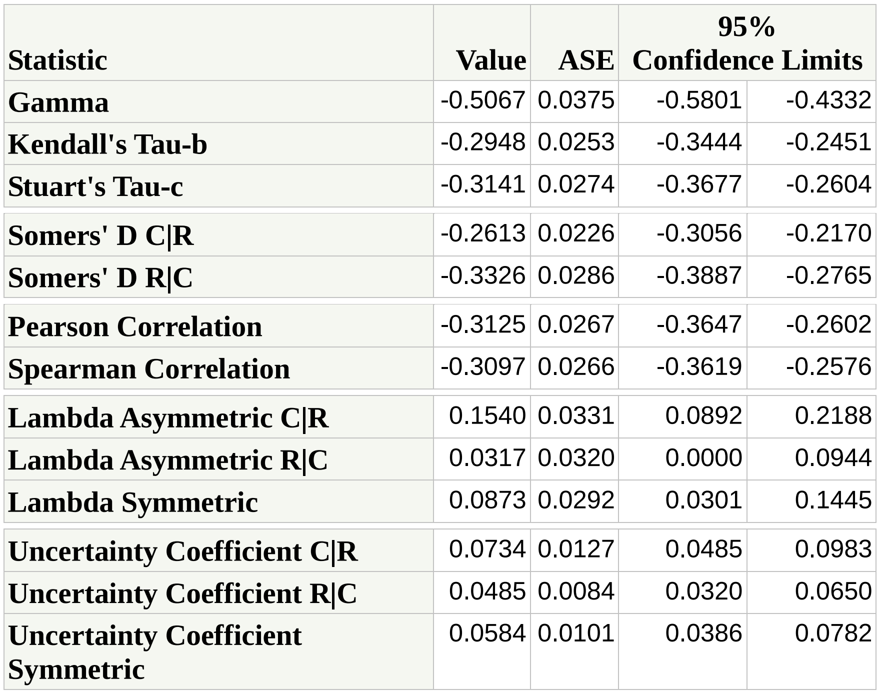
The results of the Mantel-Haenszel chi-square test are shown below.



Statistic	DF	Value	Prob
Chi-Square	2	127.8592	<.0001
Likelihood Ratio Chi-Square	2	127.7655	<.0001
Mantel-Haenszel Chi-Square	1	127.7093	<.0001
Phi Coefficient		0.3125	
Contingency Coefficient		0.2983	
Cramer's V		0.3125	

Because the *p*-value of the Mantel-Haenszel chi-square is <.0001, you can conclude at the 0.05 significance level that there is evidence of an ordinal association between **Survived** and **Class**.

The Spearman correlation statistic and the 95% confidence bounds are shown below.



Statistic	Value	ASE	95% Confidence Limits	
Gamma	-0.5067	0.0375	-0.5801	-0.4332
Kendall's Tau-b	-0.2948	0.0253	-0.3444	-0.2451
Stuart's Tau-c	-0.3141	0.0274	-0.3677	-0.2604
Somers' D C|R	-0.2613	0.0226	-0.3056	-0.2170
Somers' D R|C	-0.3326	0.0286	-0.3887	-0.2765
Pearson Correlation	-0.3125	0.0267	-0.3647	-0.2602
Spearman Correlation	-0.3097	0.0266	-0.3619	-0.2576
Lambda Asymmetric C|R	0.1540	0.0331	0.0892	0.2188
Lambda Asymmetric R|C	0.0317	0.0320	0.0000	0.0944
Lambda Symmetric	0.0873	0.0292	0.0301	0.1445
Uncertainty Coefficient C|R	0.0734	0.0127	0.0485	0.0983
Uncertainty Coefficient R|C	0.0485	0.0084	0.0320	0.0650
Uncertainty Coefficient Symmetric	0.0584	0.0101	0.0386	0.0782

The Spearman Correlation (-0.3097) indicates that there is a moderate, negative ordinal relationship between **Class** and **Survived** (that is, as **Class** levels increase, **Survived** tends to decrease).

The ASE is the asymptotic standard error (0.0266), which is an appropriate measure of the standard error for larger samples.

Because the 95% confidence interval (-0.3619, -0.2576) for the Spearman correlation statistic does not contain 0, the relationship is significant at the 0.05 significance level.

The confidence bounds are valid only if your sample size is large. A general guideline is to have a sample size of at least 25 for each degree of freedom in the Pearson chi-square statistic.

Exercises

1. Performing Tests and Measures of Association

An insurance company wants to relate the safety of vehicles to several other variables. A score is given to each vehicle model, using the frequency of insurance claims as a basis. The data are in the **sasuser.safety** data set.

The variables in the data set are as follows:

Unsafe dichotomized safety score (`1`=`Below Average`, `0`=`Average or Above`)

Type type of car (`Large`, `Medium`, `Small`, `Sport/Utility`, `Sports`)

Region manufacturing region (`Asia`, `N America`)

Weight weight in 1000s of pounds

Size trichotomized version of **Type** (`1`=`Small or Sports`, `2`=`Medium`, `3`=`Large or Sport/Utility`).

a. Invoke the FREQ procedure and create one-way frequency tables for the categorical variables.

1) What is the measurement scale of each variable?

<u>Variable</u> <u>Measurement Scale</u>

Unsafe ________________

Type ________________

Region ________________

Weight ________________

Size ________________

2) What is the proportion of cars made in North America?

3) For the variables **Unsafe**, **Size**, **Region**, and **Type**, are there any unusual data values that warrant further investigation?

b. Use PROC FREQ to examine the crosstabulation of the variables **Region** by **Unsafe**. Generate a temporary format to clearly identify the values of **Unsafe**. Along with the default output, generate the expected frequencies, the chi-square test of association, and the odds ratio.

Use the following code for the format:

```
proc format;
   value safefmt 0='Average or Above'
                 1='Below Average';
run;
```

1) For the cars made in Asia, what percentage had a below-average safety score?

2) For the cars with an average or above safety score, what percentage was made in North America?

3) Do you see a statistically significant (at the 0.05 level) association between **Region** and **Unsafe**?

4) What does the odds ratio compare and what does this one say about the difference in odds between Asian and North American cars?

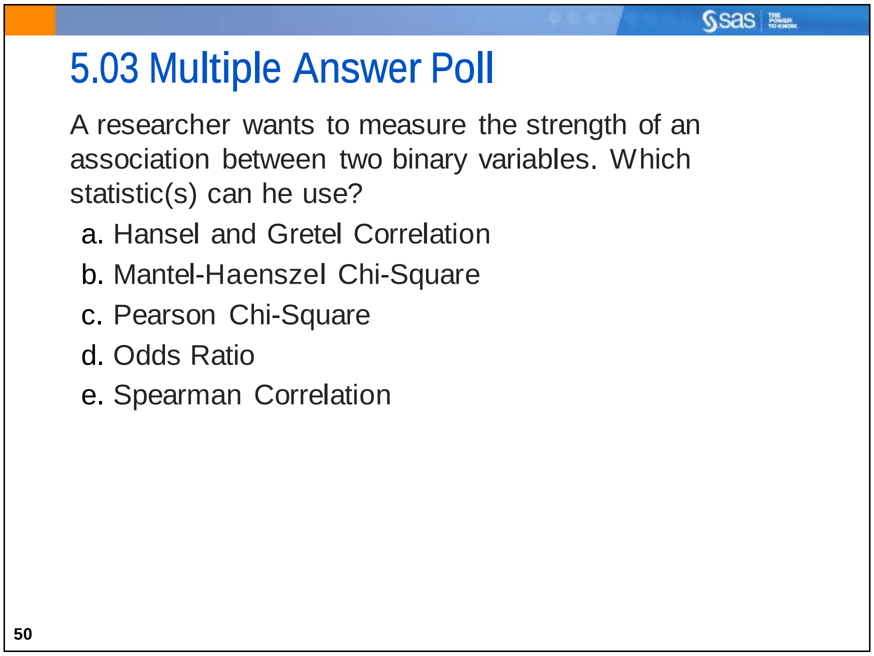
c. Use the variable named **Size**. Examine the ordinal association between **Size** and **Unsafe**. Use PROC FREQ.

1) What statistic should you use to detect an ordinal association between **Size** and **Unsafe**?

2) Do you reject or fail to reject the null hypothesis at the 0.05 level?

3) What is the strength of the ordinal association between **Size** and **Unsafe**?

4) What is the 95% confidence interval around that statistic?



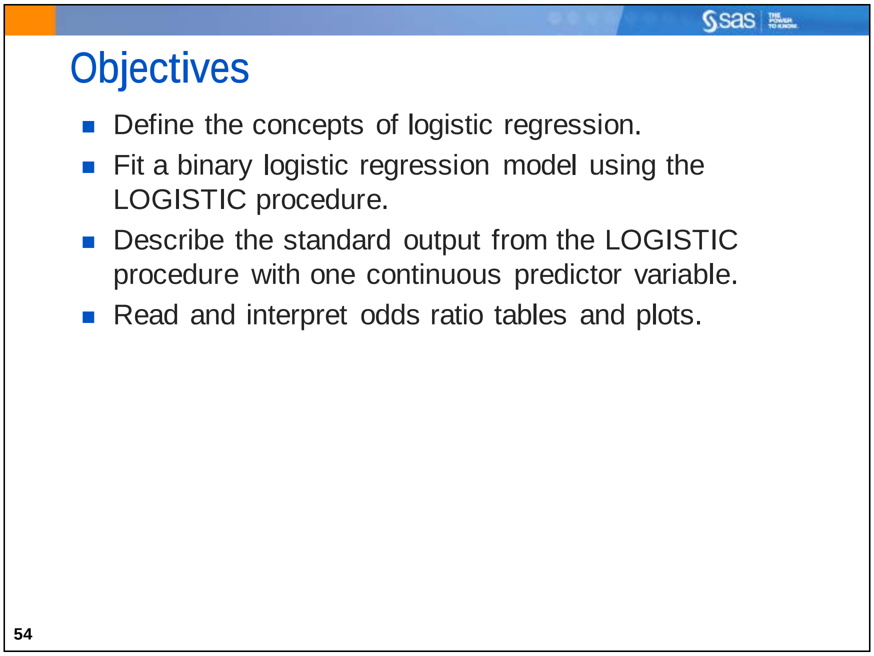
5.03 Multiple Answer Poll

A researcher wants to measure the strength of an association between two binary variables. Which statistic(s) can he use?

a. Hansel and Gretel Correlation
b. Mantel-Haenszel Chi-Square
c. Pearson Chi-Square
d. Odds Ratio
e. Spearman Correlation

50

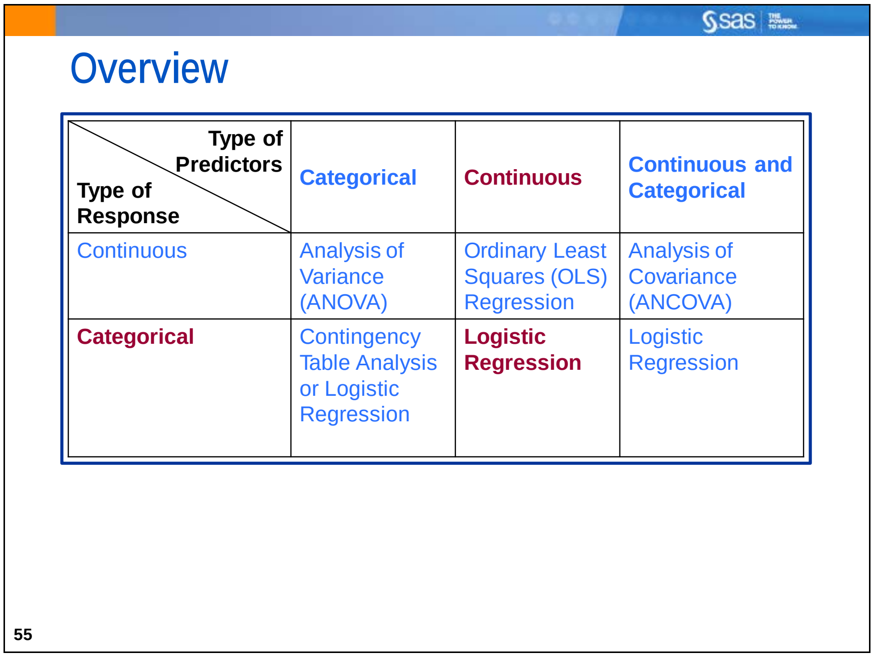
5.3 Introduction to Logistic Regression



Objectives

- Define the concepts of logistic regression.
- Fit a binary logistic regression model using the LOGISTIC procedure.
- Describe the standard output from the LOGISTIC procedure with one continuous predictor variable.
- Read and interpret odds ratio tables and plots.

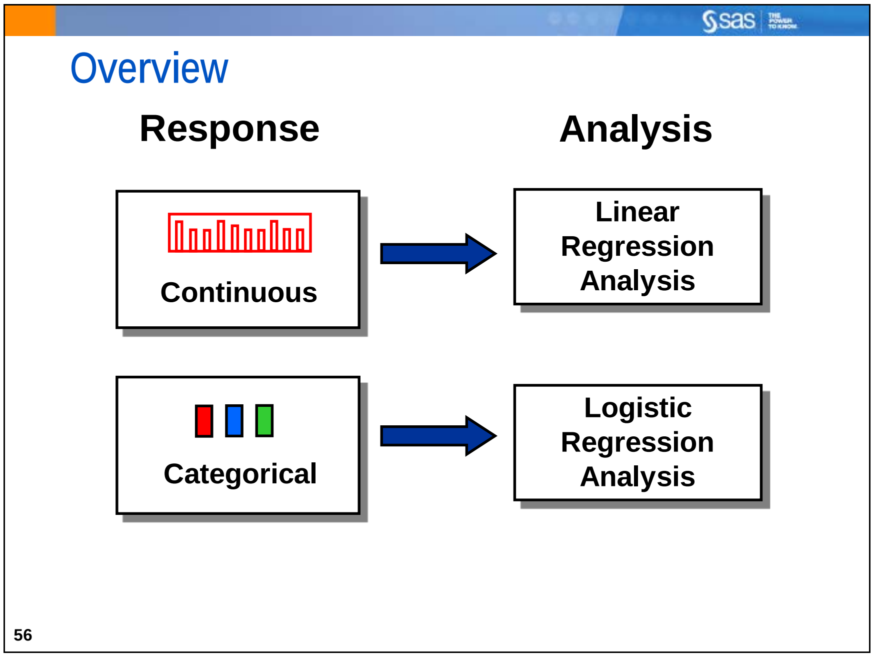
54



Overview

Type of Predictors Type of Response	**Categorical**	**Continuous**	**Continuous and Categorical**
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	**Logistic Regression**	Logistic Regression

55



Overview

Response

Analysis

Continuous

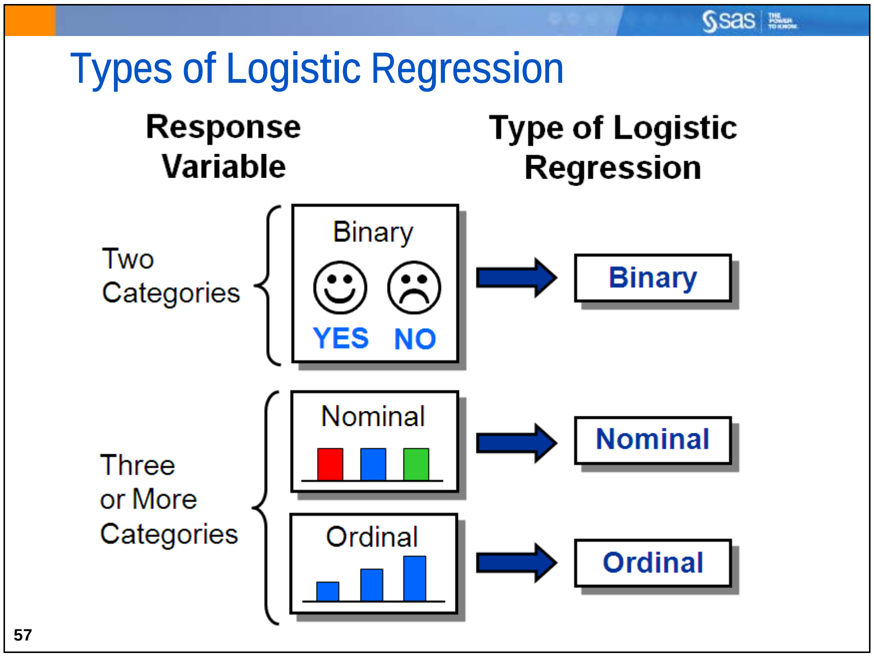
Linear Regression Analysis

Categorical

Logistic Regression Analysis

56

Regression analysis enables you to characterize the relationship between a response variable and one or more predictor variables. In linear regression, the response variable is continuous. In *logistic regression*, the response variable is categorical.



Types of Logistic Regression

Response Variable

Type of Logistic Regression

Two Categories

Binary

☺ ☹

YES NO

→ **Binary**

Three or More Categories

Nominal → **Nominal**

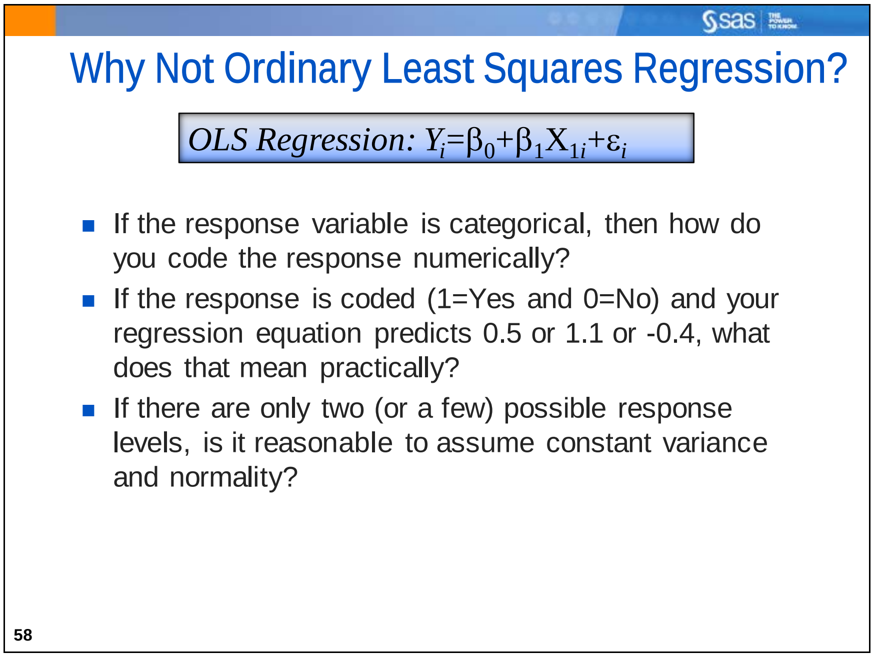
Ordinal → **Ordinal**

57

If the response variable is dichotomous (two categories), the appropriate logistic regression model is binary logistic regression.

If you have more than two categories (levels) within the response variable, then there are two possible logistic regression models:

1. If the response variable is nominal, you fit a nominal logistic regression model.
2. If the response variable is ordinal, you fit an ordinal logistic regression model.



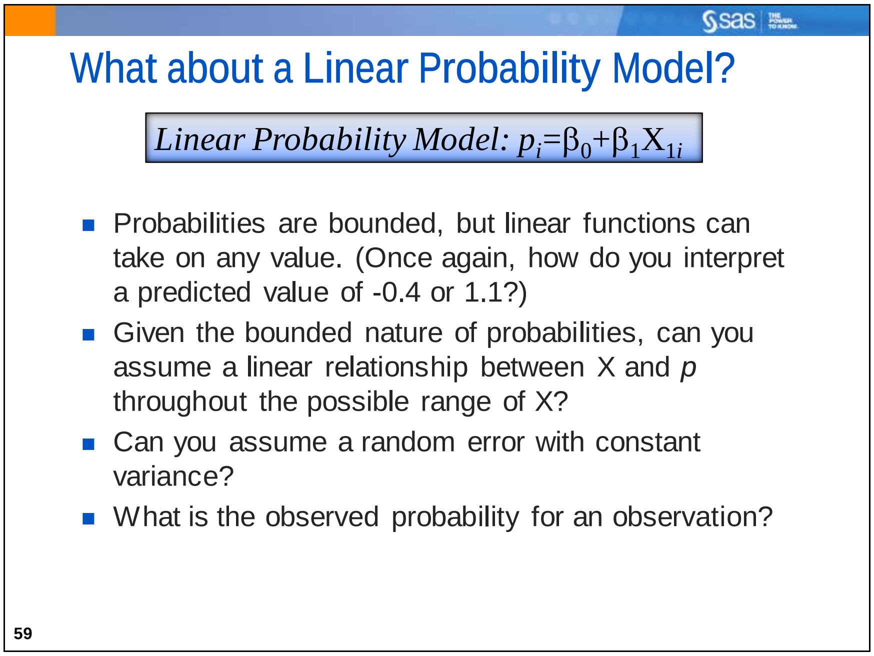
Why Not Ordinary Least Squares Regression?

OLS Regression: $Y_i=\beta_0+\beta_1 X_{1i}+\varepsilon_i$

- If the response variable is categorical, then how do you code the response numerically?
- If the response is coded (1=Yes and 0=No) and your regression equation predicts 0.5 or 1.1 or -0.4, what does that mean practically?
- If there are only two (or a few) possible response levels, is it reasonable to assume constant variance and normality?

58

You might be tempted to analyze a regression model with a binary response variable using PROC REG. However, there are problems with that. Besides the arbitrary nature of the coding, there is the problem that the predicted values will take on values that have no intrinsic meaning, with regard to your response variable. There is also the mathematical inconvenience of not being able to assume normality and constant variance when the response variable has only two values.



What about a Linear Probability Model?

Linear Probability Model: $p_i=\beta_0+\beta_1X_{1i}$

- Probabilities are bounded, but linear functions can take on any value. (Once again, how do you interpret a predicted value of -0.4 or 1.1?)
- Given the bounded nature of probabilities, can you assume a linear relationship between X and *p* throughout the possible range of X?
- Can you assume a random error with constant variance?
- What is the observed probability for an observation?

59

Instead of modeling the zeros and ones directly, another way of thinking about modeling a binary variable is to model the probability of either the zero or the one. If you can model the probability of the one (called *p*), then you also modeled the probability of the zero, which would be (1–*p*). Probabilities are truly continuous and so this line of thinking might sound compelling at first.

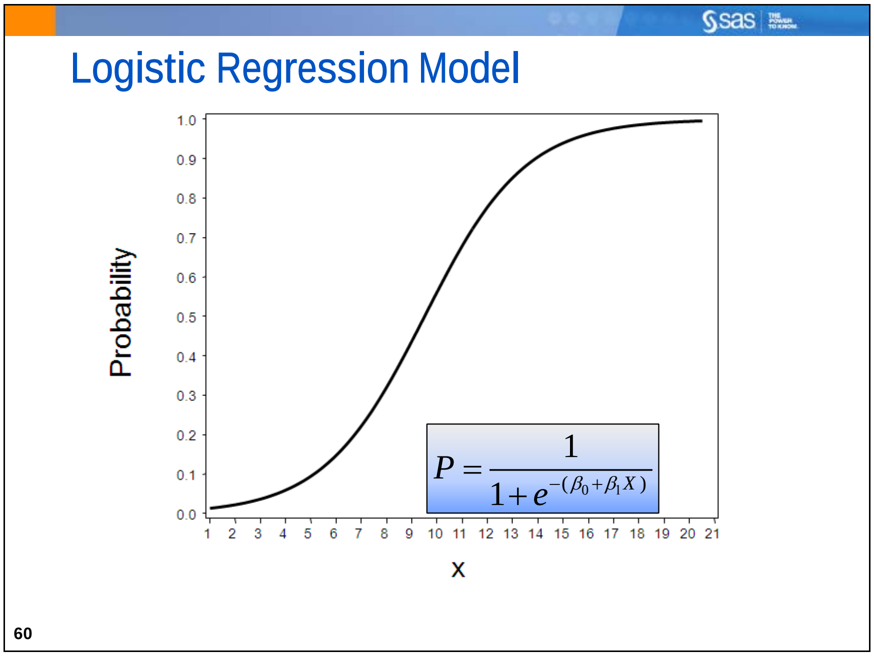
One problem is that the predicted values from a linear model can assume, theoretically, any value. However, probabilities are by definition bounded between 0 and 1.

Another problem is that the relationship between the probability of the outcome and a predictor variable is usually nonlinear rather than linear. In fact, the relationship often resembles an S-shaped curve (a “sigmoidal” relationship).

Probabilities do not have a random normal error associated with them, but rather a binomial error of *p**(1-*p*). That error is greatest at probabilities close to 0.5 and lowest near 0 and 1.

As mentioned above, probabilities have a binomial error of the form $p*(1-p)=(p-p^2)$. Taking the derivative of this expression with respect to *p* yields the expression $1-2*p$. Setting the derivative equal to zero and solving for *p* returns a value of 0.5. This binomial error equation is a downward facing parabola, which means that the greatest value is at 0.5 and lowest values are near 0 and 1.

Finally, there is no such thing as an “observed probability” and therefore least squares methods cannot be used. The response variable is always either 0 or 1 and therefore the probability of the event is either 0% or 100%. This is another reason why it is untenable to assume a normal distribution of error.



Logistic Regression Model

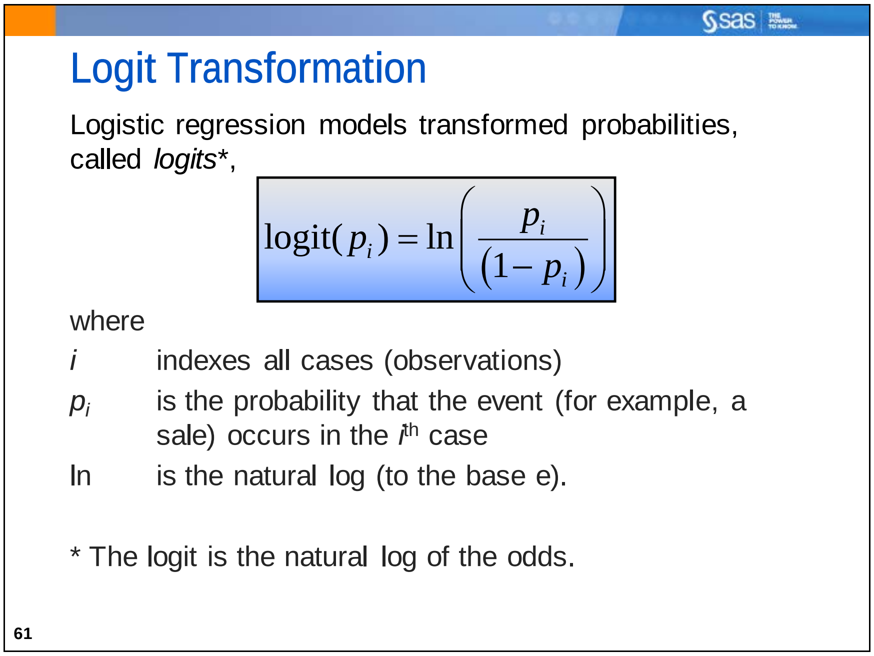
$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

60

This plot shows a model of the relationship between a continuous predictor and the probability of an event or outcome. The linear model clearly does not fit if this is the true relationship between X and the probability. In order to model this relationship directly, you must use a nonlinear function. One such function is displayed. The S-shape of the function is known as a *sigmoid*.

The rate of change parameter of this function (β_1) determines the rate of increase or decrease of the curve. When the parameter value is greater than 0, the probability of the outcome increases as the predictor variable values increase. When the parameter is less than 0, the probability decreases as the predictor variable values increase. As the absolute value of the parameter increases, the curve has a steeper rate of change. When the parameter value is equal to 0, the curve can be represented by a straight, horizontal line that shows an equal probability of the event for everyone.

The β values for this model cannot be computed in PROC REG because this is not a linear model.



Logit Transformation

Logistic regression models transformed probabilities, called *logits**,

$$\boxed{\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)}$$

where

i indexes all cases (observations)

p_i is the probability that the event (for example, a sale) occurs in the *i*th case

ln is the natural log (to the base e).

* The logit is the natural log of the odds.

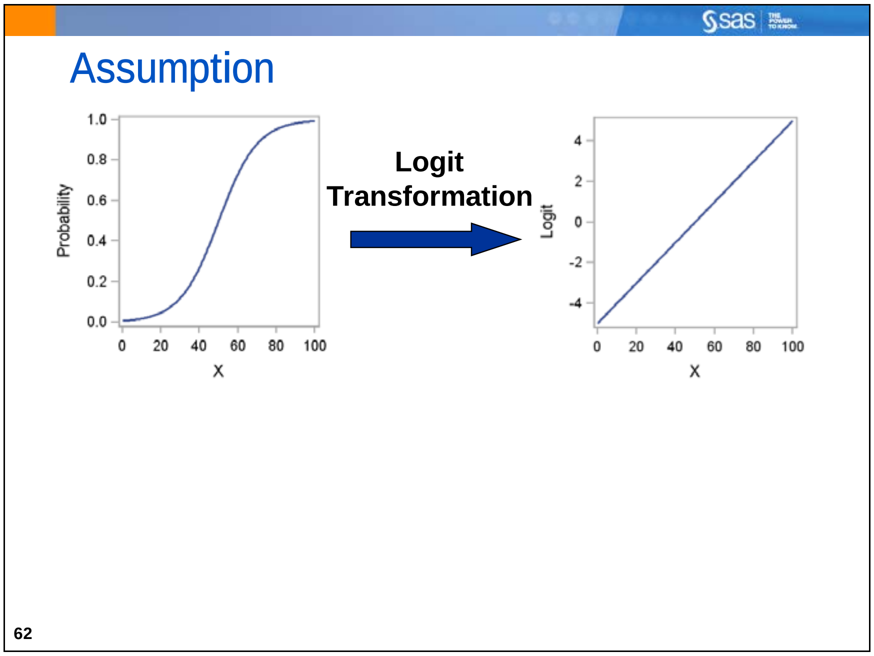
61

A logistic regression model applies a logit transformation to the probabilities. Two of the problems that you saw with modeling the probability directly were that probabilities were bounded between 0 and 1, and that there was not likely a straight line relationship between predictors and probabilities.

First, deal with the problem of restricted range of the probability. What about the range of a logit? As *p* approaches its maximum value of 1, the value ln(*p*/(1–*p*)) goes to infinity. As *p* approaches its minimum value of 0, *p*/(1–*p*) approaches 0. The natural log of something approaching 0 is something that goes to negative infinity. So, the logit has no upper or lower bounds.

If you can model the logit, then simple algebra enables you to model the odds or the probability. The logit transformation ensures that the model generates estimated probabilities between 0 and 1.

The logit is the natural log of the odds. The odds and odds ratios were discussed in a previous section. This relationship between the odds and the logit will become important later in this section.



Assumption

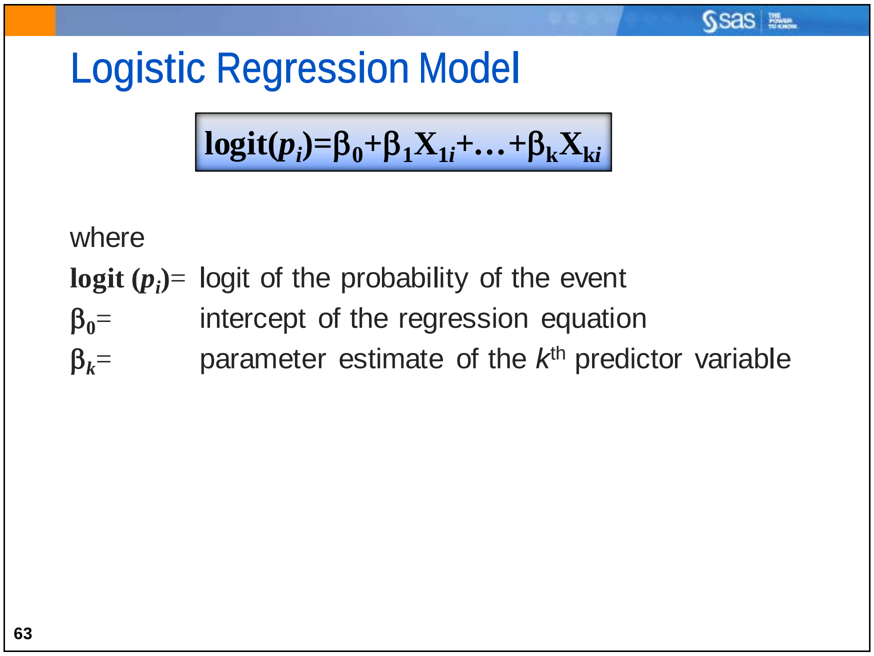
Logit Transformation

62

Assumption in logistic regression: The logit has a linear relationship with the predictor variables.

If the hypothesized nature of the direct relationship between X and *p* are correct, then the logit has a linear relationship with X through the parameters. In other words, a linear function of X, additive in relation to the parameters, can be used to model the logit. In that way, you can indirectly model the probability.

To verify this assumption, it would be useful to plot the logits by the predictor variable. (Logit plots are illustrated in a later section.)



Logistic Regression Model

$$\boxed{\mathrm{logit}(p_i)=\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki}}$$

where

logit (p_i)= logit of the probability of the event

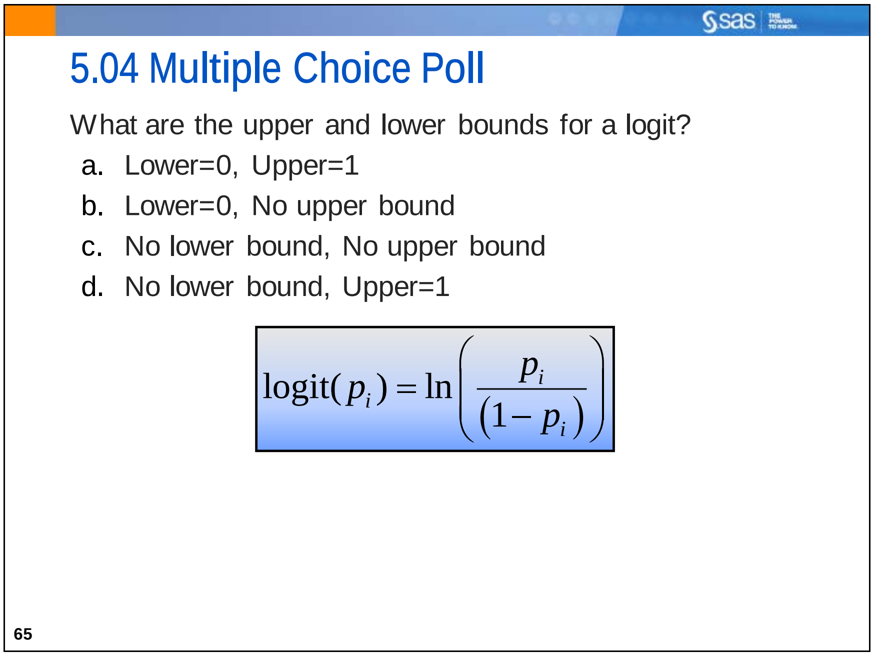
β_0= intercept of the regression equation

β_k= parameter estimate of the k^{th} predictor variable

63

For a binary response variable, the linear logistic model with one predictor variable has the form above.

Unlike linear regression, the logit is not normally distributed and the variance is not constant. Therefore, logistic regression requires a more computationally complex estimation method, named the *Method of Maximum Likelihood*, to estimate the parameters. This method finds the values of the parameters that make the observed data most likely. This is accomplished by maximizing the *likelihood function* that expresses the probability of the observed data as a function of the unknown parameters.



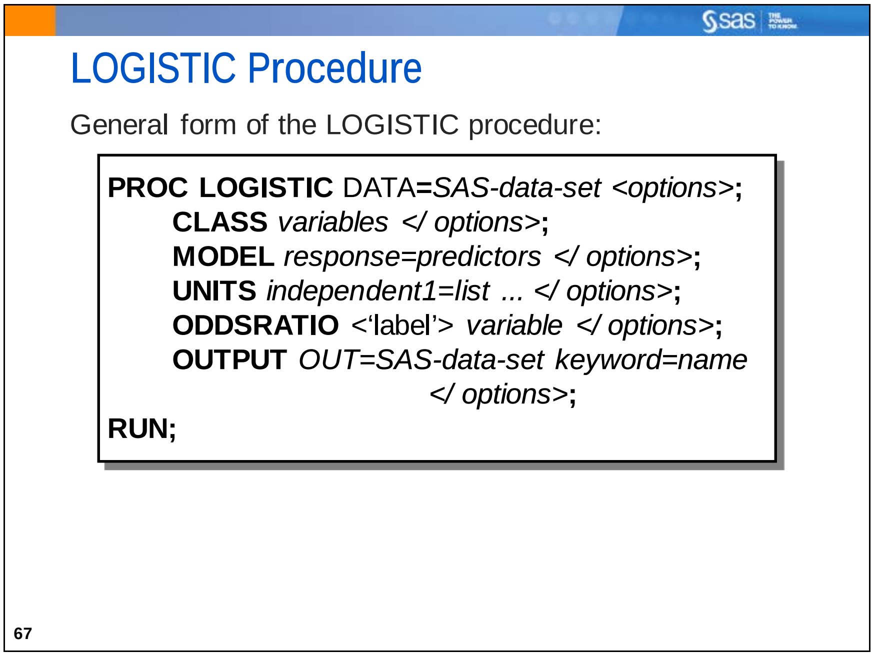
5.04 Multiple Choice Poll

What are the upper and lower bounds for a logit?

a. Lower=0, Upper=1
b. Lower=0, No upper bound
c. No lower bound, No upper bound
d. No lower bound, Upper=1

$$\boxed{\mathrm{logit}(p_i)=\ln\left(\frac{p_i}{(1-p_i)}\right)}$$

65



LOGISTIC Procedure

General form of the LOGISTIC procedure:

```
PROC LOGISTIC DATA=SAS-data-set <options>;
      CLASS variables </ options>;
      MODEL response=predictors </ options>;
      UNITS independent1=list ... </ options>;
      ODDSRATIO <'label'> variable </ options>;
      OUTPUT OUT=SAS-data-set keyword=name
                        </ options>;
RUN;
```

67

Selected LOGISTIC procedure statements:

CLASS names the classification variables to be used in the analysis. The CLASS statement must precede the MODEL statement. By default, these variables will be analyzed using effects coding parameterization. This can be changed with the PARAM= option.

MODEL specifies the response variable and the predictor variables.

OUTPUT creates an output data set containing all the variables from the input data set and any requested statistics.

UNITS enables you to obtain an odds ratio estimate for a specified change in a predictor variable. The unit of change can be a number, standard deviation (SD), or a number times the standard deviation (for example, 2*SD).

ODDSRATIO produces odds ratios for variables even when the variables are involved in interactions with other covariates, and for classification variables that use any parameterization. You can specify several ODDSRATIO statements.

Simple Logistic Regression Model

Example: Fit a binary logistic regression model in PROC LOGISTIC. Select **Survived** as the outcome variable and **Age** as the predictor variable. Use the EVENT= option to model the probability of surviving and request profile likelihood confidence intervals around the estimated odds ratios.

```
/*st105d05.sas*/
proc logistic data=sasuser.Titanic alpha=.05
              plots(only)=(effect oddsratio);
   model Survived(event='1')=Age / clodds=pl;
   title1 'LOGISTIC MODEL (1):Survived=Age';
run;
```

Selected PLOTS options:

EFFECT requests a plot of the predicted probability on the Y axis by the predictor on the X axis. If there is more than one predictor variable in the model, the partial effect plot can be requested using the option (X=<*variable*>).

ODDSRATIO requests a plot of the odds ratios, along with its (1-ALPHA) confidence limits. The width of the confidence limits can be changed from the default of 95% using an ALPHA= option in the PROC LOGISTIC statement. The chosen alpha level applies to all confidence intervals produced in all tables and plots in that run of PROC LOGISTIC.

Selected MODEL statement options:

(EVENT=) specifies the event category for the binary response model. PROC LOGISTIC models the probability of the event category. You can specify the value (formatted if a format is applied) of the event category in quotation marks or you can specify one of the following keywords. The default is EVENT=FIRST.

FIRST designates the first ordered category as the event.

LAST designates the last ordered category as the event.

CLODDS=PL requests profile likelihood confidence intervals for the odds ratios of all predictor variables, which are desirable for small sample sizes. The CLODDS= option also enables production of the ODDSRATIO plot.

SAS Output

Model Information	
Data Set	SASUSER.TITANIC
Response Variable	Survived
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

The Model Information table describes the data set, the response variable, the number of response levels, the type of model, the algorithm used to obtain the parameter estimates, and the number of observations read and used.

The Optimization Technique is the iterative numerical technique that PROC LOGISTIC uses to estimate the model parameters.

The model is assumed to be “binary logit” when there are exactly two response levels.

Number of Observations Read	1309
Number of Observations Used	1046

The Number of Observations Used is the count of all observations that are nonmissing for all variables specified in the MODEL statement. The ages of 263 of these 1309 passengers cannot be determined and cannot be used to estimate the model.

Response Profile		
Ordered Value	**Survived**	**Total Frequency**
1	0	619
2	1	427

The Response Profile table shows the response variable values listed according to their ordered values. By default, PROC LOGISTIC orders the response variable alphanumerically so that it bases the logistic regression model on the probability of the smallest value. Because you used the EVENT=option in this example, the model is based on the probability of surviving (**Survived**=1). The Response Profile table also shows frequencies of response values.

Probability modeled is Survived=1.

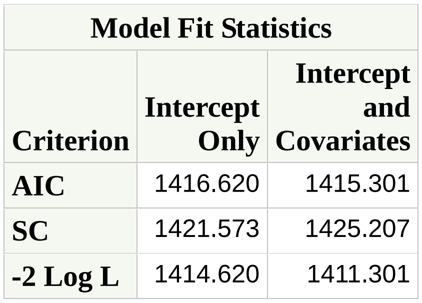
It is advisable to check that the modeled response level is the one that you intended.

Note: 263 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

The Model Convergence Status simply informs you that the convergence criterion was met. There are a number of options to control the convergence criterion.

The optimization technique does not always converge to a maximum likelihood solution. When this is the case, the output after this point cannot be trusted. Always check to see that the Convergence criterion is satisfied.

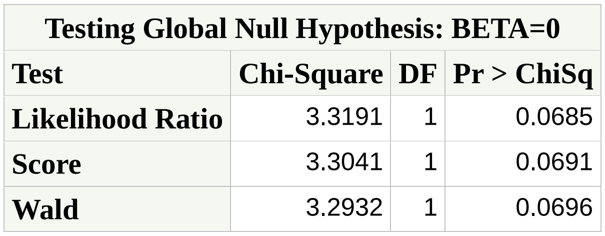


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1416.620	1415.301
SC	1421.573	1425.207
-2 Log L	1414.620	1411.301

The Model Fit Statistics provides three tests:

- AIC is Akaike's 'A' information criterion.
- SC is the Schwarz criterion.
- –2 Log L is –2 times the natural log of the likelihood.

-2 Log L, AIC, and SC are goodness-of-fit measures that you can use to compare one model to another. ***These statistics measure relative fit among models, but they do not measure absolute fit of any single model.*** Smaller values for all of these measures indicate better fit. However, -2 Log L can be reduced by simply adding more regression parameters to the model. Therefore, it is not used to compare the fit of models that use different numbers of parameters. AIC adjusts for the number of predictor variables, and SCs adjust for the number of predictor variables and the number of observations. SC uses a bigger penalty for extra variables and therefore favors more parsimonious models.



Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.3191	1	0.0685
Score	3.3041	1	0.0691
Wald	3.2932	1	0.0696

The Testing Global Null Hypothesis: BETA=0 table provides three statistics to test the null hypothesis that all regression coefficients of the model are 0.

A significant *p*-value for these tests provides evidence that at least one of the regression coefficients for an explanatory variable is significantly different from 0. In this way, they are similar to the overall *F* test in linear regression. The Likelihood Ratio Chi-Square is calculated as the difference between the -2 Log L value of the baseline model (Intercept Only) and the -2 Log L value of the hypothesized model (Intercept and Covariates). The statistic is a distributed asymptotically chi-square with degrees of freedom equal to the difference in number of parameters between the hypothesized model and the baseline model. The Score and Wald tests are also used to test whether all the regression coefficients are 0. The likelihood ratio test is the most reliable, especially for small sample sizes (Agresti 1996). All three tests are asymptotically equivalent and often give very similar values.

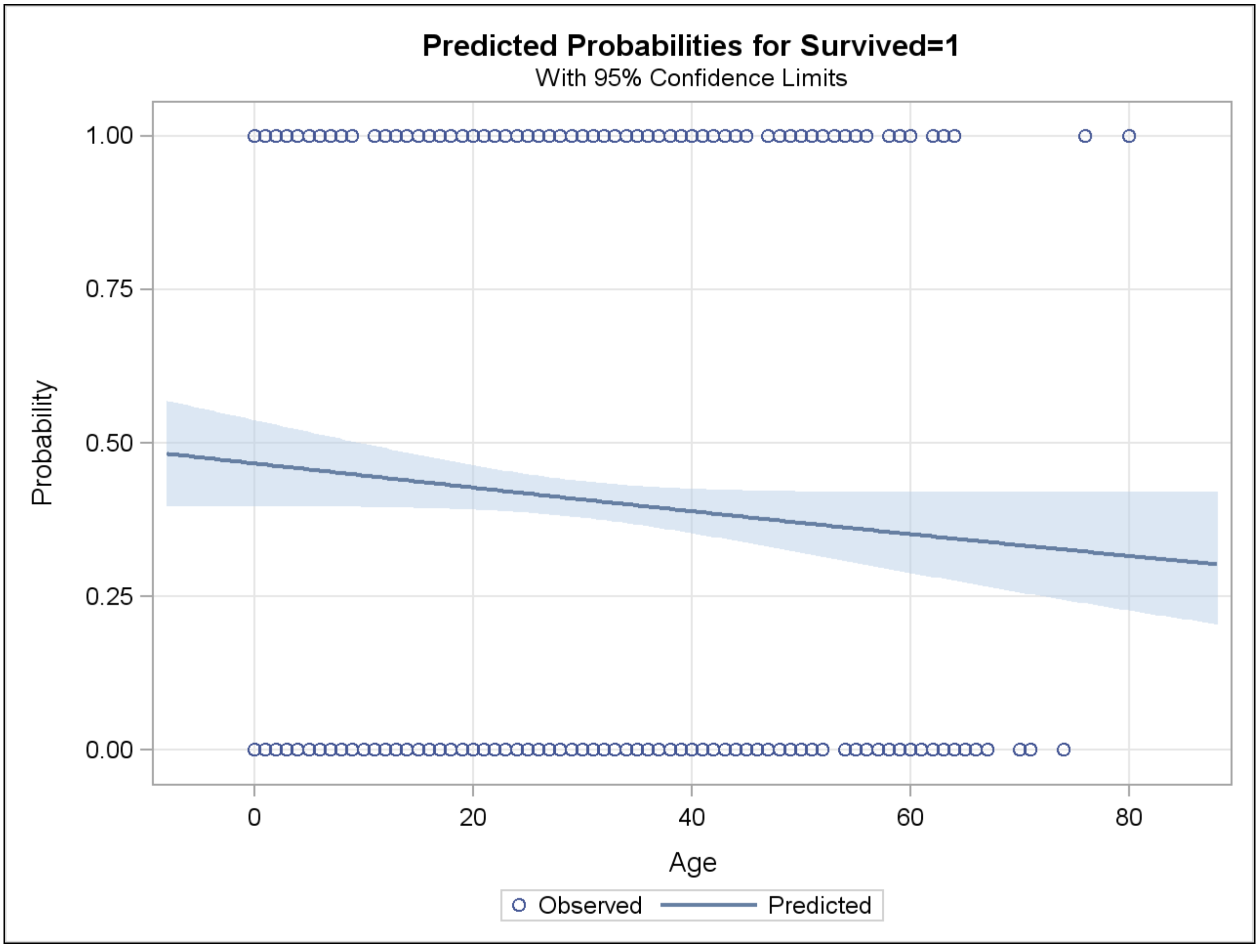
Wald statistics (*p*-values and confidence limits) require fewer computations to perform and are therefore the default for most output in PROC LOGISTIC.

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-0.1335	0.1448	0.8501	0.3565
Age	1	-0.00800	0.00441	3.2932	0.0696

The Analysis of Maximum Likelihood Estimates table lists the estimated model parameters, their standard errors, Wald Chi-Square values, and *p*-values.

The parameter estimates are the estimated coefficients of the fitted logistic regression model. The logistic regression equation is logit($\hat{p}$)=–0.1335+(-0.00800)***Age** for this example.

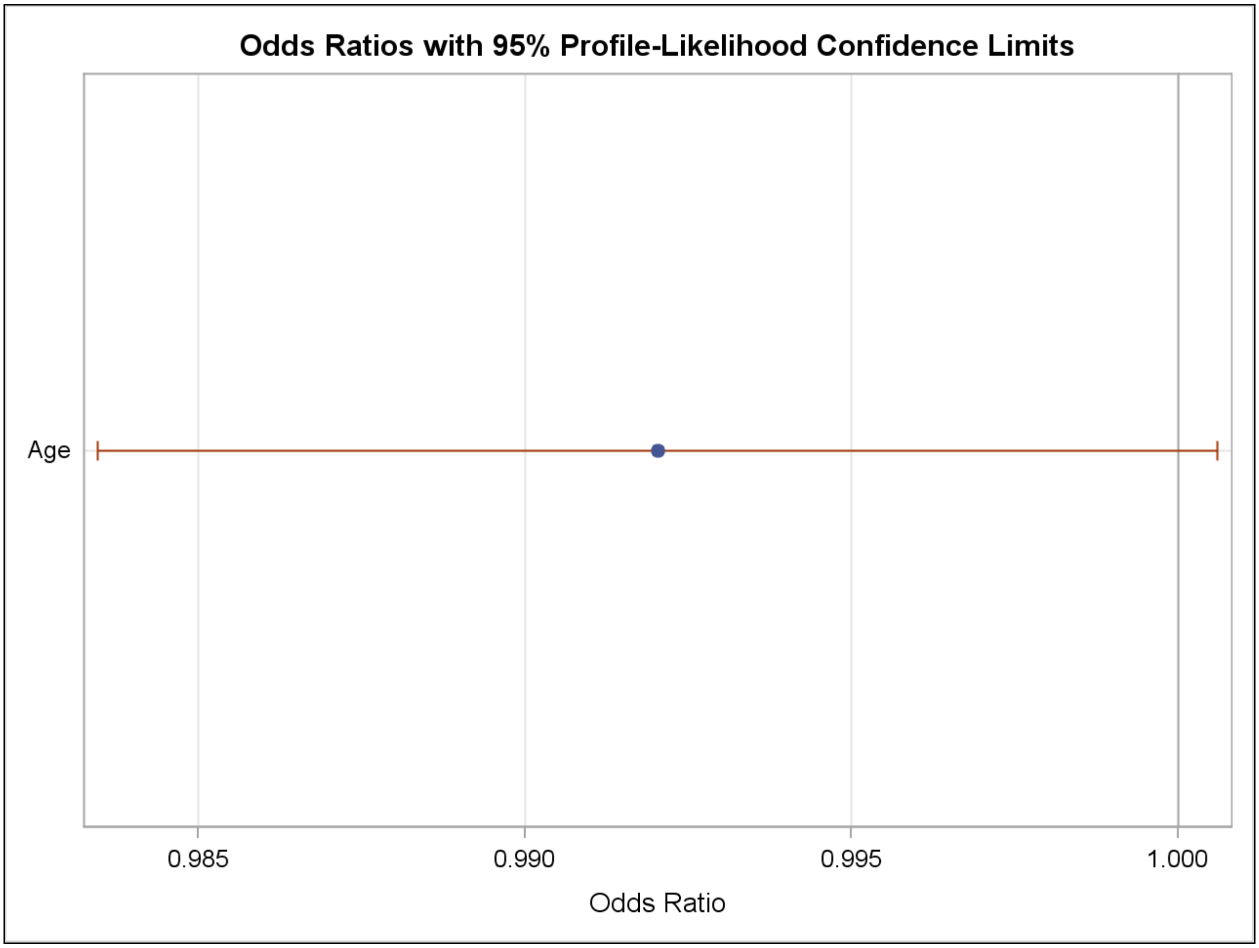
The Wald chi-square and its associated *p*-value tests whether the parameter estimate is significantly different from 0. For this example, the *p*-values for the variable **Age** is not significant at the 0.05 significance level (p=0.0696). It cannot be concluded that **Age** is not important in a multivariate model.



The estimated model is displayed on the probability scale in the Effect plot. The observed values are plotted at probabilities 1.00 and 0.00.

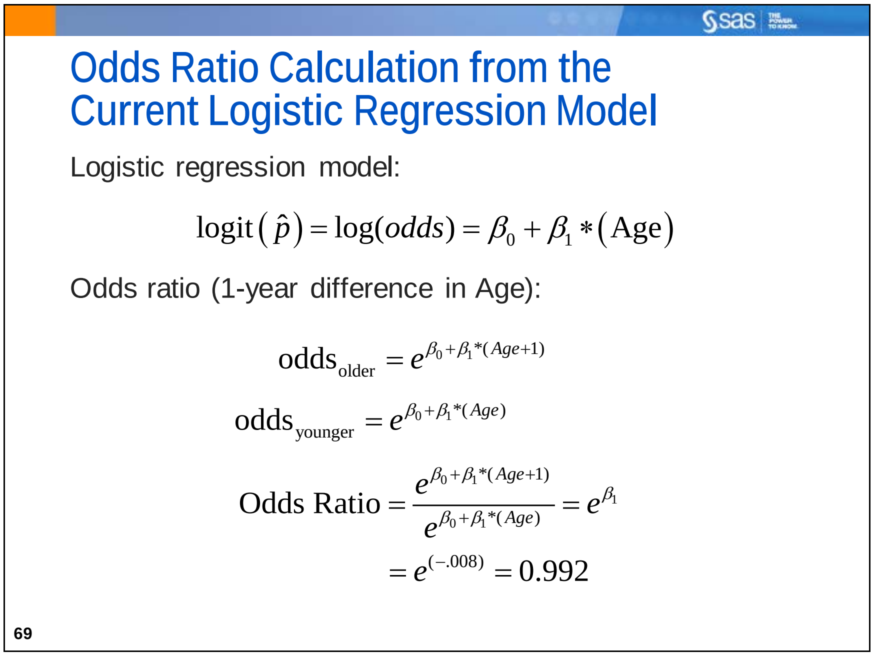
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	51.3	Somers' D	0.050
Percent Discordant	46.4	Gamma	0.051
Percent Tied	2.3	Tau-a	0.024
Pairs	264313	c	0.525

Profile Likelihood Confidence Interval for Odds Ratios				
Effect	Unit	Estimate	95% Confidence Limits	
Age	1.0000	0.992	0.983	1.001



Odds Ratios with 95% Profile-Likelihood Confidence Limits

The above tables and plots are described in detail in the next slides.



Odds Ratio Calculation from the Current Logistic Regression Model

Logistic regression model:

$$\text{logit}\left(\hat{p}\right) = \log(odds) = \beta_0 + \beta_1 * \left(\text{Age}\right)$$

Odds ratio (1-year difference in Age):

$$\text{odds}_{\text{older}} = e^{\beta_0 + \beta_1 * (Age+1)}$$

$$\text{odds}_{\text{younger}} = e^{\beta_0 + \beta_1 * (Age)}$$

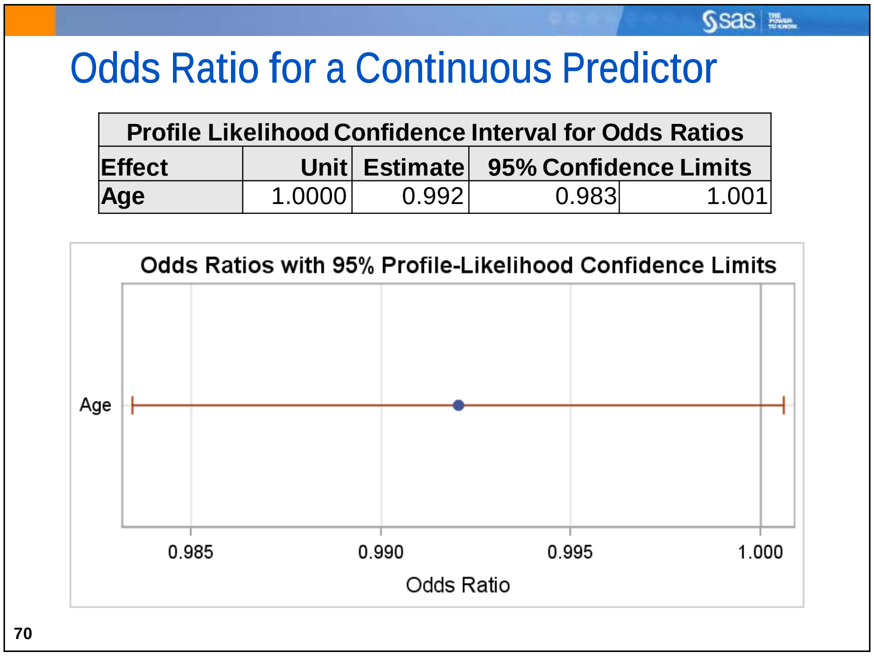
$$\text{Odds Ratio} = \frac{e^{\beta_0 + \beta_1 * (Age+1)}}{e^{\beta_0 + \beta_1 * (Age)}} = e^{\beta_1}$$

$$= e^{(-.008)} = 0.992$$

69

The odds ratio for a continuous predictor calculates the estimated relative odds for subjects that are one unit apart on the continuous measure. For example, in the Titanic example, **Age** is the continuous measure. If you remember, the logit is the natural log of the odds. Because you can calculate an estimated logit from the logistic model, the odds can be calculated by simply exponentiating that value. An odds ratio for a one-unit difference is then the ratio of the exponentiated predicted logits for two people who are one unit apart.

The odds ratio for age indicates that the odds of surviving decrease by 0.8% for each year older.



Odds Ratio for a Continuous Predictor

Profile Likelihood Confidence Interval for Odds Ratios				
Effect	Unit	Estimate	95% Confidence Limits	
Age	1.0000	0.992	0.983	1.001

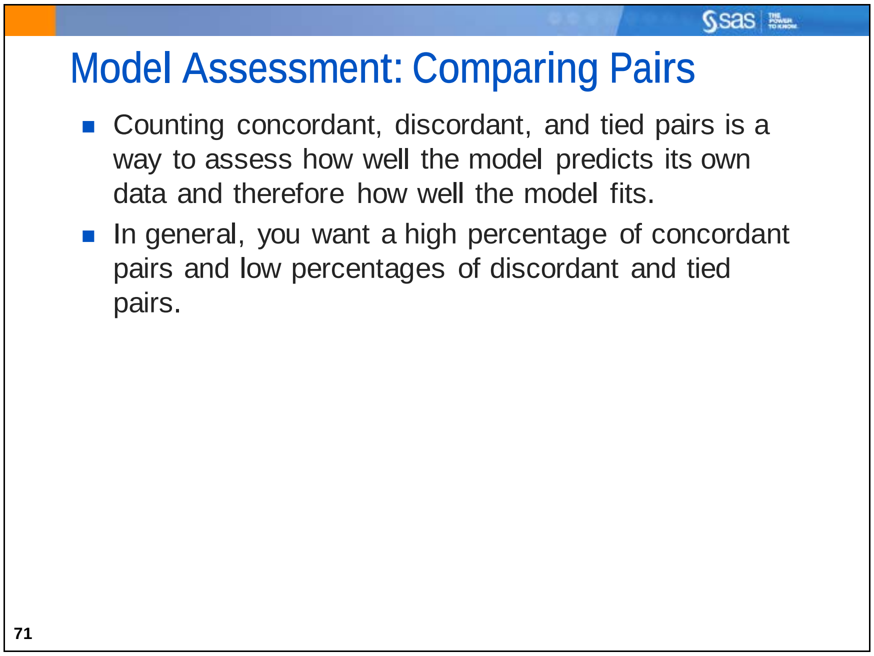
Odds Ratios with 95% Profile-Likelihood Confidence Limits

The 95% confidence limits indicate that you are 95% confident that the true odds ratio is between 0.983 and 1.001. Because the 95% confidence interval includes 1.000, the odds ratio is not significant at the .05 alpha level.

If you want a different significance level for the confidence intervals, you can use the ALPHA= option in the MODEL statement. The value must be between 0 and 1. The default value of .05 results in the calculation of a 95% confidence interval.

The profile likelihood confidence intervals are different from the Wald-based confidence intervals. This difference is because the Wald confidence intervals use a normal error approximation, whereas the profile likelihood confidence intervals are based on the value of the log-likelihood. These likelihood-ratio confidence intervals require a much greater number of computations, but are generally preferred to the Wald confidence intervals, especially for sample sizes less than 50 (Allison 1999).

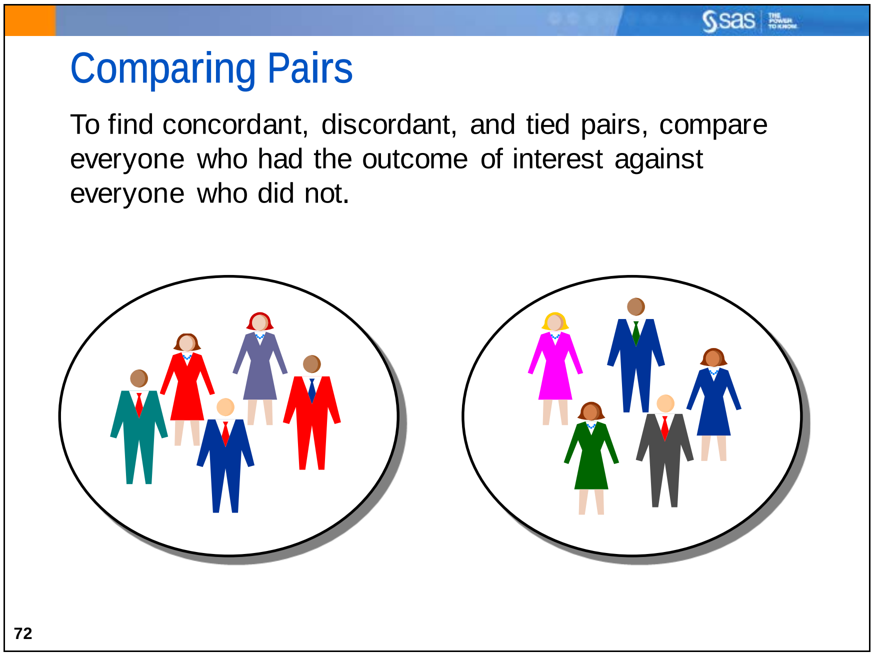
The Odds Ratio plot displays the results of the Odds Ratio table graphically. A reference line shows the null hypothesis. When the confidence interval crosses the reference line, the effect of the variable is not significant.



Model Assessment: Comparing Pairs

- Counting concordant, discordant, and tied pairs is a way to assess how well the model predicts its own data and therefore how well the model fits.
- In general, you want a high percentage of concordant pairs and low percentages of discordant and tied pairs.

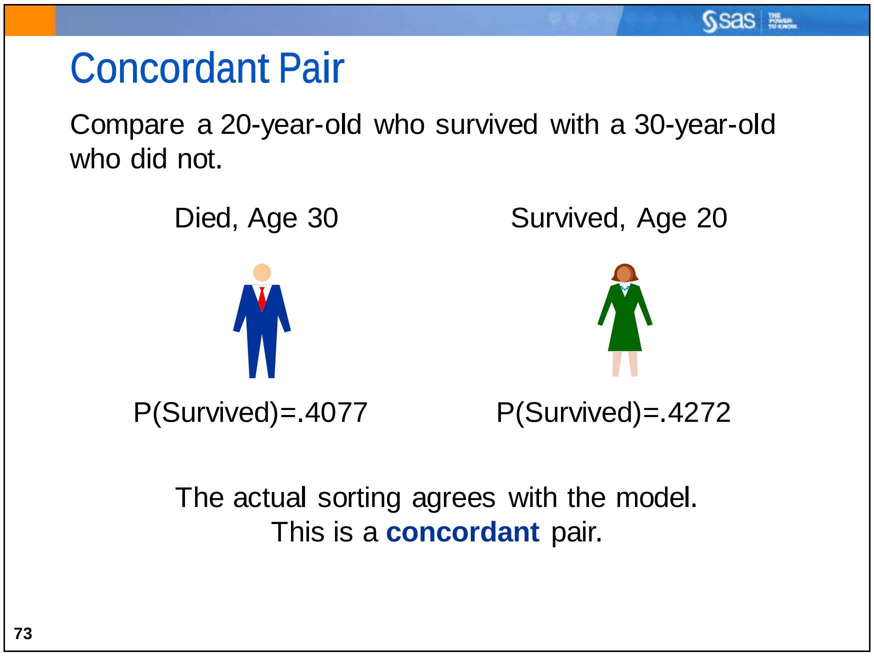
71



Comparing Pairs

To find concordant, discordant, and tied pairs, compare everyone who had the outcome of interest against everyone who did not.

72



Concordant Pair

Compare a 20-year-old who survived with a 30-year-old who did not.

Died, Age 30

P(Survived)=.4077

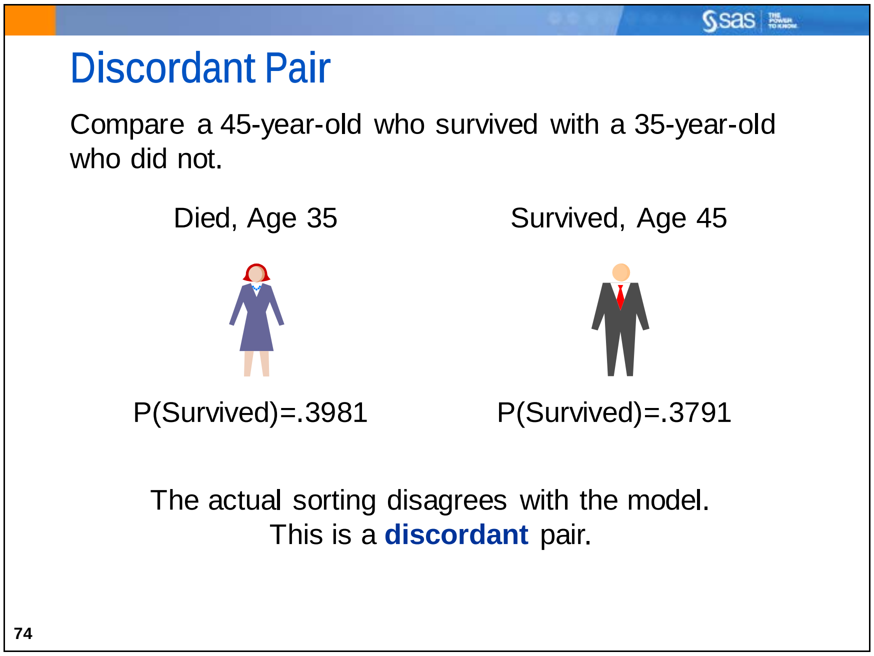
Survived, Age 20

P(Survived)=.4272

The actual sorting agrees with the model.
This is a **concordant** pair.

73

For all pairs of observations with different values of the response variable, a pair is *concordant* if the observation with the outcome has a ***higher*** predicted outcome probability (based on the model) than the observation without the outcome.



Discordant Pair

Compare a 45-year-old who survived with a 35-year-old who did not.

Died, Age 35

P(Survived)=.3981

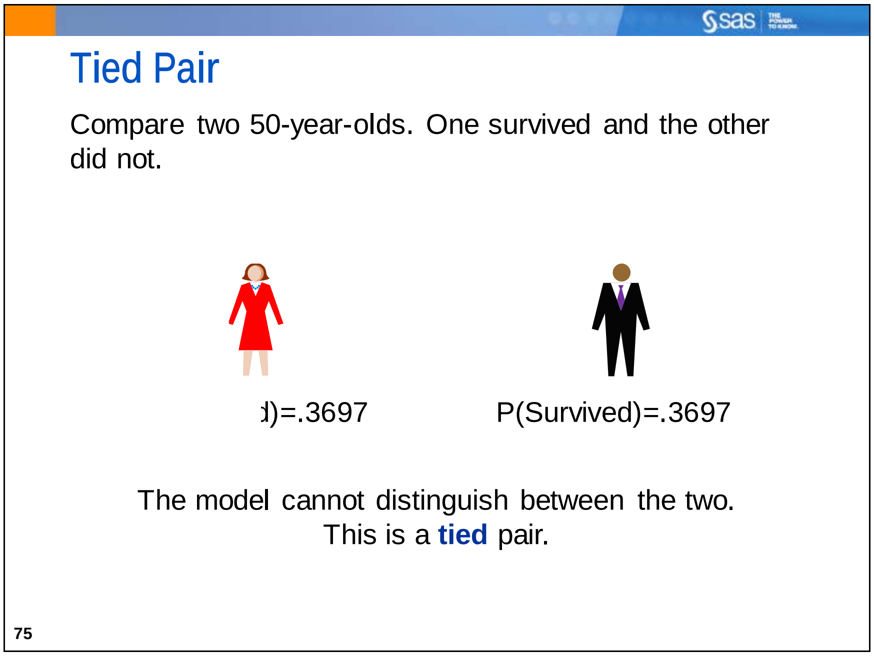
Survived, Age 45

P(Survived)=.3791

The actual sorting disagrees with the model.
This is a **discordant** pair.

74

A pair is *discordant* if the observation with the outcome has a ***lower*** predicted outcome probability than the observation without the outcome.



Tied Pair

Compare two 50-year-olds. One survived and the other did not.

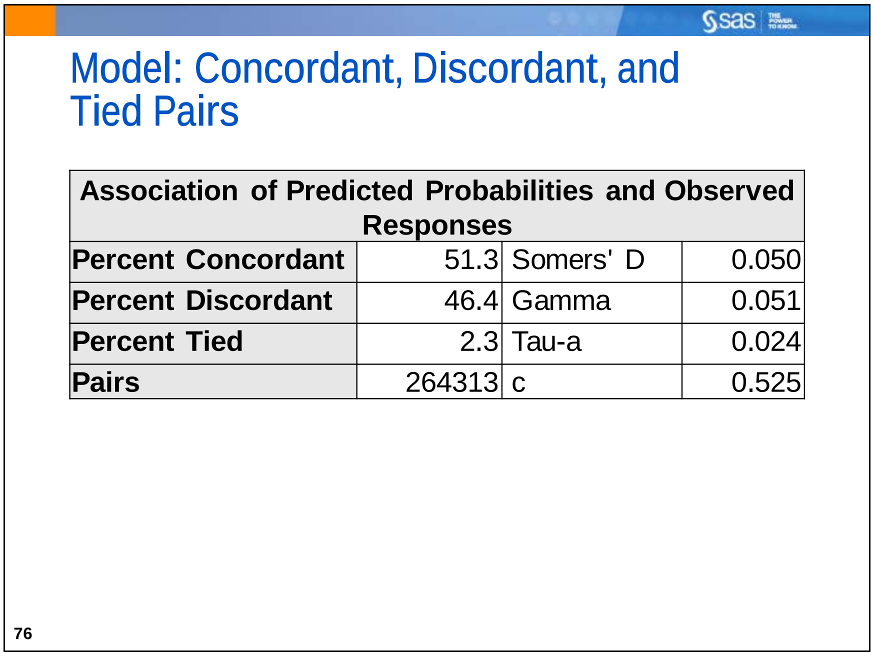
d)=.3697

P(Survived)=.3697

The model cannot distinguish between the two.
This is a **tied** pair.

75

A pair is *tied* if it is neither concordant nor discordant. (The probabilities are the same.)



Model: Concordant, Discordant, and Tied Pairs

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	51.3	Somers' D	0.050
Percent Discordant	46.4	Gamma	0.051
Percent Tied	2.3	Tau-a	0.024
Pairs	264313	c	0.525

76

The Association of Predicted Probabilities and Observed Responses table lists several measures of association to help you assess the predictive ability of the logistic model.

The number of pairs used to calculate the values of this table is equal to the product of the counts of observations with positive responses and negative responses. In this example, that value is 427*619=264,313.

You can use these percentages as goodness-of-fit measures to compare one model to another. In general, higher percentages of concordant pairs and lower percentages of discordant pairs indicate a more desirable model.

The four rank correlation indices (Somer's D, Gamma, Tau-a, and *c*) are computed from the numbers of concordant, discordant, and tied pairs of observations. In general, a model with higher values for these indices has better predictive ability than a model with lower values for these indices.

The *c* (concordance) statistic estimates the probability of an observation with the outcome having a higher predicted probability than an observation without the outcome. It is calculated as the percent concordant plus one half the percent tied. The range of possible values is 0.500 (no better predictive power than flipping a fair coin) to 1.000 (perfect prediction). The value of 0.525 shows a very weak ability of **Age** to discriminate between those who survived and those who did not.

Exercises

2. Performing a Logistic Regression Analysis

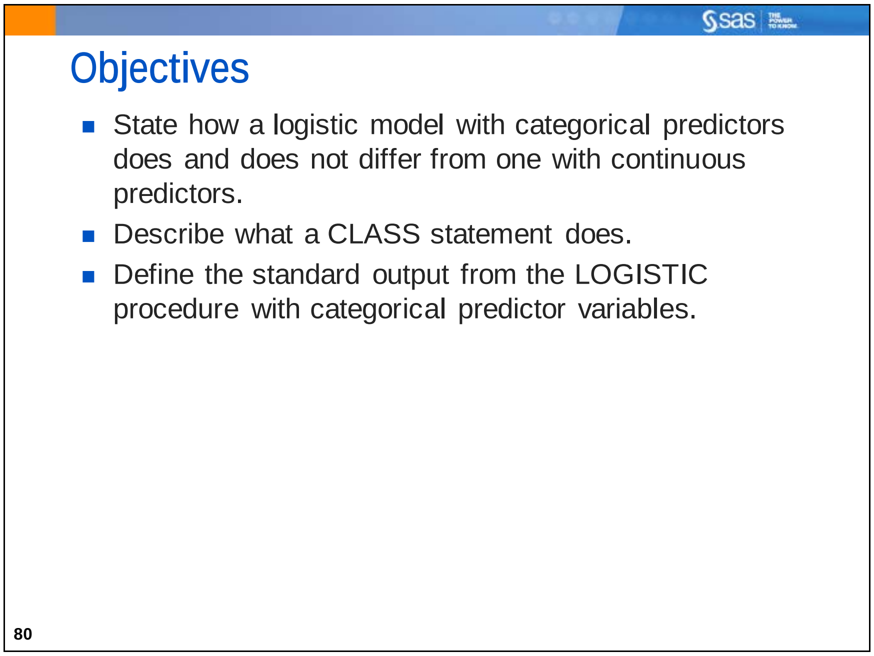
Fit a simple logistic regression model using **sasuser.safety** with **Unsafe** as the outcome variable and **Weight** as the predictor variable. Use the EVENT= option to model the probability of below-average safety scores. Request Profile Likelihood confidence limits and an odds ratio plot along with an effect plot.

a. Do you reject or fail to reject the global null hypothesis that all regression coefficients of the model are 0?

b. Write the logistic regression equation.

c. Interpret the odds ratio for **Weight**.

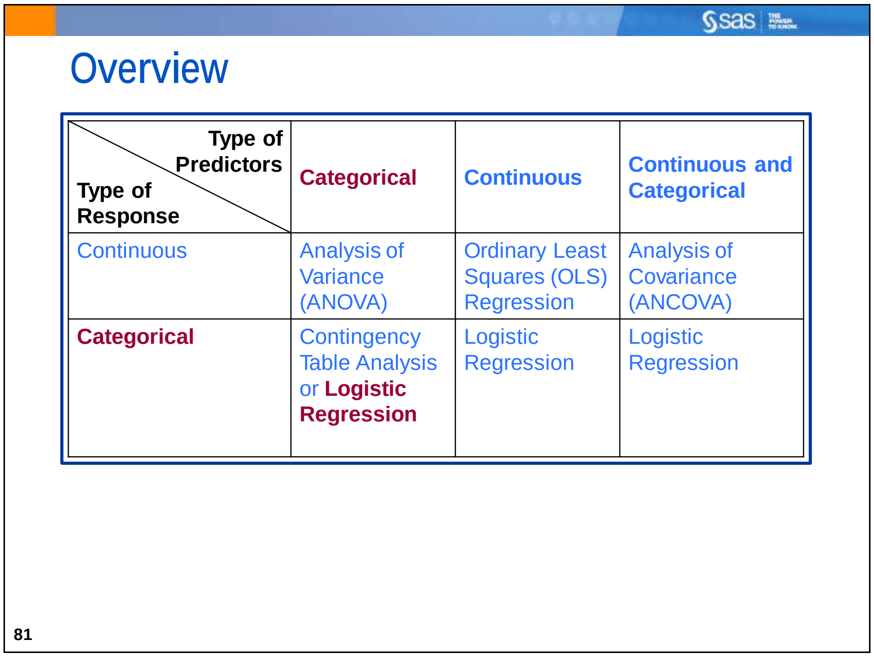
5.4 Logistic Regression with Categorical Predictors



Objectives

- State how a logistic model with categorical predictors does and does not differ from one with continuous predictors.
- Describe what a CLASS statement does.
- Define the standard output from the LOGISTIC procedure with categorical predictor variables.

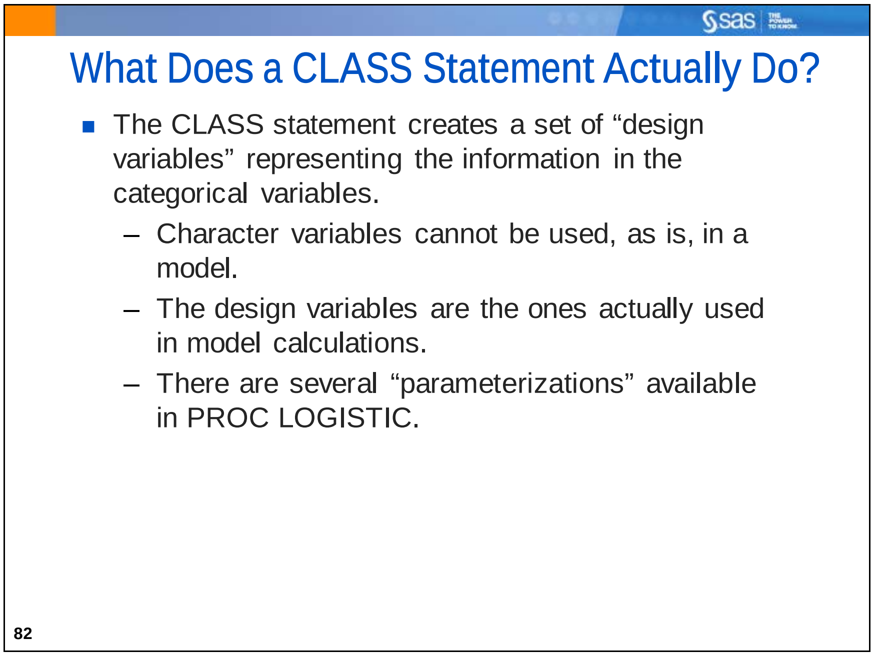
80



Overview

Type of Predictors / Type of Response	**Categorical**	**Continuous**	**Continuous and Categorical**
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or **Logistic Regression**	Logistic Regression	Logistic Regression

81



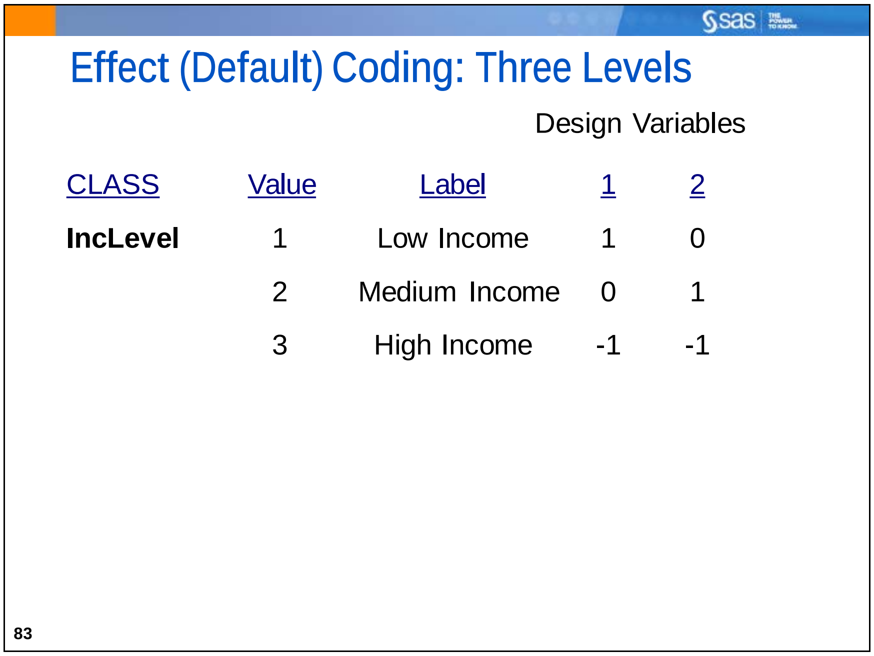
What Does a CLASS Statement Actually Do?

- The CLASS statement creates a set of “design variables” representing the information in the categorical variables.
  - Character variables cannot be used, as is, in a model.
  - The design variables are the ones actually used in model calculations.
  - There are several “parameterizations” available in PROC LOGISTIC.

82

The CLASS statement creates a set of “design variables” representing the information contained in any categorical variables. These design variables are incorporated into the model calculations rather than the original categorical variables. Character variables cannot be used, as is, in the model. SAS cannot use a variable with values such as ‘yes’ or ‘no’ adequately in the determination of a model.

Even if categorical variables are represented by numbers such as 1, 2, 3, the CLASS statement tells SAS to set up design variables to represent the categories. This is necessary because the numeric values that are assigned to the levels of the categorical variable are generally arbitrary and might not truly reflect distances between levels.

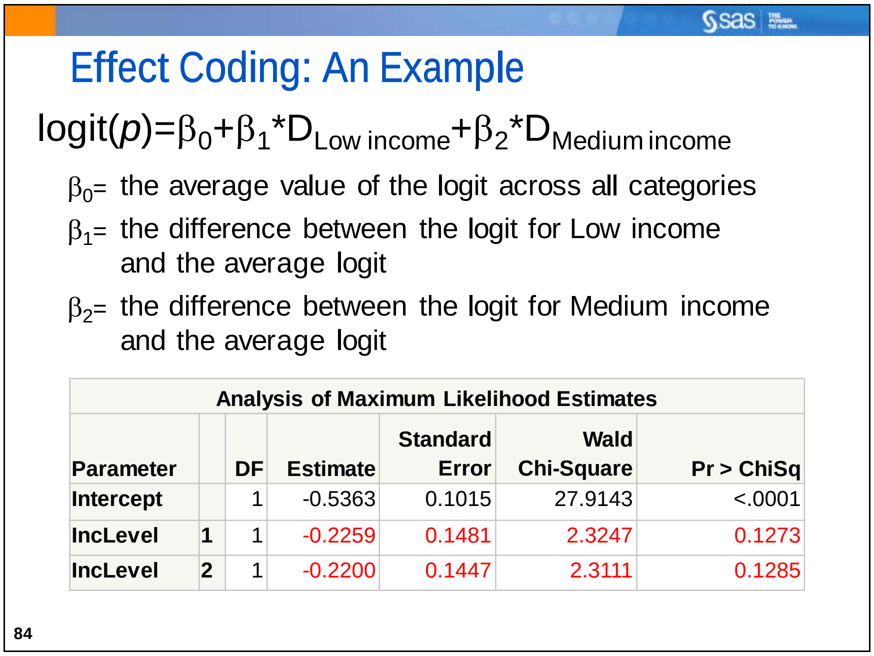


Effect (Default) Coding: Three Levels

			Design	Variables
CLASS	Value	Label	1	2
IncLevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	-1	-1

83

For *effect coding* (also called *deviation from the mean coding*), the number of design variables created is the number of levels of the CLASS variable minus 1. For example, because the variable **IncLevel** has three levels, two design variables were created. For the last level of the CLASS variable (**`High Income`**), all the design variables have a value of –1. Parameter estimates of the CLASS main effects using this coding scheme estimate the **difference** between the effect of each level and the average effect over all levels.



Effect Coding: An Example

$$\text{logit}(p)=\beta_0+\beta_1{*}D_{\text{Low income}}+\beta_2{*}D_{\text{Medium income}}$$

β_0= the average value of the logit across all categories

β_1= the difference between the logit for Low income and the average logit

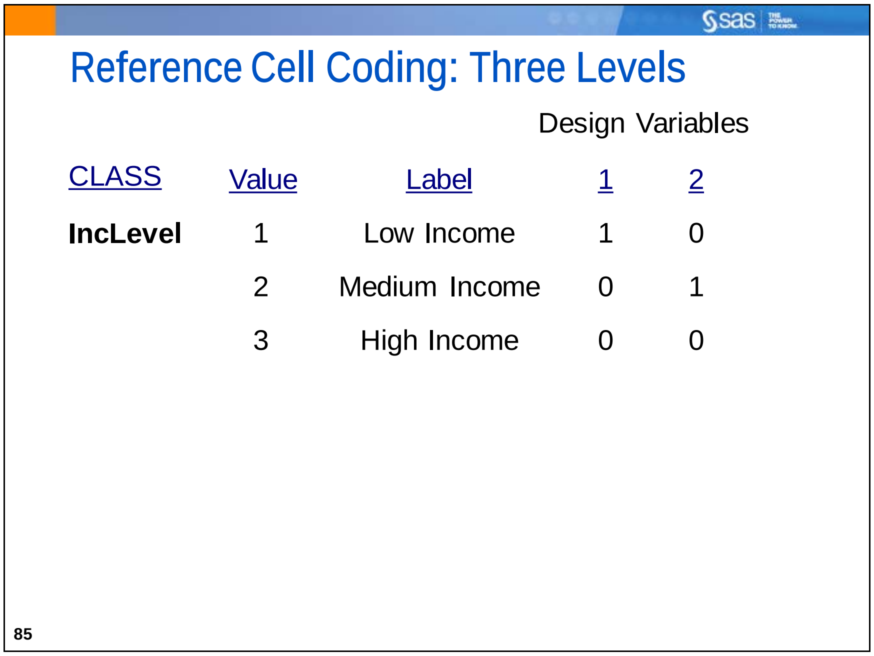
β_2= the difference between the logit for Medium income and the average logit

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.5363	0.1015	27.9143	<.0001
IncLevel	1	1	-0.2259	0.1481	2.3247	0.1273
IncLevel	2	1	-0.2200	0.1447	2.3111	0.1285

84

If you use Effect Coding for a CLASS variable, then the parameter estimates and *p*-values reflect differences from the mean logit value over all levels. So, for **IncLevel**, the Estimate shows the estimated difference in logit values between **IncLevel**=1 (Low Income) and the average logit across all income levels.

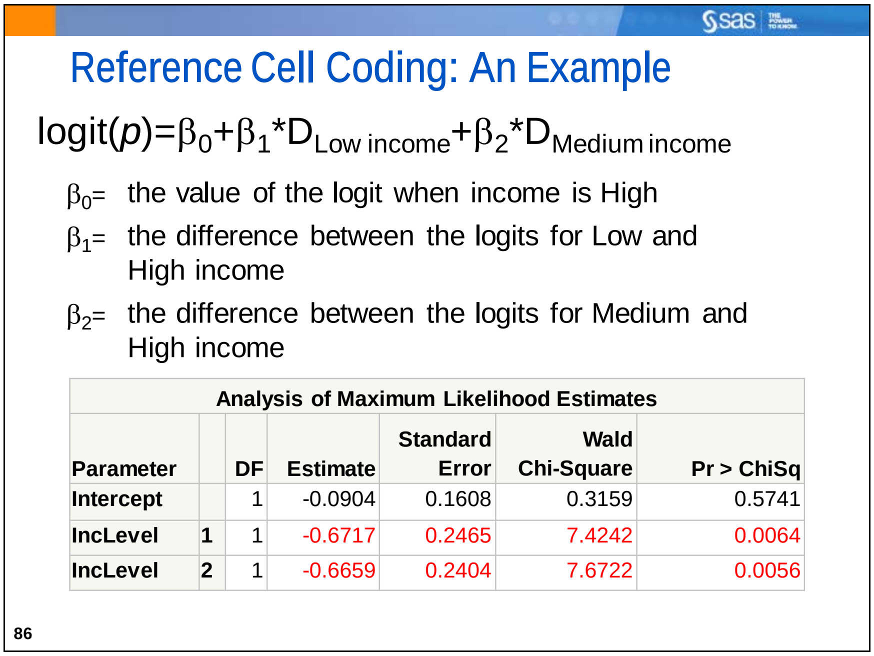


Reference Cell Coding: Three Levels

CLASS	Value	Label	Design Variables 1	Design Variables 2
IncLevel	1	Low Income	1	0
	2	Medium Income	0	1
	3	High Income	0	0

85

For *reference cell coding*, parameter estimates of the CLASS main effects estimate the difference between the effect of each level and the last level, called the *reference level*. For example, the effect for the level **`Low`** estimates the logit difference between **`Low`** and **`High`**. You can choose the reference level in the CLASS statement.



Reference Cell Coding: An Example

$$\text{logit}(p)=\beta_0+\beta_1*D_{\text{Low income}}+\beta_2*D_{\text{Medium income}}$$

β_0= the value of the logit when income is High

β_1= the difference between the logits for Low and High income

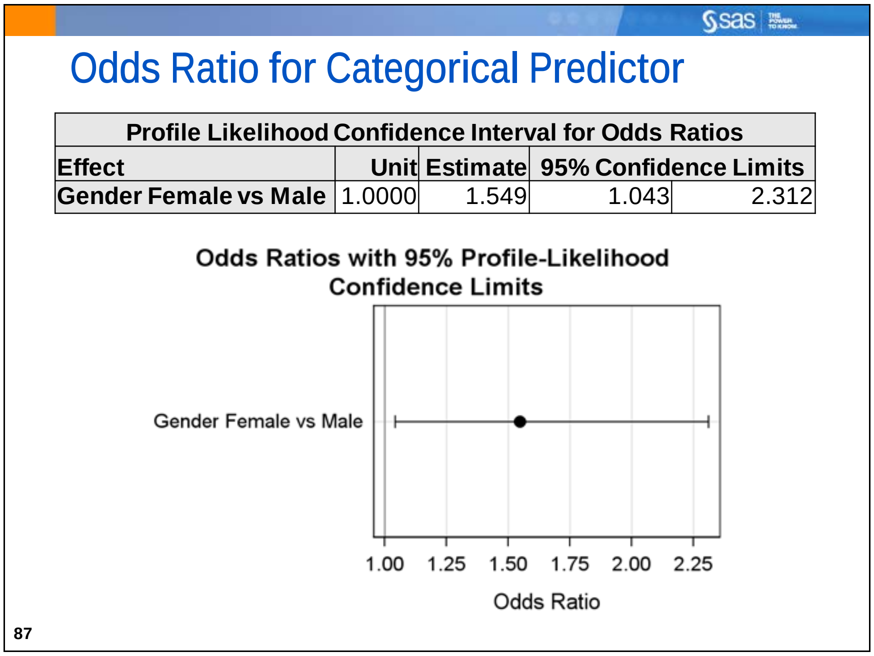
β_2= the difference between the logits for Medium and High income

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.0904	0.1608	0.3159	0.5741
IncLevel	**1**	1	-0.6717	0.2465	7.4242	0.0064
IncLevel	**2**	1	-0.6659	0.2404	7.6722	0.0056

Notice the difference between this table and the previous parameter estimates table. Because you used Reference Cell Coding, instead of Effect Coding, the meanings of the parameter estimates and *p*-values are different. Now, the parameter estimate and *p*-value for **IncLevel**=1 reflect the difference between **IncLevel**=1 and **Inclevel**=3 (the reference level).

It is important to know what type of parameterization you are using in order to interpret and report the results of this table.



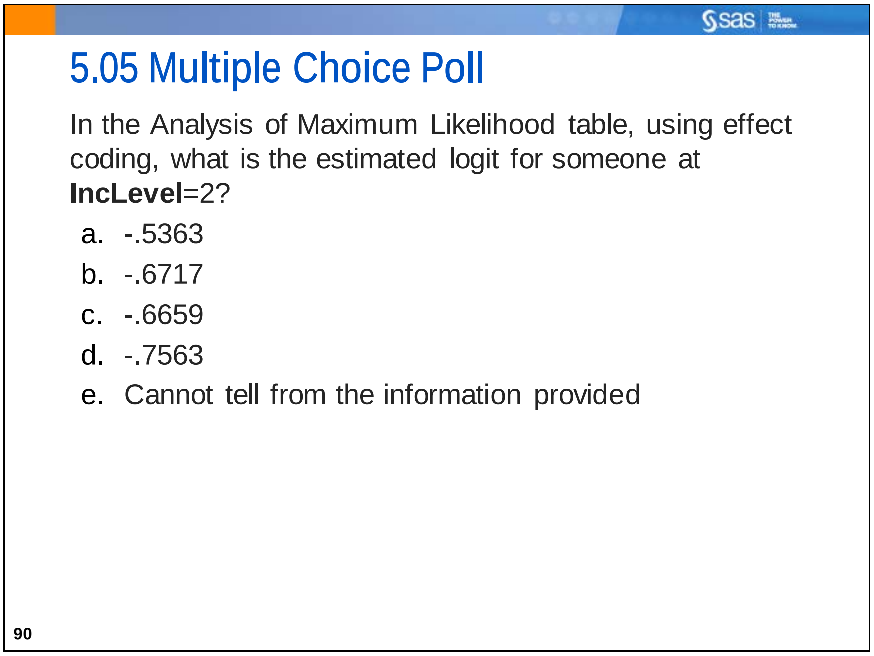
Odds Ratio for Categorical Predictor

Profile Likelihood Confidence Interval for Odds Ratios				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
Gender Female vs Male	1.0000	1.549	1.043	2.312

Odds Ratios with 95% Profile-Likelihood Confidence Limits

87

Odds ratios for categorical predictors are reported for bi-group comparisons in PROC LOGISTIC, no matter which parameterization is chosen. Thus, even if Effect Coding is selected for the **Gender** variable, the odds ratio tables display odds comparisons between females and males (and not females versus the average of both).

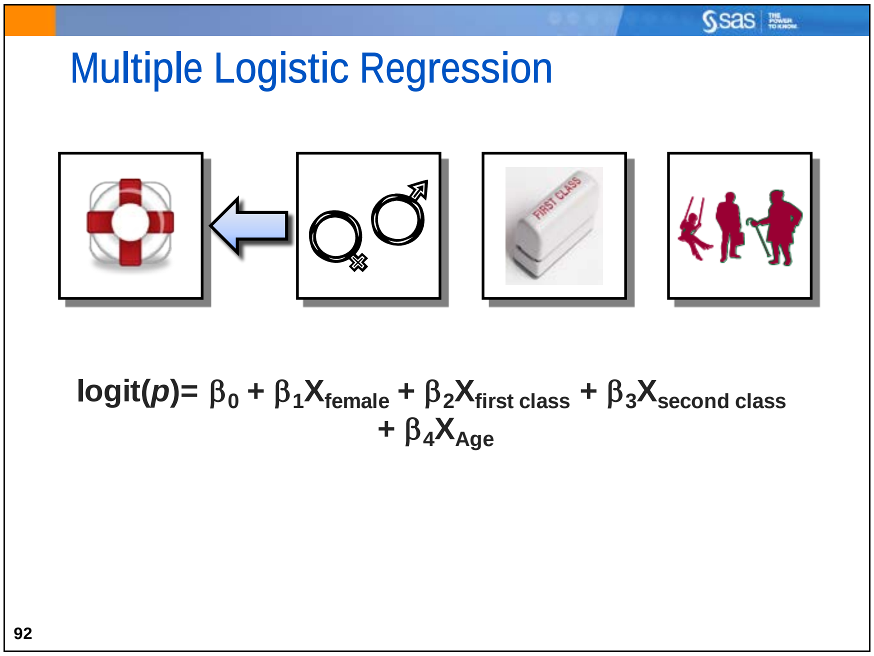


5.05 Multiple Choice Poll

In the Analysis of Maximum Likelihood table, using effect coding, what is the estimated logit for someone at **IncLevel**=2?

a. -.5363
b. -.6717
c. -.6659
d. -.7563
e. Cannot tell from the information provided

90



Multiple Logistic Regression

$$\text{logit}(p) = \beta_0 + \beta_1 X_{\text{female}} + \beta_2 X_{\text{first class}} + \beta_3 X_{\text{second class}} + \beta_4 X_{\text{Age}}$$

92

Each design variable is assigned its own beta value. The number of parameters in the logistic model take into account the intercept, the number of continuous predictors, and the number of design variables assigned to CLASS variables.

Multiple Logistic Regression with Categorical Predictors

Example: Fit a binary logistic regression model in PROC LOGISTIC. Select **Survived** as the outcome variable and **Age**, **Gender**, and **Class** as the predictor variables. Specify reference cell coding and specify **male** as the reference group for **Gender** and **3** as the reference level for **Class**. Also use the EVENT= option to model the probability of surviving and request profile likelihood confidence intervals around the estimated odds ratios.

```
/*st105d06.sas*/
proc logistic data=sasuser.Titanic plots(only)=(effect oddsratio);
   class Gender(ref='male') Class(ref='3') / param=ref;
   model Survived(event='1')=Age Gender Class / clodds=pl;
   units age=10;
   title1 'LOGISTIC MODEL (2):Survived=Age Gender Class';
run;
```

Selected PROC LOGISTIC statement:

UNITS enables you to specify units of change for the continuous explanatory variables so that customized odds ratios can be estimated.

Selected CLASS statement options:

(REF='level') specifies the event category chosen as the reference level when using Reference or Effect parameterization. You can specify the value (formatted if a format is applied) of the reference category in quotation marks or you can specify one of the following keywords. The default is REF=LAST.

FIRST designates the first ordered category as the reference level.

LAST designates the last ordered category as the reference level.

PARAM= specifies the parameterization. This value can be specified for each CLASS variable by typing it within parentheses after the variable name, or for all CLASS variables, by typing it after the options slash (/) at the end of the list of CLASS variables.

If there are numerous levels in the CLASS variable, you might want to use subject-matter knowledge to reduce the number of levels. This is especially important when the levels have few or no observations.

Model Information	
Data Set	SASUSER.TITANIC
Response Variable	Survived
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1309
Number of Observations Used	1046

Response Profile		
Ordered Value	**Survived**	**Total Frequency**
1	0	619
2	1	427

Probability modeled is Survived=1.

Class Level Information			
Class	**Value**	**Design Variables**	
Gender	**female**	1	
	male	0	
Class	**1**	1	0
	2	0	1
	3	0	0

The Class Level Information table includes the predictor variable in the CLASS statement. Because you used the PARAM=REF and REF='`male`' options, this table reflects your choice of **Gender**='`male`' as the reference level. The design variable is 1 when **Gender**='`female`' and 0 when **Gender**='`male`'. The reference level for **Class** is 3, so there or two design variables, each coded 0 for observations where **Class**=3.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1416.620	992.315
SC	1421.573	1017.079
-2 Log L	1414.620	982.315

The SC value in the **Age** only model was 1425.207. Here it is 1017.079. Recalling that smaller values imply better fit, you can conclude that this model is better fitting.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	432.3052	4	<.0001
Score	386.1522	4	<.0001
Wald	277.3202	4	<.0001

This model is statistically significant, indicating at least one of the predictors in the model is useful in predicting survival.

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Age	1	29.6314	<.0001
Gender	1	226.2235	<.0001
Class	2	103.3575	<.0001

The Type 3 Analysis of Effects table is generated when a predictor variable is used in the CLASS statement. This analysis is similar to the individual tests in the GLM procedure parameter estimates table. Just as in PROC GLM and PROC REG, these are adjusted effects.

All effects, including the **Age** effect, which was not statistically significant in the univariate model, are statistically significant.

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-1.2628	0.2030	38.7108	<.0001
Age		1	-0.0345	0.00633	29.6314	<.0001
Gender	**female**	1	2.4976	0.1661	226.2235	<.0001
Class	**1**	1	2.2907	0.2258	102.8824	<.0001
Class	**2**	1	1.0093	0.1984	25.8849	<.0001

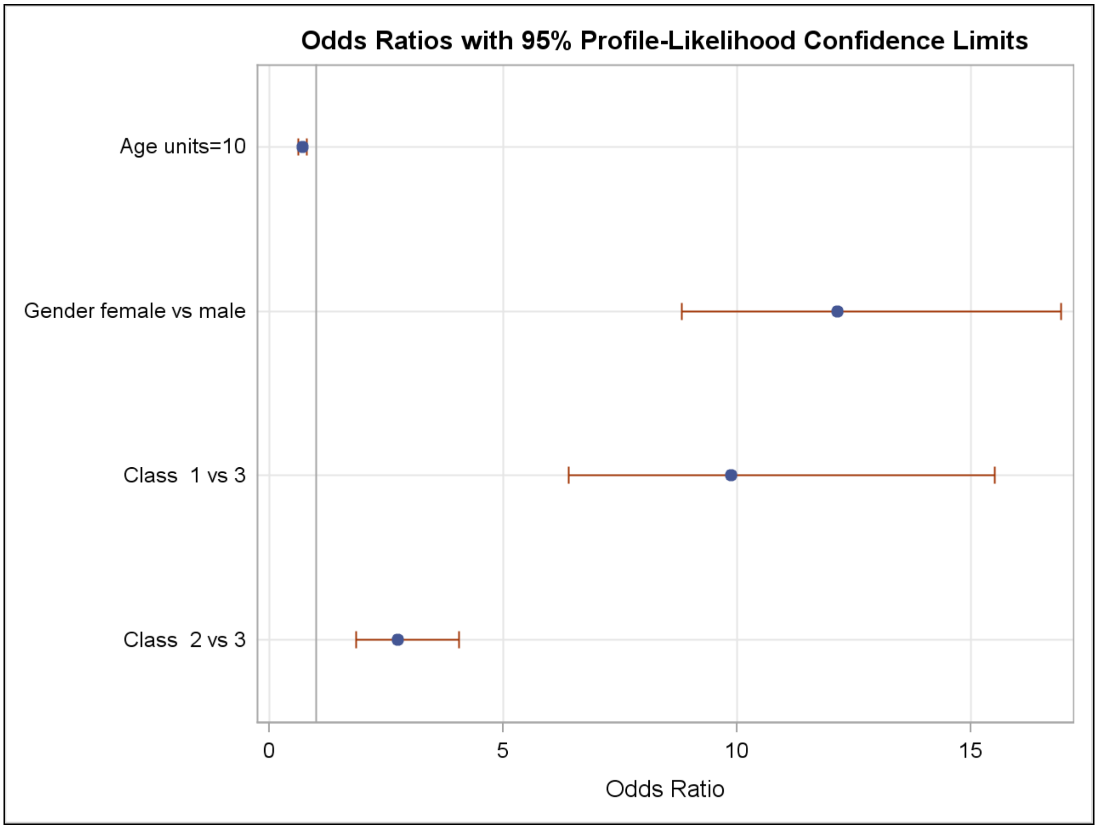
For CLASS variables, effects are displayed for each of the design variables. Because reference cell coding was used, each effect is measured against the reference level. For example, the estimate for **Gender | female** shows the difference in logits between females and males. **Class | 1** shows the logit difference between first-class passengers and third-class passengers and **Class | 2** shows the difference in logits between second class and third class. All of these contrasts are statistically significant.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	83.8	**Somers' D**	0.680
Percent Discordant	15.8	**Gamma**	0.683
Percent Tied	0.4	**Tau-a**	0.329
Pairs	264313	**c**	0.840

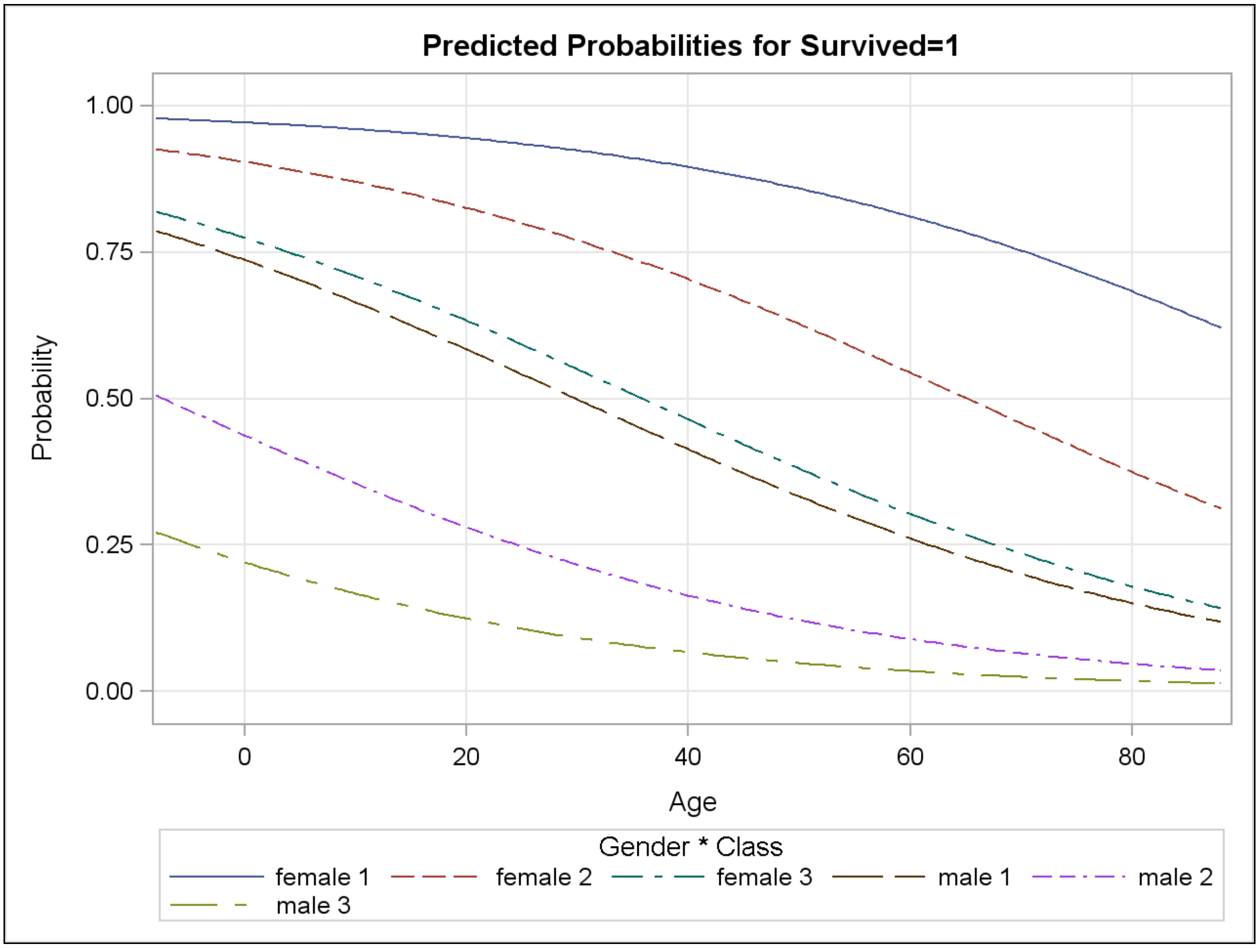
The c (Concordance) statistic value is 0.840 for this model, indicating that 84% of the positive and negative response pairs are correctly sorted using **Age**, **Gender**, and **Class**.

Profile Likelihood Confidence Interval for Odds Ratios				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
Age	10.0000	0.708	0.625	0.801
Gender female vs male	1.0000	12.153	8.823	16.925
Class 1 vs 3	1.0000	9.882	6.395	15.513
Class 2 vs 3	1.0000	2.744	1.863	4.059

The odds ratios show that, adjusting for the other predictor variables, females had 12.153 times the male odds of surviving. First-class passengers had nearly 10 times the odds (9.882) of third-class passengers and second-class passengers had 174.4% greater odds than third-class passengers. The UNITS statement applies to the odds ratio table requested by the CLODDS=PL option. The table shows that a 10-year-older age is associated with a 29.2% decrease in survival odds. The ODDSRATIO plot displays these values graphically.



Finally, the Effects plot shows the probability of survival across all combinations of categories and levels of all three predictor variables.



This plot is obtained by applying the parameter estimates from the logistic model to values of the predictors and then converting the predictions to the probability scale.

The plot indicates that at every age, all classes of women are predicted to survive at a greater rate than all classes of men. Holding **Age** and **Gender** constant, the descending order of predicted survival probabilities are first class, second class, and third class.

Exercises

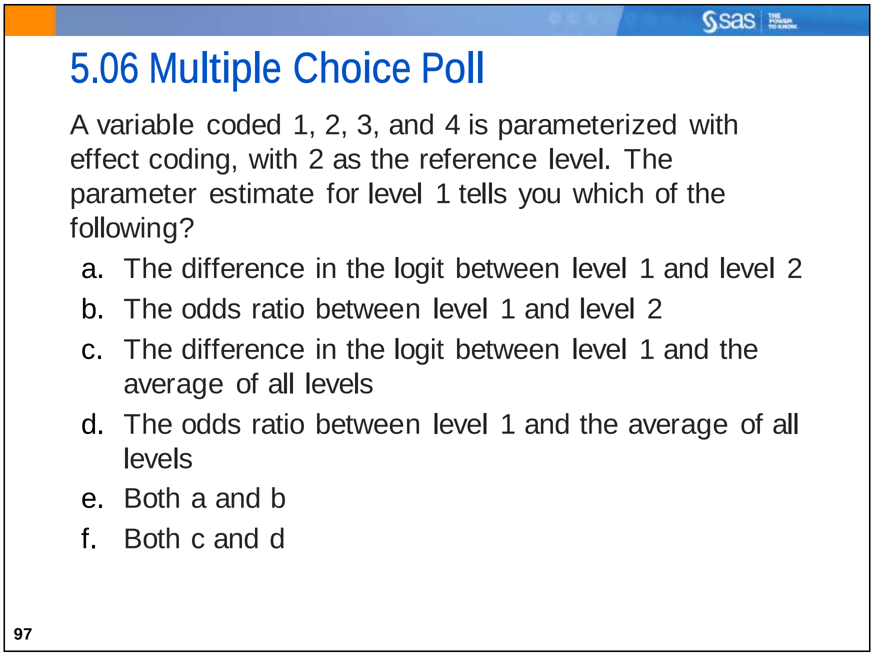
3. Performing a Multiple Logistic Regression Analysis Including Categorical Variables

Fit a logistic regression model using **sasuser.safety** with **Unsafe** as the outcome variable and **Weight**, **Region**, and **Size** as the predictor variables. Request reference cell coding with **Asia** as the reference level for **Region** and **3** (large cars) as the reference level for **Size**. Use the EVENT= option to model the probability of below-average safety scores. Request Profile Likelihood confidence limits and an odds ratio plot along with an effect plot.

a. Do you reject or fail to reject the null hypothesis that all regression coefficients of the model are 0?

b. If you do reject the global null hypothesis, then which predictors significantly predict safety outcome?

c. Interpret the odds ratio for significant predictors.



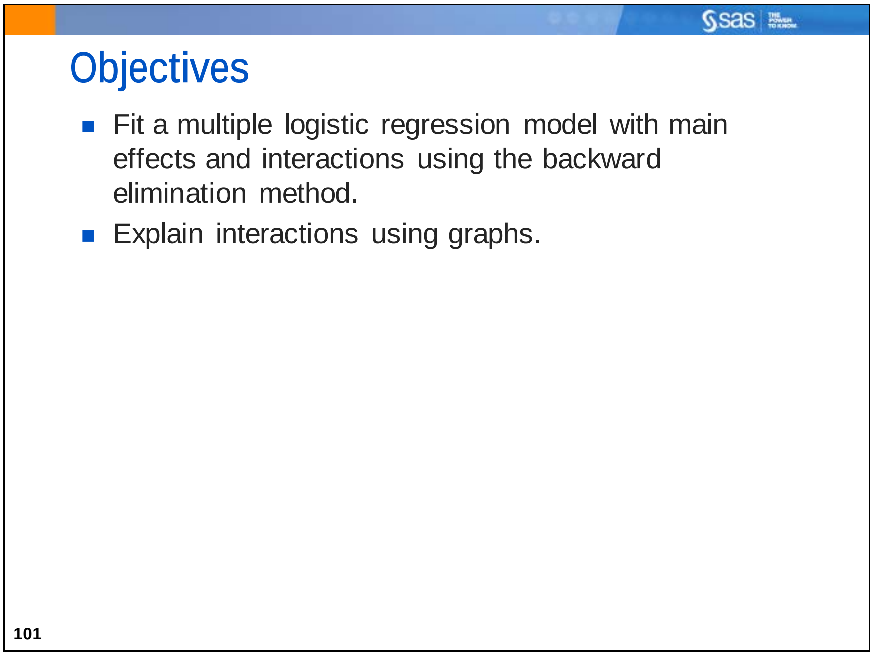
5.06 Multiple Choice Poll

A variable coded 1, 2, 3, and 4 is parameterized with effect coding, with 2 as the reference level. The parameter estimate for level 1 tells you which of the following?

a. The difference in the logit between level 1 and level 2
b. The odds ratio between level 1 and level 2
c. The difference in the logit between level 1 and the average of all levels
d. The odds ratio between level 1 and the average of all levels
e. Both a and b
f. Both c and d

97

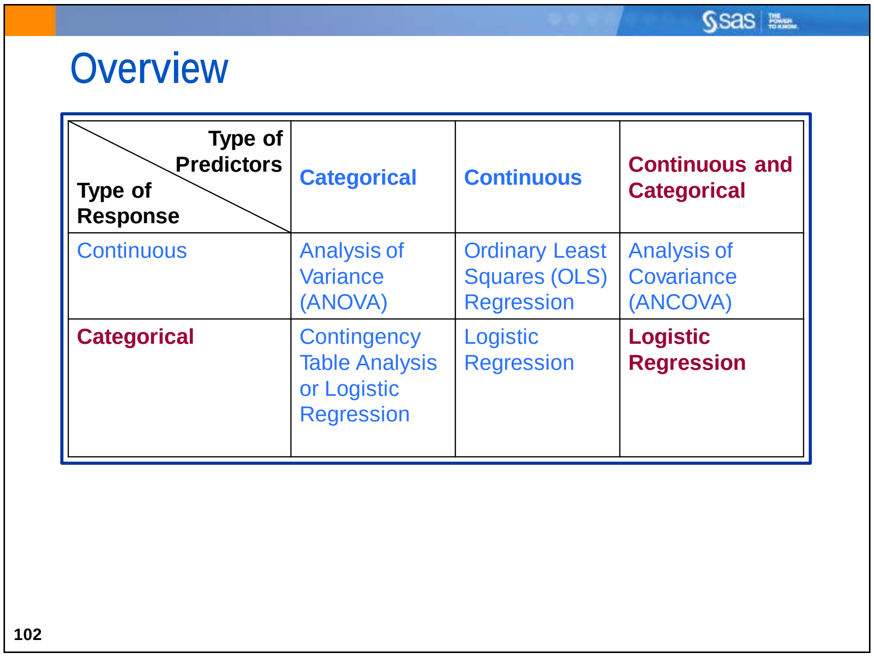
5.5 Stepwise Selection with Interactions



Objectives

- Fit a multiple logistic regression model with main effects and interactions using the backward elimination method.
- Explain interactions using graphs.

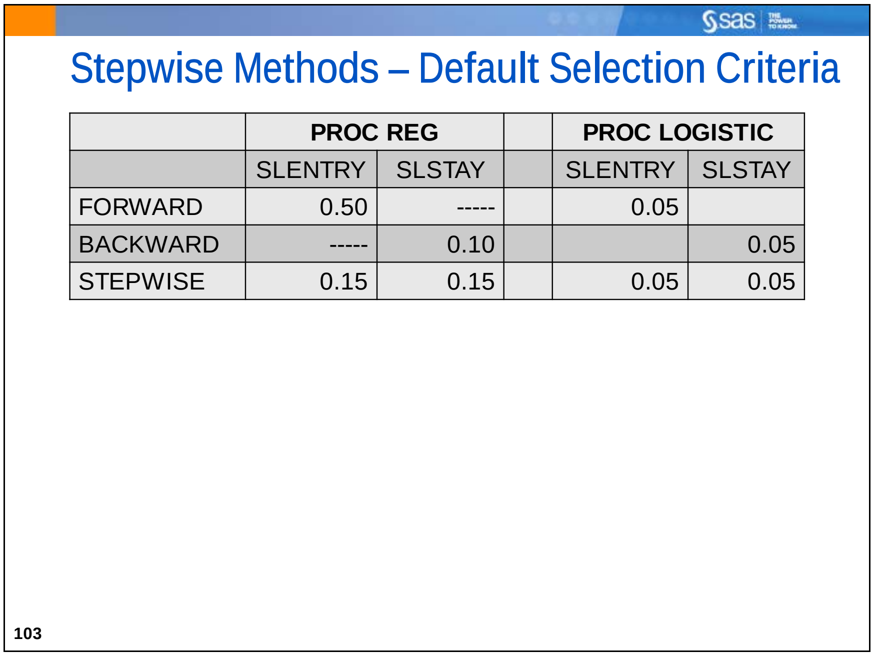
101



Overview

Type of Predictors / Type of Response	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

102

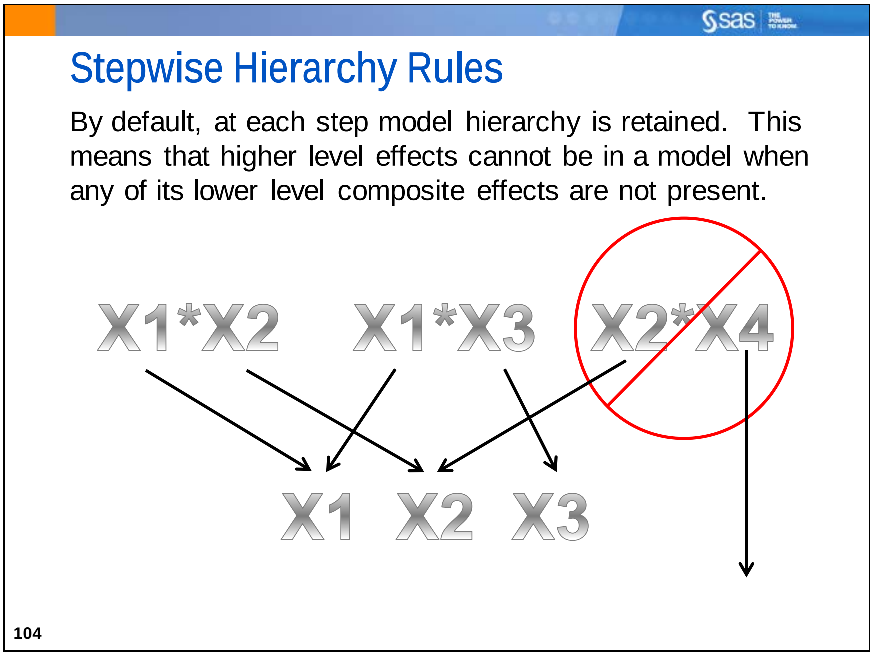


Stepwise Methods – Default Selection Criteria

	PROC REG			PROC LOGISTIC	
	SLENTRY	SLSTAY		SLENTRY	SLSTAY
FORWARD	0.50	-----		0.05	
BACKWARD	-----	0.10			0.05
STEPWISE	0.15	0.15		0.05	0.05

If you are doing exploratory analysis and want to find a best subset model, PROC LOGISTIC provides the three stepwise methods that are available in PROC REG. However, the default selection criteria are not the same as in PROC REG. Remember that you can always change the selection criteria using the SLENTRY= and SLSTAY= options in the MODEL statement.

If you have a large number of variables, you might first need to try a variable reduction method such as variable clustering.



Stepwise Hierarchy Rules

By default, at each step model hierarchy is retained. This means that higher level effects cannot be in a model when any of its lower level composite effects are not present.

X1*X2 X1*X3 X2*X4

X1 X2 X3

104

Model hierarchy refers to the requirement that, for any term to be in the model, all effects contained in the term must be present in the model. For example, in order for the interaction X2*X4 to enter the model, the main effects X2 and X4 must be in the model. Likewise, neither effect X2 nor X4 can leave the model while the interaction X2*X4 is in the model.

When you use the backward elimination method with interactions in the model, PROC LOGISTIC begins by fitting the full model with all the main effects and interactions. PROC LOGISTIC then eliminates the nonsignificant interactions one at a time, starting with the least significant interaction (the one with the largest *p*-value). Next, PROC LOGISTIC eliminates the nonsignificant main effects not involved in any significant interactions. The final model should consist of only significant interactions, the main effects involved in those interactions, and any other significant main effects.

✎ For a more customized analysis, the HIERARCHY= option specifies whether the hierarchy is maintained and whether a single effect or multiple effects are allowed to enter or leave the model in one step for forward, backward, and stepwise selection.

The default is HIERARCHY=SINGLE. You can change this option by inserting the HIERARCHY= option in the MODEL statement. See the *SAS/STAT® 9.3 User's Guide* in the SAS online documentation for more information about using this option. In the LOGISTIC procedure, HIERARCHY=SINGLE is the default, meaning that SAS will not remove a main effect before first removing all interactions involving that main effect.

Logistic Regression: Backward Elimination with Interactions

Example: Fit a multiple logistic regression model using the backward elimination method. The full model should include all the main effects and two-way interactions.

```
/*st105d07.sas*/  /*Part A*/
proc logistic data=sasuser.Titanic plots(only)=(effect oddsratio);
   class Gender(ref='male') Class(ref='3') / param=ref;
   model Survived(event='1')=Age|Gender|Class @2 /
         selection=backward clodds=pl slstay=0.01;
   units age=10;
   title1 'LOGISTIC MODEL (3): Backward Elimination '
          'Survived=Age|Gender|Class';
run;
```

✎ The bar notation with the @2 constructs a model with all the main effects and the two-factor interactions. If you increase it to @3, then you construct a model with all of the main effects, the two-factor interactions, and the three-factor interaction. However, the three-factor interaction might be more difficult to interpret.

✎ A more conservative significance level criterion was used because the sample size was relatively large and the interactions were not hypothesized a priori. Under these circumstances, there is an elevated risk of creating a model with effects that are significant only by chance. Lowering the significance criteria provides some protection.

Selected MODEL statement option:

SELECTION= specifies the method to select the variables in the model. BACKWARD requests backward elimination, FORWARD requests forward selection, NONE fits the complete model specified in the MODEL statement, STEPWISE requests stepwise selection, and SCORE requests best subset selection. The default is NONE.

Model Information	
Data Set	SASUSER.TITANIC
Response Variable	Survived
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1309
Number of Observations Used	1046

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	619
2	1	427

Probability modeled is Survived=1.

Note: 263 observations were deleted due to missing values for the response or explanatory variables.

All information to this point is the same as that from the previous model.

Backward Elimination Procedure

Class Level Information			
Class	Value	Design Variables	
Gender	female	1	
	male	0	
Class	1	1	0
	2	0	1
	3	0	0

The Model Fit Statistics and Testing Global Null Hypothesis tables at Step 0 are presented.

Step 0. The following effects were entered:

Intercept Age Gender Age*Gender Class Age*Class Gender*Class

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1416.620	937.714
SC	1421.573	987.241
-2 Log L	1414.620	917.714

Step 1. Effect Age*Gender is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
		Intercept
	Intercept	and
Criterion	Only	Covariates
AIC	1416.620	940.064
SC	1421.573	984.638
-2 Log L	1414.620	922.064

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	492.5567	8	<.0001
Score	424.3790	8	<.0001
Wald	221.7387	8	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
4.3665	1	0.0367

Step 2. Effect Age*Class is removed:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
		Intercept
	Intercept	and
Criterion	Only	Covariates
AIC	1416.620	945.832
SC	1421.573	980.501
-2 Log L	1414.620	931.832

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	482.7886	6	<.0001
Score	422.4668	6	<.0001
Wald	237.1963	6	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
13.2931	3	0.0040

No (additional) effects met the 0.01 significance level for removal from the model.

The procedure stops after the two interactions involving **Age** are removed.

Summary of Backward Elimination					
Step	**Effect Removed**	**DF**	**Number In**	**Wald Chi-Square**	**Pr > ChiSq**
1	**Age*Gender**	1	5	4.3264	0.0375
2	**Age*Class**	2	4	8.8477	0.0120

The *p*-values associated with the two removed effects were below 0.05 and would therefore be retained, if the default SLSTAY criterion were used.

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Age	1	32.5663	<.0001
Gender	1	40.0553	<.0001
Class	2	44.4898	<.0001
Gender*Class	2	43.9289	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter			**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept			1	-0.6552	0.2113	9.6165	0.0019
Age			1	-0.0385	0.00674	32.5663	<.0001
Gender	**female**		1	1.3970	0.2207	40.0553	<.0001
Class	**1**		1	1.5770	0.2525	38.9980	<.0001
Class	**2**		1	-0.0242	0.2720	0.0079	0.9292
Gender*Class	**female**	**1**	1	2.4894	0.5403	21.2279	<.0001
Gender*Class	**female**	**2**	1	2.5599	0.4562	31.4930	<.0001

Notice that when a CLASS statement is used containing multiple variables, new rows are added to the parameter estimates table. These represent design variables that SAS creates in order to test the interactions.

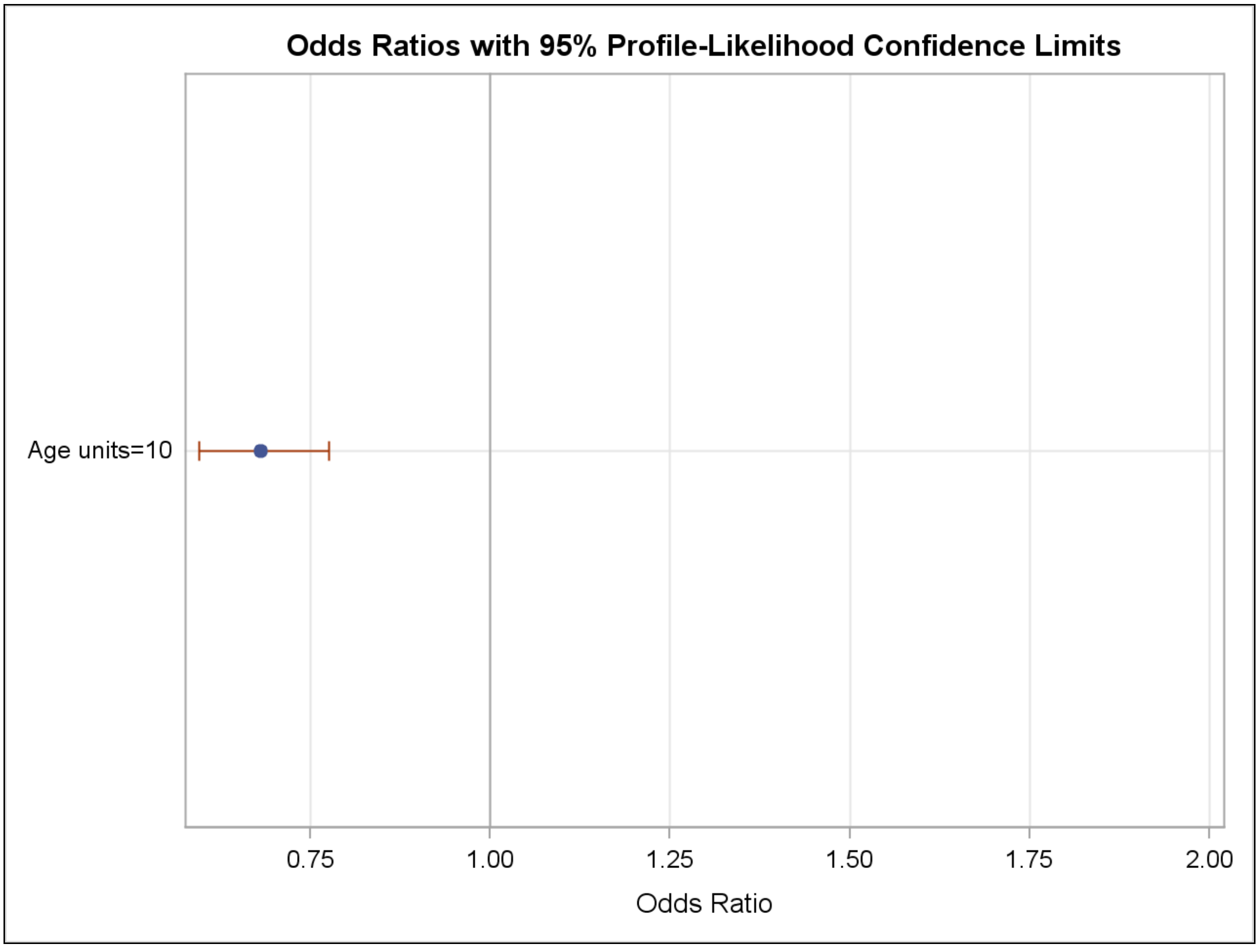
As described in the ANOVA chapter, an interaction between two variables means that the effect of one variable is different at different values of the other variable. This makes the model more complex to interpret.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.0	**Somers' D**	0.703
Percent Discordant	14.7	**Gamma**	0.706
Percent Tied	0.4	**Tau-a**	0.340
Pairs	264313	**c**	0.852

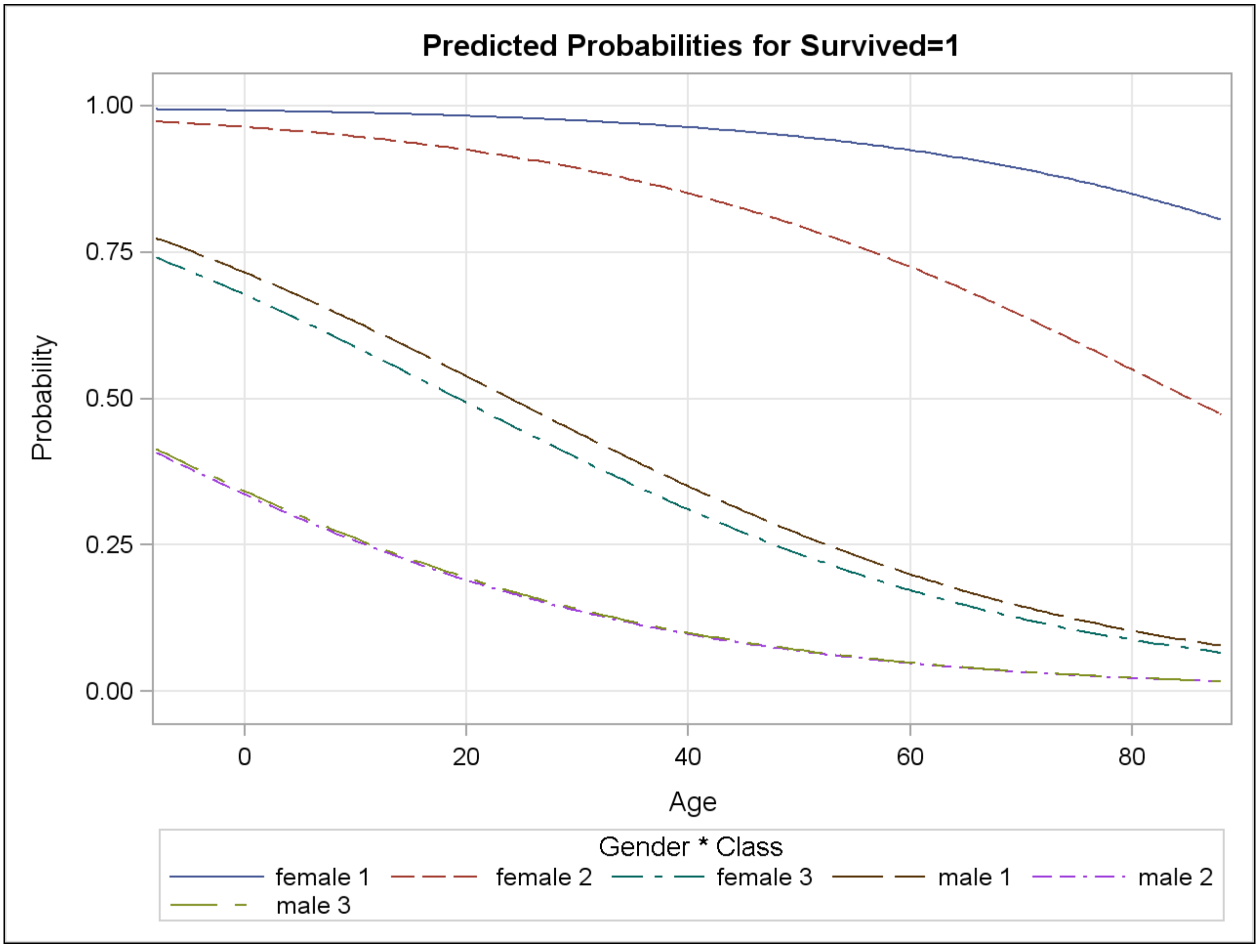
The c value is a slight improvement over the previous model (c=0.840) that only included the main effects.

Odds ratios are not calculated for effects involved in interactions. Any single odds ratio for **Class** or for **Gender** would be misleading because the effects vary for each at different levels of the other variable.

Profile Likelihood Confidence Interval for Odds Ratios				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
Age	10.0000	0.681	0.595	0.775



The odds ratio for **Age** in this model changed only slightly with the addition of the interaction term.



The effect plot shows the interaction. **Male** second class is closer to **male** third class (they have almost identical probabilities at every age) than **male** first class. **Female** second class is much closer to **female** first class than to **female** third class.

In order to estimate and plot odds ratios for the effects involved in an interaction, an ODDSRATIO statement can be used. An EFFECTSPLOT statement can help display the interaction, as well.

```
/*st105d07.sas*/  /*Part B*/
```

```
proc logistic data=sasuser.Titanic
              plots(only)=oddsratio(range=clip);
   class Gender(ref='male') Class(ref='3') / param=ref;
   model Survived(event='1')=Age Gender|Class;
   units age=10;
   oddsratio Gender / at (Class=ALL) cl=pl;
   oddsratio Class / at (Gender=ALL) cl=pl;
   oddsratio Age / cl=pl;
   title1 'LOGISTIC MODEL (3.1): Survived=Age Gender|Class';
run;
```

Selected PROC LOGISTIC statement PLOTS option:

RANGE= with suboptions **(<*min*><,*max*>) | CLIP**, specifies the range of the displayed odds ratio axis. The RANGE=CLIP option has the same effect as specifying the minimum odds ratio as ***min*** and the maximum odds ratio as ***max***. By default, all odds ratio confidence intervals are displayed. This option is helpful when one or more odds ratio confidence intervals are so large that the smaller ones become difficult to see on the scale required to show the larger ones.

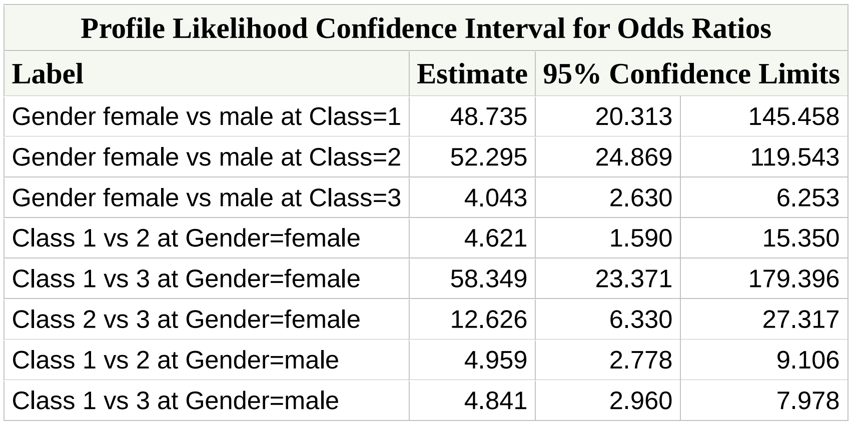
Selected statement:

ODDSRATIO produces odds ratios for a variable even when the variable is involved in interactions with other covariates, and for classification variables that use any parameterization. You can also specify variables on which constructed effects are based, in addition to the names of COLLECTION or MULTIMEMBER effects.

Selected options for the ODDSRATIO statement:

AT specifies fixed levels of the interacting covariates. If a specified covariate does not interact with the variable, then its AT list is ignored. For continuous interacting covariates, you can specify one or more numbers in the value-list. For classification covariates, you can specify one or more formatted levels of the covariate enclosed in single quotation marks (for example, A=‘cat’ ‘dog’), you can specify the keyword REF to select the reference-level, or you can specify the keyword ALL to select all levels of the classification variable. By default, continuous covariates are set to their means, while CLASS covariates are set to ALL.

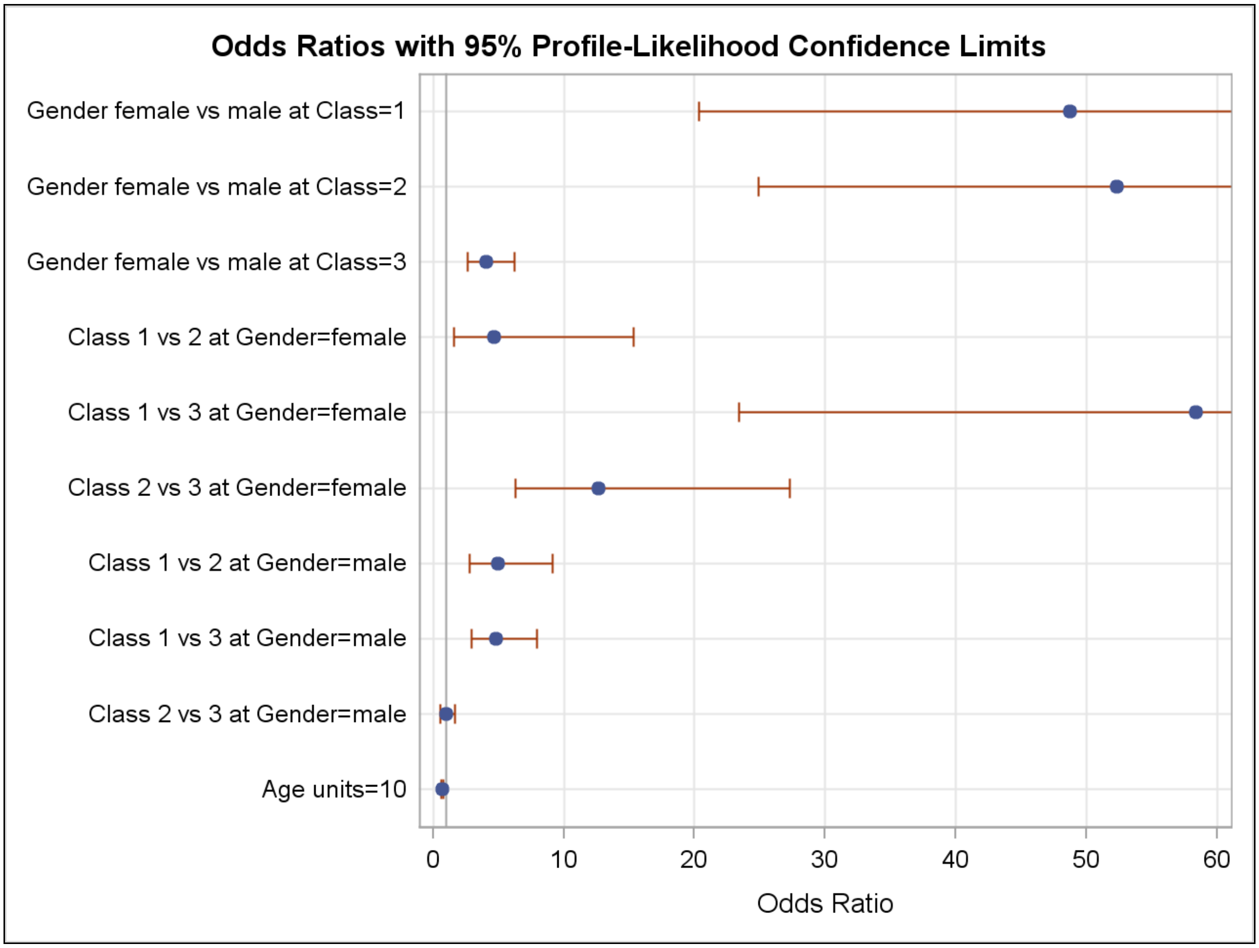
Partial PROC LOGISTIC Output



Profile Likelihood Confidence Interval for Odds Ratios

Label	**Estimate**	**95% Confidence Limits**	
Gender female vs male at Class=1	48.735	20.313	145.458
Gender female vs male at Class=2	52.295	24.869	119.543
Gender female vs male at Class=3	4.043	2.630	6.253
Class 1 vs 2 at Gender=female	4.621	1.590	15.350
Class 1 vs 3 at Gender=female	58.349	23.371	179.396
Class 2 vs 3 at Gender=female	12.626	6.330	27.317
Class 1 vs 2 at Gender=male	4.959	2.778	9.106
Class 1 vs 3 at Gender=male	4.841	2.960	7.978

Profile Likelihood Confidence Interval for Odds Ratios			
Label	**Estimate**	**95% Confidence Limits**	
Class 2 vs 3 at Gender=male	0.976	0.564	1.645
Age units=10	0.681	0.595	0.775



Notice the effect of the RANGE=CLIP suboption. The Odds Ratio axis is clipped just beyond the odds ratio estimate of Class 1 versus 3 at **Gender**=`female`. The upper bound of the associated 95% confidence interval is 179.396. Even with the range clipped, the **Age** confidence interval is barely discernible.

From this plot it is clear that the gender effect is different at different classes. (There is a stark difference in first and second classes, but not nearly as much in third class.) The class differences are also evidently different at each gender. For example, the odds ratio comparing first and third classes for females is far greater than the odds ratio comparing those same classes for males.

Exercises

4. Performing Backward Elimination

Using the **sasuser.safety** data set, run PROC LOGISTIC and use backward elimination. Start with a model using ***only main effects***.

Use **Unsafe** as the outcome variable and **Weight**, **Size**, and **Region** as the predictor variables.
Use the EVENT= option to model the probability of below-average safety scores.

Use the SIZEFMT format for the variable **Size**.

Specify **Region** and **Size** as classification variables using reference cell coding and specify **`Asia`** as the reference level for **Region** and **`Small`** as the reference level for **Size**.

Use a UNITS statement with -1 as the units for weight, so that you can see the odds ratio for lighter cars over heavier cars. Request any relevant plots.

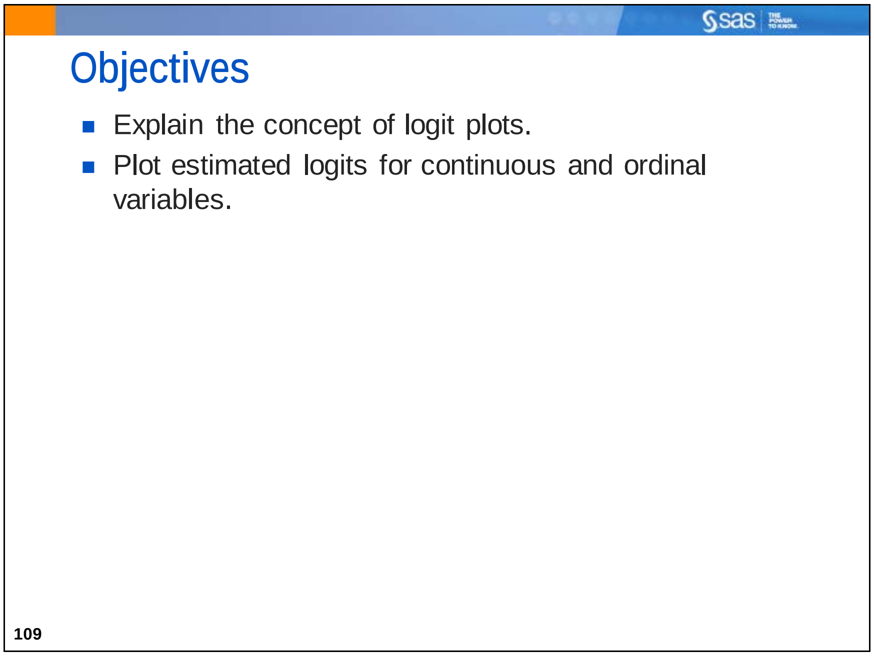
a. Which terms appear in the final model?

b. Do you think this is a better model than the one fit with only **Region**?

✎ The variable **Size** is coded (1, 2, 3), but the applied format requires that the formatted value be used in the CLASS statement for the REF= category.

```
   value sizefmt 1='Small'
                 2='Medium'
                 3='Large';
```

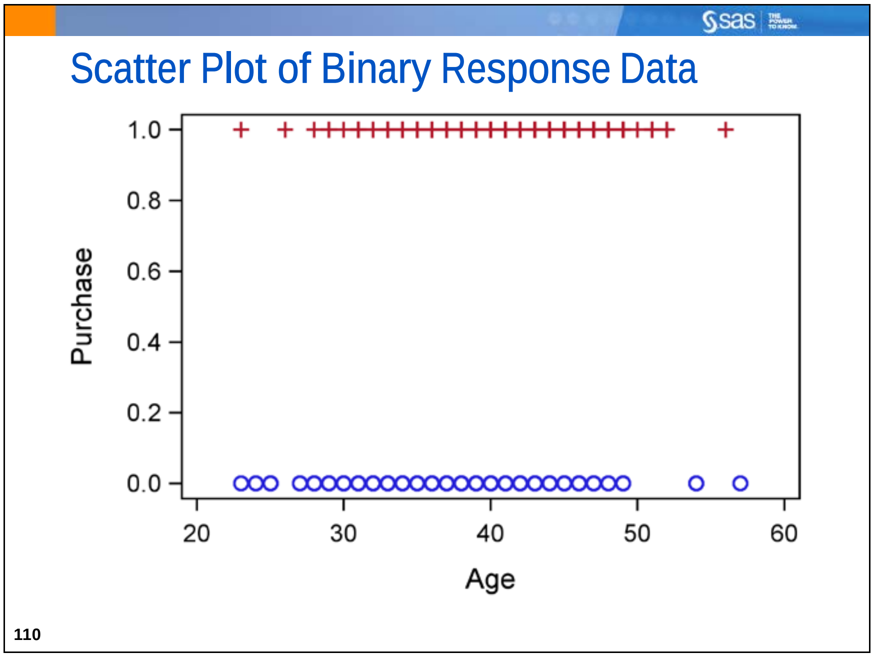
5.6 Logit Plots (Self-Study)



Objectives

- Explain the concept of logit plots.
- Plot estimated logits for continuous and ordinal variables.

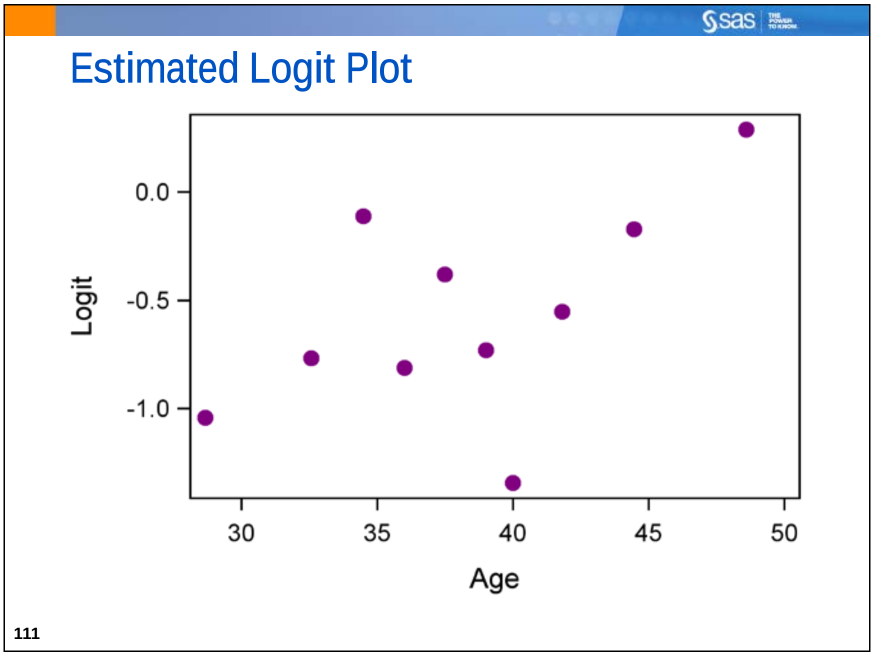
109



Scatter Plot of Binary Response Data

110

For continuous data, a recommended step before building a regression model is to analyze the bivariate relationships between the regressors and the response variables. The goal is not only to detect outliers, but also to analyze the shape of the relationships to determine whether there might be some nonlinear trend that should be modeled in the analysis. For binary response variables, a scatter plot contributes little to these ends.

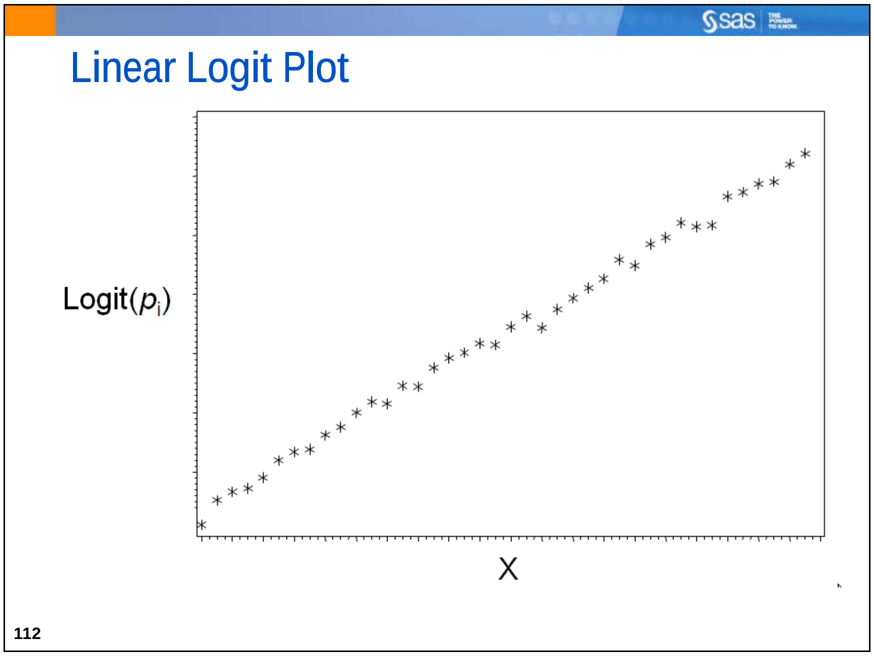


Estimated Logit Plot

111

The logistic model asserts a linear relationship with the logit (not with the actual binary values). However, a logit for one observation will be infinite in either the positive or negative direction (ln(*p*/(1–*p*))=ln(1/0) or ln(0/1)). A recommendation, however, is to group the data into approximately equally sized bins, based on the values of the predictor variable. The bin size should be adequate in number of observations to reduce the sample variability of the logits. You can then assume that the average probability within each bin is approximately the value of the proportion in the bin with the event. The estimated logit is then approximately equal to ln(proportion/(1–proportion)).

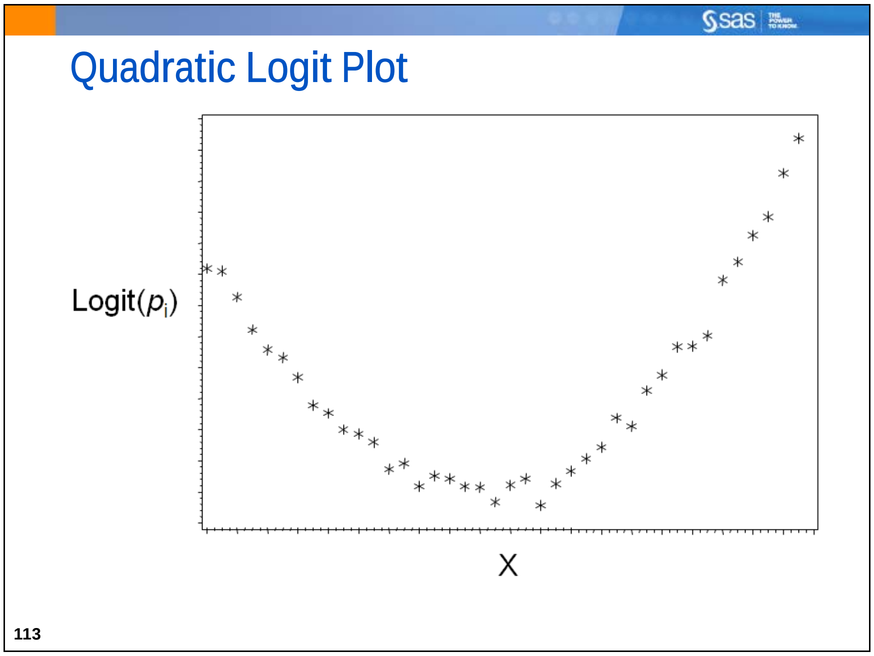
✎ If the predictor variable is a nominal variable, then there is no need to create a logit plot.



Linear Logit Plot

112

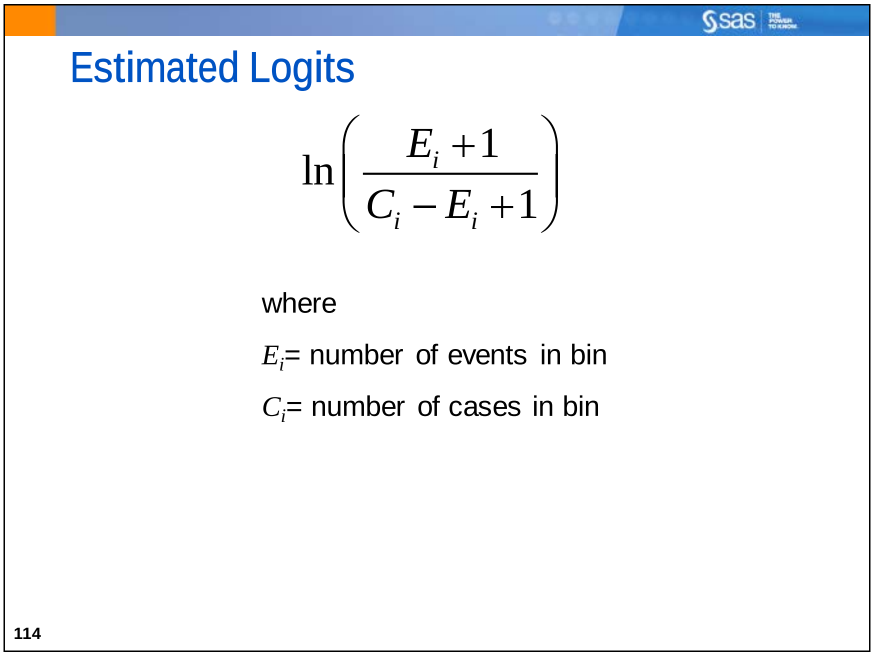
If the standard logistic regression model adequately fits the data, the logit plots should be fairly linear. The above graph shows a predictor variable that meets the assumption of linearity in the logit.



Quadratic Logit Plot

113

The logit plot can also show serious nonlinearities between the outcome variable and the predictor variable. The above graph reveals a quadratic relationship between the outcome and predictor variables. Adding a polynomial term or binning the predictor variable into three groups (two dummy variables would model the quadratic relationship) and treating it as a classification variable can improve the model fit.



Estimated Logits

$$\ln\left(\frac{E_i+1}{C_i-E_i+1}\right)$$

where

E_i= number of events in bin

C_i= number of cases in bin

114

A common approach when computing logits is to take the log of the odds. The path from the definition of a logit to the formula above is shown below. *C* represents the total number in the bin and *E* represents the total number of positive events in the bin.

$$\left(\frac{P_i}{(1-P_i)}\right)=\left(\frac{\frac{E_i}{C_i}}{\left(\frac{C_i}{C_i}-\frac{E_i}{C_i}\right)}\right)=\left(\frac{E_i}{(C_i-E_i)}\right)$$

The logit is undefined for any bin in which the outcome rate is 100% or 0%. To eliminate this problem and reduce the variability of the logits, a common recommendation is to add a small constant to the numerator and denominator of the formula that computes the logit (Santner and Duffy 1989).

Plotting Estimated Logits

Example: Plot the estimated logits of the outcome variable **Survived** versus the predictor variable **Class**. To construct the estimated logits, the number of passengers who survived and the total number of customers by each level of **Class** must be computed.

```
/*st105d08.sas*/  /*Part A*/
proc means data=sasuser.Titanic noprint nway;
   class Class;
   var Survived;
   output out=bins sum(Survived)=NEvent n(Survived)=NCases;
run;

data bins;
   set bins;
   Logit=log((NEvent+1)/(NCases-NEvent+1));
run;

proc sgplot data=bins;
   reg Y=Logit X=Class /
       markerattrs=(symbol=asterisk color=blue size=15);
   pbspline Y=Logit X=Class / nomarkers;
   xaxis integer;
   title "Estimated Logit Plot of Passenger Class";
run;
quit;
```

Selected PROC MEANS statement option:

NWAY causes the output data set to have only one observation for each level of the class variable.

Selected PROC SGPLOT statements:

REG creates a fitted regression line or curve.

PBSPLINE creates a fitted penalized B-spline curve.

Selected options for REG or PBSPLINE statements:

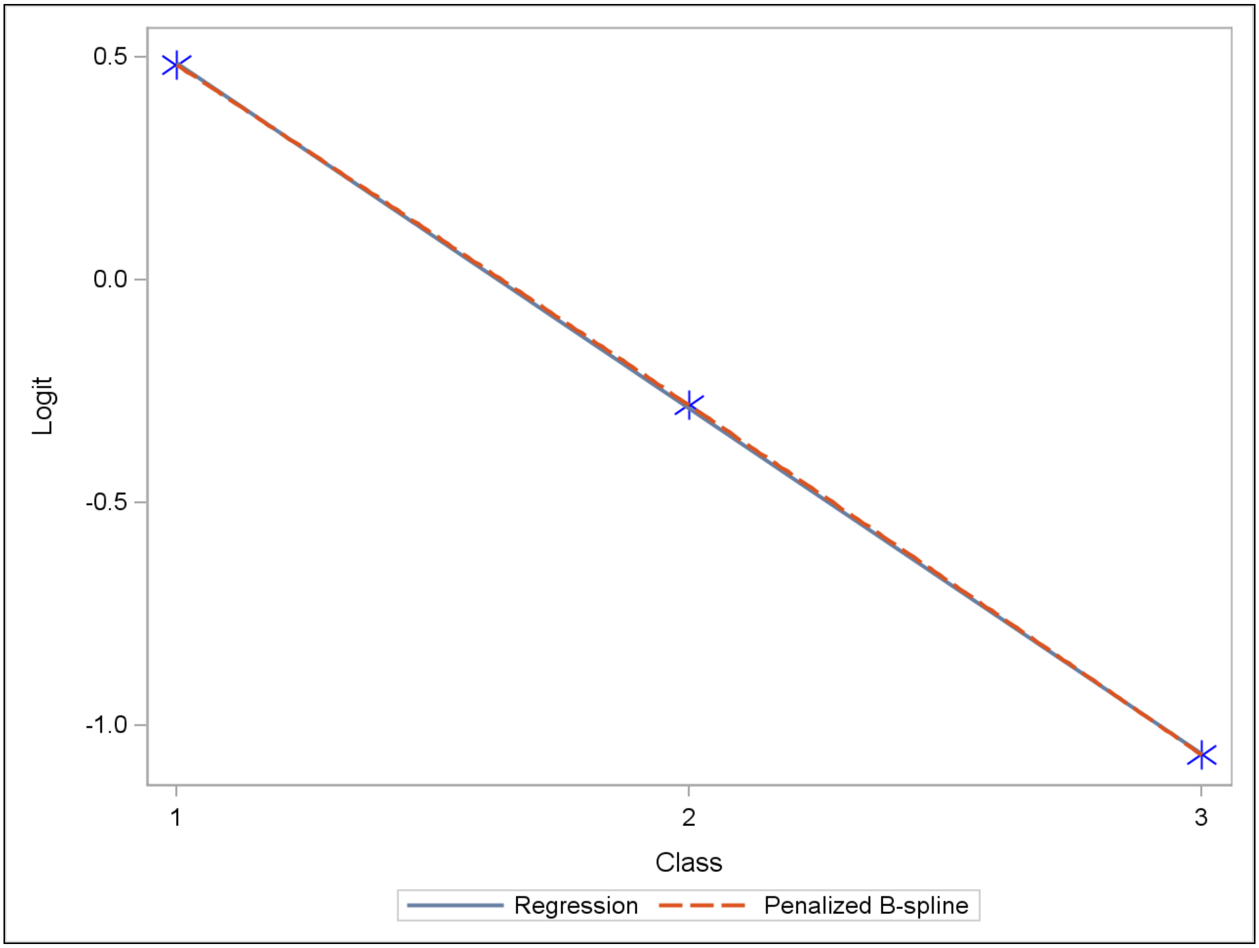
MARKERATTRS controls the display of the marker values for data points on the plot. SIZE is measured in pixels.

NOMARKERS removes the scatter markers from the plot.

PROC MEANS creates a data set that contains a separate value for the requested statistics for each level of the CLASS variable. Because **Survived** is coded 0/1, SUM(**Survived**) returns the value for the count of ones within each level of **CLASS**. N(**Survived**) returns the number of nonmissing values of **Survived**, which is the total effective sample size within each level.

The logit is created in the DATA step, using the formula seen in the slide shown previously. In this case, **C** is represented by **NCases** and **E** is represented by **NEvent**.

PROC SGPLOT shows the data, a regression line, and a penalized B-spline curve. The regression line and the curve can be compared to each other to assess the linearity of the relationship between the logit and the predictor. If the curve approximates the regression line, this gives evidence of linearity.



The logit plot for this ordinal variable is almost perfectly linear.

When a linear pattern is detected in a logit plot for an ordinal variable, the variable can be removed from the CLASS statement, implying that it would be considered the same as a continuous variable. The statistical advantage of doing so would be to increase model power, due to obtaining almost the same information using fewer degrees of freedom. However, theoretical justifications should always supersede such data-driven considerations.

Example: Plot the estimated logits of the outcome variable **Survived** versus the predictor variable **Age**. Because **Age** is a continuous variable, bin the observations into 50 groups to ensure that an adequate number of observations are used to compute the estimated logit.

```
/*st105d08.sas*/  /*Part B*/
proc rank data=sasuser.Titanic groups=50 out=Ranks;
   var Age;
   ranks Rank;
run;

proc means data=Ranks noprint nway;
   class Rank;
   var Survived Age;
   output out=Bins sum(Survived)=NEvent n(Survived)=NCases
          mean(Age)=Age;
run;

data bins;
   set bins;
   Logit=log((NEvent+1)/(NCases-NEvent+1));
run;

proc sgplot data=bins;
   reg Y=Logit X=Age /
       markerattrs=(symbol=asterisk color=blue size=15);
   pbspline Y=Logit X=Age / nomarkers;
   title "Estimated Logit Plot of Passenger's Age";
run;
quit;
```

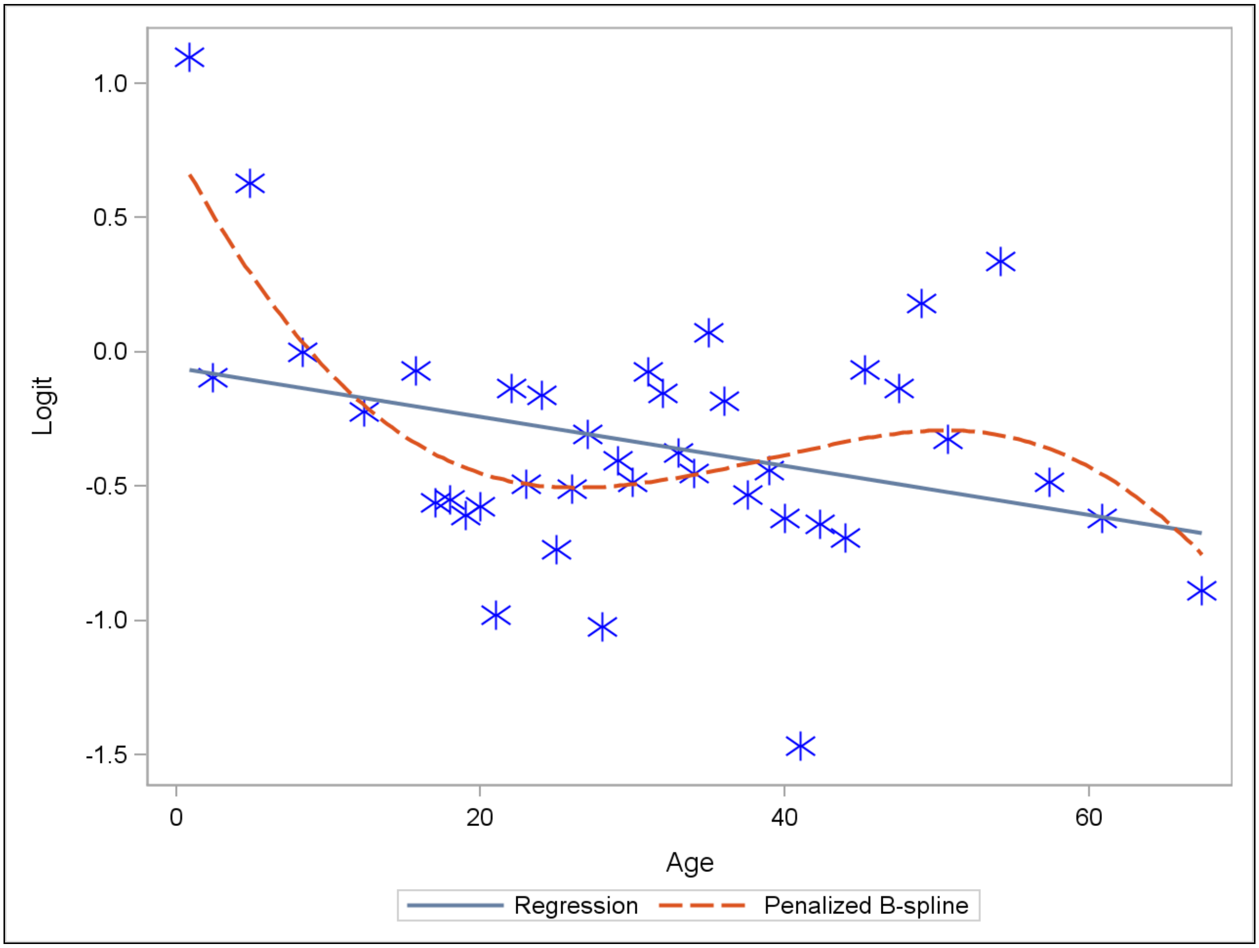
Selected PROC RANK statement option:

GROUPS=*n*   bins the variables into *n* groups.

Selected RANK procedure statement:

RANKS   names the group indicators in the OUT= data set. If the RANKS statement is omitted, then the group indicators replace the VAR variables in the OUT= data set.

In the case of **Age**, you do not have a made-to-order bin variable, so you must create one. You can use the RANK procedure for this purpose. You have 1309 observations. It is recommended that you have approximately 20 to 30 observations per bin. At approximately 26 per bin, you could create 1309/26~50 bins. That will be the option value of GROUPS=.



The estimated logit plot shows a deviation from linearity. One possibility is to add a quadratic (squared) or cubic (cubed) term for **Age**.

The estimated logit plot is a univariate plot and therefore can be misleading in the presence of interactions and partial associations. (Association between the response variable and the predictor variable changes with the addition of another predictor variable in the model.) If an interaction is suspected, a model with the interaction term and main effects should be evaluated before any variable is eliminated. Estimated logit plots should never be used to eliminate variables from consideration for a multiple logistic regression model.

5.7 Solutions

Solutions to Exercises

1. **Performing Tests and Measures of Association**

   An insurance company wants to relate the safety of vehicles to several other variables. A score is given to each vehicle model, using the frequency of insurance claims as a basis. The data are in the **sasuser.safety** data set.

   a. Invoke the FREQ procedure and create one-way frequency tables for the categorical variables.

```
/*st105s01.sas*/  /*Part A*/
ods graphics off;
proc freq data=sasuser.safety;
   tables Unsafe Type Region Size;
   title "Safety Data Frequencies";
run;
ods graphics on;
```

Unsafe	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	66	68.75	66	68.75
1	30	31.25	96	100.00

Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Large	16	16.67	16	16.67
Medium	29	30.21	45	46.88
Small	20	20.83	65	67.71
Sport/Utility	16	16.67	81	84.38
Sports	15	15.63	96	100.00

Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	35	36.46	35	36.46
N America	61	63.54	96	100.00

Size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	35	36.46	35	36.46
2	29	30.21	64	66.67
3	32	33.33	96	100.00

1) What is the measurement scale of each variable?

Variable	Measurement Scale
Unsafe	**Nominal, Ordinal, Binary**
Type	**Nominal**
Region	**Nominal**
Weight	**Ratio (Continuous)**
Size	**Ordinal**

2) What is the proportion of cars made in North America?

63.54 %

3) For the variables **Unsafe**, **Size**, **Region**, and **Type**, are there any unusual data values that warrant further investigation?

No.

b. Use PROC FREQ to examine the crosstabulation of the variables **Region** by **Unsafe**. Generate a temporary format to clearly identify the values of **Unsafe**. Along with the default output, generate the expected frequencies, the chi-square test of association and the odds ratio.

```
/*st105s01.sas*/  /*Part B*/
proc format;
   value safefmt 0='Average or Above'
                 1='Below Average';
run;

proc freq data=sasuser.safety;
   tables Region*Unsafe / expected chisq relrisk;
   format Unsafe safefmt.;
   title "Association between Unsafe and Region";
run;
```

Table of Region by Unsafe			
Region	**Unsafe**		
Frequency **Expected** **Percent** **Row Pct** **Col Pct**	**Average or Above**	**Below Average**	**Total**
Asia	20 24.063 20.83 57.14 30.30	15 10.938 15.63 42.86 50.00	35 36.46
N America	46 41.938 47.92 75.41 69.70	15 19.063 15.63 24.59 50.00	61 63.54
Total	66 68.75	30 31.25	96 100.00

Statistics for Table of Region by Unsafe

Statistic	DF	Value	Prob
Chi-Square	1	3.4541	0.0631
Likelihood Ratio Chi-Square	1	3.3949	0.0654
Continuity Adj. Chi-Square	1	2.6562	0.1031
Mantel-Haenszel Chi-Square	1	3.4181	0.0645
Phi Coefficient		-0.1897	
Contingency Coefficient		0.1864	
Cramer's V		-0.1897	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	20
Left-sided Pr <= F	0.0525
Right-sided Pr >= F	0.9809
Table Probability (P)	0.0334
Two-sided Pr <= P	0.0718

Estimates of the Relative Risk (Row1/Row2)			
Type of Study	**Value**	**95% Confidence Limits**	
Case-Control (Odds Ratio)	0.4348	0.1790	1.0562
Cohort (Col1 Risk)	0.7578	0.5499	1.0443
Cohort (Col2 Risk)	1.7429	0.9733	3.1210

1) For the cars made in Asia, what percentage had a below-average safety score?

Region is a row variable, so look at the Row Pct value in the `Below Average` cell of the `Asia` row. That value is 42.86.

2) For the cars with an average or above safety score, what percentage was made in North America?

The Col Pct value for the cell for `North America` in the column for `Average or Above` is 69.70.

3) Do you see a statistically significant (at the 0.05 level) association between **Region** and **Unsafe**?

The association is not statistically significant at the 0.05 alpha level. The *p*-value is 0.0631.

4) What does the odds ratio compare and what does this one say about the difference in odds between Asian and North American cars?

The odds ratio compares the odds of below average safety for North America versus Asia. The odds ratio of 0.4348 means that cars made in North America have 56.52 percent lower odds for being unsafe than cars made in Asia.

Recall that odds ratios given in the Estimates of Relative Risk table are calculated comparing row1/row2 for column1. In this problem, this comparison is **`Asia`** to **`N America`** whose outcome is **`Average or Above`** in safety. The value 0.4348 is interpreted as the odds of having an **`Average or Above`** car made in **`Asia`** is 0.4348 times the odds for American-made cars. If you wished to compare **`N America`** to **`Asia`**, still using **`Average or Above`** for safety, the odds ratio would be the inverse of 0.4348, or approximately 2.3. This is interpreted as cars made in North America have 2.3 times the odds for being safe than cars made in Asia. This single inversion would also create the odds ratio for comparing **`Asia`** to **`N America`** but **`Below Average`** in safety. If you wished to compare **`N America`** to **`Asia`** using **`Below Average`** in safety, you would invert your odds ratio twice returning to the value 0.4348.

c. Use the variable named **Size**. Examine the ordinal association between **Size** and **Unsafe**. Use PROC FREQ.

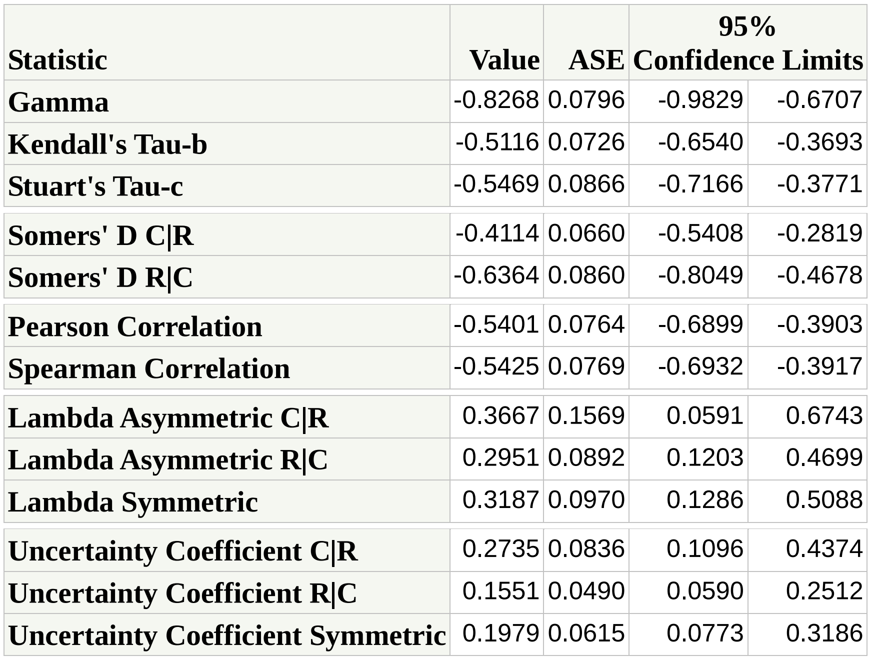
```
/*st105s01.sas*/  /*Part C*/
proc freq data=sasuser.safety;
   tables Size*Unsafe / chisq measures cl;
   format Unsafe safefmt.;
   title "Association between Unsafe and Size";
run;
```

Table of Size by Unsafe

Size	Unsafe		
Frequency Percent Row Pct Col Pct	Average or Above	Below Average	Total
1	12 12.50 34.29 18.18	23 23.96 65.71 76.67	35 36.46
2	24 25.00 82.76 36.36	5 5.21 17.24 16.67	29 30.21
3	30 31.25 93.75 45.45	2 2.08 6.25 6.67	32 33.33
Total	66 68.75	30 31.25	96 100.00

Statistics for Table of Size by Unsafe

Statistic	DF	Value	Prob
Chi-Square	2	31.3081	<.0001
Likelihood Ratio Chi-Square	2	32.6199	<.0001
Mantel-Haenszel Chi-Square	1	27.7098	<.0001
Phi Coefficient		0.5711	
Contingency Coefficient		0.4959	
Cramer's V		0.5711	



Statistic	Value	ASE	95% Confidence Limits	
Gamma	-0.8268	0.0796	-0.9829	-0.6707
Kendall's Tau-b	-0.5116	0.0726	-0.6540	-0.3693
Stuart's Tau-c	-0.5469	0.0866	-0.7166	-0.3771
Somers' D C|R	-0.4114	0.0660	-0.5408	-0.2819
Somers' D R|C	-0.6364	0.0860	-0.8049	-0.4678
Pearson Correlation	-0.5401	0.0764	-0.6899	-0.3903
Spearman Correlation	-0.5425	0.0769	-0.6932	-0.3917
Lambda Asymmetric C|R	0.3667	0.1569	0.0591	0.6743
Lambda Asymmetric R|C	0.2951	0.0892	0.1203	0.4699
Lambda Symmetric	0.3187	0.0970	0.1286	0.5088
Uncertainty Coefficient C|R	0.2735	0.0836	0.1096	0.4374
Uncertainty Coefficient R|C	0.1551	0.0490	0.0590	0.2512
Uncertainty Coefficient Symmetric	0.1979	0.0615	0.0773	0.3186

1) What statistic should you use to detect an ordinal association between **Size** and **Unsafe**?

   The Mantel-Haenszel Chi-Square

2) Do you reject or fail to reject the null hypothesis at the 0.05 level?

   Reject

3) What is the strength of the ordinal association between **Size** and **Unsafe**?

   The Spearman correlation is -0.5425.

4) What is the 95% confidence interval around that statistic?

   The CI is (-0.6932, -0.3917).

2. Performing a Logistic Regression Analysis

Fit a simple logistic regression model using **sasuser.safety** with **Unsafe** as the outcome variable and **Weight** as the predictor variable. Use the EVENT= option to model the probability of below-average safety scores. Request Profile Likelihood confidence limits and an odds ratio plot along with an effect plot.

```
/*st105s02.sas*/
proc logistic data=sasuser.safety plots(only)=(effect oddsratio);
   model Unsafe(event='1')=Weight / clodds=pl;
   title1 'LOGISTIC MODEL (1):Unsafe=Weight';
run;
```

Model Information	
Data Set	SASUSER.SAFETY
Response Variable	Unsafe
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	96
Number of Observations Used	96

Response Profile		
Ordered Value	**Unsafe**	**Total Frequency**
1	0	66
2	1	30

Probability modeled is Unsafe=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

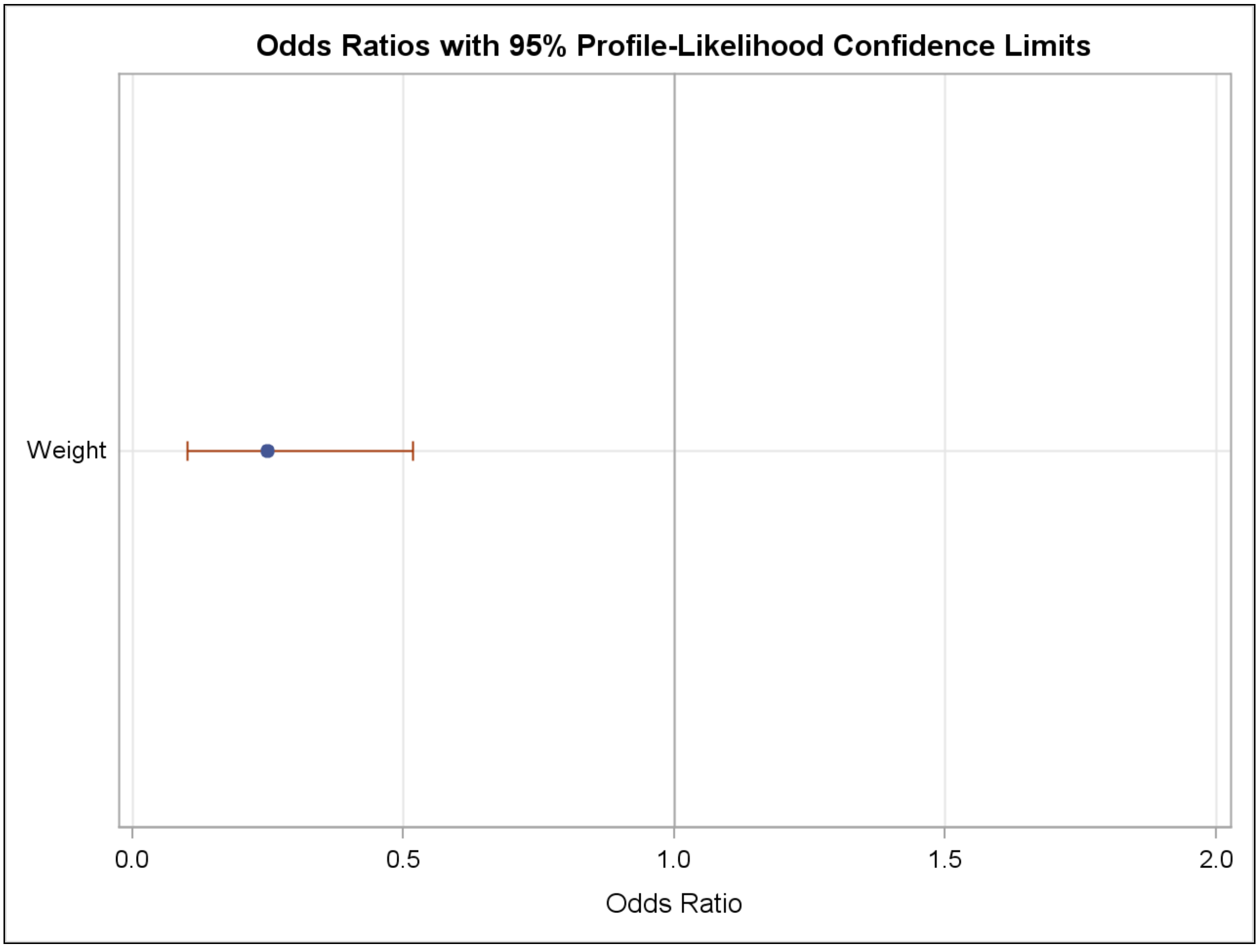
Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	121.249	106.764
SC	123.813	111.893
-2 Log L	119.249	102.764

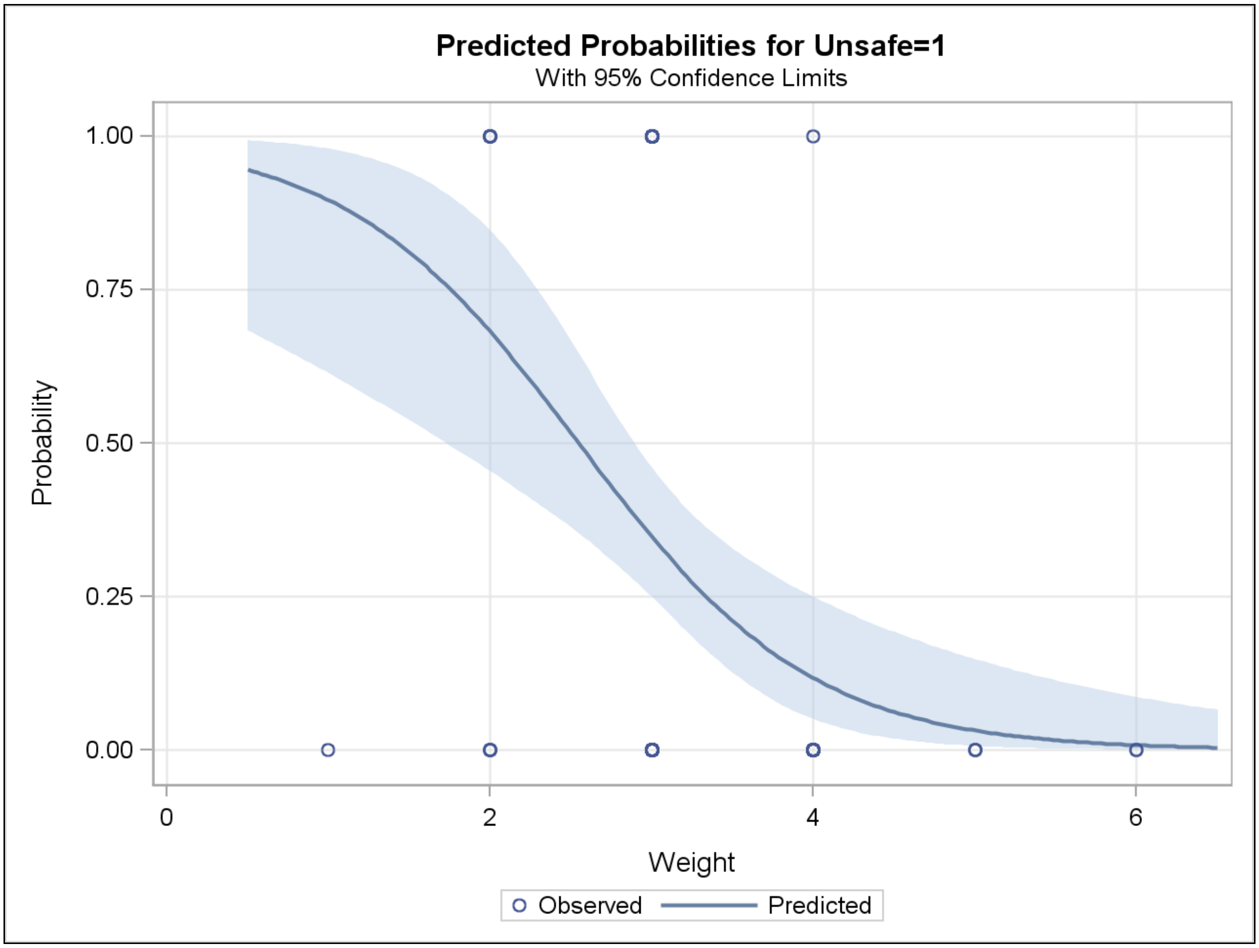
Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	16.4845	1	<.0001
Score	13.7699	1	0.0002
Wald	11.5221	1	0.0007

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	3.5422	1.2601	7.9023	0.0049
Weight	1	-1.3901	0.4095	11.5221	0.0007

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	55.2	**Somers' D**	0.474
Percent Discordant	7.7	**Gamma**	0.754
Percent Tied	37.1	**Tau-a**	0.206
Pairs	1980	**c**	0.737

Profile Likelihood Confidence Interval for Odds Ratios				
Effect	**Unit**	**Estimate**	**95% Confidence Limits**	
Weight	1.0000	0.249	0.102	0.517





a. Do you reject or fail to reject the global null hypothesis that all regression coefficients of the model are 0?

The *p*-value for the Likelihood Ratio test is <.0001 and therefore the global null hypothesis is rejected.

b. Write the logistic regression equation.

The regression equation is as follows:

Logit(Unsafe)=3.5422 + (-1.3901)*Weight.

c. Interpret the odds ratio for **Weight**.

The odds ratio for Weight (0.249) says that the odds for being unsafe (having a below average safety rating) are 75.1% lower for each thousand pound increase in weight. The confidence interval (0.102 , 0.517) does not contain 1, indicating that that the odds ratio is statistically significant.

3. Performing a Multiple Logistic Regression Analysis Including Categorical Variables

Fit a logistic regression model using **sasuser.safety** with **Unsafe** as the outcome variable and **Weight**, **Region**, and **Size** as the predictor variables. Request reference cell coding with **Asia** as the reference level for **Region** and **3** (large cars) as the reference level for **Size**. Use the EVENT= option to model the probability of below-average safety scores. Request Profile Likelihood confidence limits and an odds ratio plot along with an effect plot.

```
/*st105s03.sas*/
proc logistic data=sasuser.safety plots(only)=(effect oddsratio);
   class Region (param=ref ref='Asia')
         Size (param=ref ref='3');
   model Unsafe(event='1')=Weight Region Size / clodds=pl;
   title1 'LOGISTIC MODEL (2):Unsafe=Weight Region Size';
run;
```

Partial PROC LOGISTIC Output

Class Level Information

Class	Value	Design Variables	
Region	Asia	0	
	N America	1	
Size	1	1	0
	2	0	1
	3	0	0

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	121.249	94.004
SC	123.813	106.826
-2 Log L	119.249	84.004

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	35.2441	4	<.0001
Score	32.8219	4	<.0001
Wald	23.9864	4	<.0001

a. Do you reject or fail to reject the null hypothesis that all regression coefficients of the model are 0?

You reject the null hypothesis with a p<.0001.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Weight	1	2.1176	0.1456
Region	1	0.4506	0.5020
Size	2	15.3370	0.0005

b. If you do reject the global null hypothesis, then which predictors significantly predict safety outcome?

Only Size is significantly predictive of Unsafe.

Analysis of Maximum Likelihood Estimates

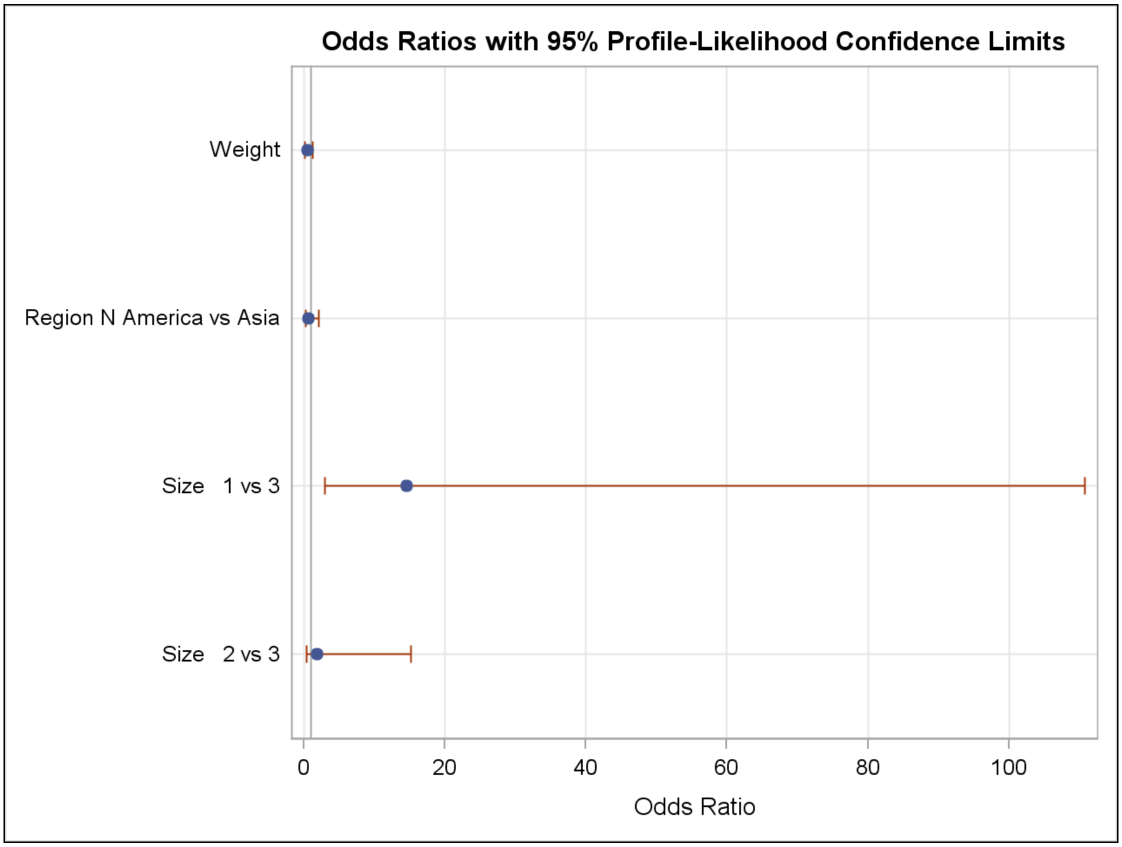
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.0500	1.8008	0.0008	0.9778
Weight		1	-0.6678	0.4589	2.1176	0.1456
Region	**N America**	1	-0.3775	0.5624	0.4506	0.5020
Size	**1**	1	2.6783	0.8810	9.2422	0.0024
Size	**2**	1	0.6582	0.9231	0.5085	0.4758

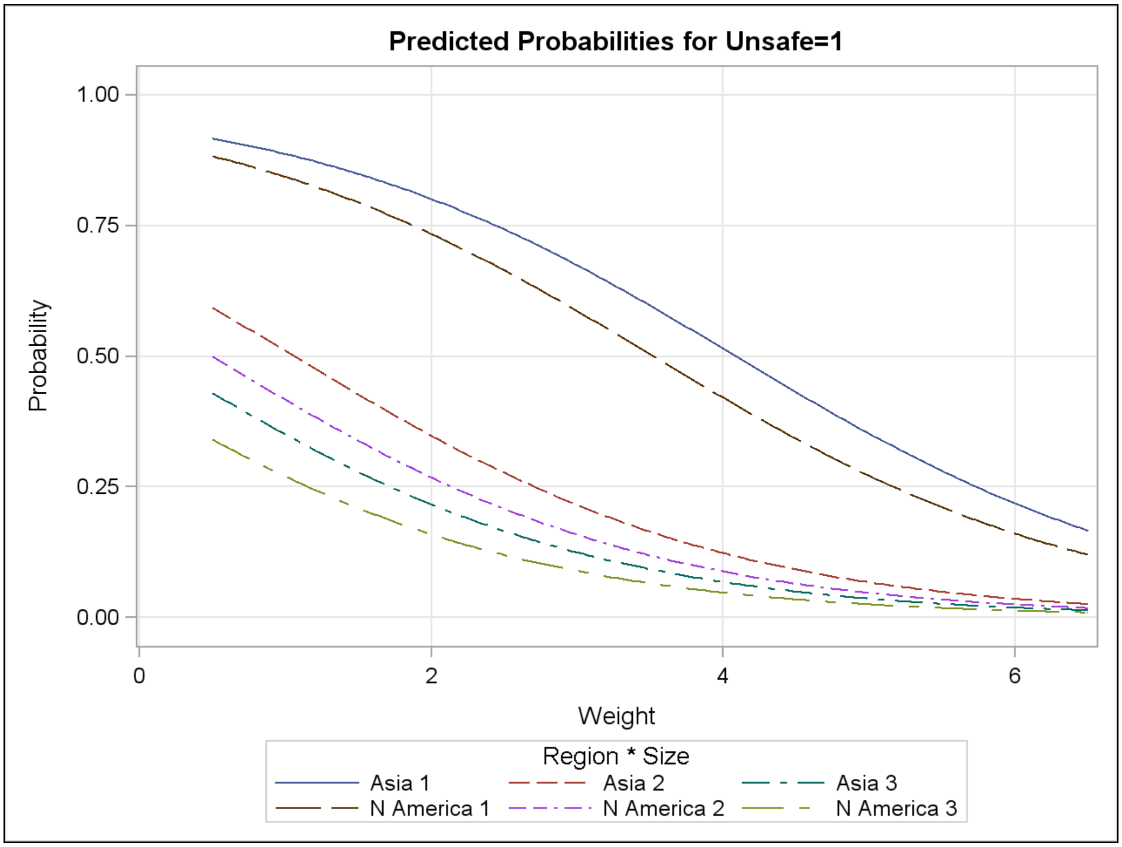
Association of Predicted Probabilities and Observed Responses

Percent Concordant	81.9	**Somers' D**	0.696
Percent Discordant	12.3	**Gamma**	0.739
Percent Tied	5.8	**Tau-a**	0.302
Pairs	1980	**c**	0.848

Profile Likelihood Confidence Interval for Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
Weight	1.0000	0.513	0.201	1.260
Region N America vs Asia	1.0000	0.686	0.225	2.081
Size 1 vs 3	1.0000	14.560	3.018	110.732
Size 2 vs 3	1.0000	1.931	0.343	15.182

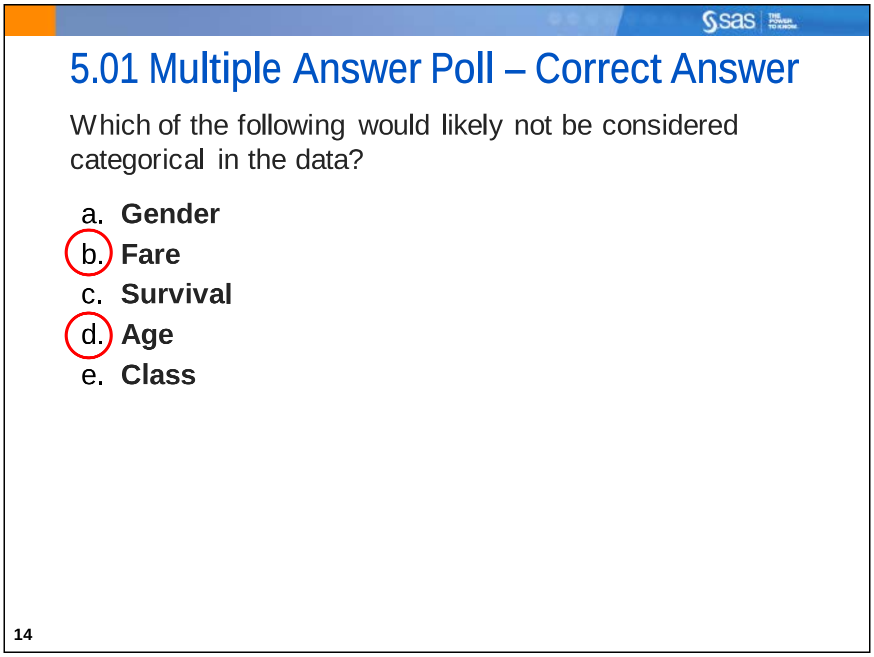

Odds Ratios with 95% Profile-Likelihood Confidence Limits
Weight
Region N America vs Asia
Size 1 vs 3
Size 2 vs 3
0
20
40
60
80
100
Odds Ratio


Predicted Probabilities for Unsafe=1
1.00
0.75
0.50
0.25
0.00
Probability
0
2
4
6
Weight
Region * Size
Asia 1
Asia 2
Asia 3
N America 1
N America 2
N America 3

c. Interpret the odds ratio for significant predictors.

Only Size is significant. The design variables show that Size=1 (Small or Sports) cars have 14.560 times the odds of having a below-average safety rating compared to the reference category, 3 (Large or Sport/Utility). The 95% confidence interval (3.018, 110.732) does not contain 1, implying that the contrast is statistically significant at the 0.05 level. The contrast from the second design variable is 1.931 (Medium versus Sport/Utility), implying a trend toward greater odds of low safety for medium cars. However, the 95% confidence interval (0.343, 15.182) contains 1 and therefore the contrast is not statistically significant.

Solutions to Student Activities (Polls/Quizzes)

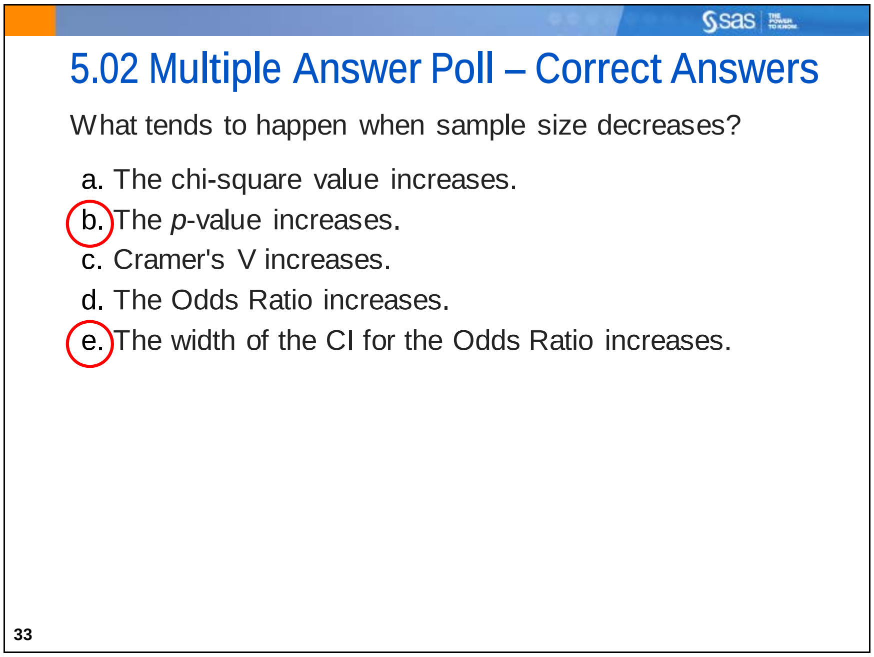


5.01 Multiple Answer Poll – Correct Answer

Which of the following would likely not be considered categorical in the data?

a. **Gender**
[✓ b.] **Fare**
c. **Survival**
[✓ d.] **Age**
e. **Class**

14

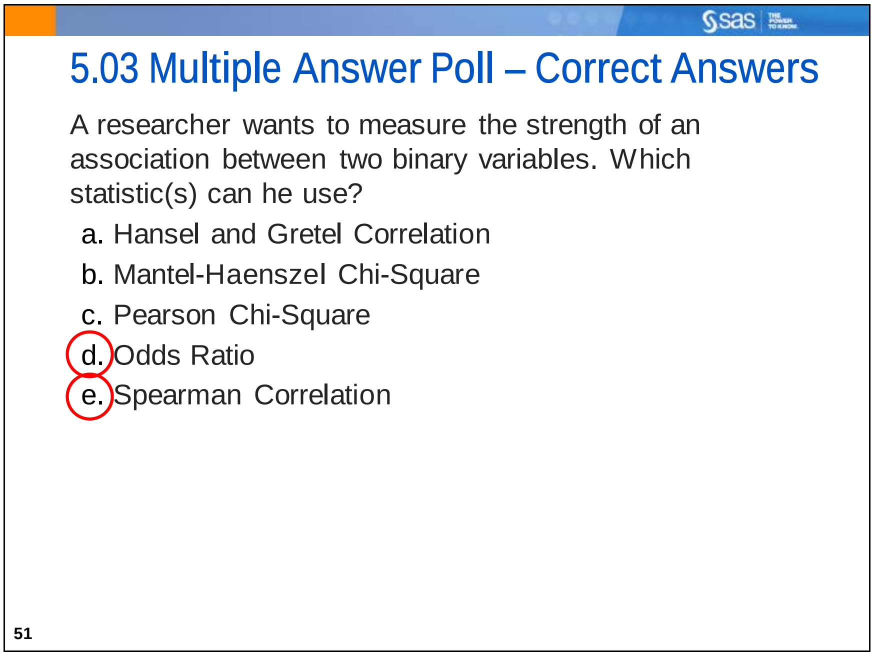


5.02 Multiple Answer Poll – Correct Answers

What tends to happen when sample size decreases?

a. The chi-square value increases.
[✓ b.] The *p*-value increases.
c. Cramer's V increases.
d. The Odds Ratio increases.
[✓ e.] The width of the CI for the Odds Ratio increases.

33



5.03 Multiple Answer Poll – Correct Answers

A researcher wants to measure the strength of an association between two binary variables. Which statistic(s) can he use?

a. Hansel and Gretel Correlation

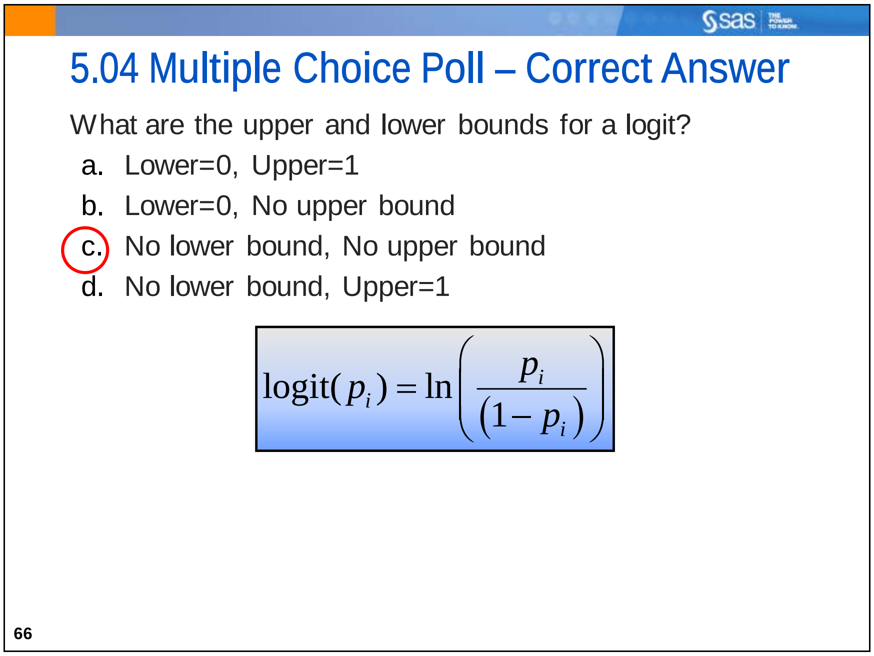
b. Mantel-Haenszel Chi-Square

c. Pearson Chi-Square

[✓ d.] Odds Ratio

[✓ e.] Spearman Correlation

51



5.04 Multiple Choice Poll – Correct Answer

What are the upper and lower bounds for a logit?

a. Lower=0, Upper=1

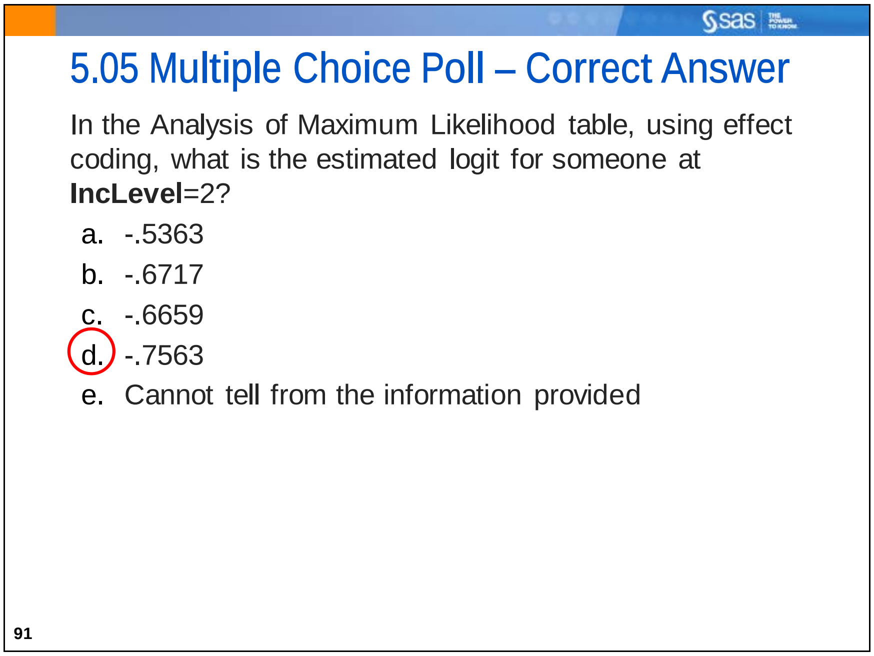
b. Lower=0, No upper bound

[✓ c.] No lower bound, No upper bound

d. No lower bound, Upper=1

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)$$

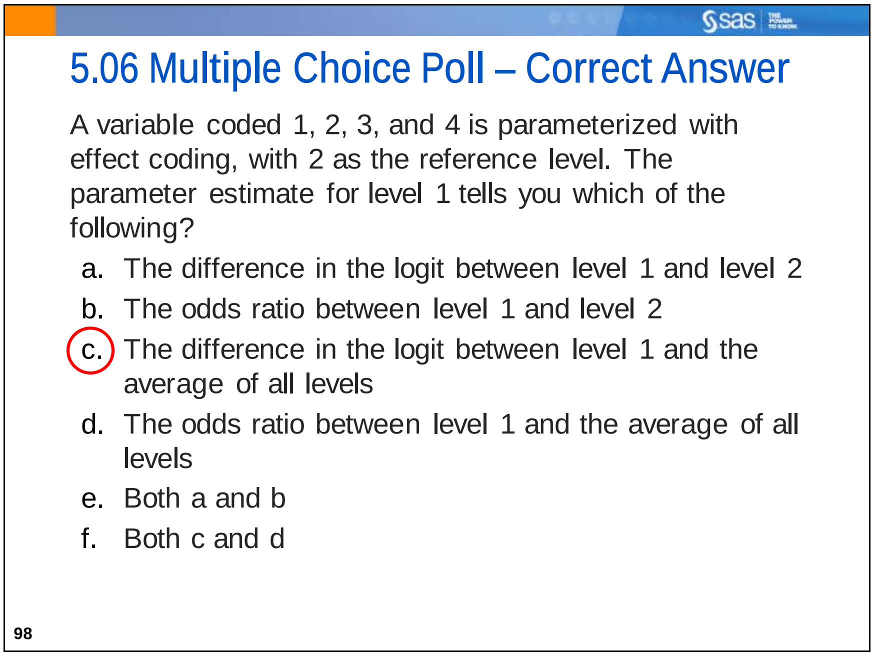
66



5.05 Multiple Choice Poll – Correct Answer

In the Analysis of Maximum Likelihood table, using effect coding, what is the estimated logit for someone at **IncLevel**=2?

a. -.5363
b. -.6717
c. -.6659
[✓ d.] -.7563
e. Cannot tell from the information provided



5.06 Multiple Choice Poll – Correct Answer

A variable coded 1, 2, 3, and 4 is parameterized with effect coding, with 2 as the reference level. The parameter estimate for level 1 tells you which of the following?

a. The difference in the logit between level 1 and level 2
b. The odds ratio between level 1 and level 2
[✓ c.] The difference in the logit between level 1 and the average of all levels
d. The odds ratio between level 1 and the average of all levels
e. Both a and b
f. Both c and d