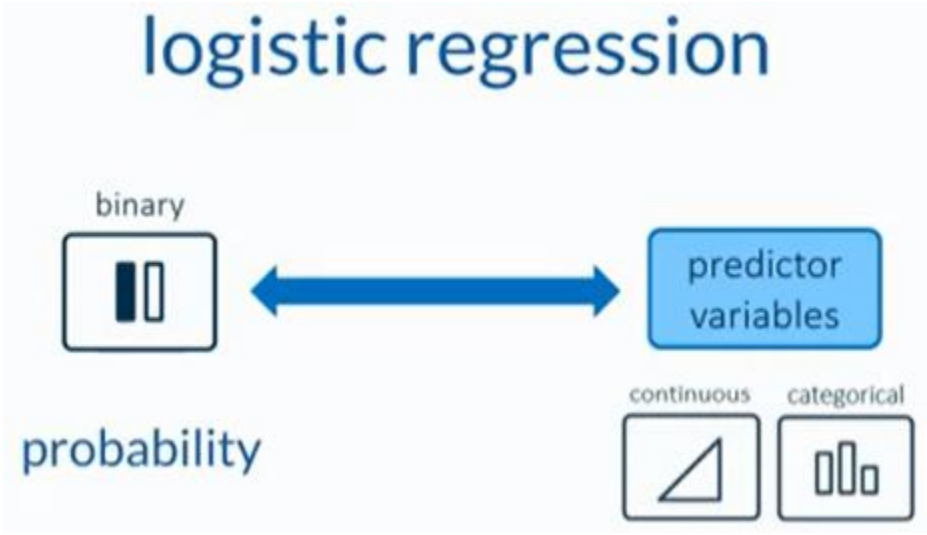
SBA Statistical Business Analyst with SAS

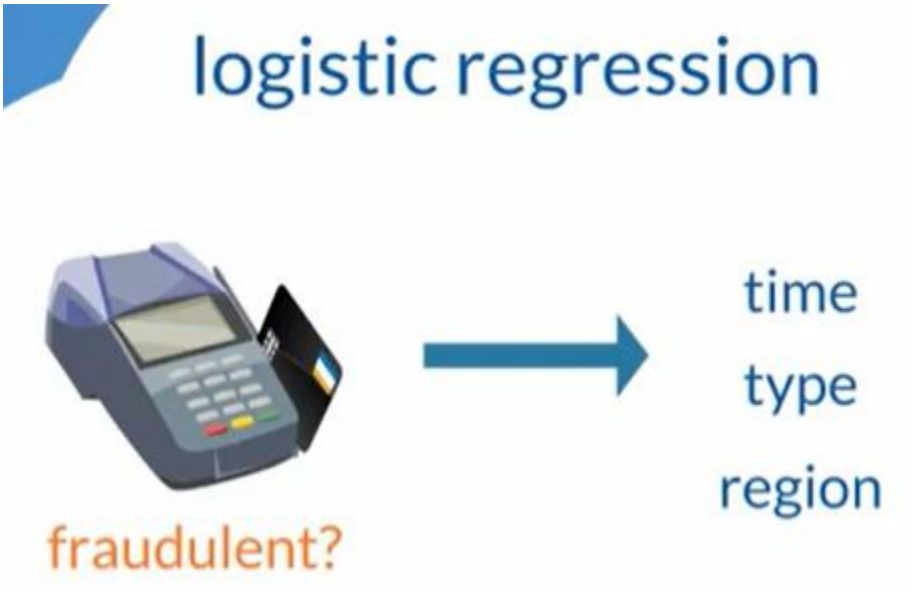
SBA2 Regression Modeling Fundamentals

SBA203 Categorical Data Analysis

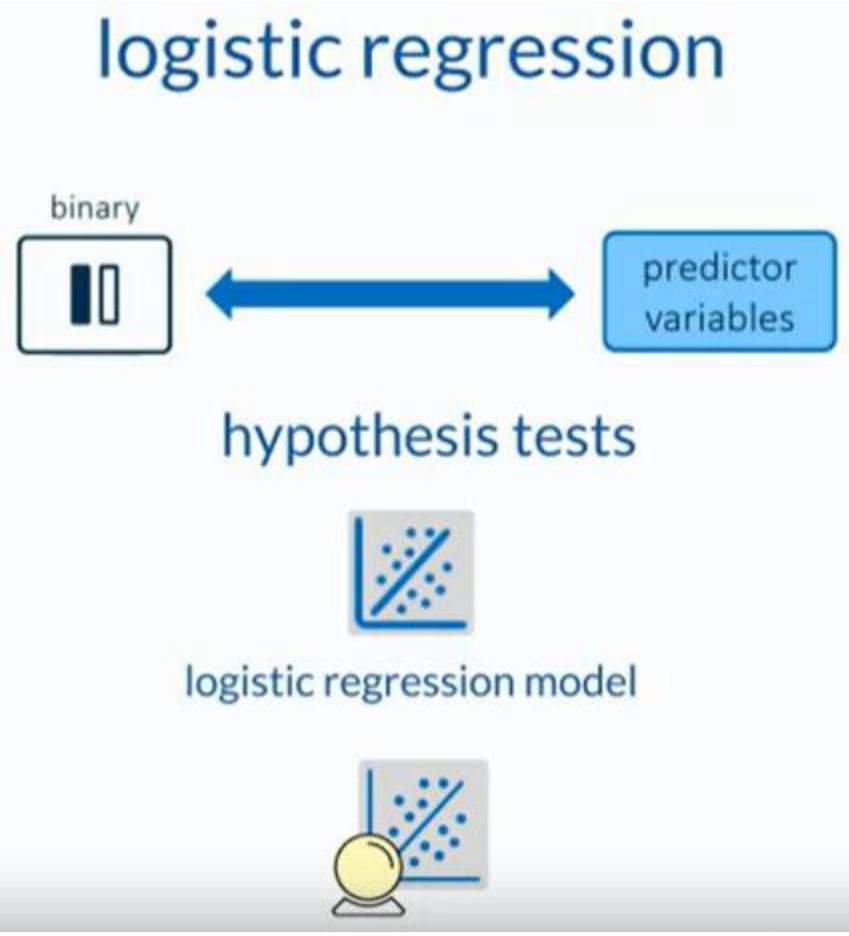
Overview




CREDIT CARD
2739 4909 8964 1234
2739 4909 8964 1234
ticket fare
age
gender
credit card
fraud


logistic regression
time
type
region
fraudulent?


logistic regression
010101010101
101010101010
010101010101
101010101010
010101010101
101010101010

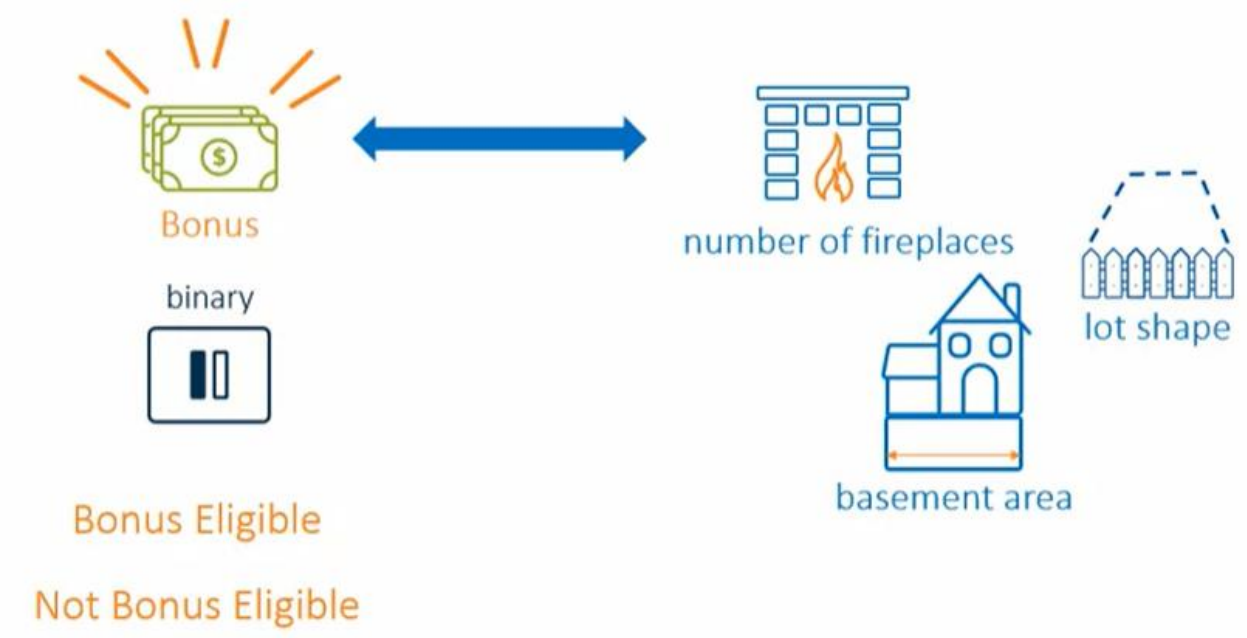

logistic regression
binary
predictor variables
hypothesis tests
logistic regression model

Describing Categorical Data

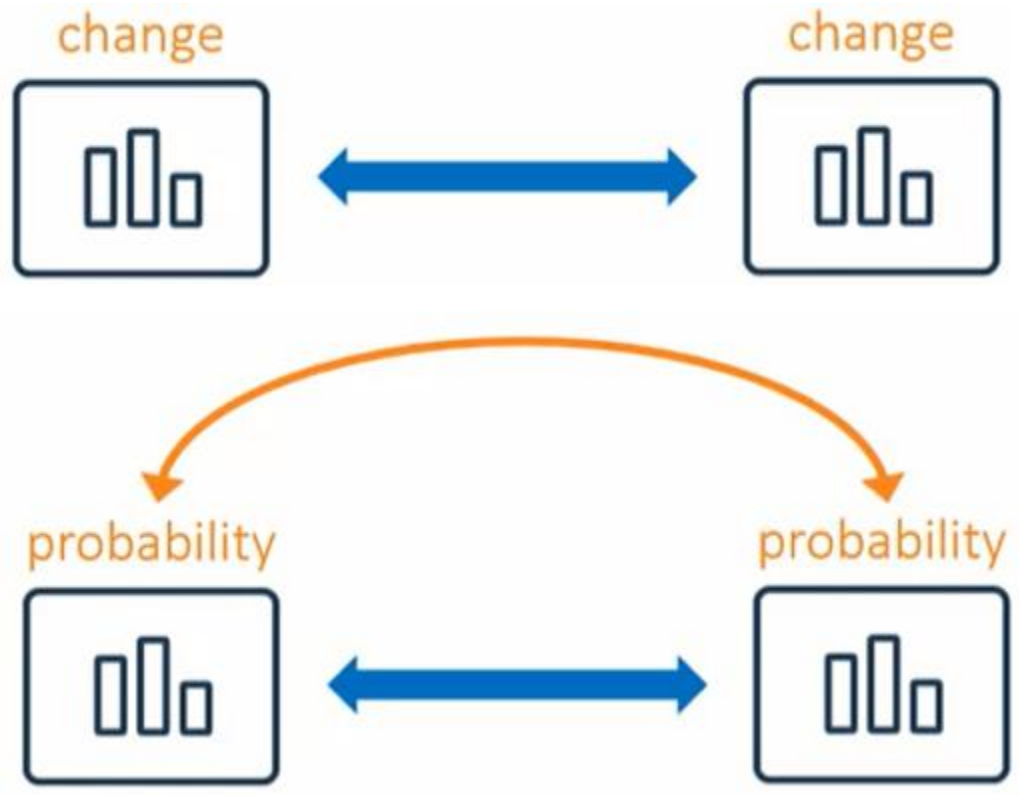
Scenario

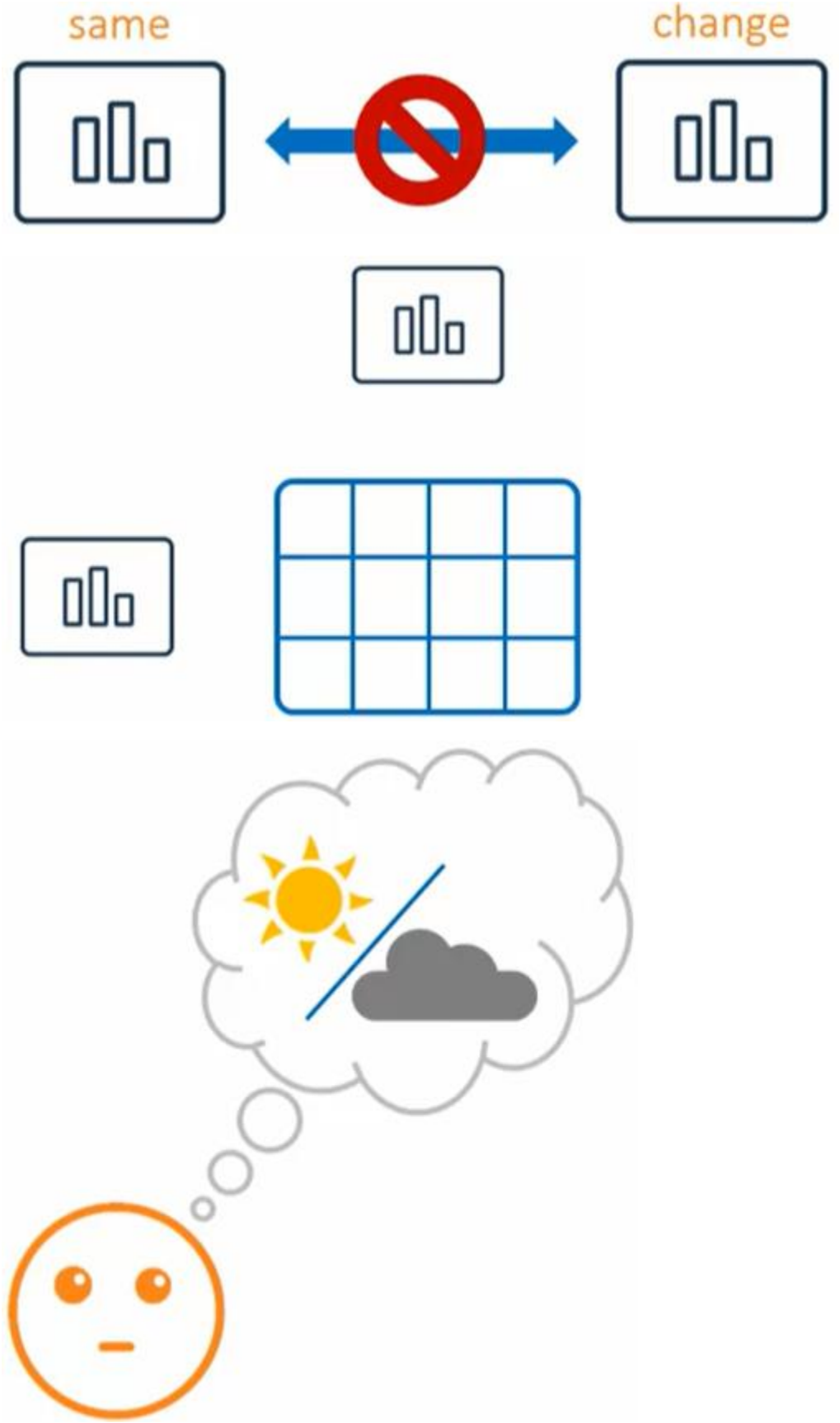

BONUS!
>$175,000
SOLD


BONUS!
>$175,000
SOLD

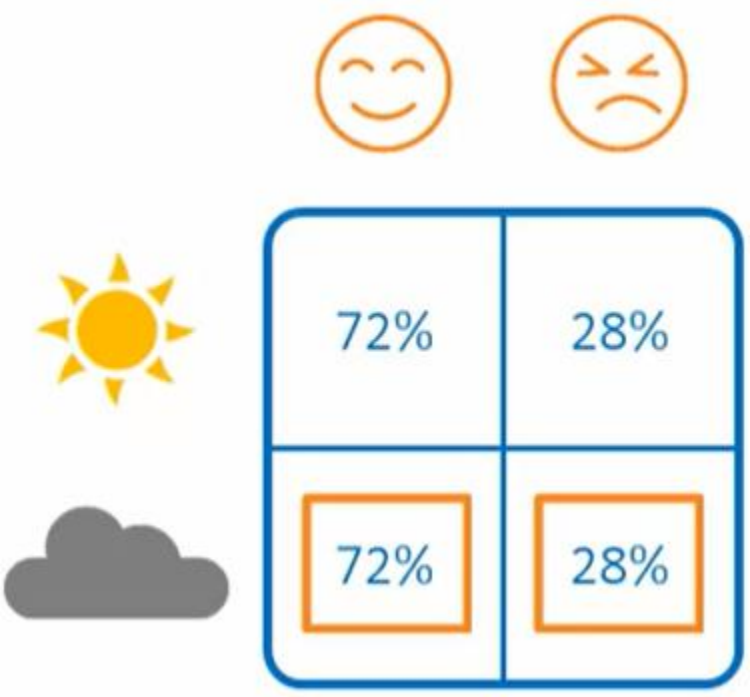


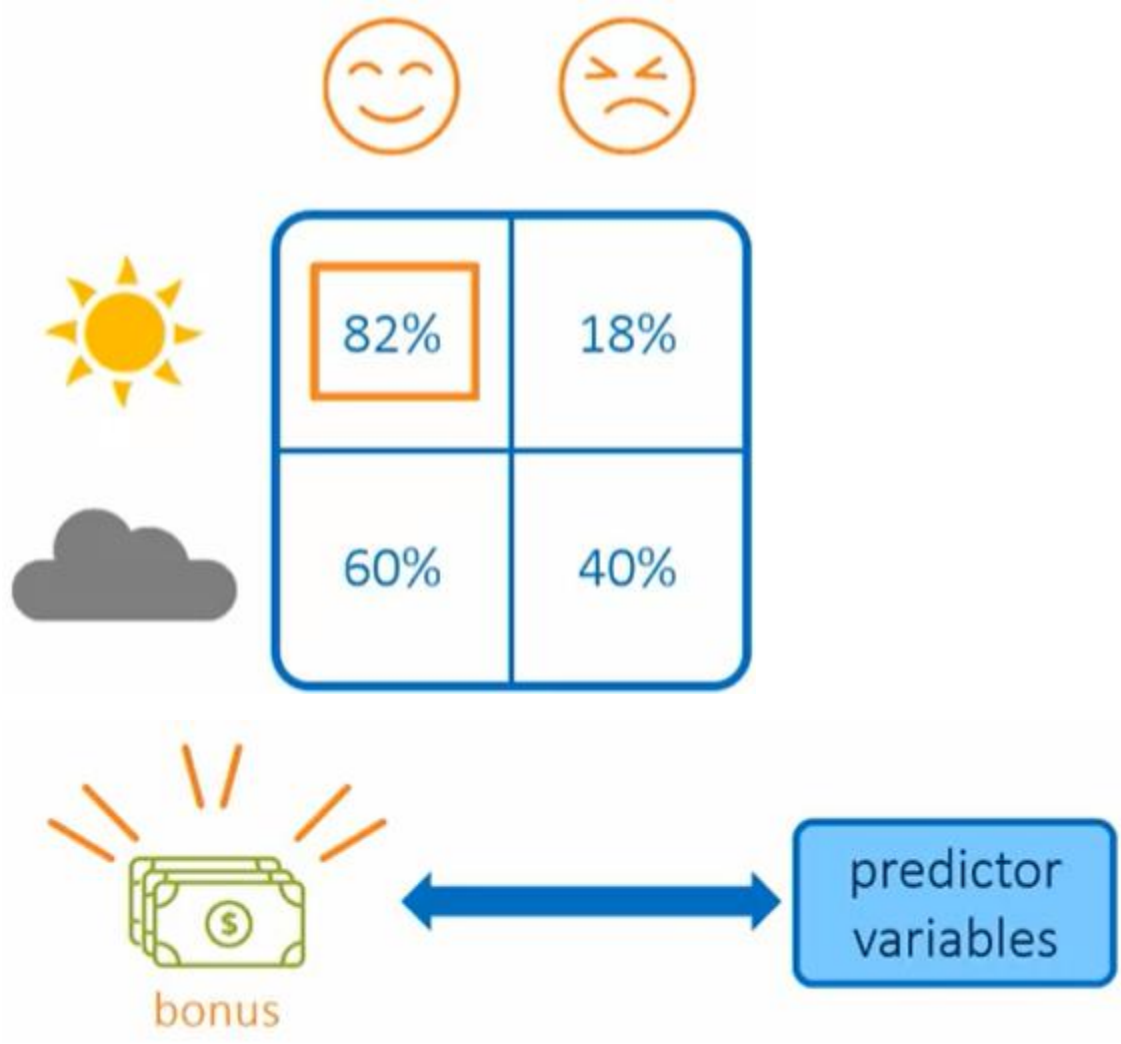
Associations between Categorical Variables

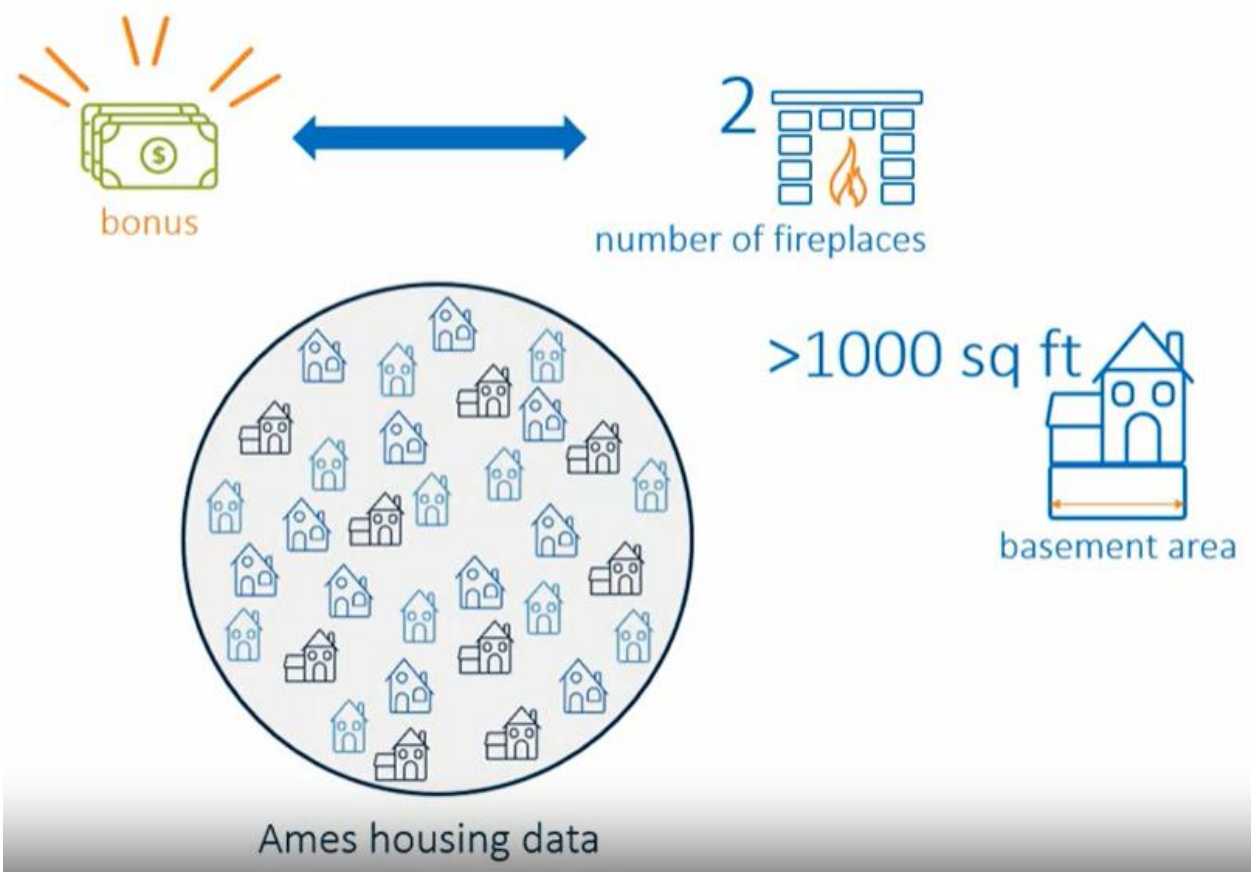



same
change

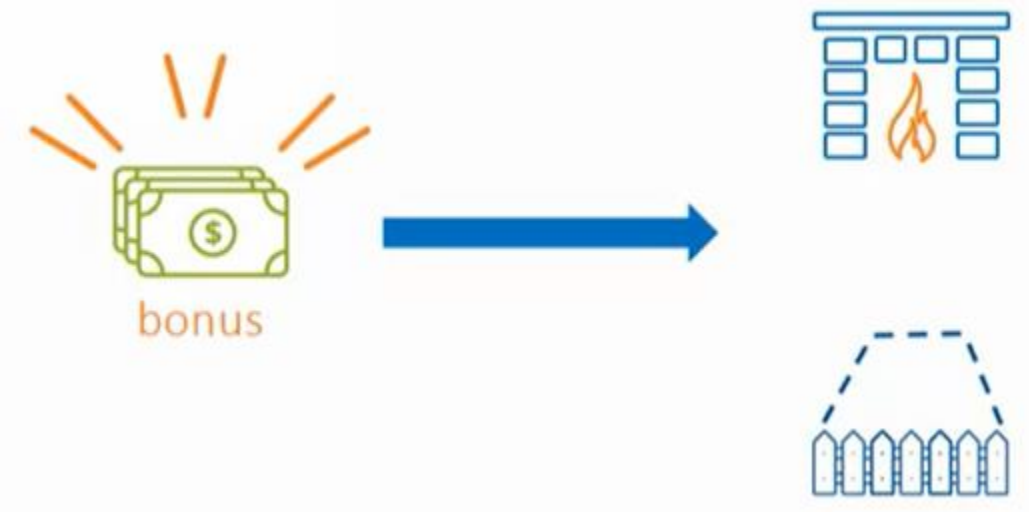

Mood
happy
grumpy
Weather
sunny
cloudy

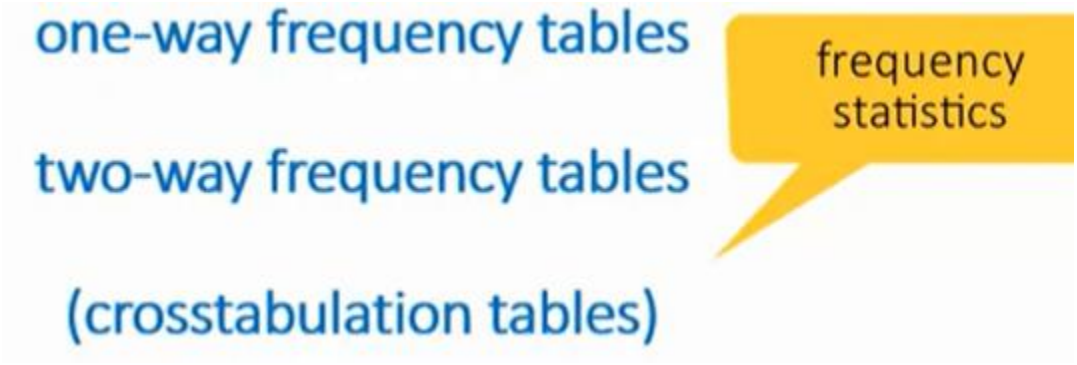

72%
28%
72%
28%


82%
18%
60%
40%
bonus
predictor variables



Demo Examining the Distribution of Categorical Variables Using PROC FREQ and PROC UNIVARIATE






```
/*st107d01.sas*/
title;
proc format;
    value bonusfmt 1 = "Bonus Eligible"
                   0 = "Not Bonus Eligible"
                  ;
run;

proc freq data=STAT1.ameshousing3;
    tables Bonus Fireplaces Lot_Shape_2
           Fireplaces*Bonus Lot_Shape_2*Bonus/
           plots(only)=freqplot(scale=percent);
    format Bonus bonusfmt.;
run;
```

PROC FREQ DATA=*SAS-data-set*;
TABLES *table-request(s) < / options>*;
<additional statements>
RUN;

```
proc univariate data=STAT1.ameshousing3 noprint;
    class Bonus;
    var Basement_Area ;
    histogram Basement_Area;
    inset mean std median min max / format=5.2 position=nw;
    format Bonus bonusfmt.;
run;
```

PROC UNIVARIATE DATA=*SAS-data-set <options>*;
VAR *variables*;
HISTOGRAM *variables < / options>*;
INSET *keywords < / options>*;
RUN;

The FREQ Procedure

Sale Price > $175,000

Bonus	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not Bonus Eligible	255	85.00	255	85.00
Bonus Eligible	45	15.00	300	100.00

Number of fireplaces

Fireplaces	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	195	65.00	195	65.00
1	93	31.00	288	96.00
2	12	4.00	300	100.00

Regular or irregular lot shape

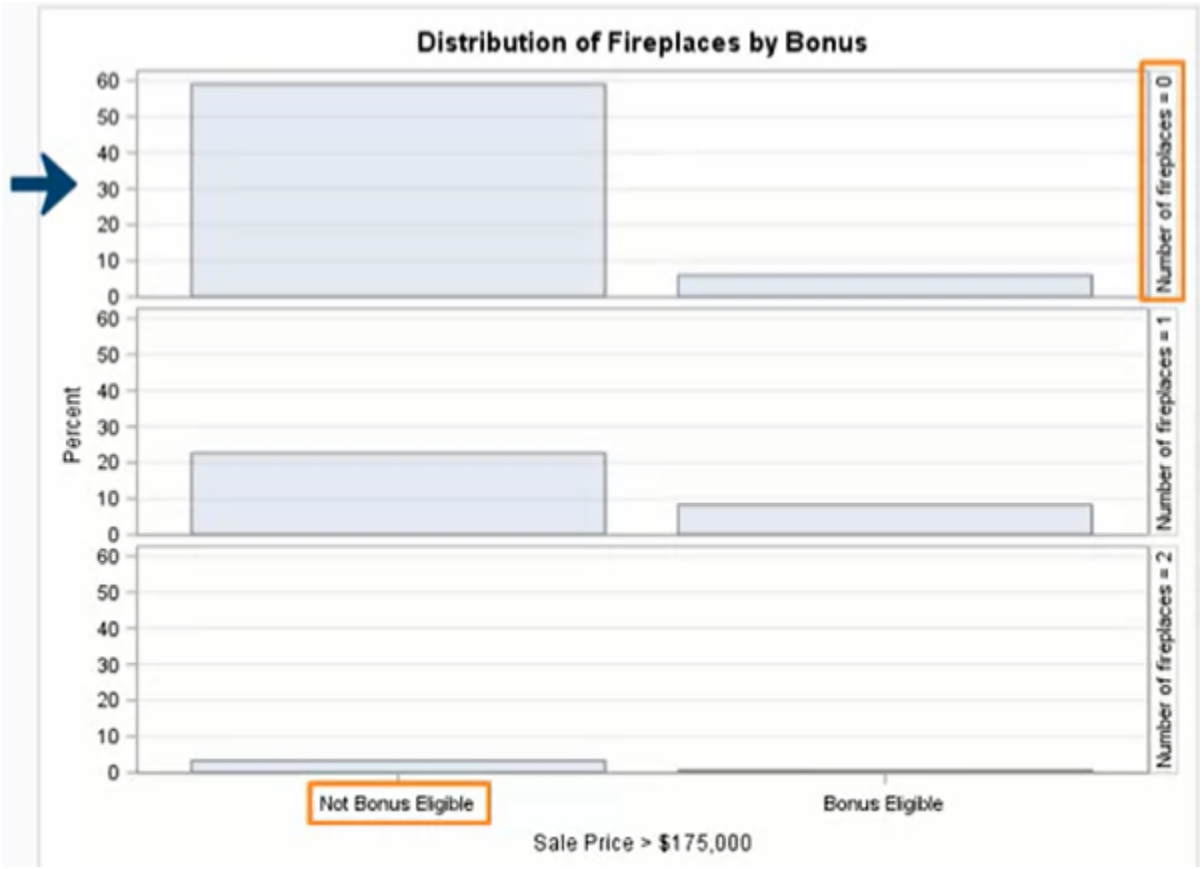
Lot_Shape_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Irregular	93	31.10	93	31.10
Regular	206	68.90	299	100.00

Frequency Missing = 1

Frequency
Percent
Row Pct
Col Pct

Table of Fireplaces by Bonus

Fireplaces(Number of fireplaces)	Bonus(Sale Price > $175,000)		
	Not Bonus Eligible	Bonus Eligible	Total
0	177 59.00 90.77 69.41	18 6.00 9.23 40.00	195 65.00
1	68 22.67 73.12 26.67	25 8.33 26.88 55.56	93 31.00
2	10 3.33 83.33 3.92	2 0.67 16.67 4.44	12 4.00
Total	255 85.00	45 15.00	300 100.00



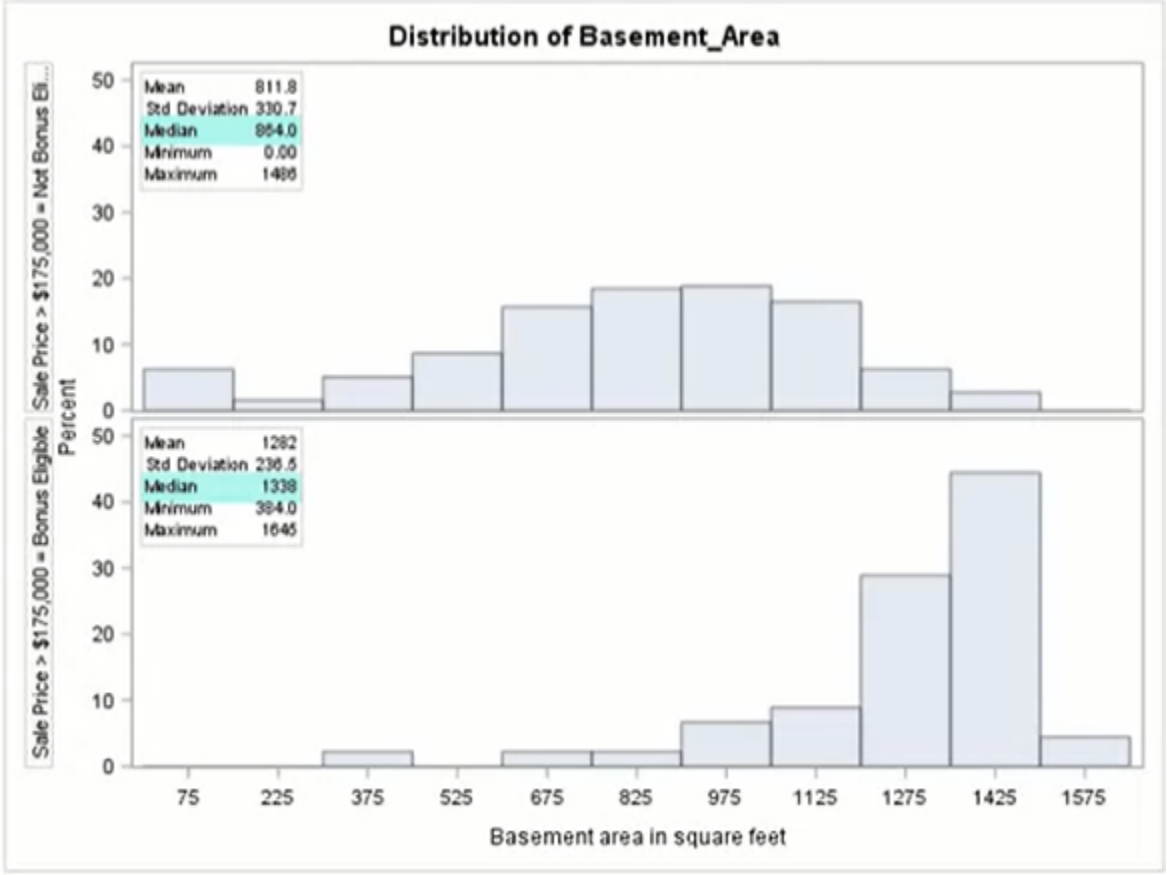
Frequency
Percent
Row Pct
Col Pct

Table of Lot_Shape_2 by Bonus

Lot_Shape_2(Regular or irregular lot shape)	Bonus(Sale Price > $175,000) Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 20.74 66.67 24.31	31 10.37 33.33 70.45	93 31.10
Regular	193 64.55 93.69 75.69	13 4.35 6.31 29.55	206 68.90
Total	255 85.28	44 14.72	299 100.00

Frequency Missing = 1

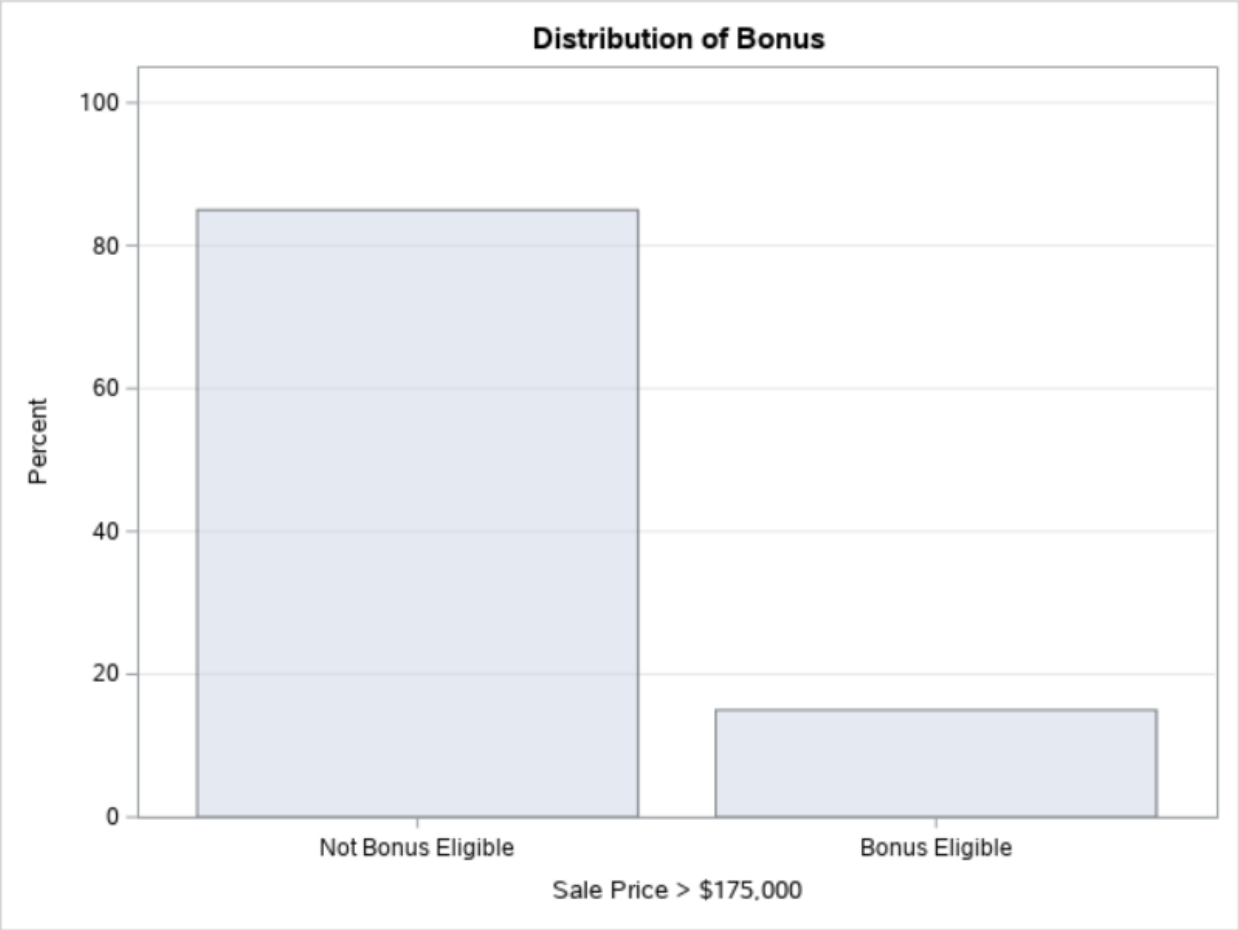
The UNIVARIATE Procedure



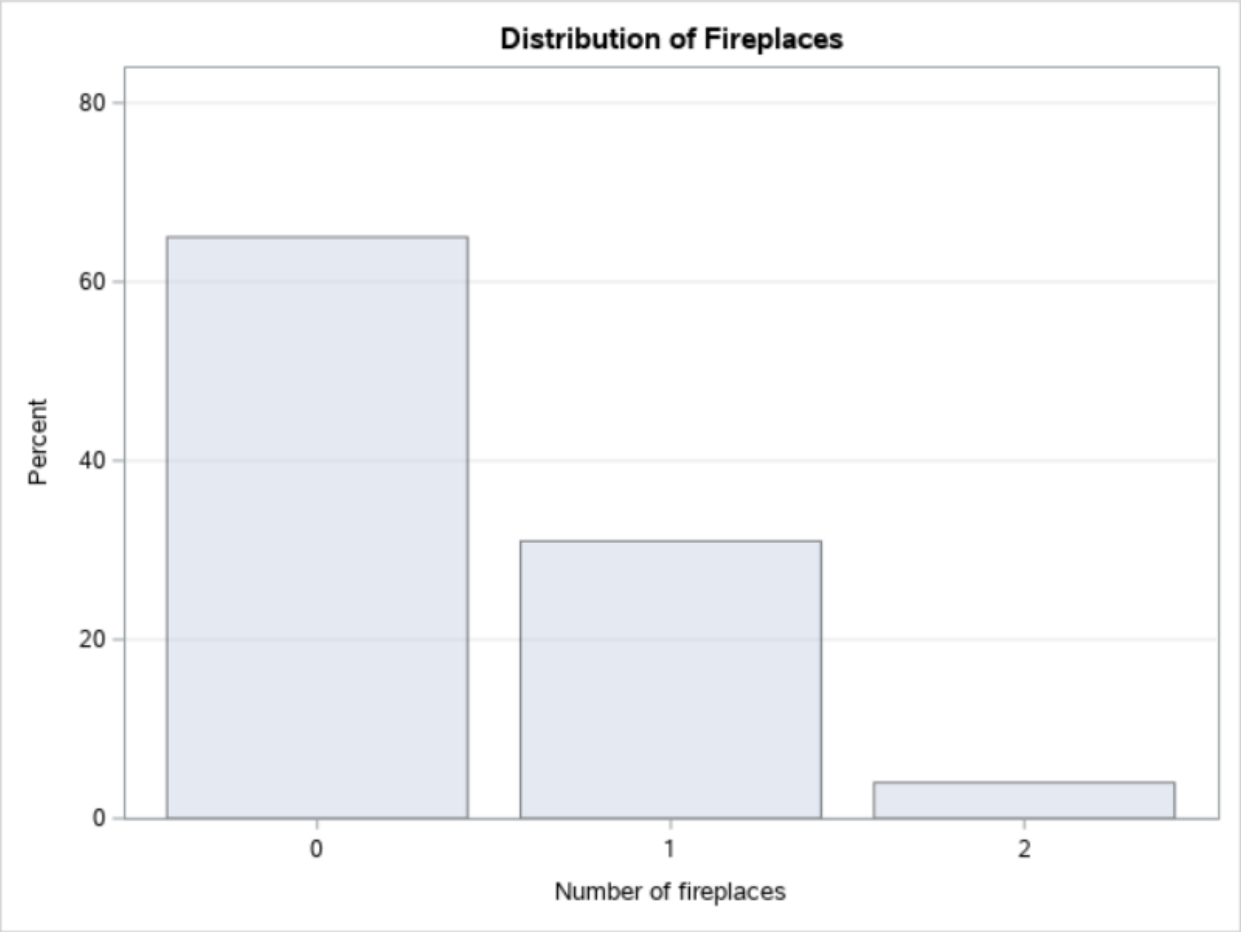
```
/*st107d01.sas*/
title;
proc format;
   value bonusfmt 1 = "Bonus Eligible"
                  0 = "Not Bonus Eligible"
                 ;
run;

proc freq data=STAT1.ameshousing3;
   tables Bonus Fireplaces Lot_Shape_2
          Fireplaces*Bonus Lot_Shape_2*Bonus/
          plots(only)=freqplot(scale=percent);
   format Bonus bonusfmt.;
run;

proc univariate data=STAT1.ameshousing3 noprint;
   class Bonus;
   var Basement_Area ;
   histogram Basement_Area;
   inset mean std median min max / format=5.2 position=nw;
   format Bonus bonusfmt.;
run;
```

The FREQ Procedure

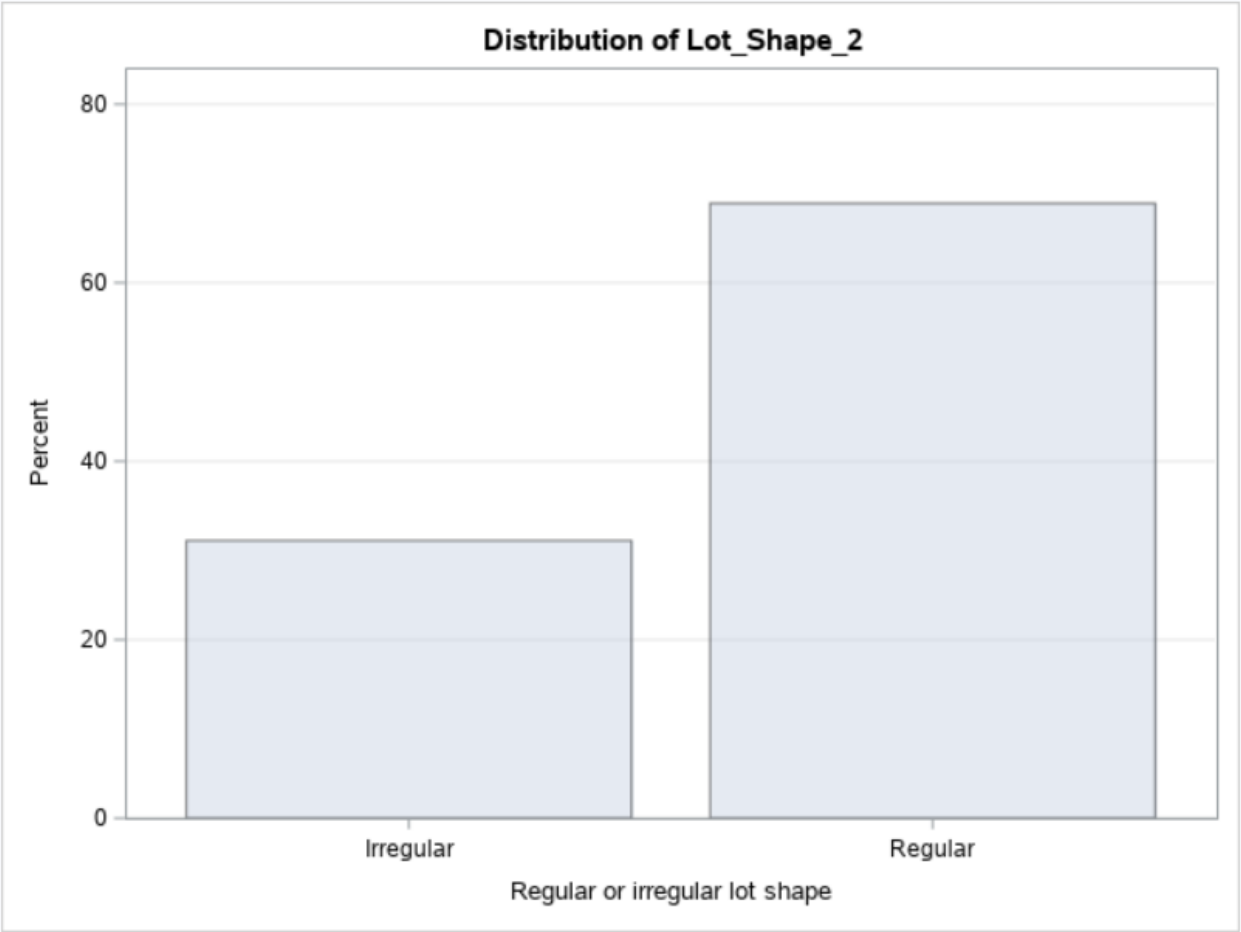
Sale Price > $175,000				
Bonus	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not Bonus Eligible	255	85.00	255	85.00
Bonus Eligible	45	15.00	300	100.00



Number of fireplaces				
Fireplaces	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	195	65.00	195	65.00
1	93	31.00	288	96.00
2	12	4.00	300	100.00

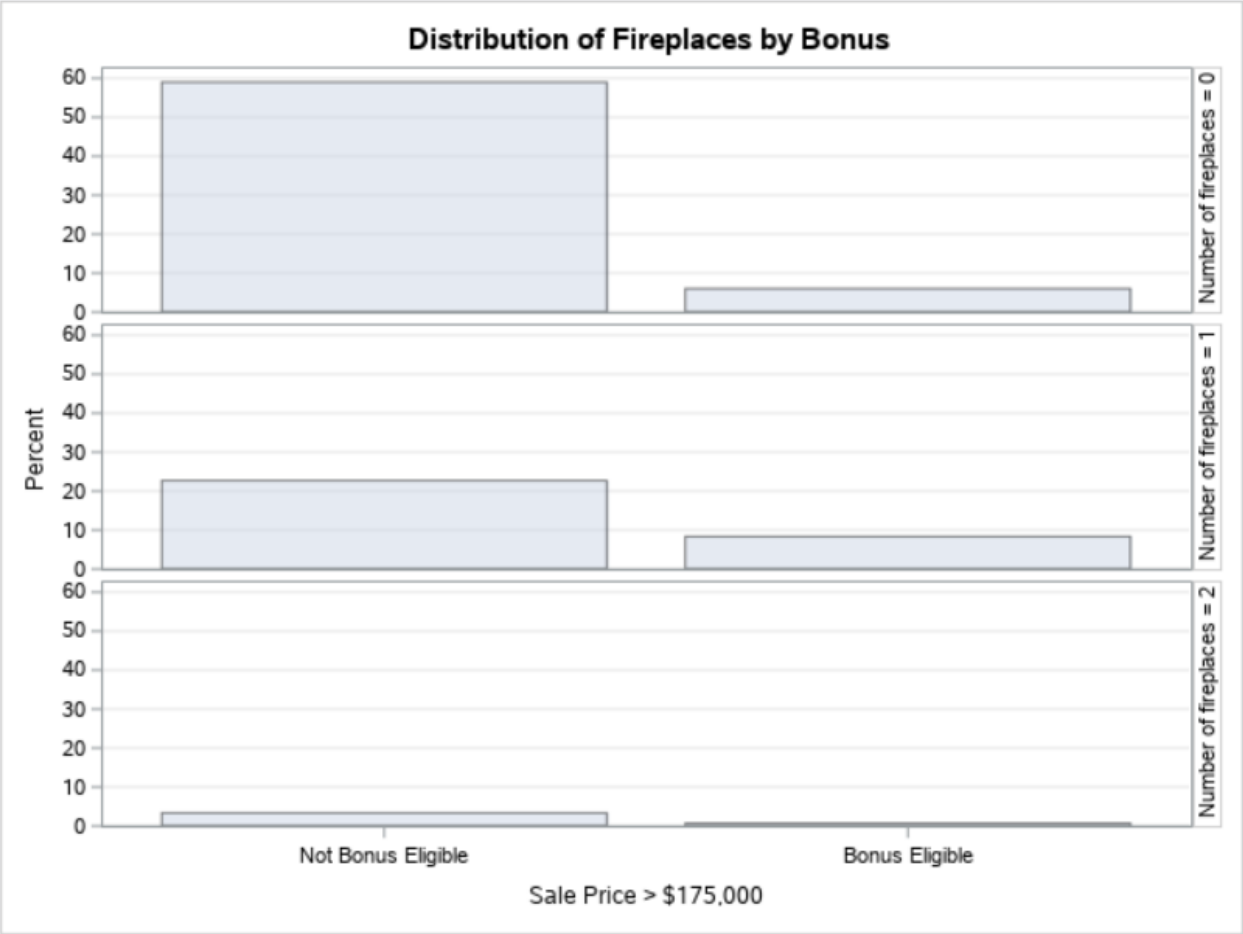


Regular or irregular lot shape				
Lot_Shape_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Irregular	93	31.10	93	31.10
Regular	206	68.90	299	100.00
Frequency Missing = 1				



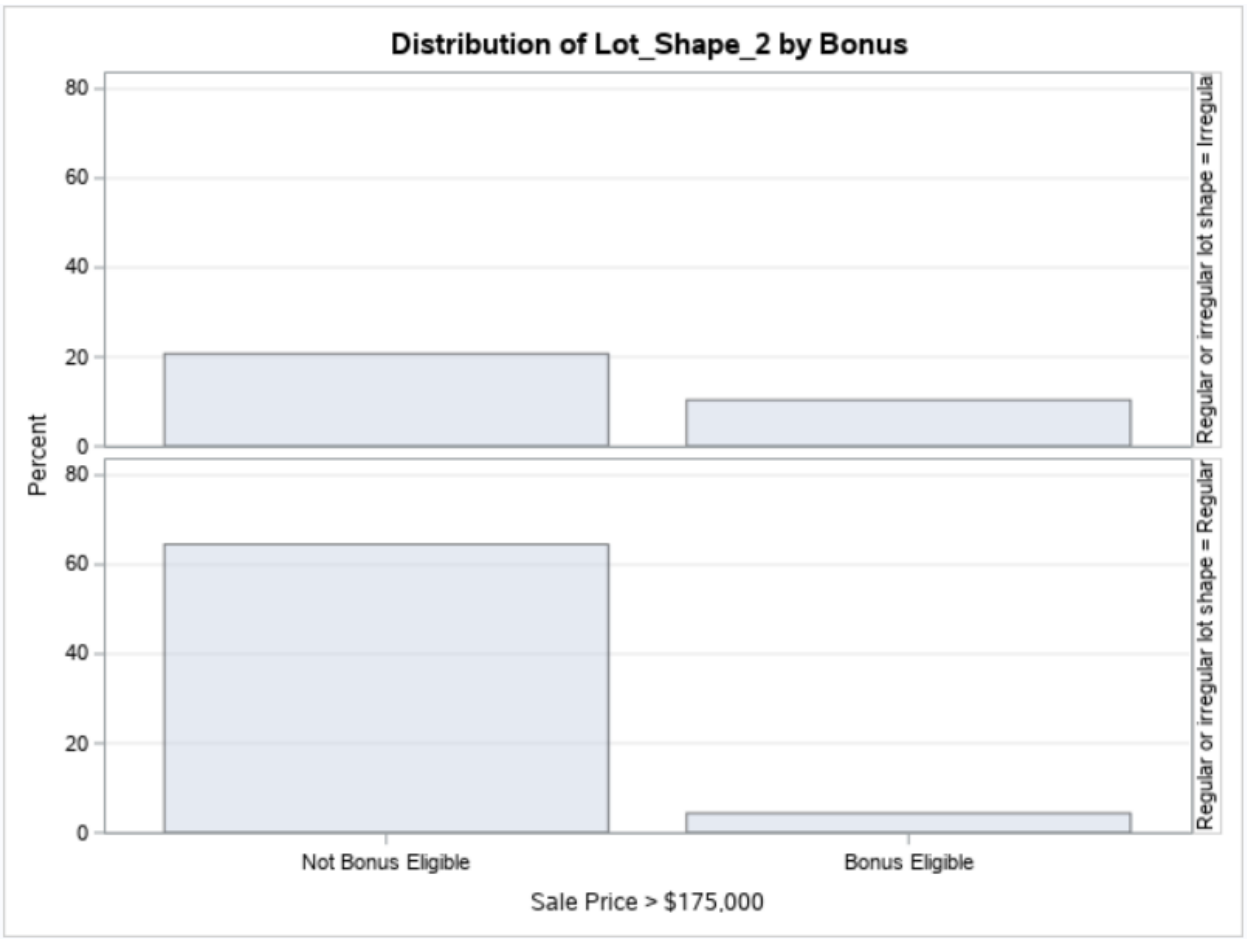
Frequency
Percent
Row Pct
Col Pct

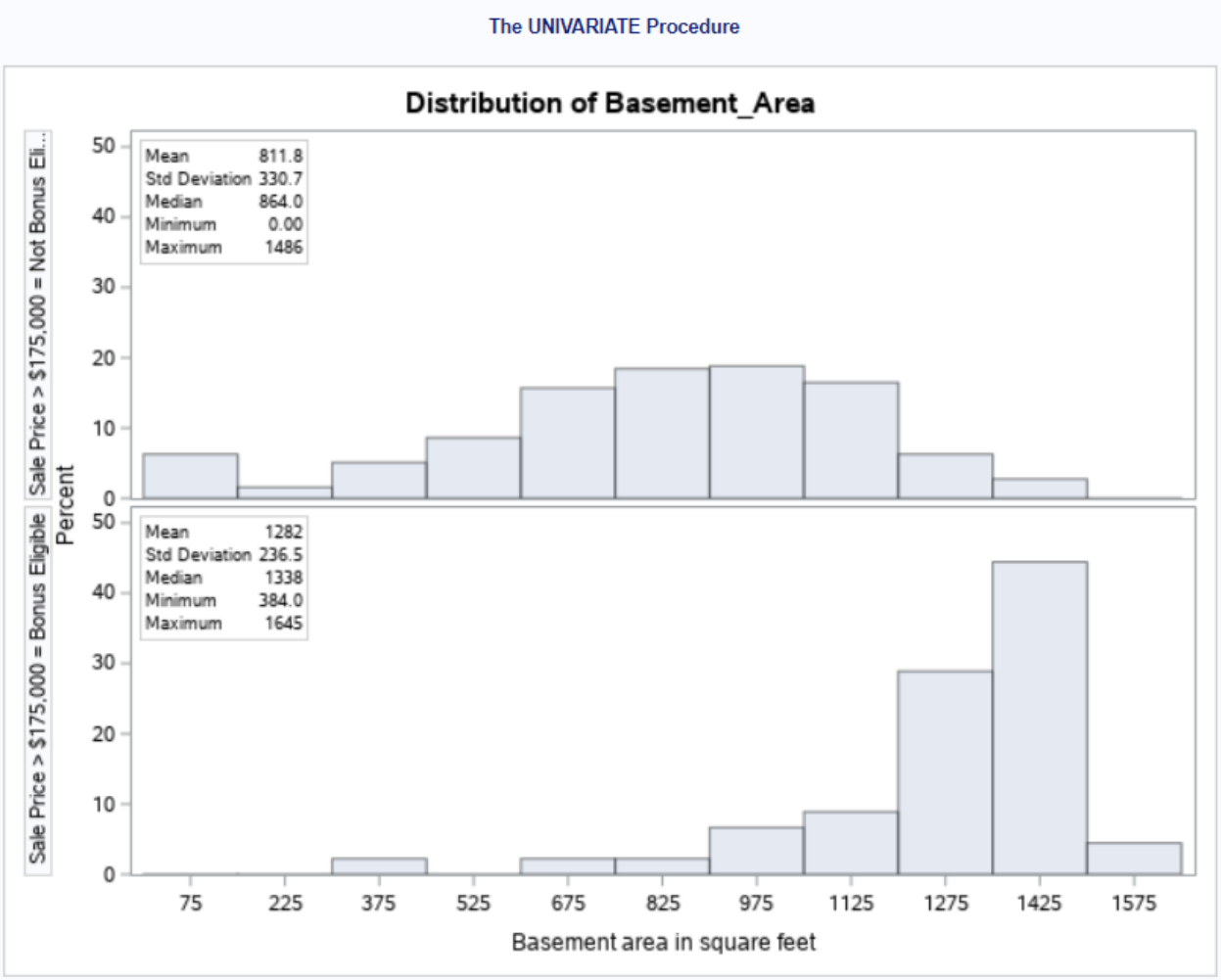
Table of Fireplaces by Bonus			
	Bonus(Sale Price > $175,000)		
Fireplaces(Number of fireplaces)	Not Bonus Eligible	Bonus Eligible	Total
0	177 59.00 90.77 69.41	18 6.00 9.23 40.00	195 65.00
1	68 22.67 73.12 26.67	25 8.33 26.88 55.56	93 31.00
2	10 3.33 83.33 3.92	2 0.67 16.67 4.44	12 4.00
Total	255 85.00	45 15.00	300 100.00



Frequency
Percent
Row Pct
Col Pct

Table of Lot_Shape_2 by Bonus			
	Bonus(Sale Price > $175,000)		
Lot_Shape_2(Regular or irregular lot shape)	Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 20.74 66.67 24.31	31 10.37 33.33 70.45	93 31.10
Regular	193 64.55 93.69 75.69	13 4.35 6.31 29.55	206 68.90
Total	255 85.28	44 14.72	299 100.00
Frequency Missing = 1			


Distribution of Lot_Shape_2 by Bonus
Percent
80
60
40
20
0
Regular or irregular lot shape = Irregula
Regular or irregular lot shape = Regular
Not Bonus Eligible
Bonus Eligible
Sale Price > $175,000



Question 7.01

In this crosstabulation table of Titanic data, what evidence indicates a possible association between the variables **Age** and **Survived**?

Table of Age by Survived			
Age	Survived		
Frequency Percent Row Pct Col Pct	no	yes	Total
adult	1438	654	2092
	65.33	29.71	95.05
	68.74	31.26	
	96.51	91.98	
child	52	57	109
	2.36	2.59	4.95
	47.71	52.29	
	3.49	8.02	
Total	1490	711	2201
	67.70	32.30	100.00

Correct

To see a possible association, you look at the row percentages. The percent of children who survived (52.29) is much higher than the percent of adults who survived (31.26).

Question 7.02

In the PROC FREQ step below, what TABLES statement would you write to create the following frequency and crosstabulation tables for categorical variables in the **entertainment.movies** data set:

- frequency tables for **Type** and **Rating**
- crosstabulation table for **Type** (as the row variable) by **Rating** (as the column variable)

```
proc freq data=entertainment.movies;
    [TABLES STATEMENT]
run;
```

In a crosstabulation table request, you specify an asterisk between the names of the variables that you want to appear in the table. The first variable represents the rows, and the second variable represents the columns.

The following TABLES statement will create the specified tables:

```
   tables Type Rating Type*Rating;
```

```
/*st107s01.sas*/  /*Part A*/
ods graphics off;
proc freq data=STAT1.safety;
   tables Unsafe Type Region Size;
   title "Safety Data Frequencies";
run;
```

Safety Data Frequencies

The FREQ Procedure

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

Practice: Using PROC FREQ to Examine Distributions

Question 1

An insurance company wants to relate the safety of vehicles to several other variables. A score was given to each vehicle model, using the frequency of insurance claims as a basis. The **stat1.safety** data set contains the data about vehicle safety.

1. Use PROC FREQ to create one-way frequency tables for the categorical variables **Unsafe**, **Type**, **Region**, and **Size**. Submit the code and view the results.
2. What is the measurement scale of each of the four variables?

```
/*st107s01.sas*/  /*Part A*/

ods graphics off;
proc freq data=STAT1.safety;
   tables Unsafe Type Region Size;
   title "Safety Data Frequencies";
run;
ods graphics on;
```

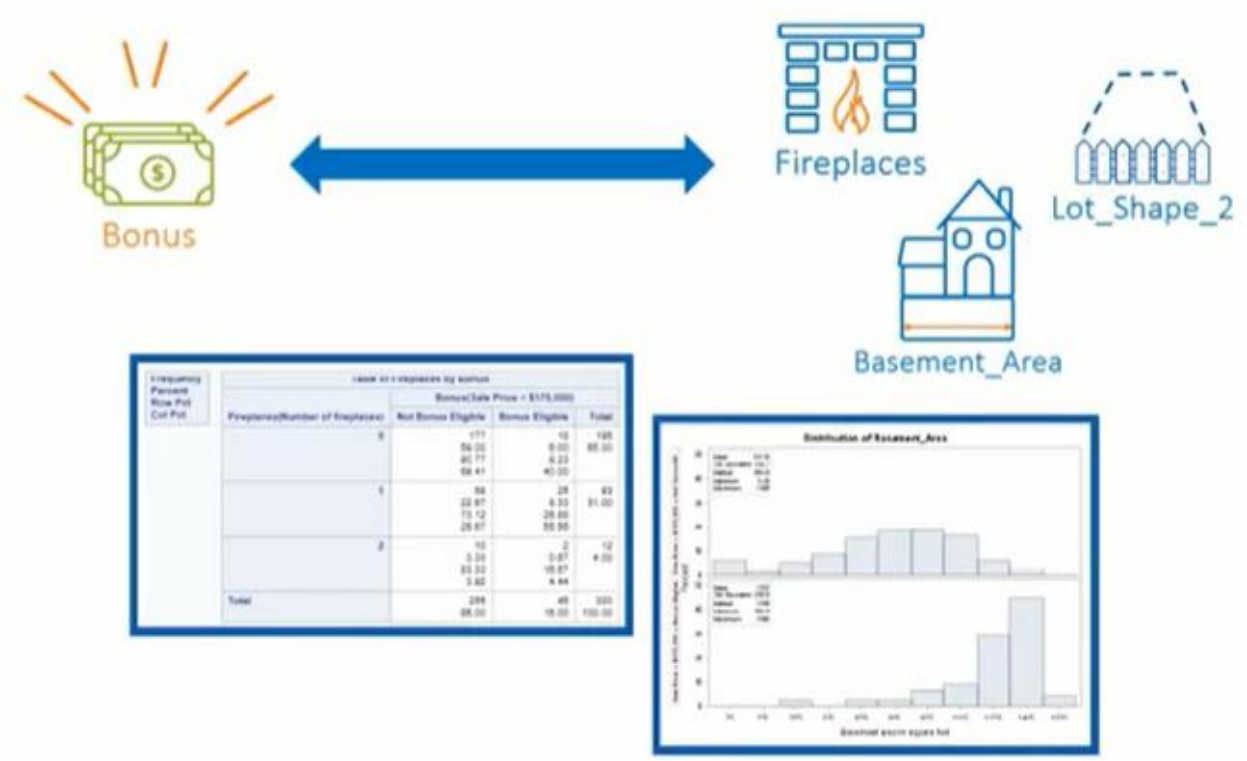
- **Unsafe** - Nominal, Binary
- **Type** - Nominal
- **Region** - Nominal
- **Size** - Ordinal
- **Weight** - Continuous

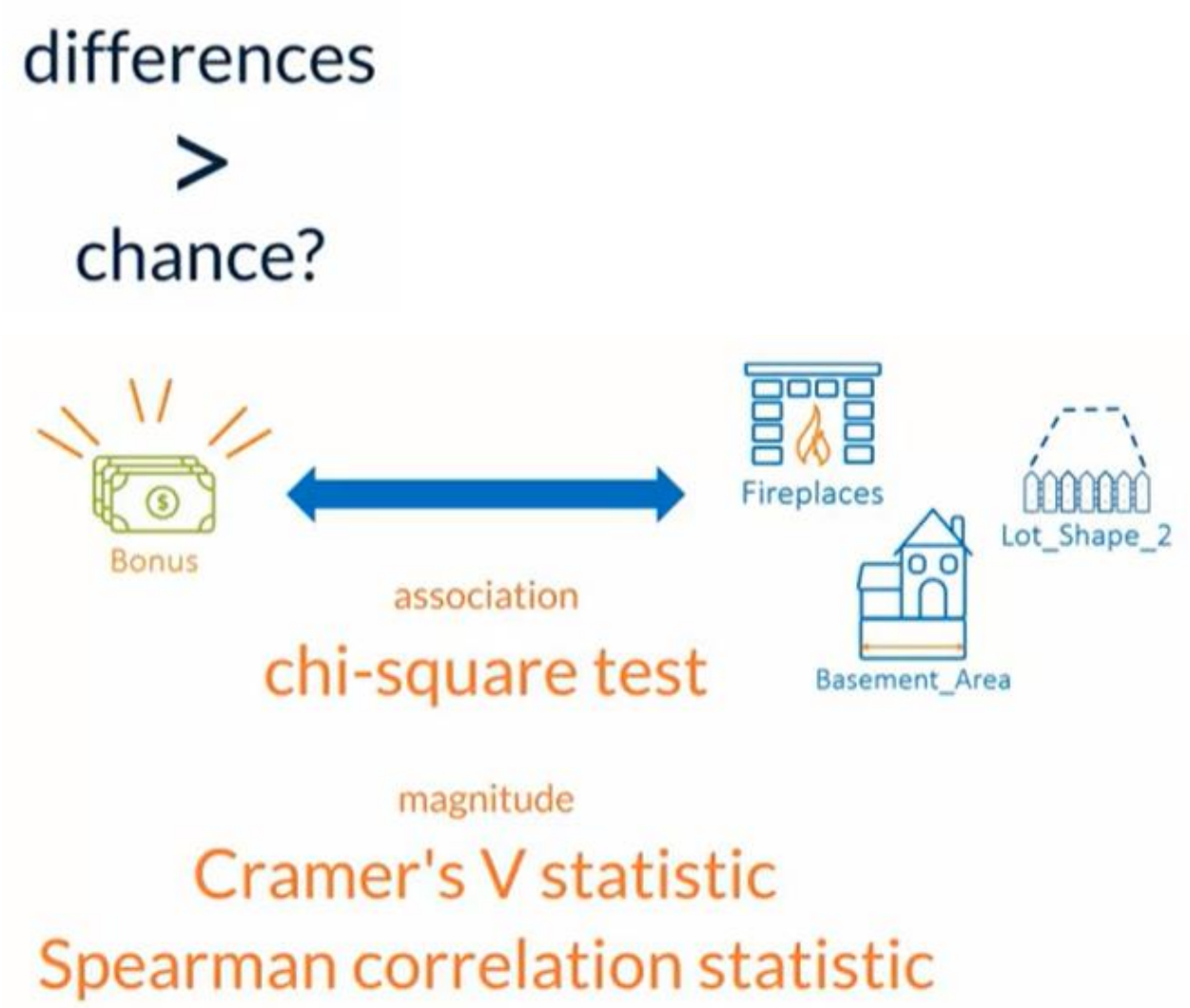
Question 2
Do the variables **Unsafe**, **Type**, **Region**, and **Size** have any unusual values that warrant further investigation?

No

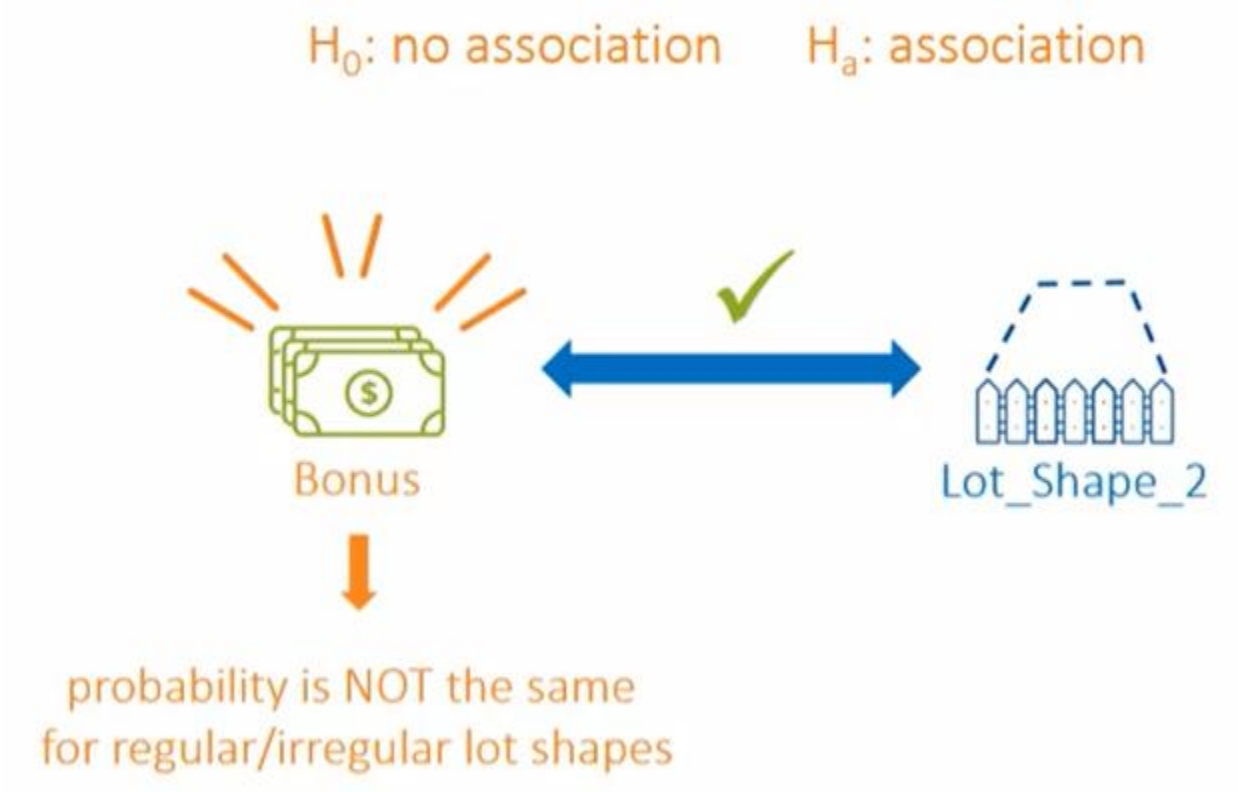
Tests of Association

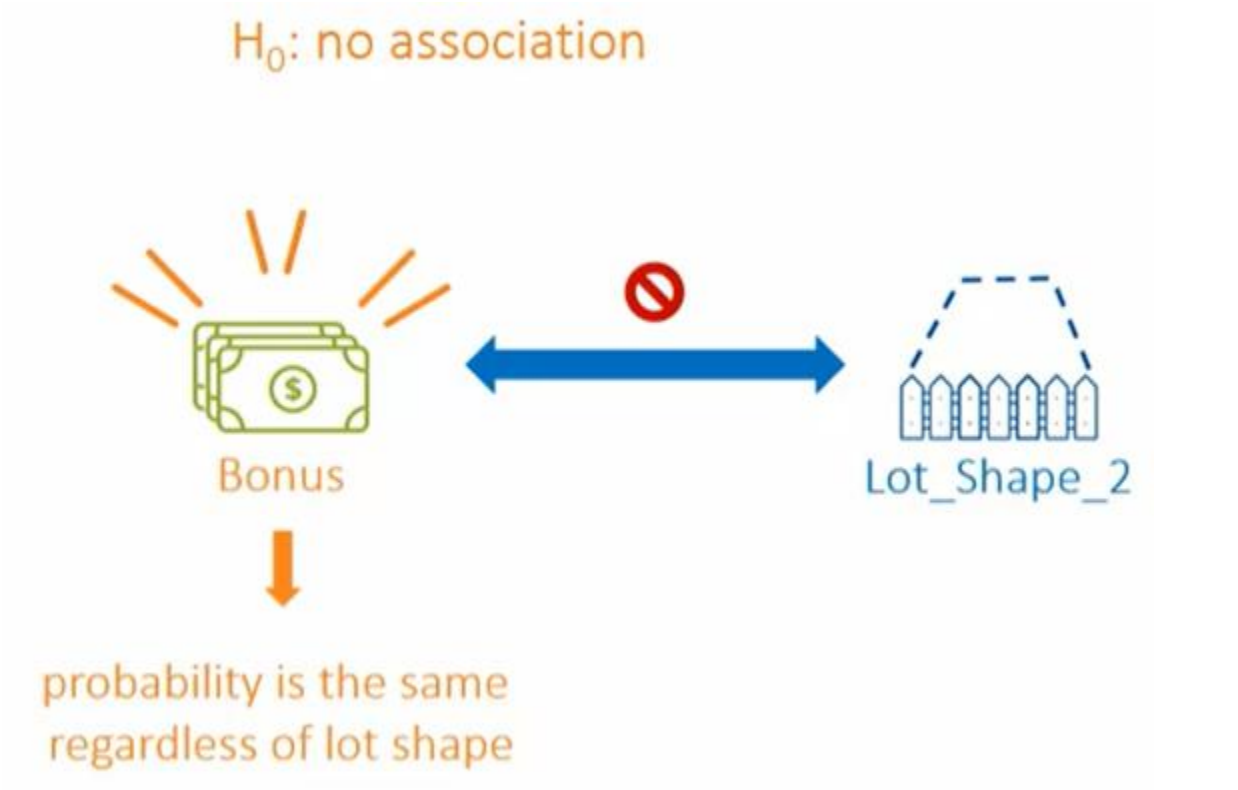
Scenario

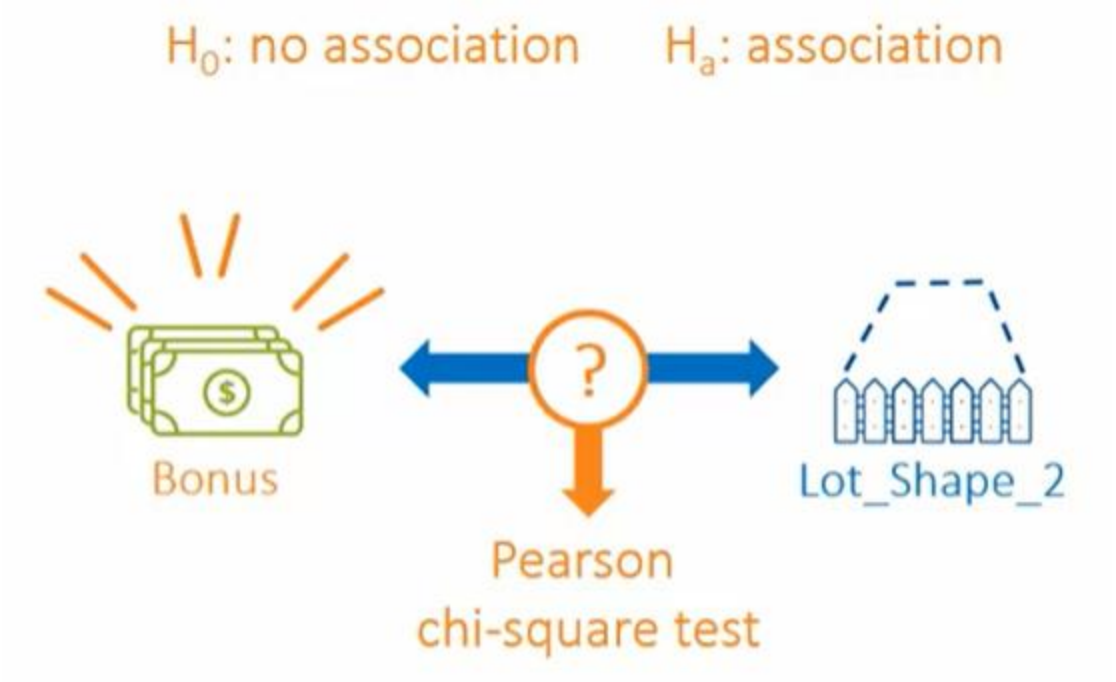


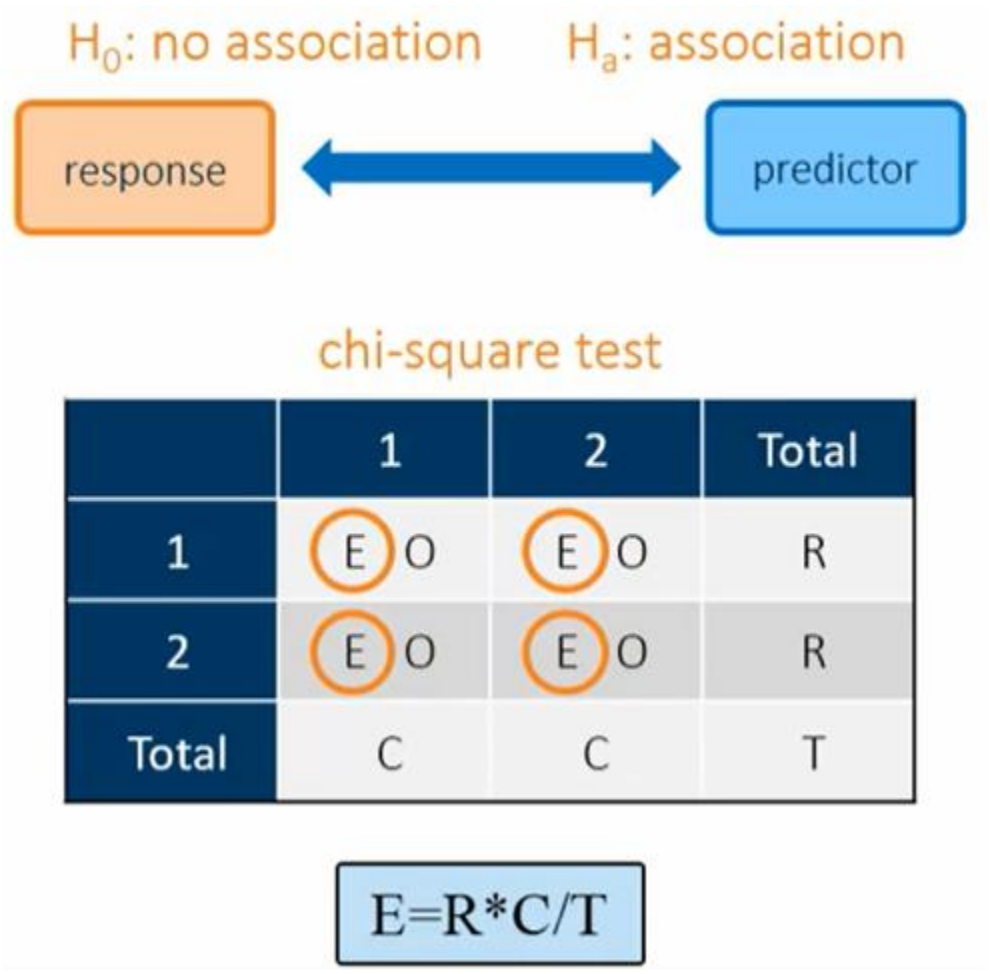

differences
>
chance?
Bonus
Fireplaces
Lot_Shape_2
Basement_Area
association
chi-square test
magnitude
Cramer's V statistic
Spearman correlation statistic

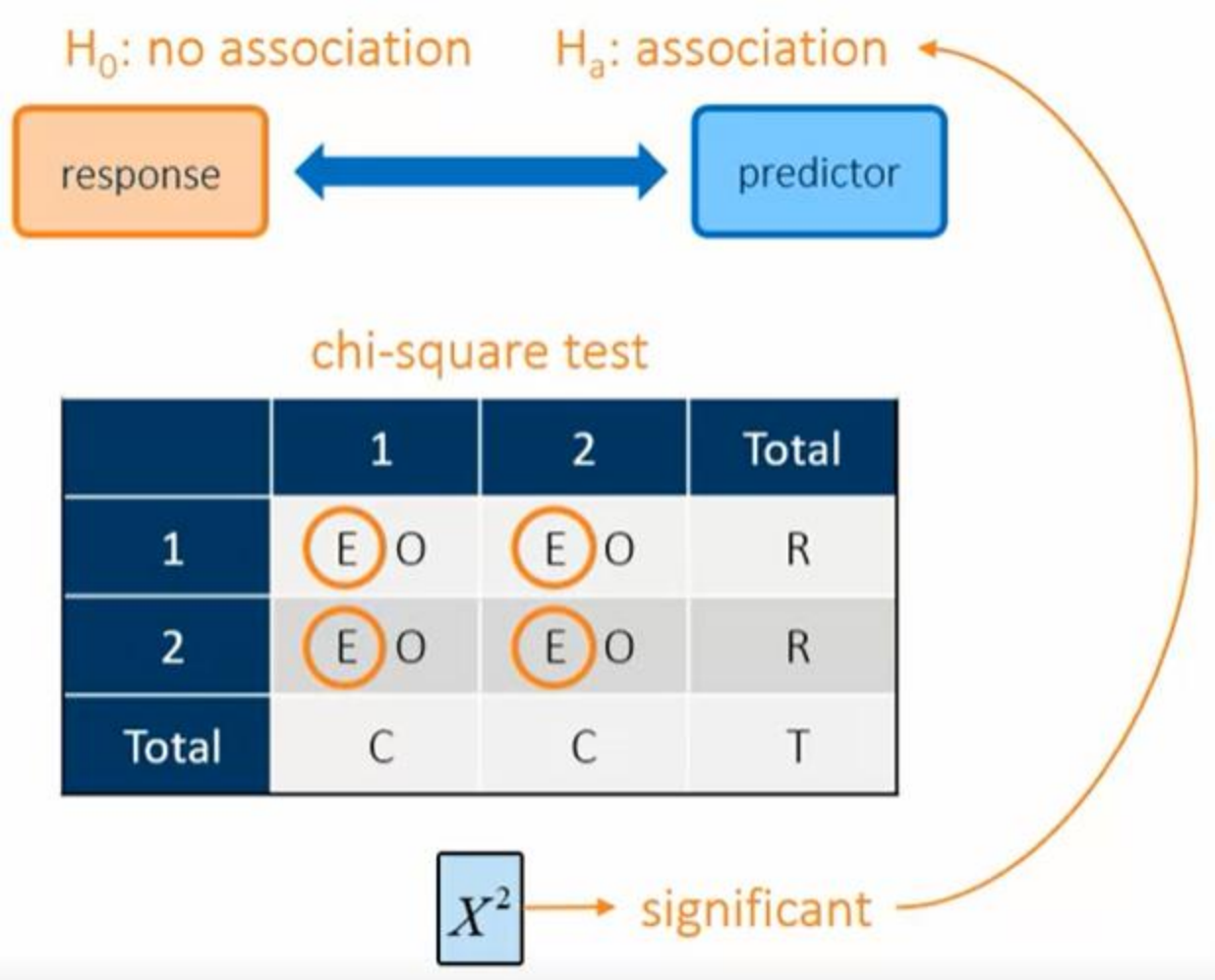
The Pearson Chi-Square Test






H_0: no association
H_a: association
Bonus
?
Lot_Shape_2
Pearson
chi-square test


H_0: no association
H_a: association
response
predictor
chi-square test
1
2
Total
1
2
Total
E O
E O
E O
E O
R
R
C
C
T
E=R*C/T



$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\sum\sum\left(\frac{(\text{observed}_{rc} - \text{expected}_{rc})^2}{\text{expected}_{rc}}\right)$$

degrees of freedom:
(number rows – 1) * (number columns – 1)

chi-square test

	1	2	Total
1	E O	E O	R
2	E O	E O	R
Total	C	C	T

degrees of freedom:
$(2 - 1) * (2 - 1)$

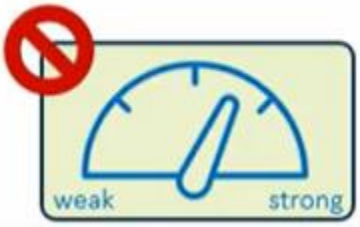
degrees of freedom:
$1 * 1 = 1$

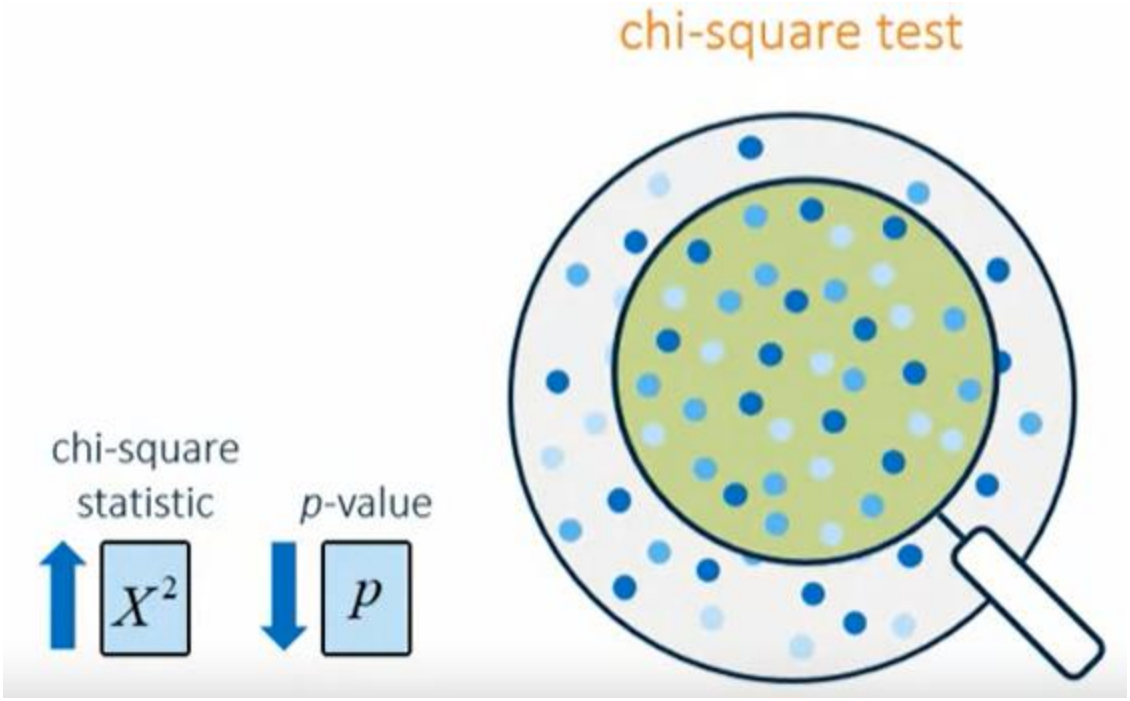
chi-square statistic

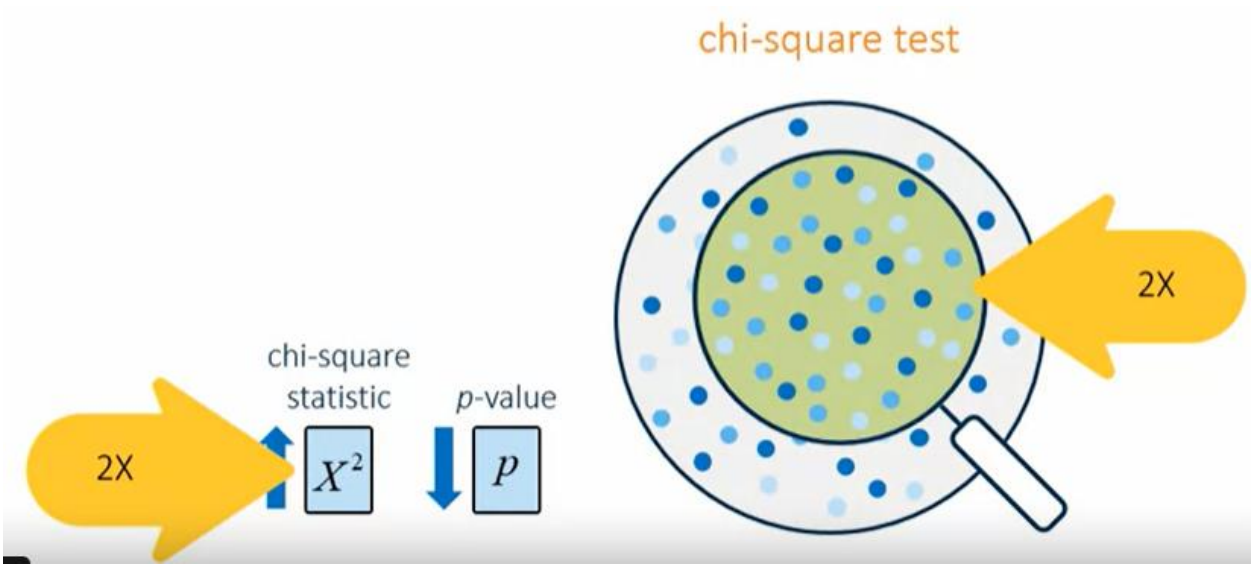
X^2

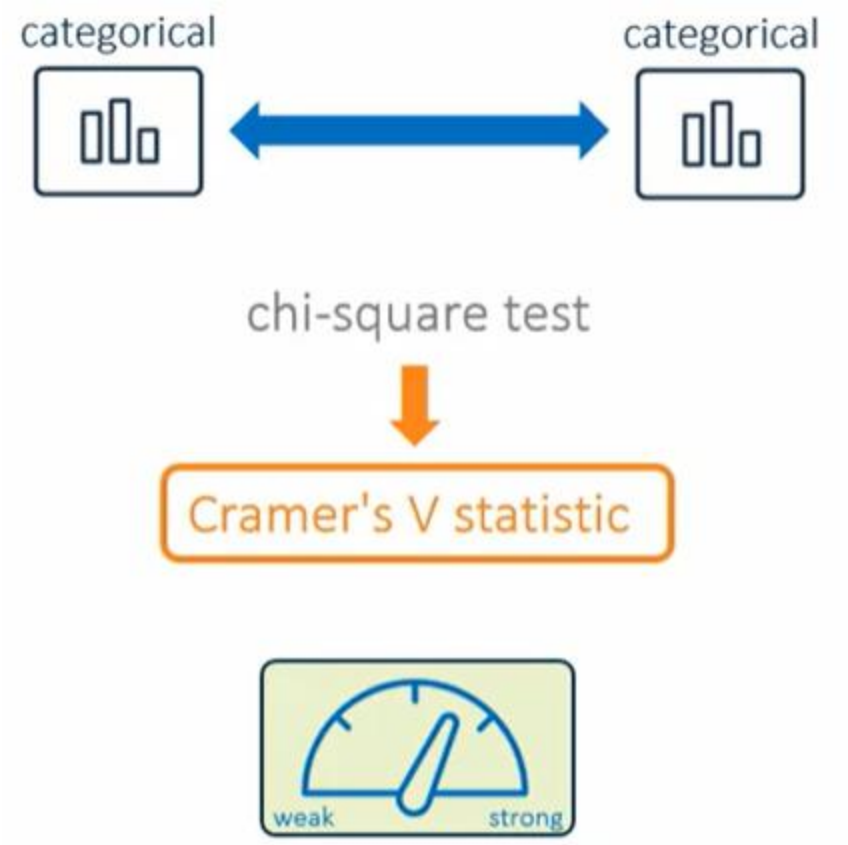
p-value

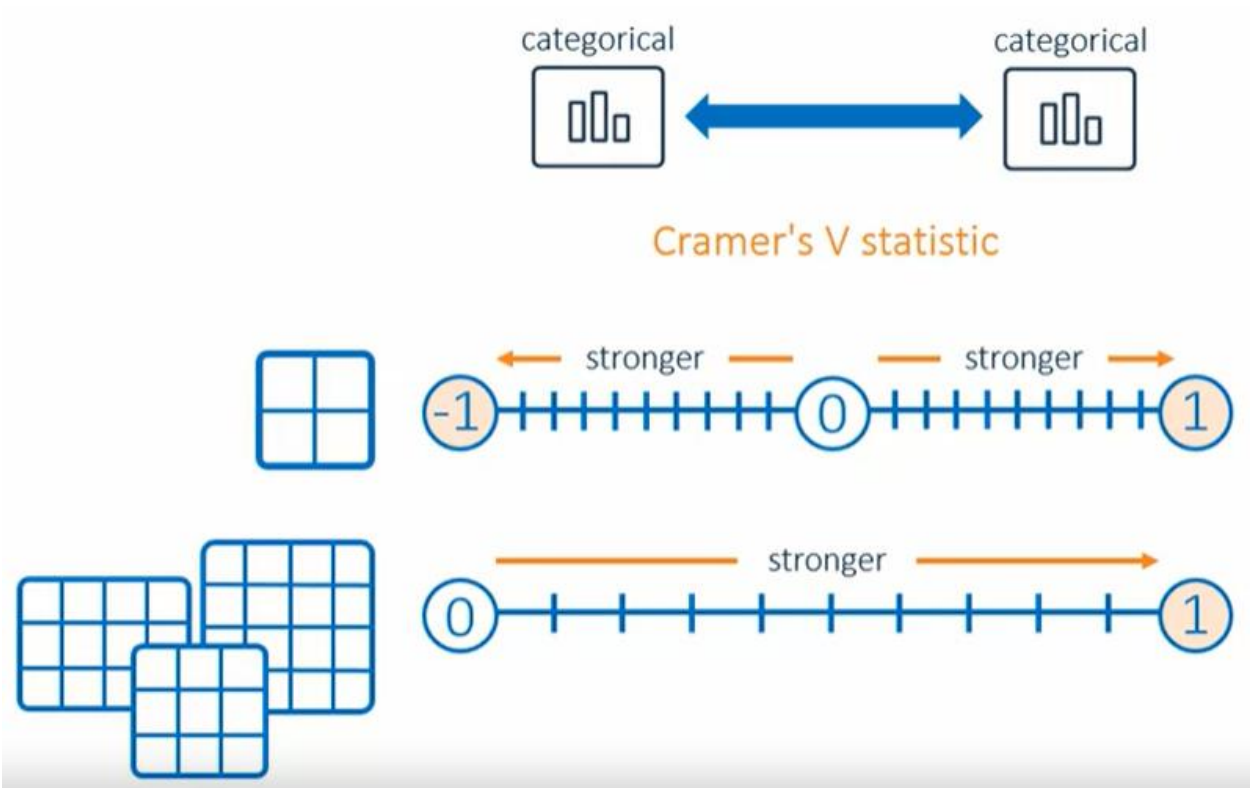
p

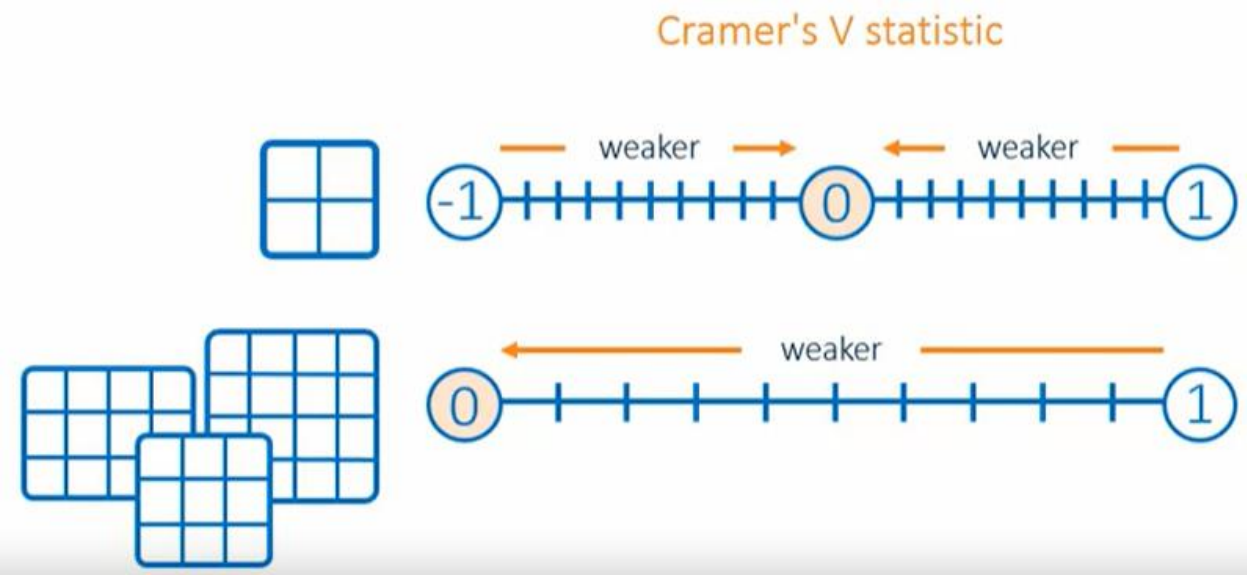



chi-square test
chi-square statistic
X^2
p-value
p

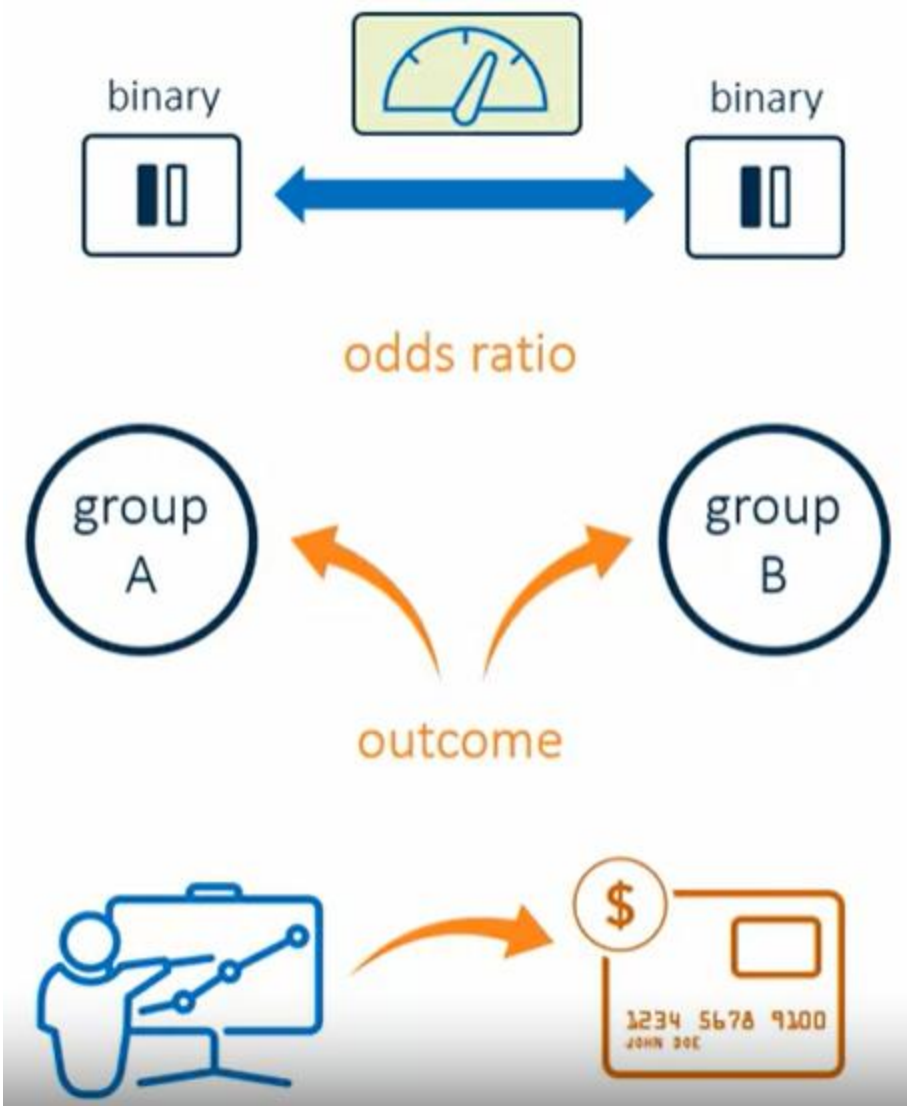

chi-square test
2X
chi-square statistic
X^2
p-value
p
2X

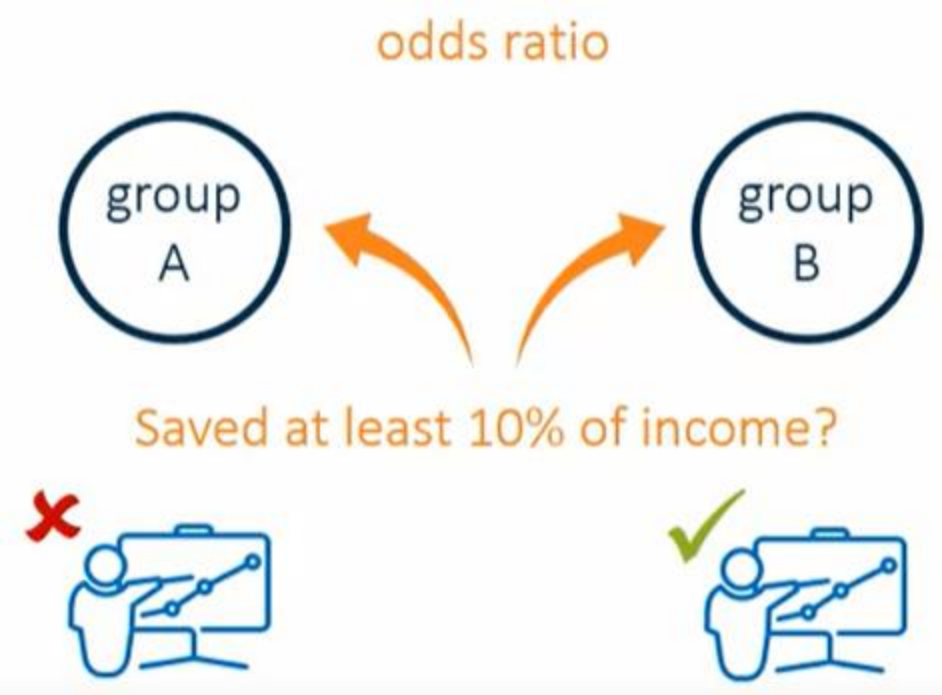

categorical
categorical
chi-square test
Cramer's V statistic
weak
strong

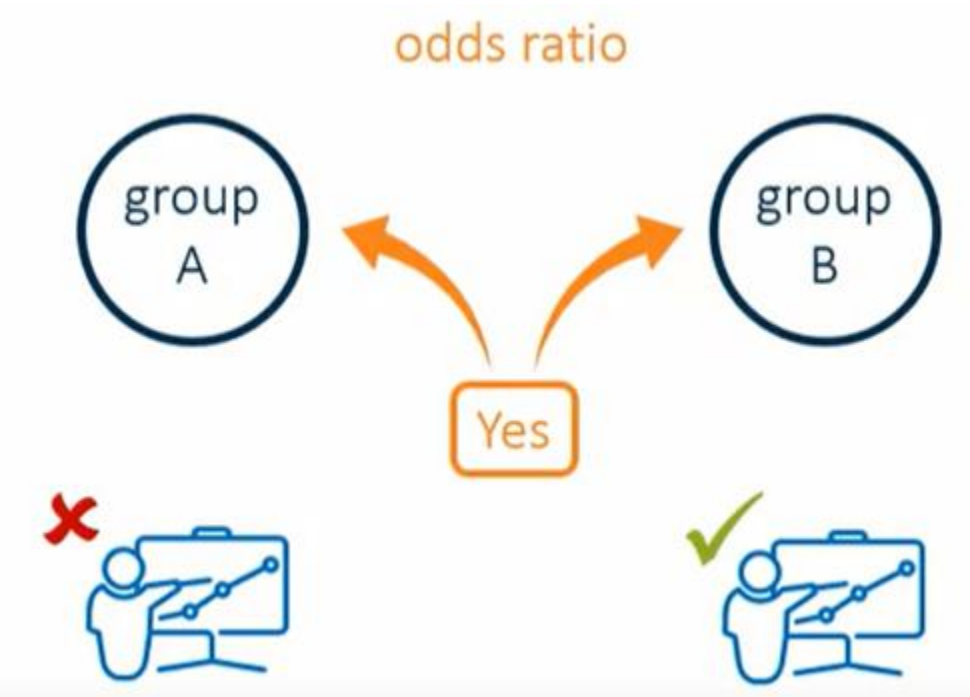

categorical
categorical
Cramer's V statistic
stronger
stronger
-1
0
1
stronger
0
1



Odds Ratios


binary
binary
odds ratio
group A
group B
outcome
$
1234 5678 9100
JOHN DOE





Odds ≠ probabilities

Odds ← probabilities

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

odds ratio

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

Probability of Yes

$90/100 = .90 = 90\%$

odds ratio

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

Probability of No

$10/100 = .10 = 10\%$

odds ratio

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

Probability of Yes

$60/80 = .75 = 75\%$

odds ratio

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

Probability of No

$20/80 = .25 = 25\%$

odds ratio

Group	Outcome		
	Yes	No	Total
A	60 .75	20 .25	80
B	90 .90	10 .10	100
Total	150	30	180

Odds of Yes

$$\frac{.90}{.10} = 9 = 9:1$$

odds ratio

Group	Outcome		
	Yes	No	Total
A	60 .75	20 .25	80
B	90 .90	10 .10	100
Total	150	30	180

Odds of Yes

$$\frac{.75}{.25} = 3 = 3:1$$

odds ratio

9/3

Group	Outcome		
	Yes	No	Total
A	60 .75	20 .25	80
B	90 .90	10 .10	100
Total	150	30	180

3 : 1

9 : 1

odds ratio

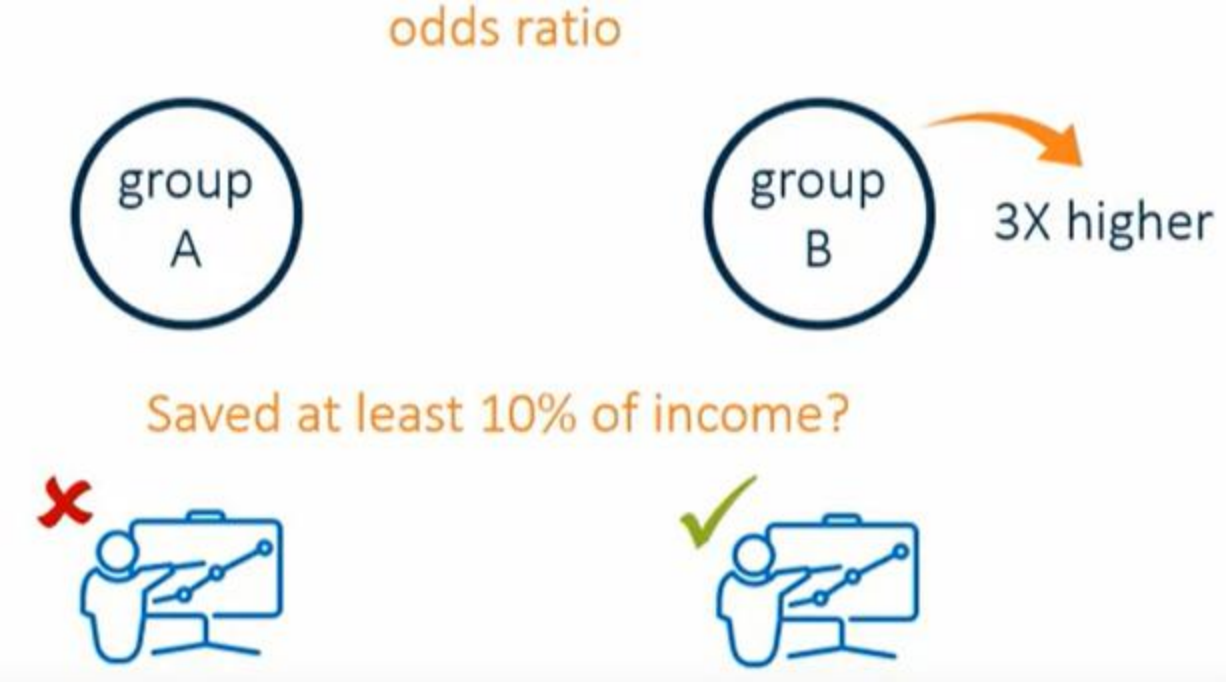
3

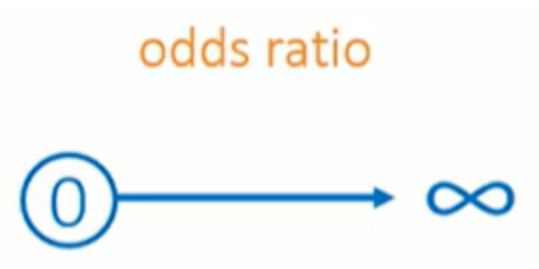
3X

Group	Outcome		
	Yes	No	Total
A	60 .75	20 .25	80
B	90 .90	10 .10	100
Total	150	30	180

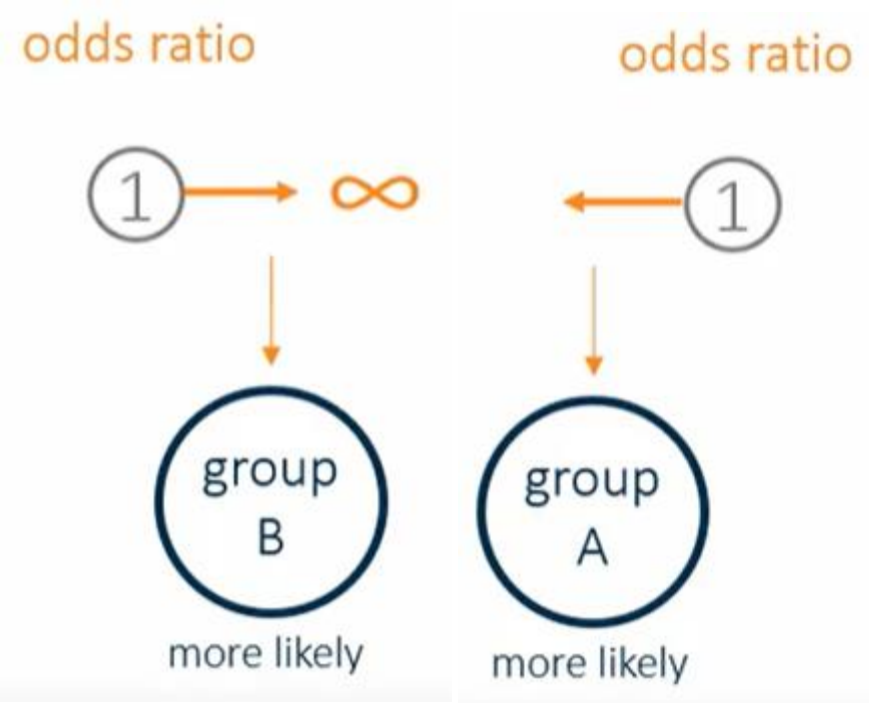
3 : 1

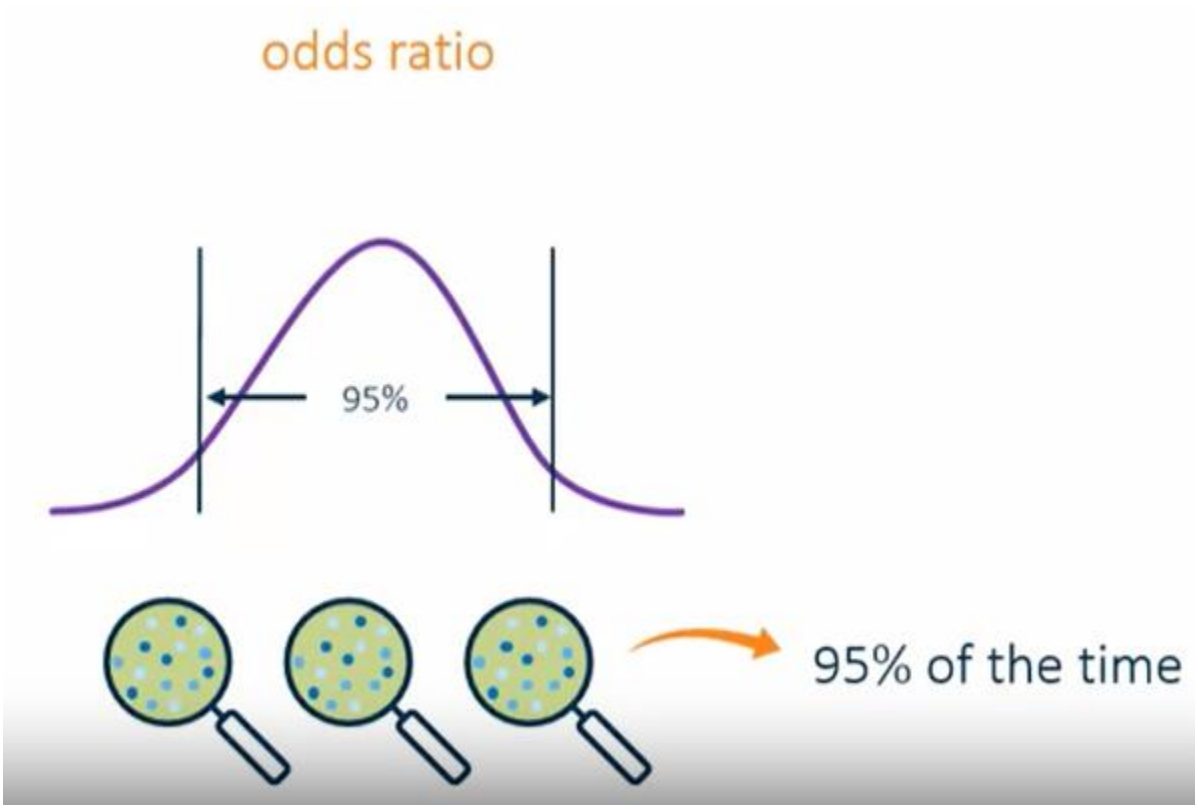
9 : 1

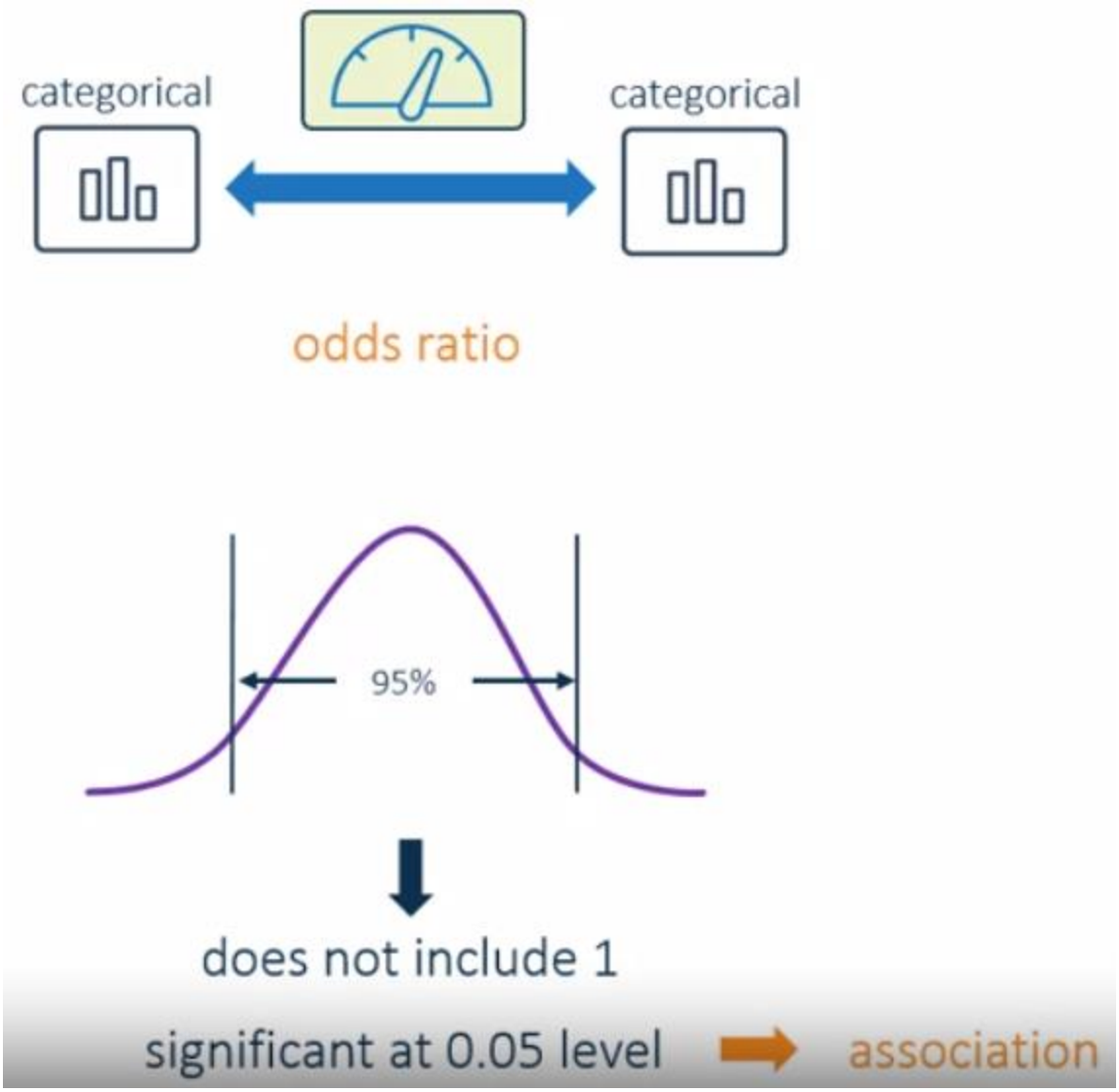

odds ratio
group A
group B
3X higher
Saved at least 10% of income?


odds ratio
0
∞


odds ratio
1
no association


odds ratio
1
∞
group B
more likely
odds ratio
1
group A
more likely


odds ratio
95%
95% of the time


categorical
categorical
odds ratio
95%
does not include 1
significant at 0.05 level
association

Question 7.03

The following crosstabulation table shows Titanic survivors by **Gender**. What is the odds ratio of survival of **females** to **males**?

Table of Gender by Survived			
Gender	Survived		
Frequency Row Pct	no	yes	Total
female	126	344	470
	26.81	73.19	
male	1364	367	1731
	78.80	21.20	
Total	1490	711	2201

The odds of survival for **females** are (344/470) / (126/470) = 2.73. The odds of survival for **males** are (367/1731) / (1364/1731) = 0.27. The odds ratio of survival of **females** to **males** is 2.73/.27 = 10.1. The odds of a **female** surviving are 10 times higher than the odds of a **male** surviving.

Demo Performing a Pearson Chi-Square Test of Association Using PROC FREQ

```
/*st107d02.sas*/
ods graphics off;
proc freq data=STAT1.ameshousing3;
    tables (Lot_Shape_2 Fireplaces)*Bonus
          / chisq expected cellchi2 nocol nopercent
            relrisk;
    format Bonus bonusfmt.;
    title 'Associations with Bonus';
run;

ods graphics on;
```

PROC FREQ DATA=*SAS-data-set***;**
TABLES *table-request(s) < / options>***;**
<additional statements>
RUN;

Associations with Bonus

The FREQ Procedure

Frequency
Expected
Cell Chi-Square
Row Pct

Table of Lot_Shape_2 by Bonus			
	Bonus(Sale Price > $175,000)		
Lot_Shape_2(Regular or irregular lot shape)	Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 79.314 3.7797 66.67	31 13.686 21.905 33.33	93
Regular	193 175.69 1.7064 93.69	13 30.314 9.8893 6.31	206
Total	255	44	299
Frequency Missing = 1			

Statistics for Table of Lot_Shape_2 by Bonus

Statistic	DF	Value	Prob
Chi-Square	1	37.2807	<.0001
Likelihood Ratio Chi-Square	1	34.4226	<.0001
Continuity Adj. Chi-Square	1	35.1587	<.0001
Mantel-Haenszel Chi-Square	1	37.1561	<.0001
Phi Coefficient		-0.3531	
Contingency Coefficient		0.3330	
Cramer's V		-0.3531	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	62
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Table of Lot_Shape_2 by Bonus			
	Bonus(Sale Price > $175,000)		
Lot_Shape_2(Regular or irregular lot shape)	Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 79.314 3.7797 66.67	31 13.686 21.905 33.33	93
Regular	193 175.69 1.7064 93.69	13 30.314 9.8893 6.31	206
Total	255	44	299
Frequency Missing = 1			

Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.1347	0.0664	0.2735
Relative Risk (Column 1)	0.7116	0.6137	0.8251
Relative Risk (Column 2)	5.2821	2.9002	9.6202

Effective Sample Size = 299
Frequency Missing = 1

1 / 0.1347 = 7.423

(Odds Ratio - 1) * 100

Table of Lot_Shape_2 by Bonus			
	Bonus(Sale Price > $175,000)		
Lot_Shape_2(Regular or irregular lot shape)	Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 79.314 3.7797 66.67	31 13.686 21.905 33.33	93
Regular	193 175.69 1.7064 93.69	13 30.314 9.8893 6.31	206
Total	255	44	299
Frequency Missing = 1			

Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.1347	0.0664	0.2735
Relative Risk (Column 1)	0.7116	0.6137	0.8251
Relative Risk (Column 2)	5.2821	2.9002	9.6202

Effective Sample Size = 299
Frequency Missing = 1

(0.1347 - 1) * 100 = 86.53%

regular lots have 86.53% lower odds of being bonus eligible compared with irregular lots

Table of Lot_Shape_2 by Bonus			
	Bonus(Sale Price > $175,000)		
Lot_Shape_2(Regular or irregular lot shape)	Not Bonus Eligible	Bonus Eligible	Total
Irregular	62 79.314 3.7797 66.67	31 13.686 21.905 33.33	93
Regular	193 175.69 1.7064 93.69	13 30.314 9.8893 6.31	206
Total	255	44	299
Frequency Missing = 1			

Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.1347	0.0664	0.2735
Relative Risk (Column 1)	0.7116	0.6137	0.8251
Relative Risk (Column 2)	5.2821	2.9002	9.6202

Effective Sample Size = 299
Frequency Missing = 1

Effective Sample Size = 299
Frequency Missing = 1

Frequency
Expected
Cell Chi-Square
Row Pct

Table of Fireplaces by Bonus

Fireplaces(Number of fireplaces)	Bonus(Sale Price > $175,000) Not Bonus Eligible	Bonus Eligible	Total
0	177 165.75 0.7636 90.77	18 29.25 4.3269 9.23	195
1	68 79.05 1.5446 73.12	25 13.95 8.7529 26.88	93
2	10 10.2 0.0039 83.33	2 1.8 0.0222 16.67	12
Total	255	45	300

Statistics for Table of Fireplaces by Bonus

Statistic	DF	Value	Prob
Chi-Square	2	15.4141	0.0004
Likelihood Ratio Chi-Square	2	14.4859	0.0007
Mantel-Haenszel Chi-Square	1	10.7458	0.0010
Phi Coefficient		0.2267	
Contingency Coefficient		0.2211	
Cramer's V		0.2267	

Sample Size = 300

```
/*st107d02.sas*/
ods graphics off;
proc freq data=STAT1.ameshousing3;
   tables (Lot_Shape_2 Fireplaces)*Bonus
        / chisq expected cellchi2 nocol nopercent
          relrisk;
   format Bonus bonusfmt.;
   title 'Associations with Bonus';
run;

ods graphics on;
```

Question 7.04

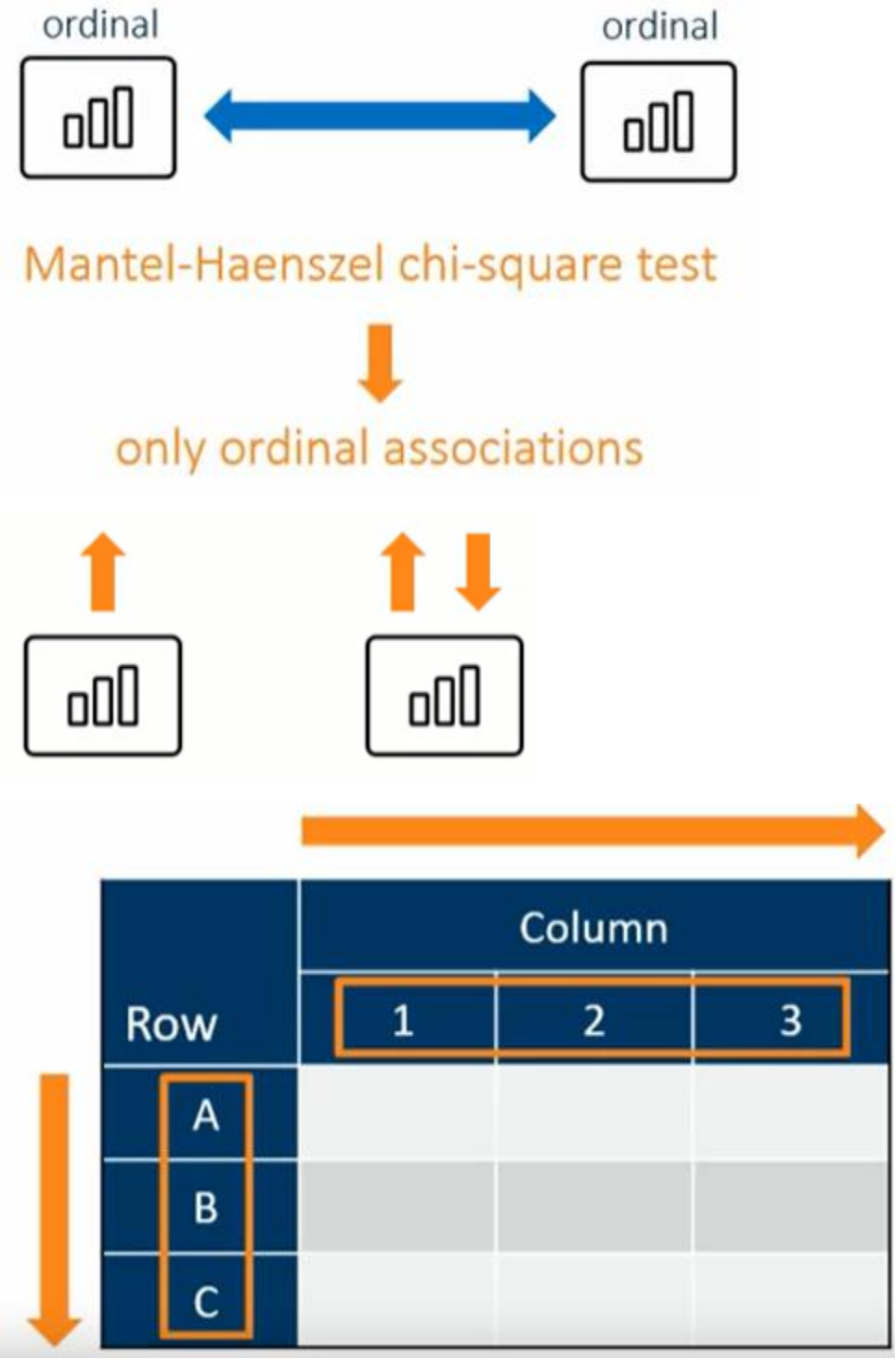
Which of the following tends to occur when a sample size decreases? Select all that apply.

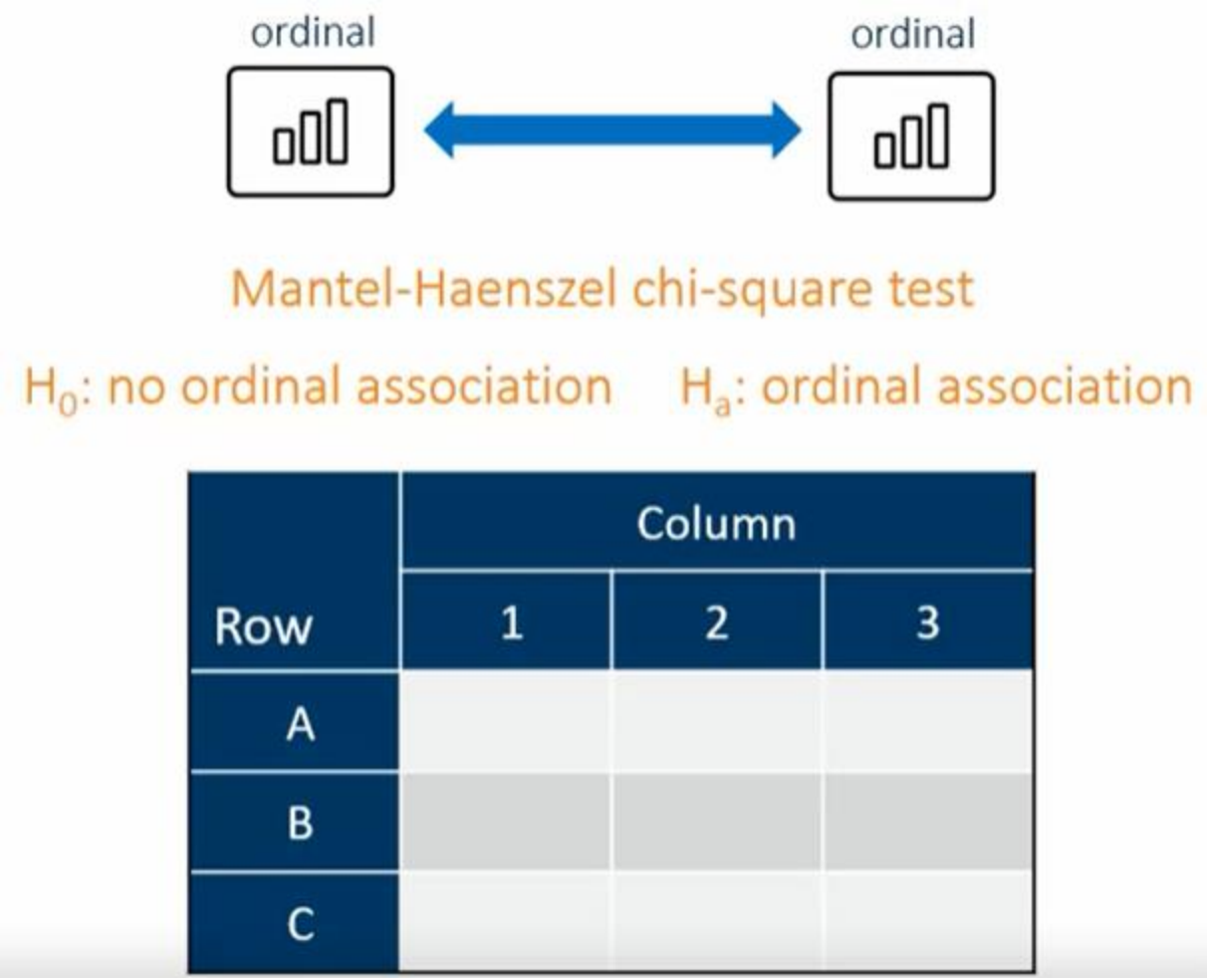
When your sample size decreases, your chi-square value decreases, causing both the *p*-value and CI width to increase because the hypothesis test becomes less significant.

The Mantel-Haenszel Chi-Square Test

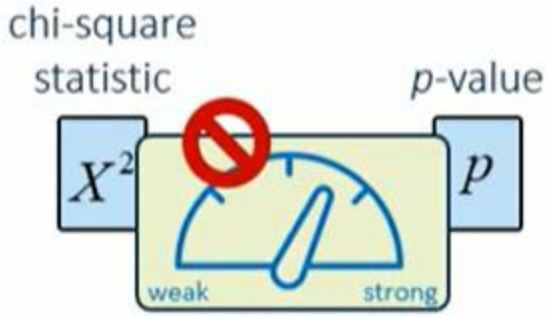
Scenario

The Mantel-Haenszel Chi-Square Test

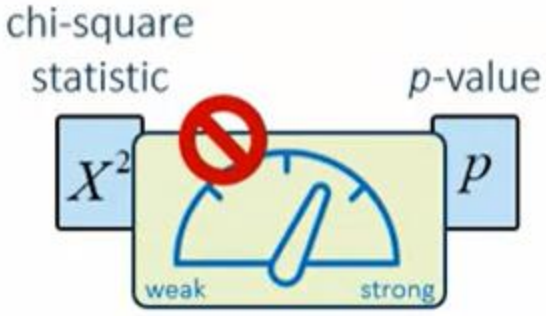



ordinal
ordinal
Mantel-Haenszel chi-square test
H_0: no ordinal association
H_a: ordinal association
Row
Column
1
2
3
A
B
C

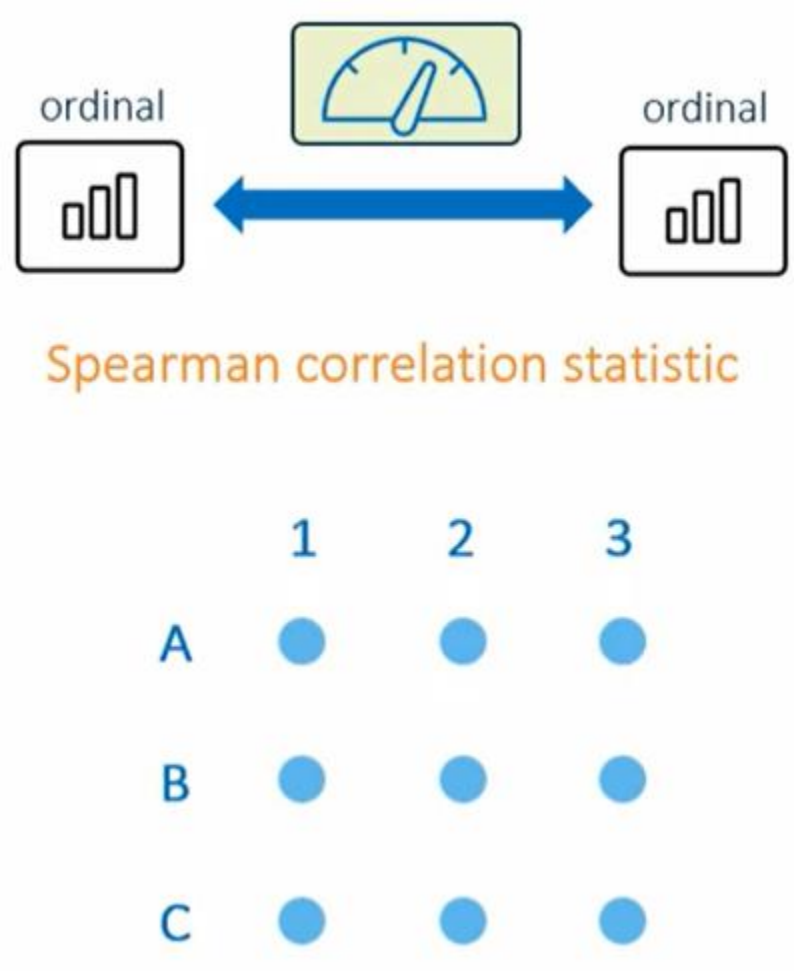
Mantel-Haenszel chi-square test

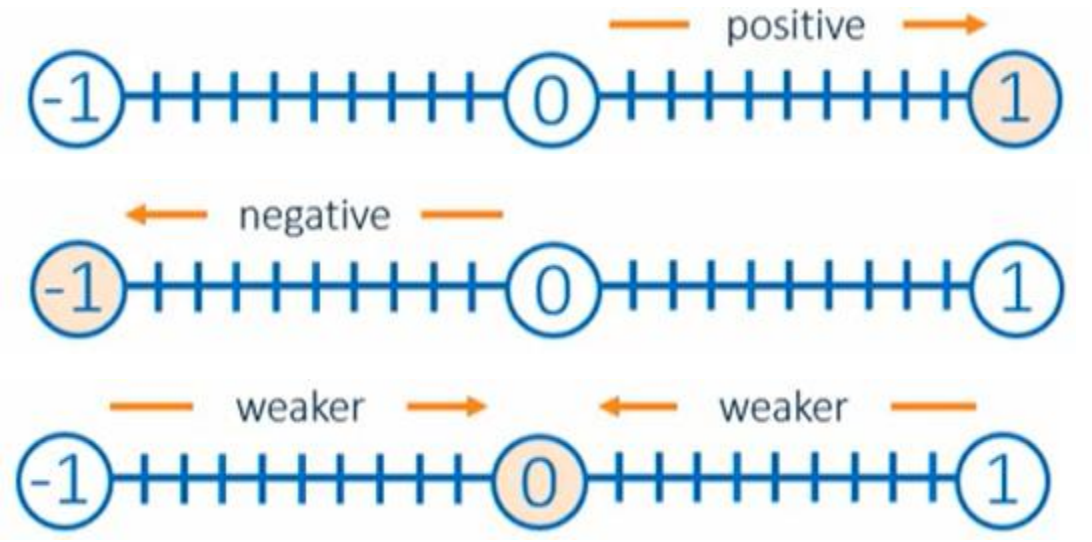


Pearson chi-square test



The Spearman Correlation Statistic





Question 7.05

Suppose you're testing for an ordinal association between people's income and their body mass index (BMI). The levels of the **Income** and **BMI** variables are shown in the table below. The Mantel-Haenszel chi-square *p*-value is 0.01 and the Spearman correlation statistic is 0.253. What can you conclude about the association between **Income** and **BMI**?

Variable	Levels
Income	1_low
	2_medium
	3_high
BMI	1_underweight
	2_normal
	3_overweight
	4_obese

There is a positive, ordinal association.

The *p*-value is significant (less than 0.05) so there is an ordinal association. The Spearman correlation statistic is positive, so the ordinal association is positive.

Demo Detecting Ordinal Associations Using PROC FREQ

```
1 /*st107d03.sas*/
2 ods graphics off;
3 proc freq data=STAT1.ameshousing3;
4     tables Fireplaces*Bonus / chisq measures cl;
5     format Bonus bonusfmt.;
6     title 'Ordinal Association between FIREPLACES and BONUS?';
7 run;
8
9 ods graphics on;
```

```
PROC FREQ DATA=SAS-data-set;
    TABLES table-request(s) < / options>;
    <additional statements>
RUN;
```

Statistics for Table of Fireplaces by Bonus

Statistic	DF	Value	Prob
Chi-Square	2	15.4141	0.0004
Likelihood Ratio Chi-Square	2	14.4859	0.0007
Mantel-Haenszel Chi-Square	1	10.7458	0.0010
Phi Coefficient		0.2267	
Contingency Coefficient		0.2211	
Cramer's V		0.2267	

Statistic	Value	ASE	95% Confidence Limits	
Gamma	0.4964	0.1111	0.2786	0.7143
Kendall's Tau-b	0.2072	0.0585	0.0926	0.3218
Stuart's Tau-c	0.1449	0.0433	0.0600	0.2298
Somers' D C|R	0.1510	0.0451	0.0626	0.2395
Somers' D R|C	0.2842	0.0786	0.1301	0.4383
Pearson Correlation	0.1896	0.0591	0.0737	0.3054
Spearman Correlation	0.2107	0.0594	0.0943	0.3272
Lambda Asymmetric C|R	0.0000	0.0000	0.0000	0.0000
Lambda Asymmetric R|C	0.0667	0.0603	0.0000	0.1849
Lambda Symmetric	0.0467	0.0424	0.0000	0.1298
Uncertainty Coefficient C|R	0.0571	0.0298	0.0000	0.1156
Uncertainty Coefficient R|C	0.0313	0.0167	0.0000	0.0640
Uncertainty Coefficient Symmetric	0.0404	0.0213	0.0000	0.0823

Sample Size = 300

```
/*st107d03.sas*/
ods graphics off;
proc freq data=STAT1.ameshousing3;
   tables Fireplaces*Bonus / chisq measures cl;
   format Bonus bonusfmt.;
   title 'Ordinal Association between FIREPLACES and BONUS?';
run;

ods graphics on;
```

Ordinal Association between FIREPLACES and BONUS?

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Fireplaces by Bonus		
	Bonus(Sale Price > $175,000)		
Fireplaces(Number of fireplaces)	0	1	Total
0	177 59.00 90.77 69.41	18 6.00 9.23 40.00	195 65.00
1	68 22.67 73.12 26.67	25 8.33 26.88 55.56	93 31.00
2	10 3.33 83.33 3.92	2 0.67 16.67 4.44	12 4.00
Total	255 85.00	45 15.00	300 100.00

Statistics for Table of Fireplaces by Bonus

Statistic	DF	Value	Prob
Chi-Square	2	15.4141	0.0004
Likelihood Ratio Chi-Square	2	14.4859	0.0007
Mantel-Haenszel Chi-Square	1	10.7458	0.0010
Phi Coefficient		0.2267	
Contingency Coefficient		0.2211	
Cramer's V		0.2267	

Statistic	Value	ASE	95% Confidence Limits	
Gamma	0.4964	0.1111	0.2786	0.7143
Kendall's Tau-b	0.2072	0.0585	0.0926	0.3218
Stuart's Tau-c	0.1449	0.0433	0.0600	0.2298
Somers' D C|R	0.1510	0.0451	0.0626	0.2395
Somers' D R|C	0.2842	0.0786	0.1301	0.4383
Pearson Correlation	0.1896	0.0591	0.0737	0.3054
Spearman Correlation	0.2107	0.0594	0.0943	0.3272
Lambda Asymmetric C|R	0.0000	0.0000	0.0000	0.0000
Lambda Asymmetric R|C	0.0667	0.0603	0.0000	0.1849
Lambda Symmetric	0.0467	0.0424	0.0000	0.1298
Uncertainty Coefficient C|R	0.0571	0.0298	0.0000	0.1156
Uncertainty Coefficient R|C	0.0313	0.0167	0.0000	0.0640
Uncertainty Coefficient Symmetric	0.0404	0.0213	0.0000	0.0823

Sample Size = 300

Question 7.06

In the PROC FREQ step below, what TABLES statement would you write to display the crosstabulation of the ordinal variables **Size** and **Severity**. Along with the default output, you want to display the Mantel-Haenszel chi-square test of association and the Spearman correlation statistic with confidence bounds.

```
proc freq data=weather.storms;
    [What TABLES statement goes here?]
run;
```

The following TABLES statement will create the specified tables:

```
tables Size*Severity / chisq measures cl;
```

Note: The options can appear in any order.

```
/*st107s01.sas*/ /*Part A*/

ods graphics off;

proc freq data=STAT1.safety;

   tables Unsafe Type Region Size;

   title "Safety Data Frequencies";

run;
```

Safety Data Frequencies

The FREQ Procedure

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

```
/*st107s01.sas*/  /*Part B*/
proc format;
  value safefmt 0='Average or Above'
                1='Below Average';
run;

proc freq data=STAT1.safety;
   tables Region*Unsafe / expected chisq relrisk;
```

```
   format Unsafe safefmt.;
   title "Association between Unsafe and Region";
run;
```

Association between Unsafe and Region

The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct	Table of Region by Unsafe		
	Unsafe		
Region	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

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.4348	0.1790	1.0562
Relative Risk (Column 1)	0.7578	0.5499	1.0443
Relative Risk (Column 2)	1.7429	0.9733	3.1210

Sample Size = 96

```
/*st107s01.sas*/  /*Part C*/
proc freq data=STAT1.safety;
   tables Size*Unsafe / chisq measures cl;
   format Unsafe safefmt.;
   title "Association between Unsafe and Size";
run;
```

Association between Unsafe and Size

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Size by Unsafe		
	Unsafe		
Size	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

Sample Size = 96

Practice: Using PROC FREQ to Perform Tests and Measures of Association

TOTAL POINTS 8

Question 1

The insurance company wants to determine whether a vehicle's safety score is associated with either the region in which it was manufactured or the vehicle's size. The **stat1.safety** data set contains the data about vehicle safety.

1. Use PROC FREQ to create the crosstabulation of the variables **Region** by **Unsafe**. Along with the default output, generate the expected frequencies, the chi-square test of association, and the odds ratio. To clearly identify the values of **Unsafe**, create and apply a temporary format. Submit the code and view the results.
2. 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. Of the cars made in Asia, 42.86% have a Below Average safety score.

```
/*st107s01.sas*/  /*Part B*/

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

```
proc freq data=STAT1.safety;
   tables Region*Unsafe / expected chisq relrisk;
   format Unsafe safefmt.;
   title "Association between Unsafe and Region";
run;
```

Question 2

For the cars with an Average or Above safety score, what percentage was made in North America?

Look at the Col Pct value in the *Average or Above* cell of the *N America* row. Of the cars with an Average or Above safety score, 69.70% were made in North America.

Question 3

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

No, The association is not statistically significant at the 0.05 alpha level. The *p*-value is 0.0631.

Question 4

What does the odds ratio compare? What does this suggest 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% lower odds for being unsafe than cars made in Asia.

Note: Recall that the odds ratios in the Estimates of Relative Risk table are calculated by comparing row1/row2 for column1. In this problem, this comparison is Asia to N America and the outcome is Average or Above in safety. The value 0.4348 is interpreted as the odds of an Average or Above car made in Asia is 0.4348 times the odds for American-made cars. If you want 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 want to compare N America to Asia using Below Average in safety, you invert your odds ratio twice and return to the value 0.4348.

Question 5

Write another PROC FREQ step to create the crosstabulation of the variables **Size** and **Unsafe**. Along with the default output, generate the measures of ordinal association. Format the values of **Unsafe**. Submit the code and view the results.

What statistic do you use to detect an ordinal association between **Size** and **Unsafe**?

The Mantel-Haenszel chi-square detects an ordinal association.

```
/*st107s01.sas*/  /*Part C*/

proc freq data=STAT1.safety;
   tables Size*Unsafe / chisq measures cl;
   format Unsafe safefmt.;
   title "Association between Unsafe and Size";
run;
```

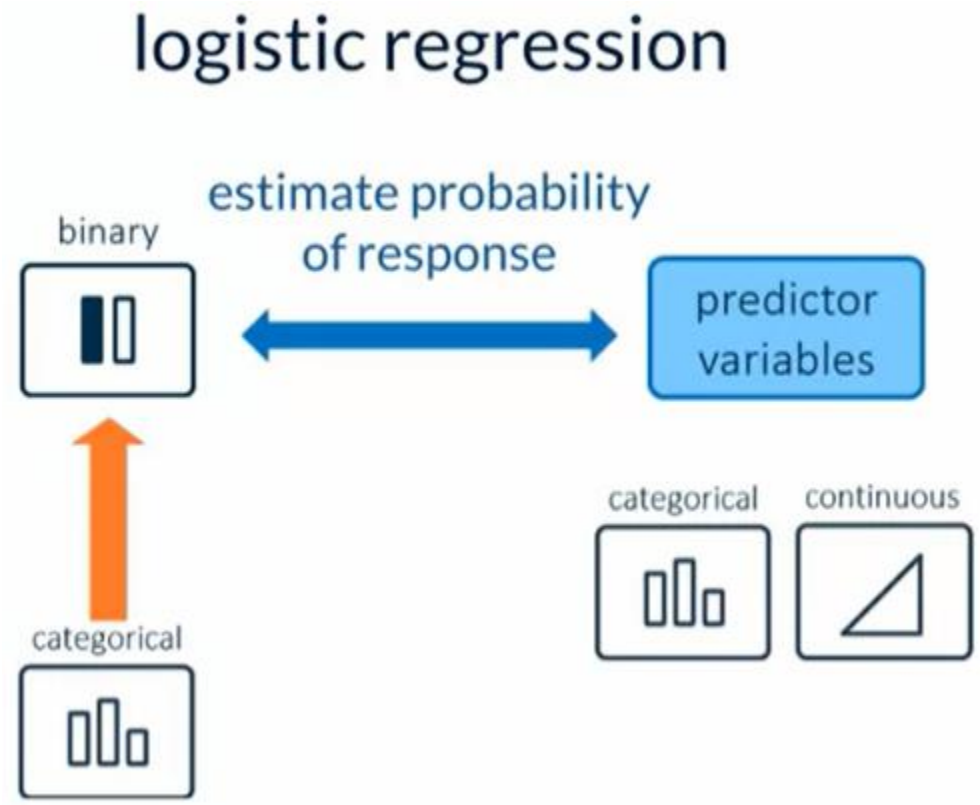
Question 6
Do you reject or fail to reject the null hypothesis at the 0.05 level?
You reject the null hypothesis at the 0.05 level.

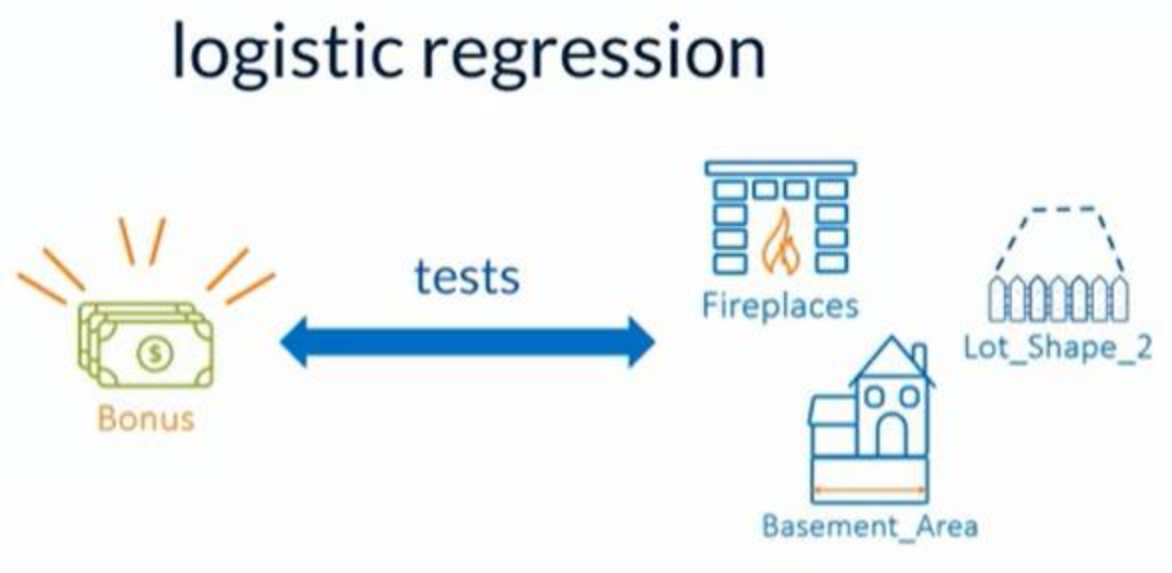
Question 7
What is the strength of the ordinal association between **Size** and **Unsafe**?
The Spearman Correlation is -0.5425.

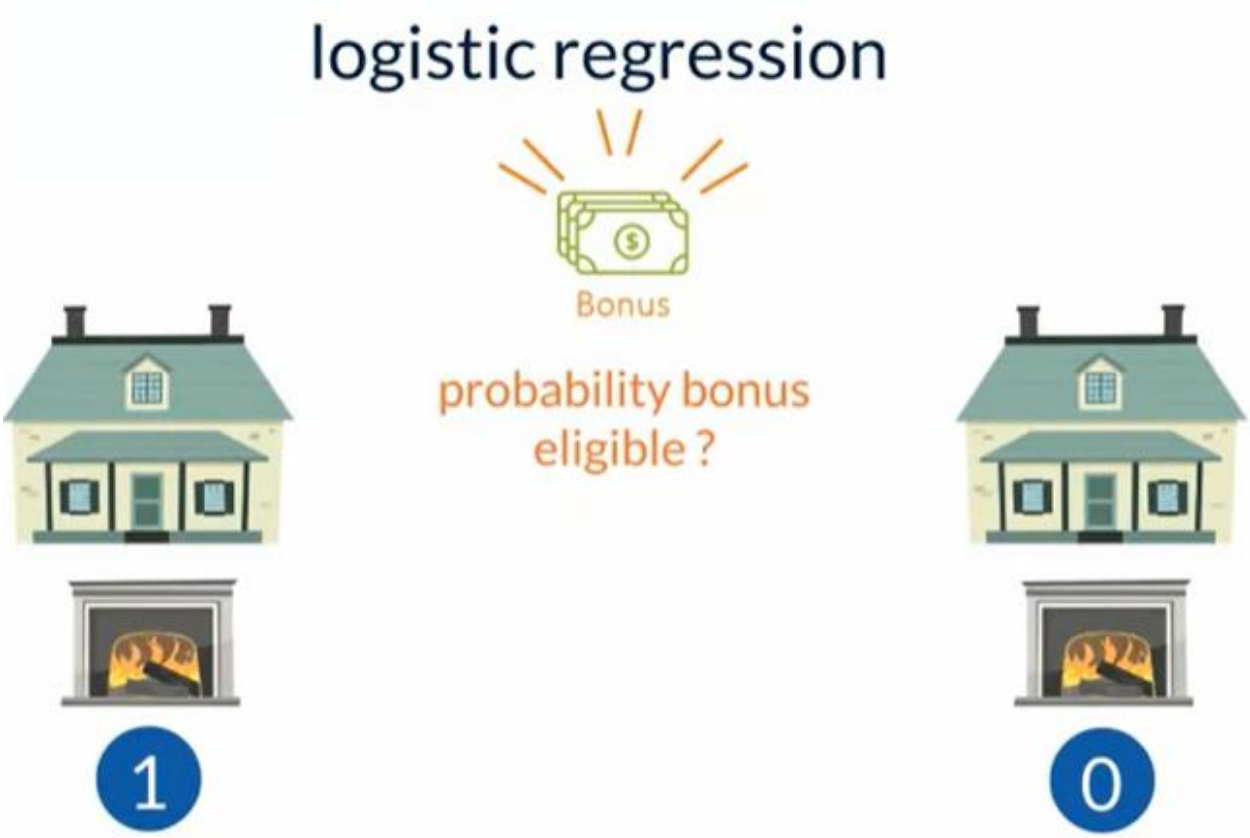
Question 8
What is the 95% confidence interval around the statistic that measures the strength of the ordinal association?
The confidence interval is (-0.6932, -0.3917).

Introduction to Logistic Regression

Scenario



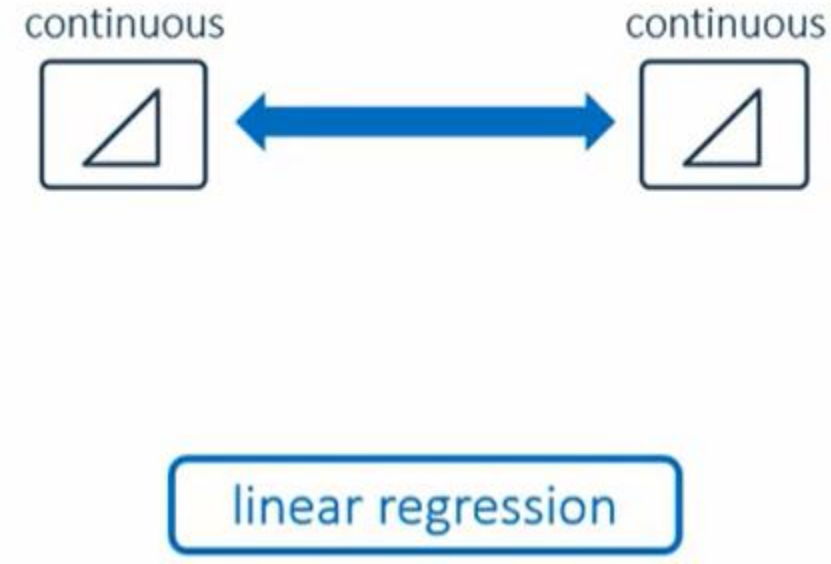

logistic regression
tests
Bonus
Fireplaces
Lot_Shape_2
Basement_Area


logistic regression
Bonus
probability bonus eligible ?
1
0

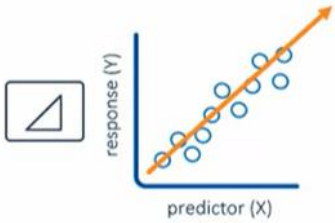

logistic regression
Bonus
probability bonus
eligible ?
irregular
regular


logistic regression
Bonus
probability bonus
eligible ?
500 sq ft
1000 sq ft

Modeling a Binary Response



linear regression



conditional mean of the response:

$$\beta_0 + \beta_1 X$$

$$-\infty \longleftrightarrow \infty$$

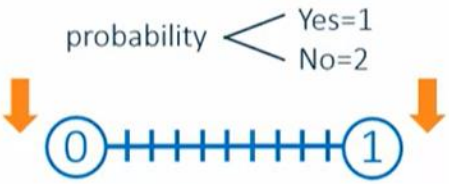
logistic regression

mean of the response:

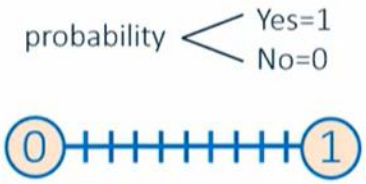
probability of a success

$$\varepsilon \sim N(0, \sigma^2)$$

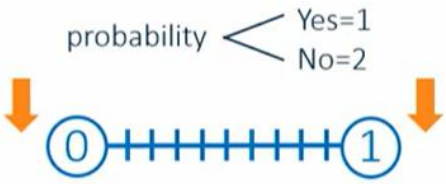
linear regression

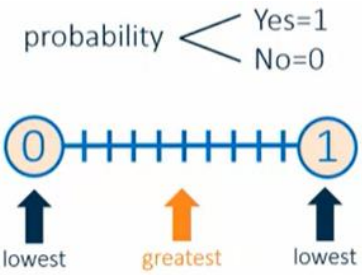


logistic regression

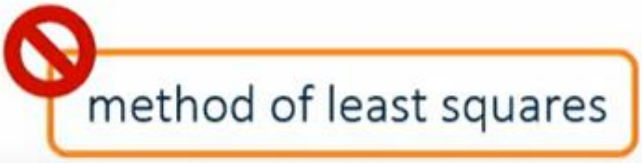


usually nonlinear





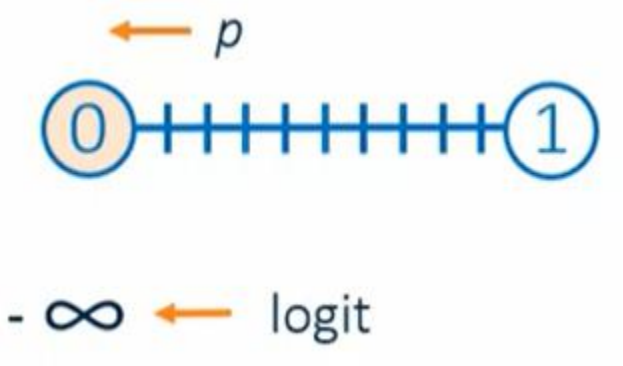
binomial error of $p*(1-p)$

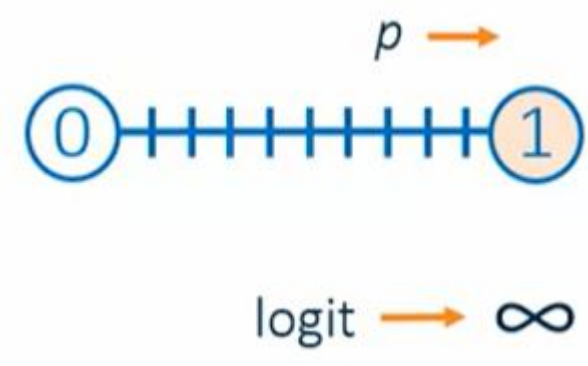


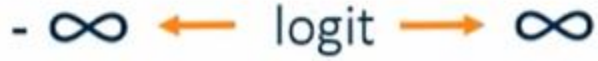
logistic regression

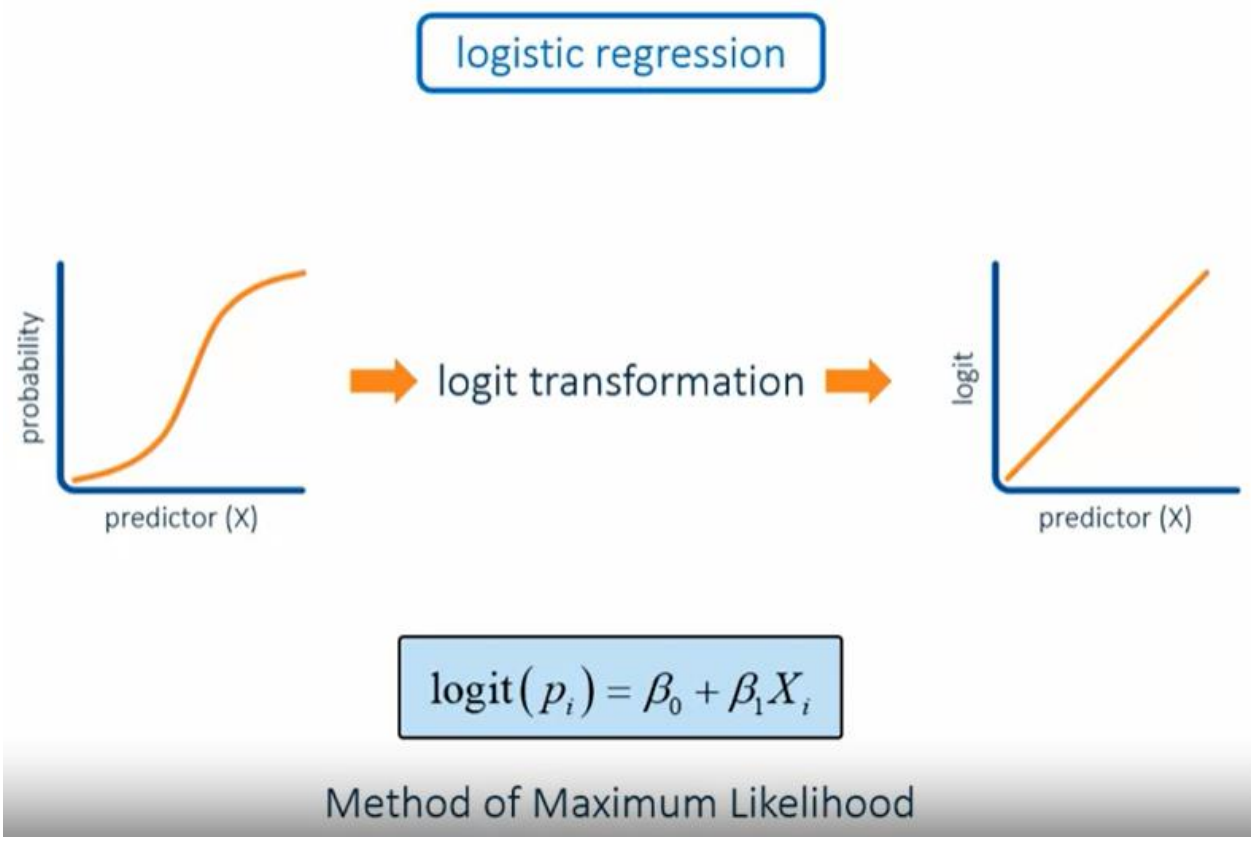
$$\log\left(\frac{p}{1-p}\right)$$

logit transformation
(log odds transformation)









Question 7.07

What are the lower and upper bounds for a logit?

A probability is bounded by 0 and 1. The logit of the probability transforms the probability into a linear function, which has no lower or upper bounds.

Demo Fitting a Binary Logistic Regression Model Using PROC LOGISTIC

```
1 /*st107d04.sas*/
2 ods graphics on;
3 proc logistic data=STAT1.ameshousing3 alpha=0.05
4             plots(only)=(effect oddsratio);
5    model Bonus(event='1')=Basement_Area / clodds=pl;
6    title 'LOGISTIC MODEL (1):Bonus=Basement_Area';
7 run;
```

PROC LOGISTIC DATA=*SAS-data-set <options>***;**
MODEL *variable <(variable_options)> = <effects> < / options>***;**
RUN;

LOGISTIC MODEL (1):Bonus=Basement_Area

The LOGISTIC Procedure

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > $175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

```
/*st107d04.sas*/
ods graphics on;
proc logistic data=STAT1.ameshousing3 alpha=0.05
              plots(only)=(effect oddsratio);
    model Bonus(event='1')=Basement_Area / clodds=pl;
    title 'LOGISTIC MODEL (1):Bonus=Basement_Area';
run;
```

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	45

Probability modeled is Bonus='1'.

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	255.625	161.838
SC	259.329	169.246
-2 Log L	253.625	157.838

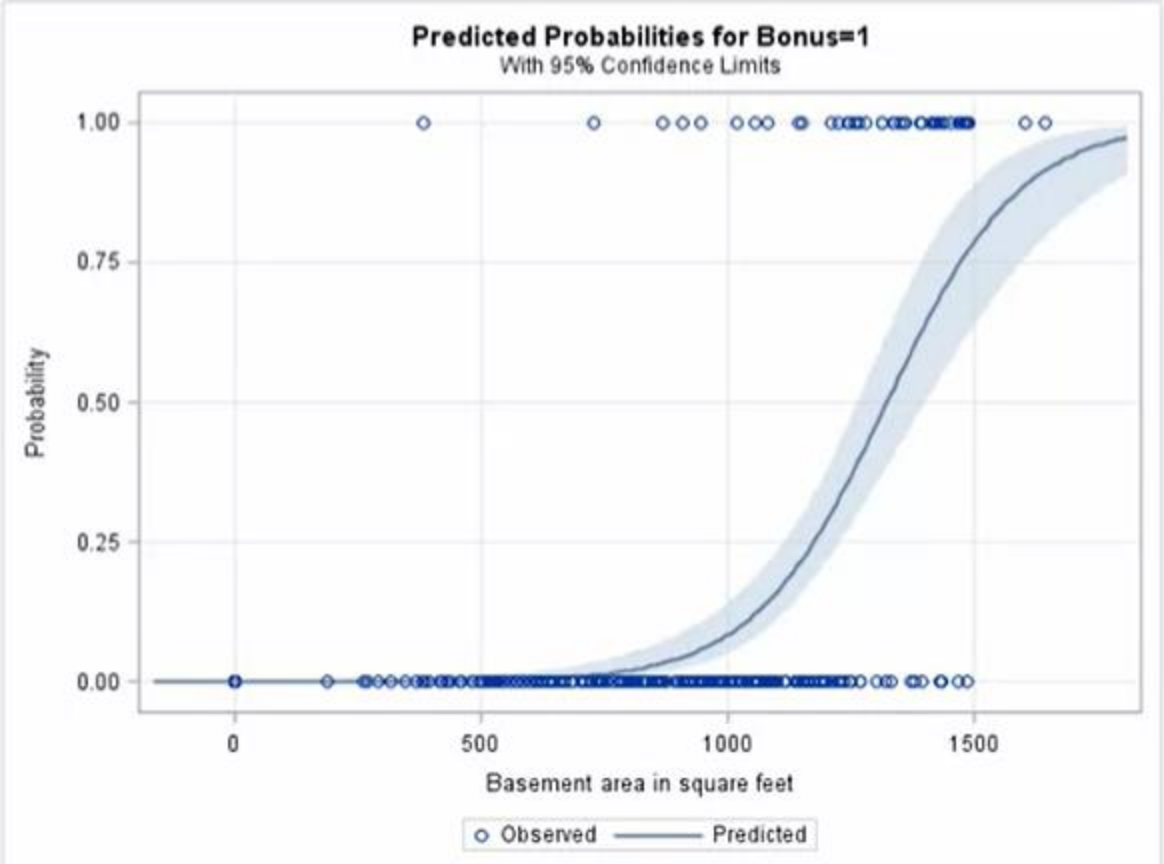
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	95.7870	1	<.0001
Score	65.5624	1	<.0001
Wald	48.0617	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-9.7854	1.2896	57.5758	<.0001
Basement_Area	1	0.00739	0.00107	48.0617	<.0001

$$\text{logit}(\hat{p}) = -9.7854 + (0.00739) * \textbf{Basement\_Area}$$

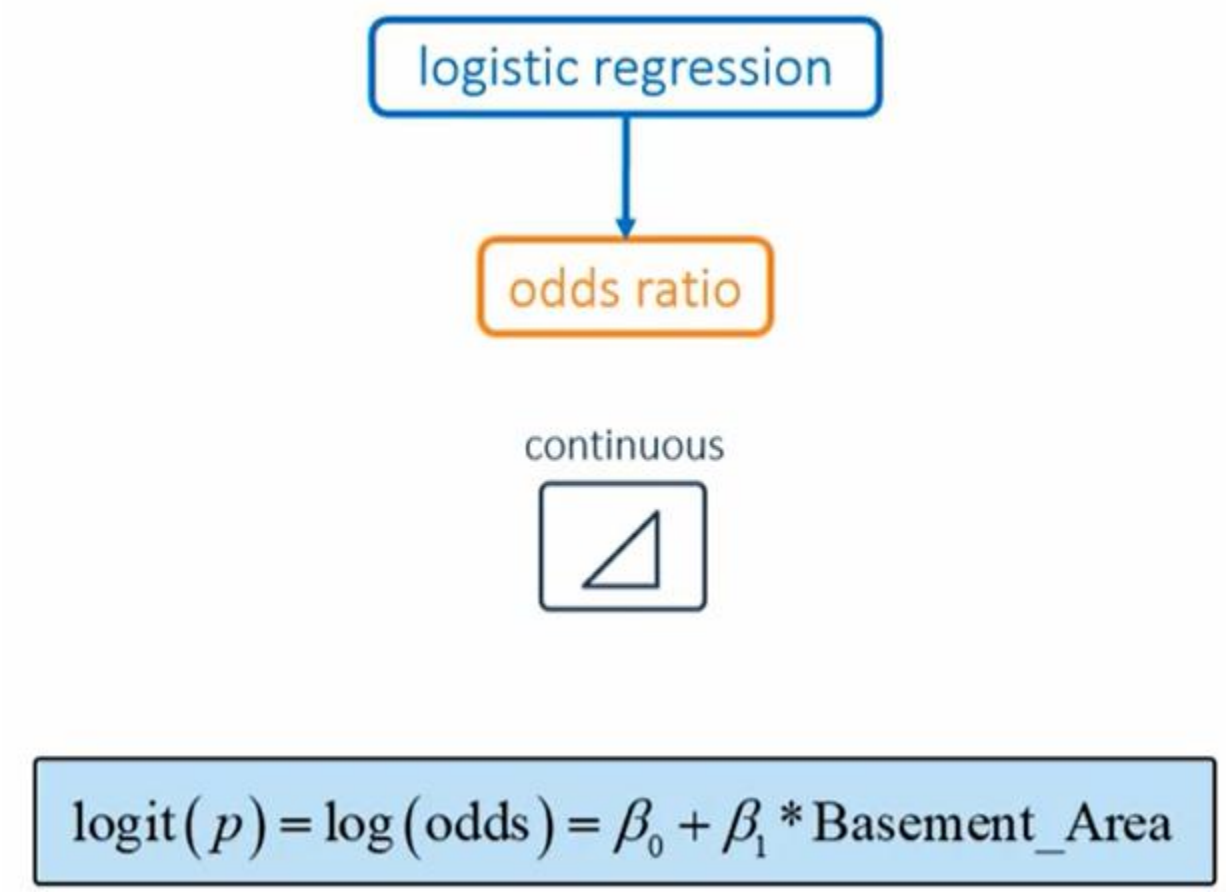
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	Somers' D	0.791
Percent Discordant	10.4	Gamma	0.792
Percent Tied	0.1	Tau-a	0.202
Pairs	11475	c	0.896

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	1.0000	1.007	1.005	1.010




```
/*st107d04.sas*/
ods graphics on;
proc logistic data=STAT1.ameshousing3 alpha=0.05
          plots(only)=(effect oddsratio);
   model Bonus(event='1')=Basement_Area / clodds=pl;
   title 'LOGISTIC MODEL (1):Bonus=Basement_Area';
run;
```

Interpreting the Odds Ratio



$$\text{logit}(p) = \log(\text{odds}) = \beta_0 + \beta_1 * \text{Basement\_Area}$$

odds ratio

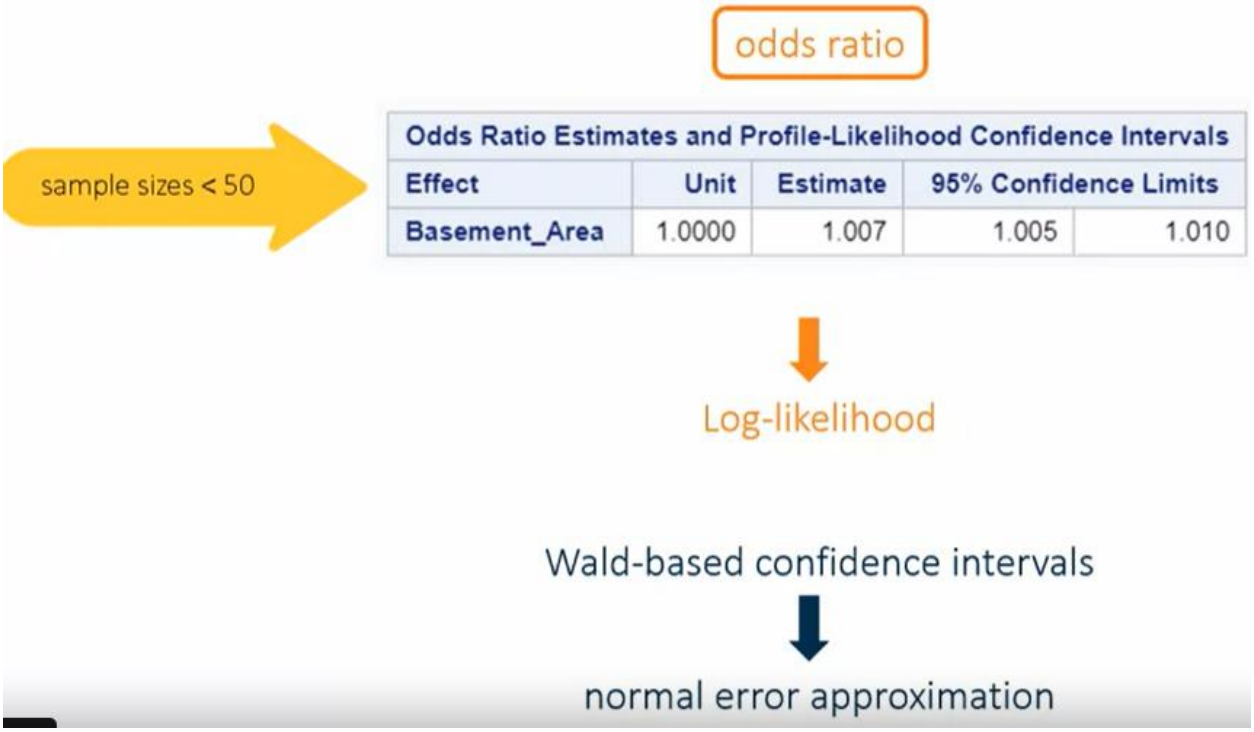
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	1.0000	1.007	1.005	1.010

Basement_Area

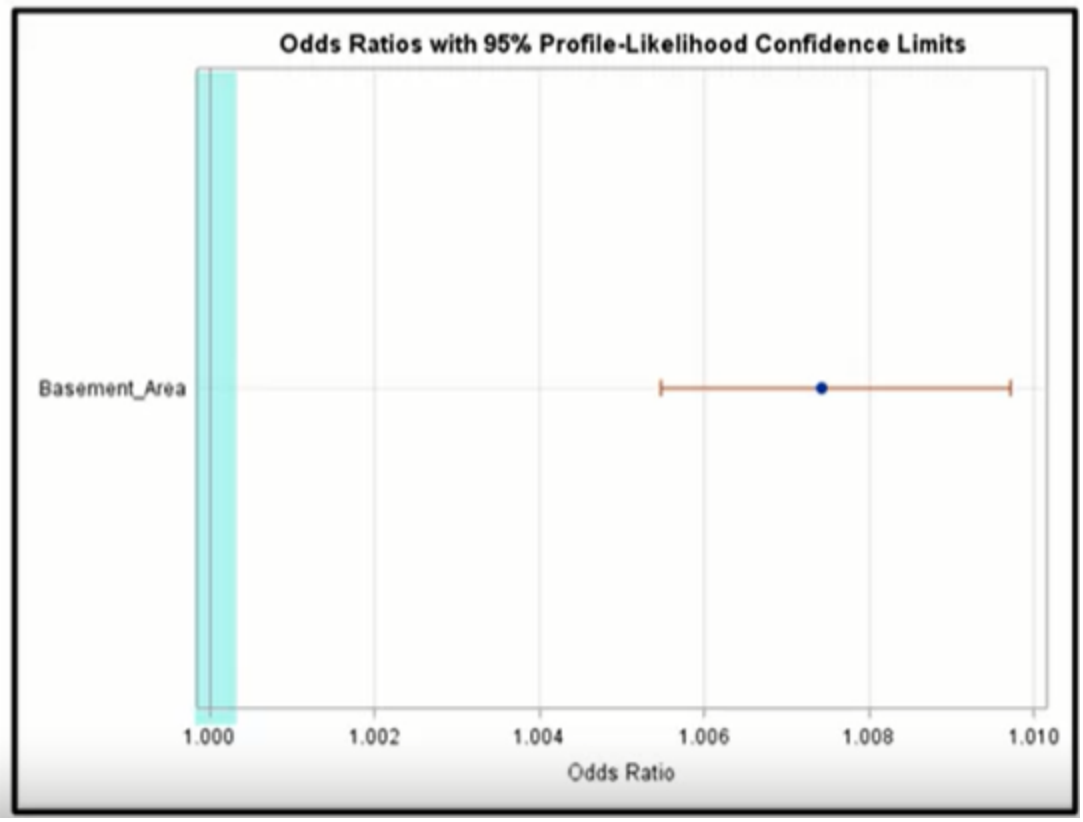
$$\text{logit}(\hat{p}) = \log(\text{odds}) = \hat{\beta}_0 + \hat{\beta}_1 * \text{Basement\_Area}$$

different from

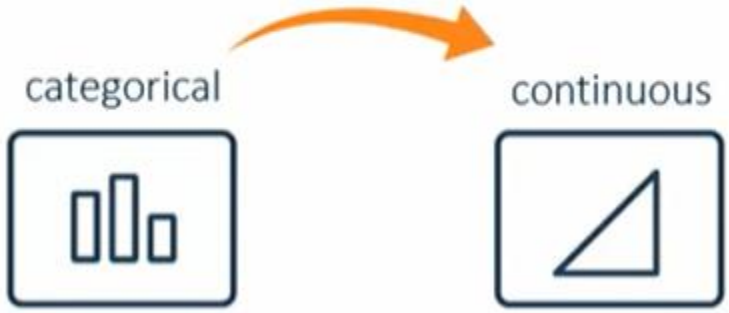
Wald-based confidence intervals

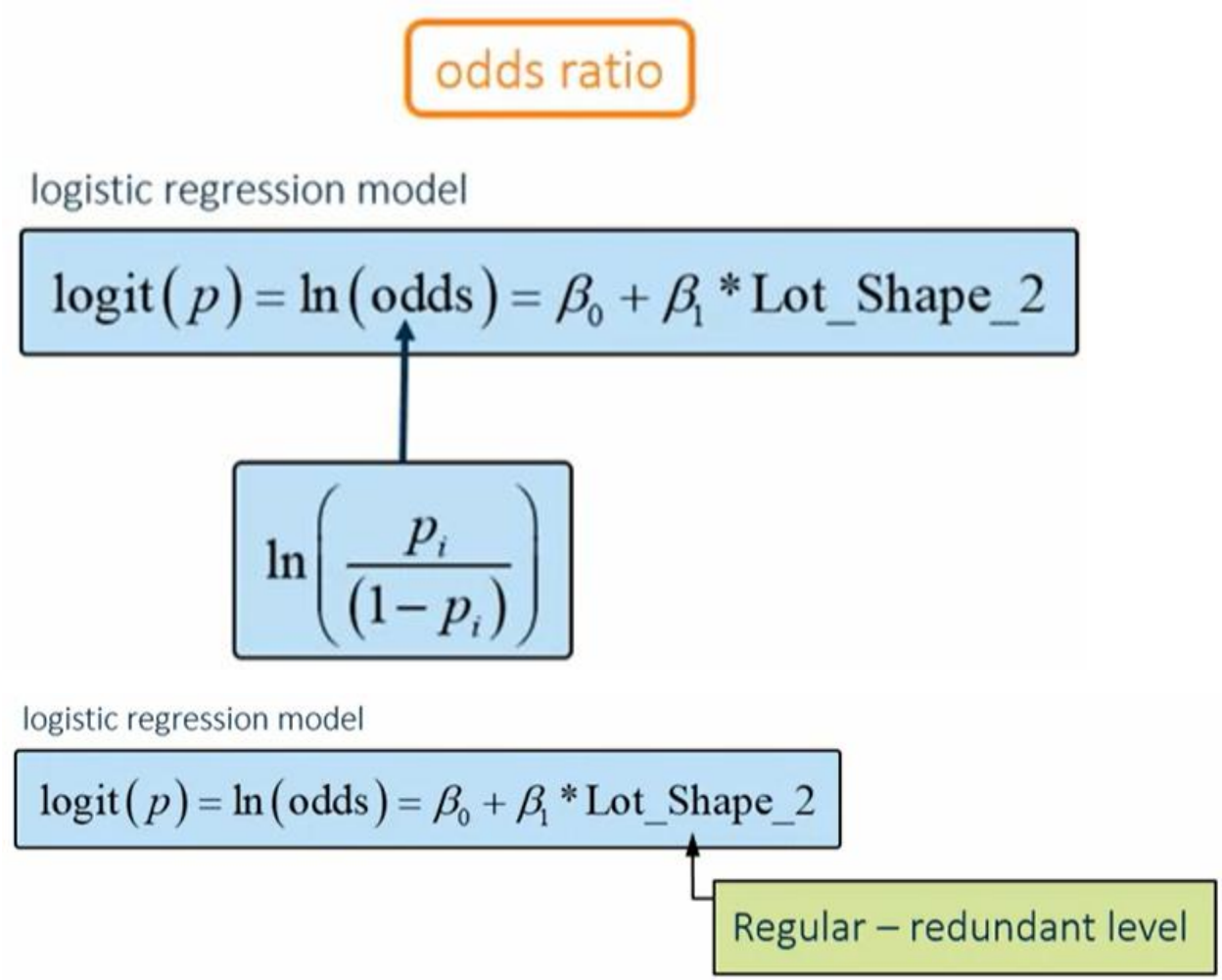

odds ratio
sample sizes < 50
Odds Ratio Estimates and Profile-Likelihood Confidence Intervals
Effect
Unit
Estimate
95% Confidence Limits
Basement_Area
1.0000
1.007
1.005
1.010
Log-likelihood
Wald-based confidence intervals
normal error approximation

odds ratio

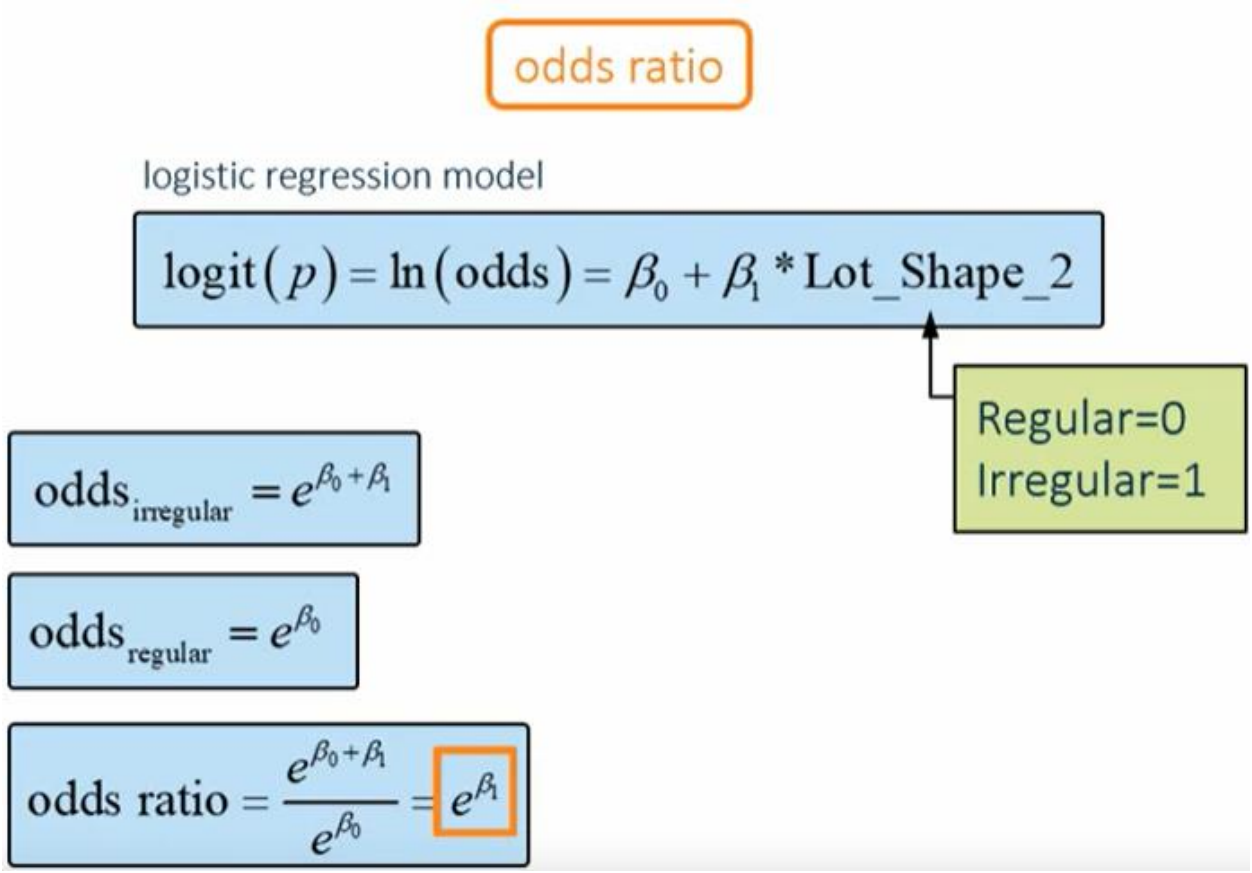


odds ratio




odds ratio
logistic regression model
$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 * \text{Lot\_Shape\_2}$
$\ln\left(\frac{p_i}{(1-p_i)}\right)$

logistic regression model
$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 * \text{Lot\_Shape\_2}$
Regular – redundant level


odds ratio
logistic regression model
$\text{logit}(p) = \ln(\text{odds}) = \beta_0 + \beta_1 * \text{Lot\_Shape\_2}$
Regular=0
Irregular=1
$\text{odds}_{\text{irregular}} = e^{\beta_0 + \beta_1}$
$\text{odds}_{\text{regular}} = e^{\beta_0}$
$\text{odds ratio} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$

Comparing Pairs to Assess the Fit of a Logistic Regression Model

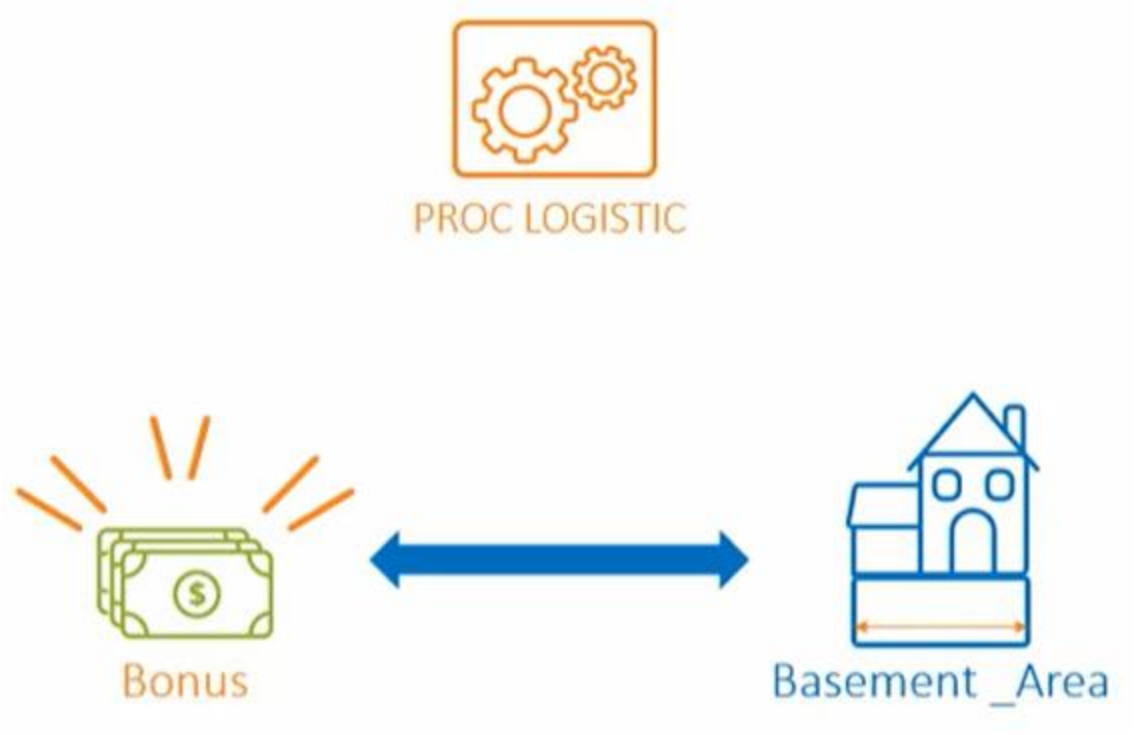
logistic regression model

$$\text{logit}(p_i) = \ln(\text{odds}) = \beta_0 + \beta_1 X_i$$

goodness-of-fit

PROC LOGISTIC

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	Somers' D	0.791
Percent Discordant	10.4	Gamma	0.792
Percent Tied	0.1	Tau-a	0.202
Pairs	11475	c	0.896


PROC LOGISTIC
Bonus
Basement _Area

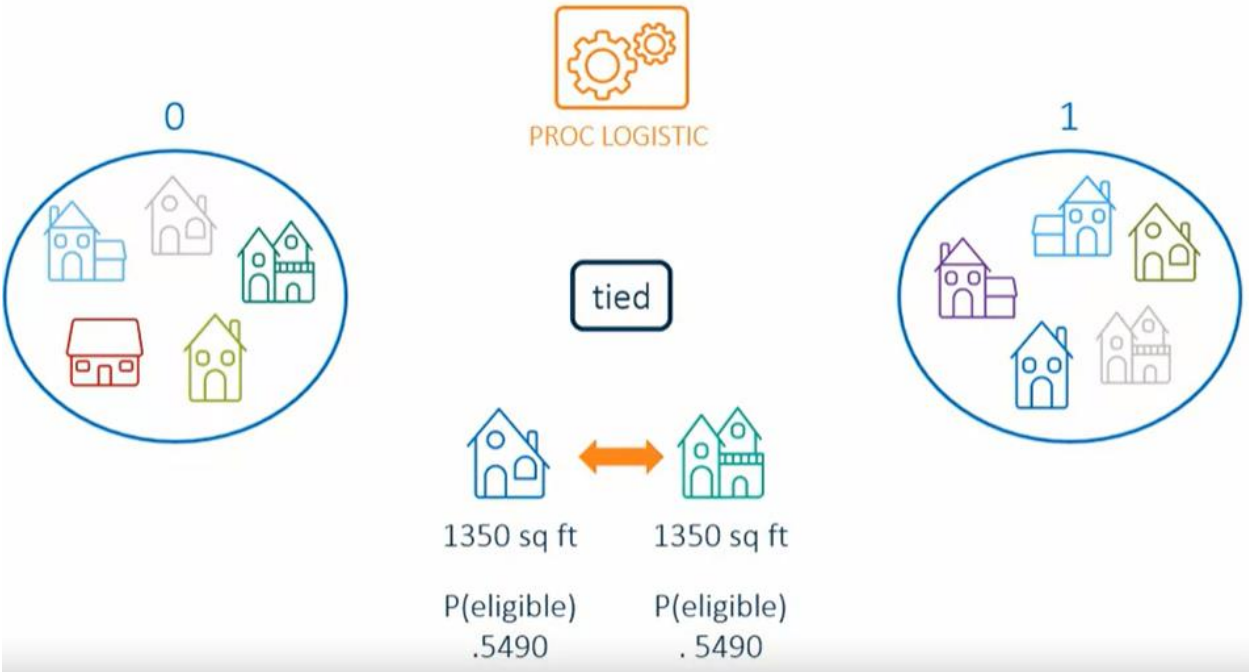

0
PROC LOGISTIC
1
Bonus
not bonus eligible
bonus eligible


0
PROC LOGISTIC
1


concordant
discordant
tied


0
PROC LOGISTIC
1
concordant
800 sq ft
P(eligible)
.0204
1200 sq ft
P(eligible)
. 2865


0
PROC LOGISTIC
1
disconcordant
1600 sq ft
1400 sq ft
P(eligible)
.8855
P(eligible)
. 6379


0
PROC LOGISTIC
1
tied
1350 sq ft
1350 sq ft
P(eligible)
.5490
P(eligible)
. 5490

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	Somers' D	0.791
Percent Discordant	10.4	Gamma	0.792
Percent Tied	0.1	Tau-a	0.202
Pairs	11475	c	0.896

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	45

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.5	Somers' D	0.791
Percent Discordant	10.4	Gamma	0.792
Percent Tied	0.1	Tau-a	0.202
Pairs	11475	c	0.896

Question 7.08

You're modeling the relationship between the variables **Gender** (with the levels *female* and *male*) and **Survived** (with the levels *yes* and *no*). How do you interpret the odds ratio in the output from this PROC LOGISTIC program?

```
proc logistic data=stat1.titanic plots(only)=(effect);
   class Gender (param=ref);
   model Survived (event='yes')=Gender;
run;
```

In the output, the odds ratio of survival for females to males is 10.147. This means that the odds of females surviving were 10 times the odds of males surviving.

```
/*st107s02.sas*/

ods graphics on;

proc logistic data=STAT1.safety plots(only)=(effect oddsratio);

   model Unsafe(event='1')=Weight / clodds=pl;

   title 'LOGISTIC MODEL (1):Unsafe=Weight';

run;
```

LOGISTIC MODEL (1):Unsafe=Weight

The LOGISTIC Procedure

Model Information	
Data Set	STAT1.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

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Weight	1.0000	0.249	0.102	0.517

Practice: Using PROC LOGISTIC to Perform a Binary Logistic Regression Analysis

Question 1

The insurance company wants to characterize the relationship between a vehicle's weight and its safety rating. The **stat1.safety** data set contains the data about vehicle safety.

1. Use PROC LOGISTIC to fit a simple logistic regression model with **Unsafe** as the response 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, an odds ratio plot, and an effect plot. Submit the code and view the results.
2. Do you reject or fail to reject the null hypothesis that all regression coefficients of the model are 0?

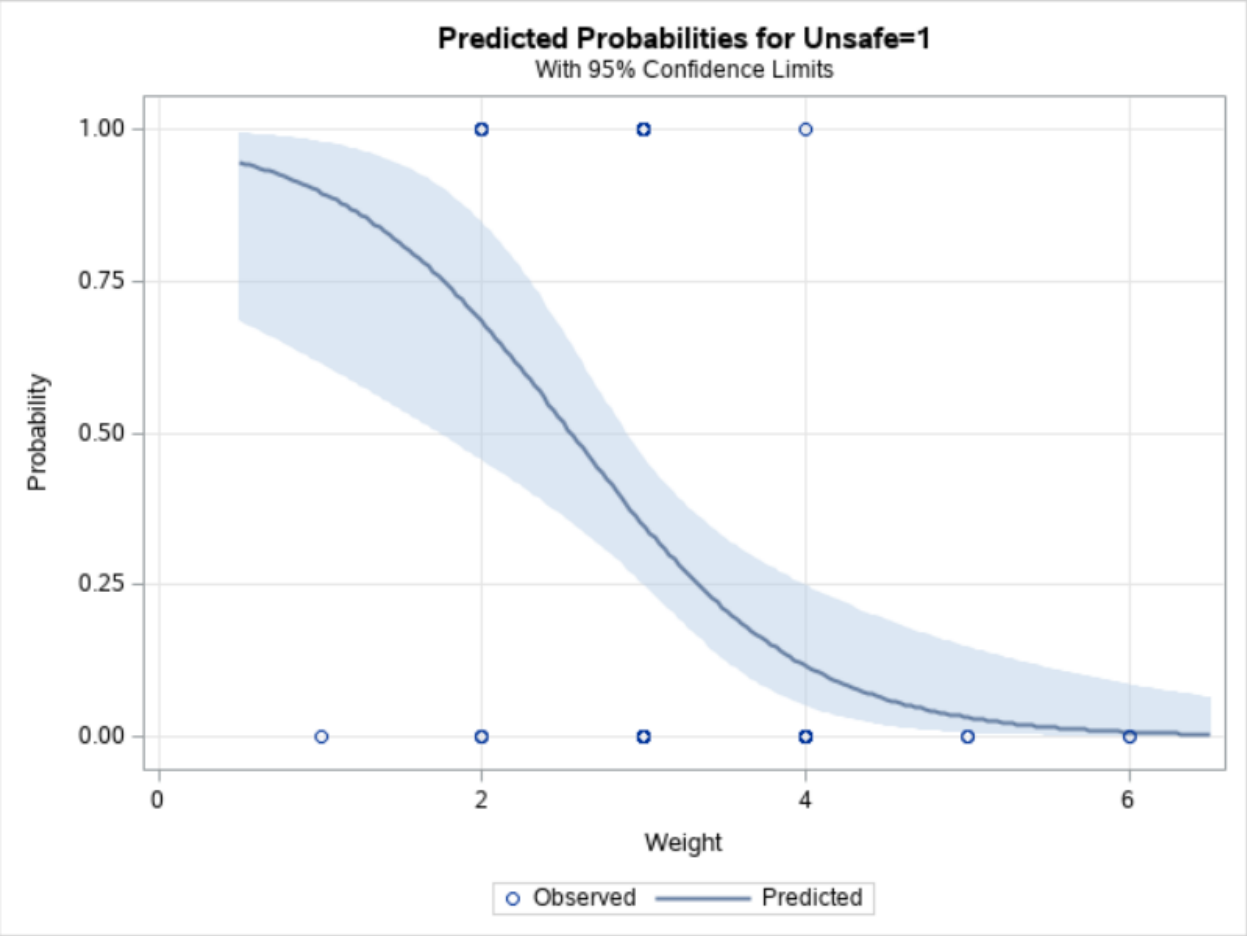
The *p*-value for the Likelihood Ratio test is <.0001, and therefore, the global null hypothesis is rejected.

```
/*st107s02.sas*/

ods graphics on;
proc logistic data=STAT1.safety plots(only)=(effect oddsratio);
```

```
   model Unsafe(event='1')=Weight / clodds=pl;
   title 'LOGISTIC MODEL (1):Unsafe=Weight';
run;
```

Question 2
Write the logistic regression equation.
The regression equation is as follows:

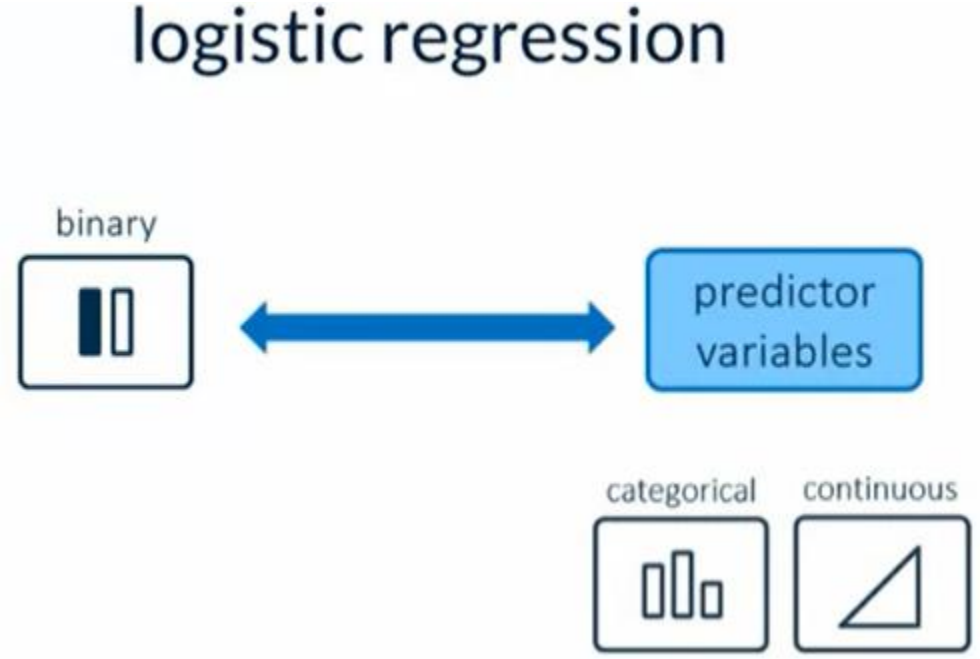
Logit(Unsafe) = 3.5422 + (-1.3901) * Weight

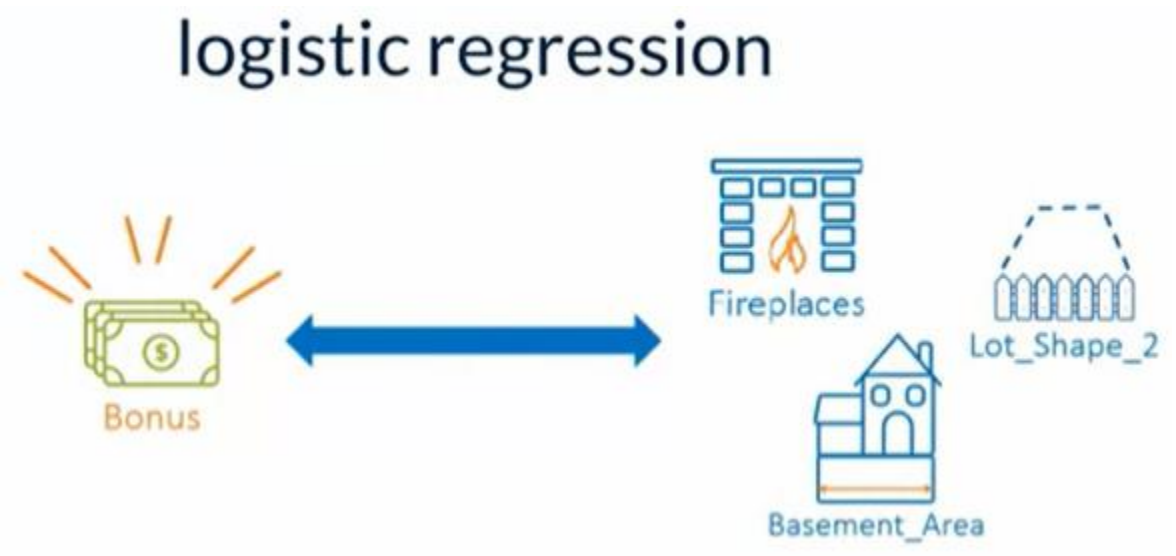


Question 3
Interpret the odds ratio for **Weight**.
The odds ratio for **Weight** (0.249) says that the odds for being unsafe (having a Below Average safety rating) are 75.1% lower for each thousand-pound increase in weight. The confidence interval (0.102 , 0.517) does not contain 1,which indicates that the odds ratio is statistically significant.

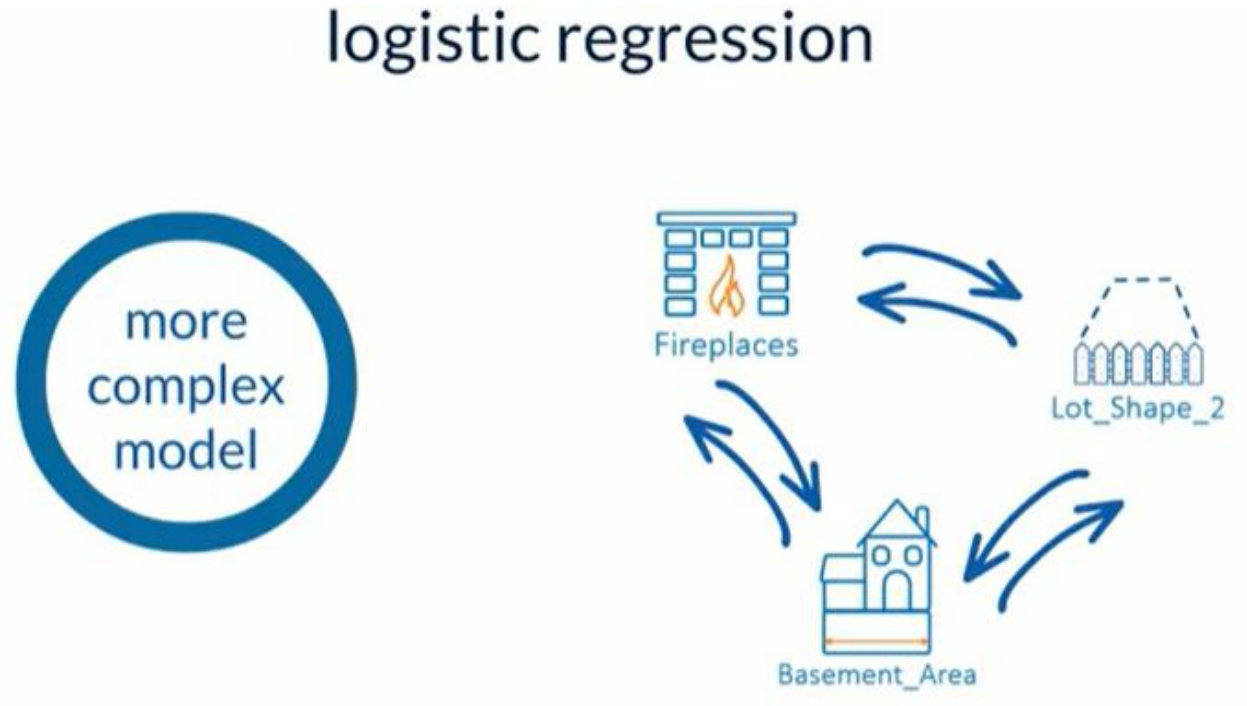
Logistic Regression with Categorical Predictors

Scenario

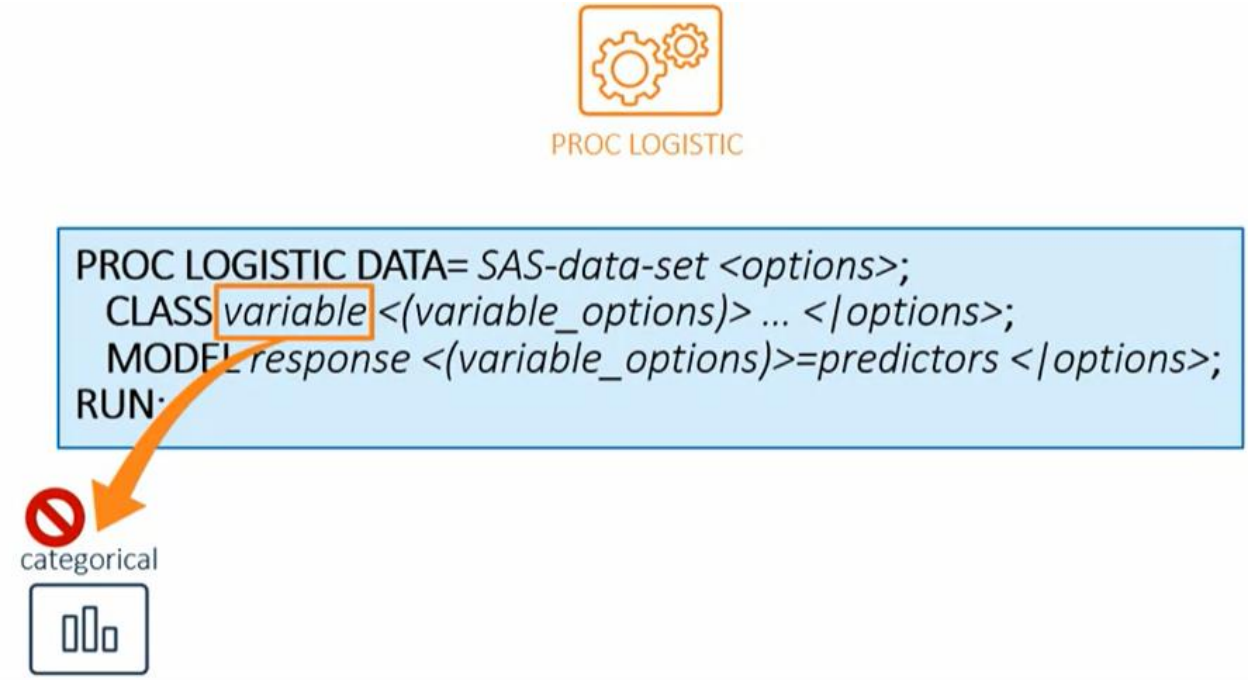


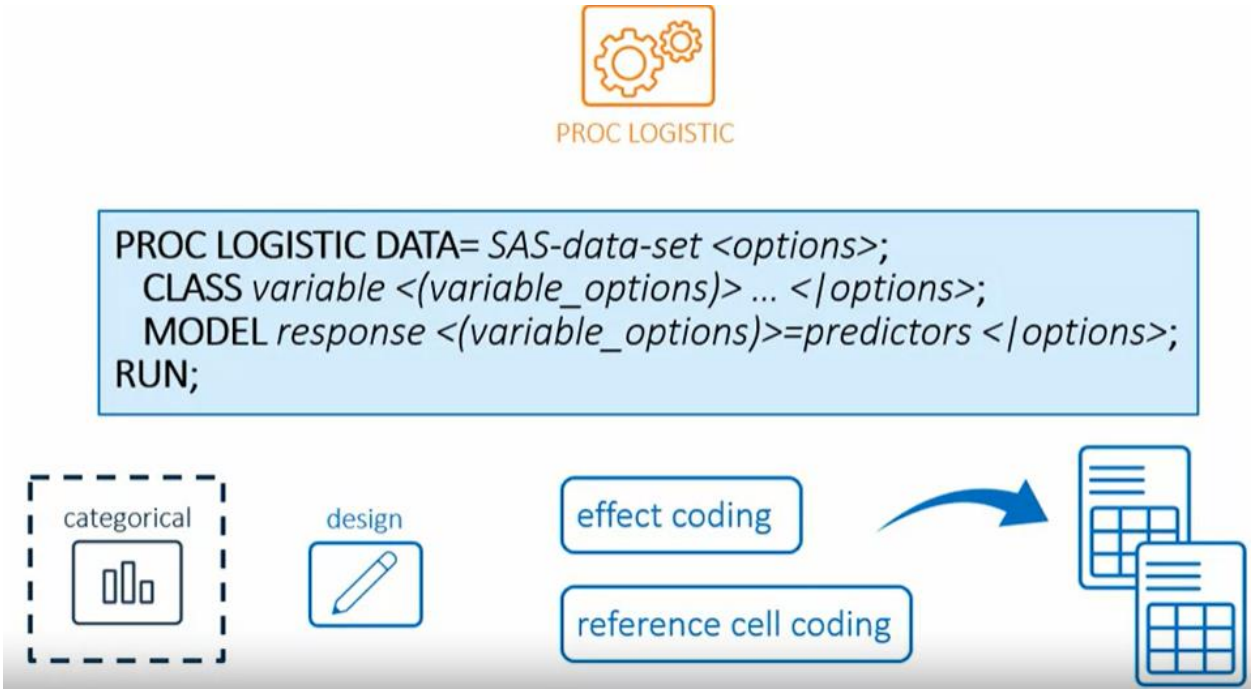


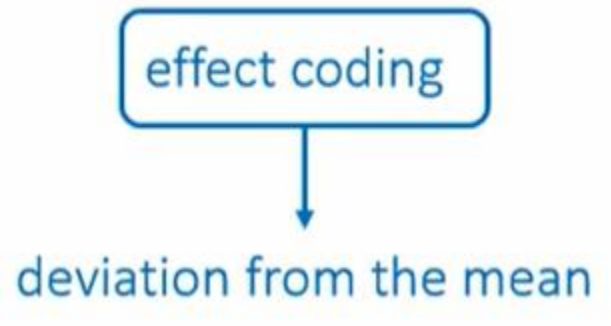

logistic regression
Bonus
more complex model
> $175,000
SOLD


logistic regression
more complex model
Fireplaces
Lot_Shape_2
Basement_Area

Specifying a Parameterization Method







Value	Label
1	low income
2	medium income
3	high income

Value	Label	D1	D2
1	low income	1	0
2	medium income	0	1
3	high income	-1	-1

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

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

reference level →

Value	Label
1	low income
2	medium income
3	high income

Value	Label	D1	D2
1	low income	1	0
2	medium income	0	1
3	high income	0	0

PROC LOGISTIC

reference cell coding

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

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

Question 7.09

The variable **Income** has the values *High*, *Low*, and *Medium*. You've parameterized the variable with reference cell coding using the default reference level. For which value of **Income** do both design variables have the value *0*?

The design variables have the value *0* for the *Medium* level, the default reference level, because it comes last in alphanumeric order.

Demo Fitting a Multiple Logistic Regression Model With Categorical Predictors Using PROC LOGISTIC

```
/*st107d05.sas*/
ods graphics on;
proc logistic data=STAT1.ameshousing3 plots(only)=(effect oddsratio);
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
    model Bonus(event='1')=Basement_Area Fireplaces Lot_Shape_2 / clodds=pl;
    units Basement_Area=100;
    title 'LOGISTIC MODEL (2):Bonus= Basement_Area Fireplaces Lot_Shape_2';
run;
```

PROC LOGISTIC DATA=*SAS-data-set <options>***;**
CLASS *variable <(options)> ... <(/ options)>*;
MODEL *variable <(variable_options)> = <effects> < / options>*;
UNITS *<independent1=list1> ... < / options>*;
RUN;

LOGISTIC MODEL (2):Bonus= Basement_Area Fireplaces Lot_Shape_2

The LOGISTIC Procedure

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > $175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	300
Number of Observations Used	299

Response Profile

Ordered Value	Bonus	Total Frequency
1	0	255
2	1	44

Probability modeled is Bonus='1'.

Note: 1 observation was deleted due to missing values for the response or explanatory variables.

Class Level Information			
Class	Value	Design Variables	
Fireplaces	0	0	0
	1	1	0
	2	0	1
Lot_Shape_2	Irregular	1	
	Regular	0	

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	140.499
SC	255.513	159.001
-2 Log L	249.812	130.499

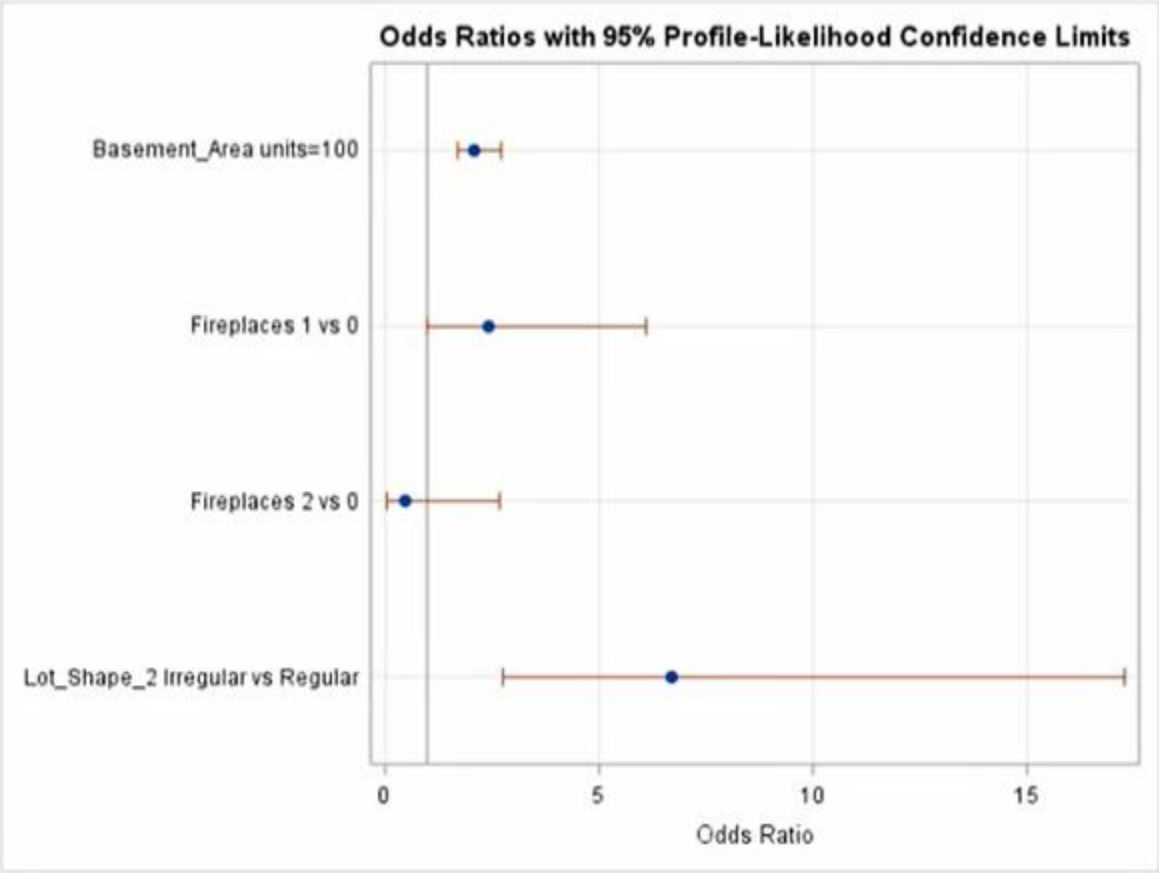
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	119.3133	4	<.0001
Score	91.7250	4	<.0001
Wald	49.8871	4	<.0001

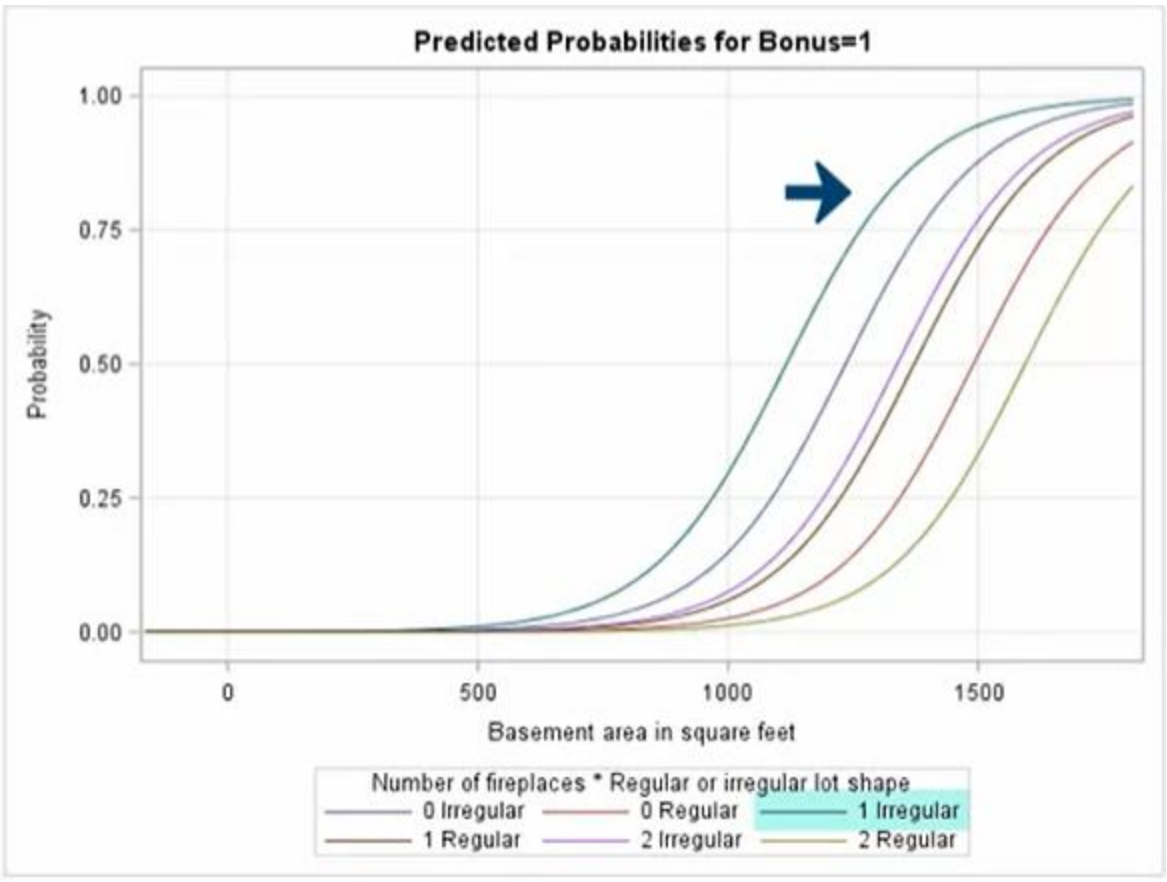
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Basement_Area	1	38.1356	<.0001
Fireplaces	2	5.2060	0.0741
Lot_Shape_2	1	16.9421	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-11.0882	1.5384	51.9467	<.0001
Basement_Area		1	0.00744	0.00120	38.1356	<.0001
Fireplaces	1	1	0.8810	0.4658	3.5770	0.0585
Fireplaces	2	1	-0.7683	0.9654	0.6335	0.4261
Lot_Shape_2	Irregular	1	1.9025	0.4622	16.9421	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	92.9	Somers' D	0.859
Percent Discordant	7.0	Gamma	0.860
Percent Tied	0.1	Tau-a	0.216
Pairs	11220	c	0.930

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Basement_Area	100.0	2.105	1.696	2.727
Fireplaces 1 vs 0	1.0000	2.413	0.973	6.127
Fireplaces 2 vs 0	1.0000	0.464	0.054	2.703
Lot_Shape_2 Irregular vs Regular	1.0000	6.703	2.786	17.301






```
/*st107d05.sas*/
ods graphics on;
proc logistic data=STAT1.ameshousing3 plots(only)=(effect oddsratio);
   class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
   model Bonus(event='1')=Basement_Area Fireplaces Lot_Shape_2 / clodds=pl;
   units Basement_Area=100;
   title 'LOGISTIC MODEL (2):Bonus= Basement_Area Fireplaces Lot_Shape_2';
run;
```

Question 7.10

Suppose you want to fit a logistic regression model to explain the relationship between two variables in the **Titanic** data set: **Age** (which has the values *adult* and *child*) and **Survived** (which has the values *no* and *yes*).

To complete this PROC LOGISTIC step, write a CLASS statement (if you think it's necessary) and a MODEL statement that models the probability of survival. Specify reference cell coding and the reference level that PROC LOGISTIC selects by default.

```
proc logistic data=stat1.titanic;
    [What CLASS and/or MODEL statements do you need here?]
run;
```

The following statements correctly complete the PROC LOGISTIC step:

```
proc logistic data=stat1.titanic;
    class Age (ref='child') / param=ref;
    model Survived(event='yes')=Age;
run;
```

Alternatively:

```
proc logistic data=stat1.titanic;
    class Age (param=ref ref='child');
    model Survived(event='yes')=Age;
run;
```

The CLASS statement is necessary because the predictor is categorical. You're specifying the level of **Age** that PROC LOGISTIC will select by default for reference cell coding, so the code will generate the same results with or without the REF= option.

```
/*st107s03.sas*/

ods graphics on;

proc logistic data=STAT1.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;

   title 'LOGISTIC MODEL (2):Unsafe=Weight Region Size';

run;
```

LOGISTIC MODEL (2):Unsafe=Weight Region Size

The LOGISTIC Procedure

Model Information	
Data Set	STAT1.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.

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

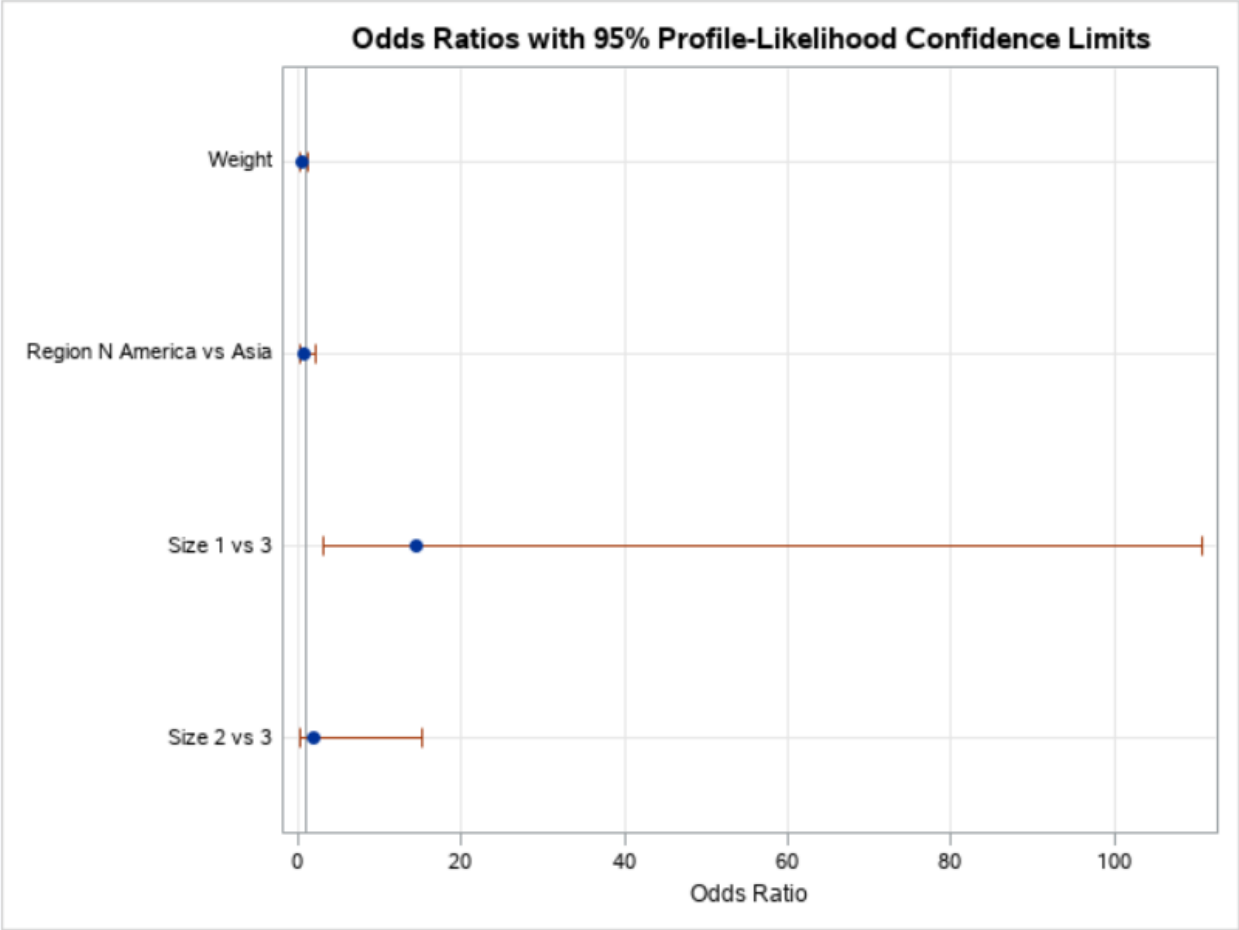
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	35.2441	4	<.0001
Score	32.8219	4	<.0001
Wald	23.9864	4	<.0001

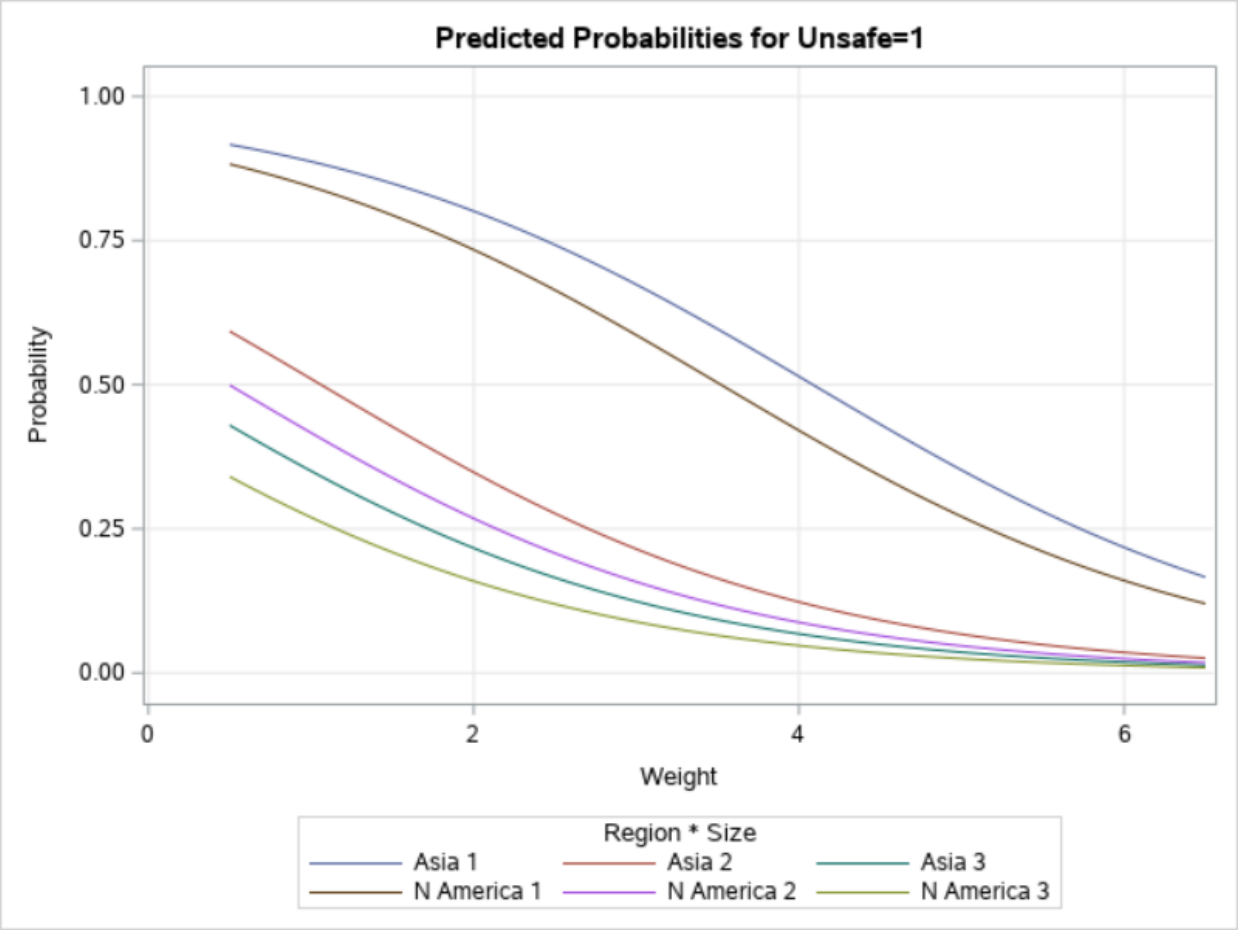
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Weight	1	2.1176	0.1456
Region	1	0.4506	0.5020
Size	2	15.3370	0.0005

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

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
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



Practice: Using PROC LOGISTIC to Perform a Multiple Logistic Regression Analysis with Categorical Variables

Question 1

The insurance company wants to model the relationship between three of a car's characteristics, weight, size, and region of manufacture, and its safety rating. The **stat1.safety** data set contains the data about vehicle safety.

1. Use PROC LOGISTIC to fit a multiple logistic regression model with **Unsafe** as the response variable and **Weight**, **Size**, and **Region** as the predictor variables.
2. Use the EVENT= option to model the probability of Below Average safety scores.
3. Specify **Region** and **Size** as classification variables and use reference cell coding. Specify *Asia* as the reference level for **Region**, and *3* (large cars) as the reference level for **Size**.
4. Request profile likelihood confidence limits, an odds ratio plot, and the effect plot.
5. Submit the code and view the results.

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

The *p*-value for the Likelihood Ratio test is <.0001, and therefore, you reject the null hypothesis.

```
/*st107s03.sas*/

ods graphics on;
proc logistic data=STAT1.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;
   title 'LOGISTIC MODEL (2):Unsafe=Weight Region Size';
run;
```

Question 2
If you reject the global null hypothesis, then which predictors significantly predict safety outcome?
size
Only **Size** is significantly predictive of **Unsafe**.

Question 3
Interpret the odds ratio for significant predictors.
14.56

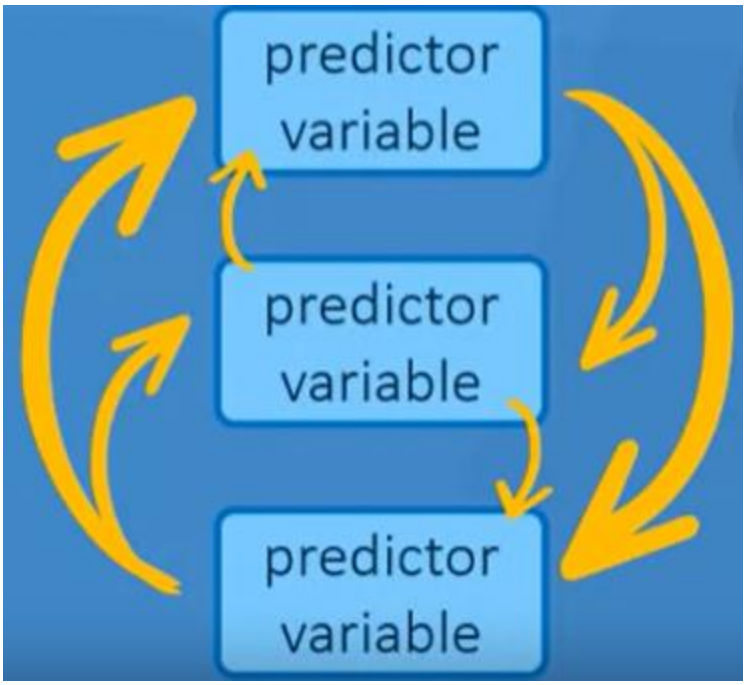
Correct

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.

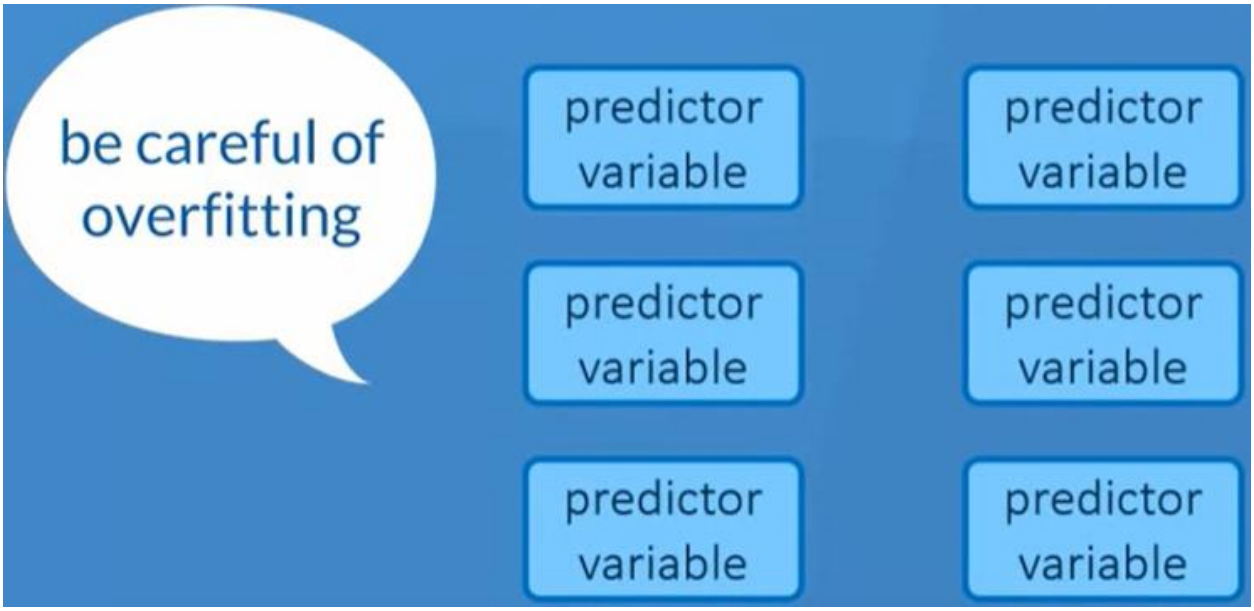
The contrast from the second design variable is 1.931 (Medium versus Sport/Utility), implying a trend toward greater odds of low safety as size decreases. However, the 95% confidence interval (0.343, 15.182) contains 1, and therefore, the contrast is not statistically significant.

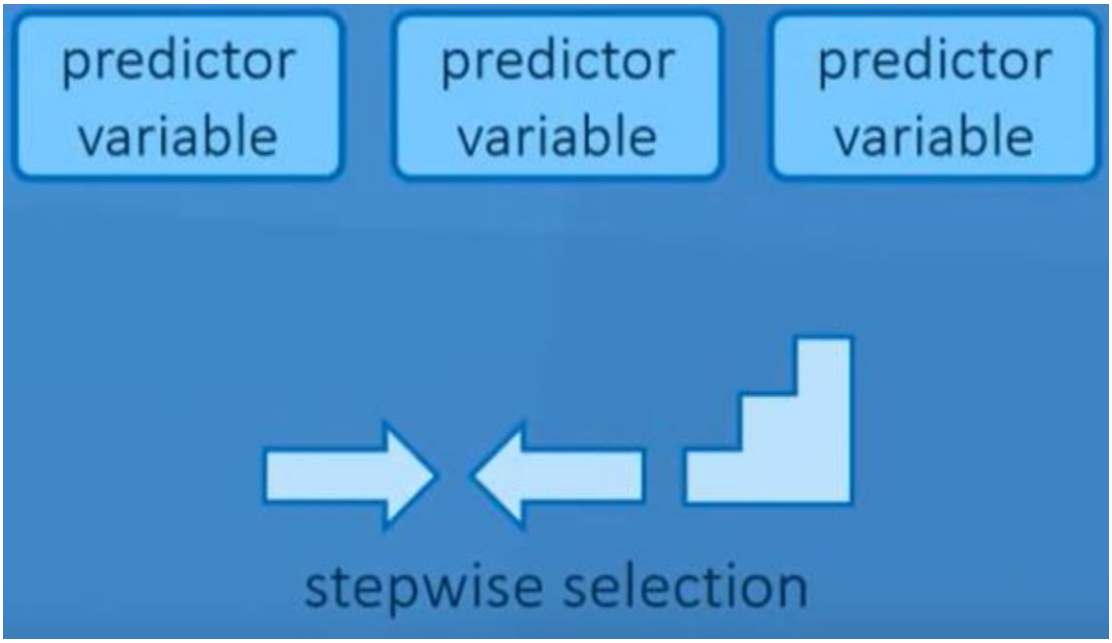
Stepwise Selection with Interactions and Predictions

Scenario

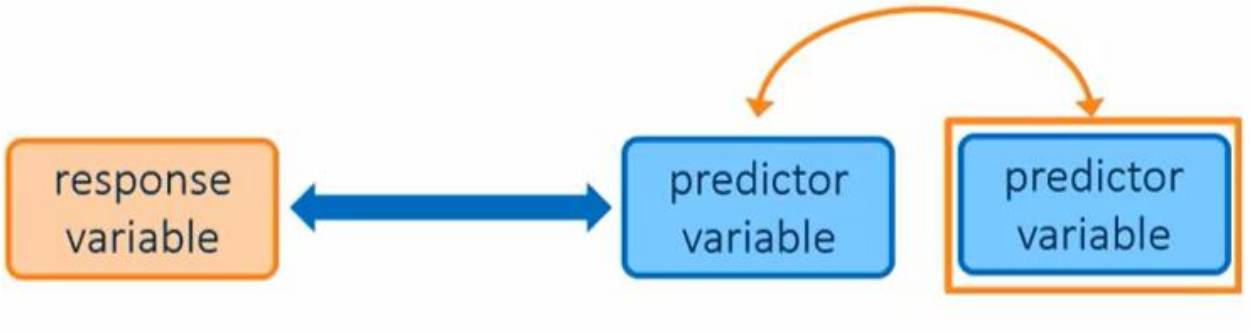

predictor variable
predictor variable
predictor variable

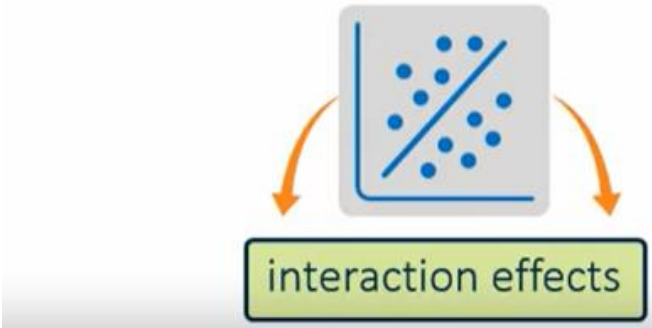

Basement_Area
Lot_Shape_2
Fireplaces

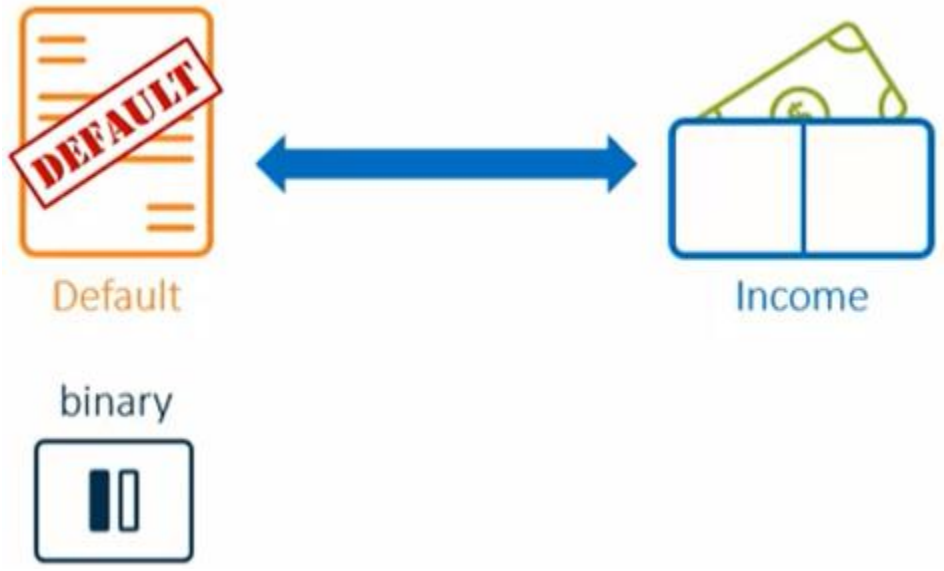

be careful of overfitting
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable

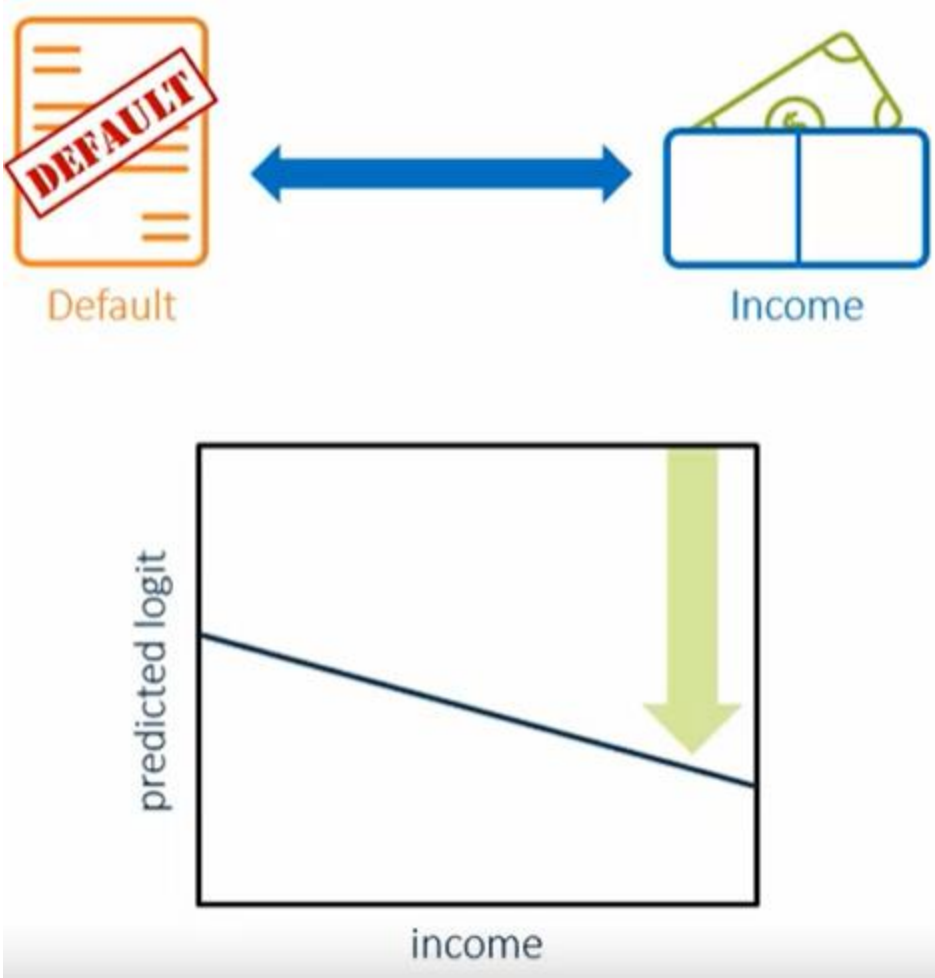

predictor variable
predictor variable
predictor variable
stepwise selection

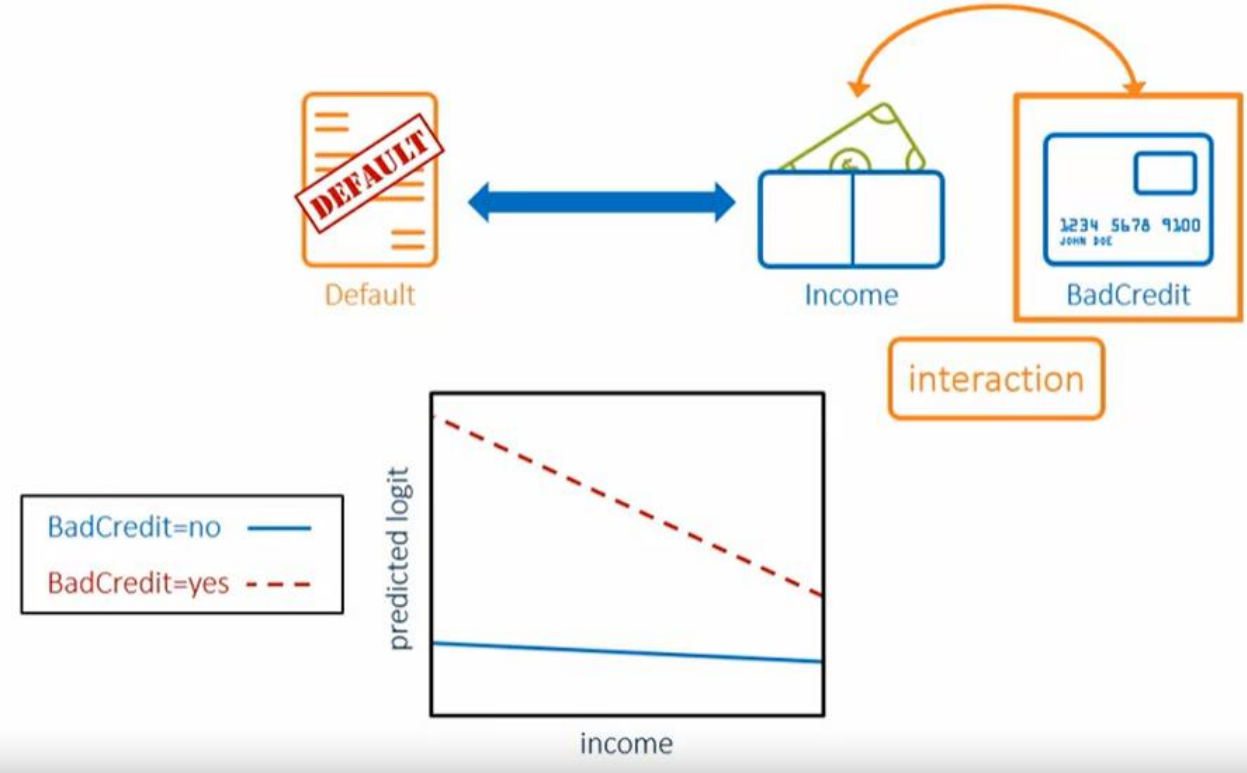
Interactions between Variables

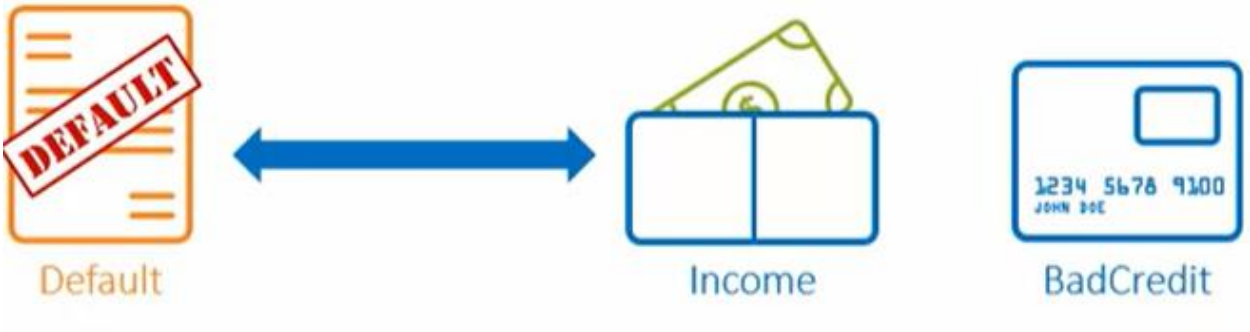


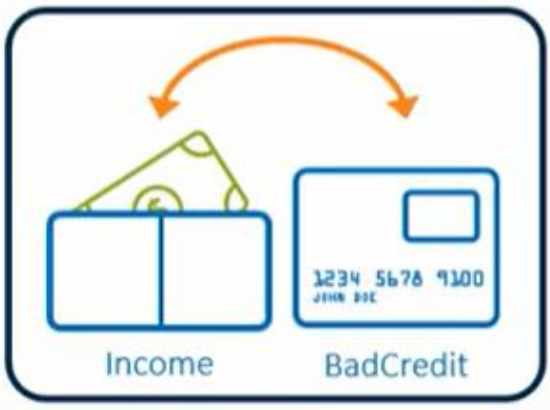


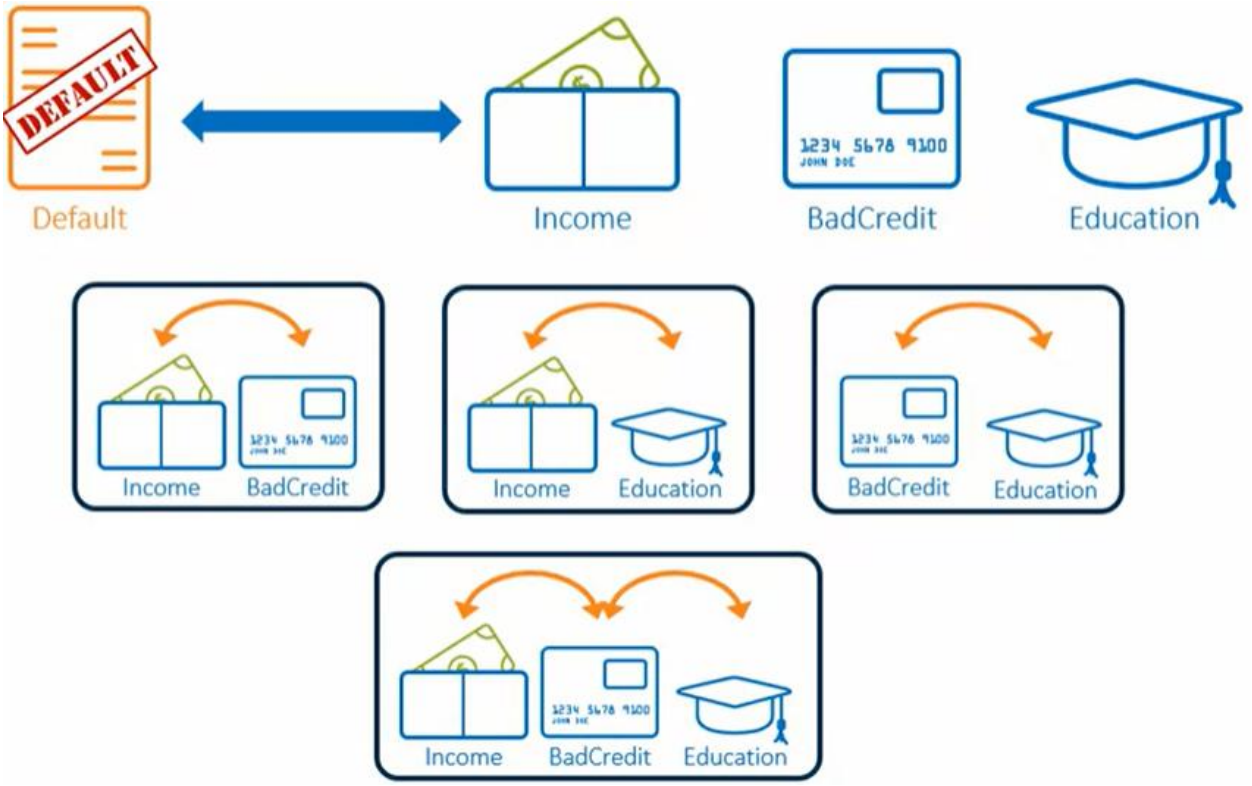



DEFAULT
Default
Income
predicted logit
income


DEFAULT
Default
Income
1234 5678 9100
JOHN DOE
BadCredit
interaction
BadCredit=no
BadCredit=yes
predicted logit
income


DEFAULT
Default
Income
1234 5678 9100
JOHN DOE
BadCredit


1234 5678 9100
JOHN DOE
Income
BadCredit



Demo Fitting a Multiple Logistic Regression Model with Interactions Using PROC LOGISTIC

```
1 /*st107d06.sas*/  /*Part A*/
2 proc logistic data=STAT1.ameshousing3 plots(only)=(effect oddsratio);
3     class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
4     model Bonus(event='1')=Basement_Area|Fireplaces|Lot_Shape_2 @2 /
5           selection=backward clodds=pl slstay=0.10;
6     units Basement_Area=100;
7     title 'LOGISTIC MODEL (3): Backward Elimination '
8           'Bonus=Basement_Area|Fireplaces|Lot_Shape_2';
9 run;
```

```
PROC LOGISTIC DATA=SAS-data-set <options>;
    CLASS variable <(options)> ... <( / options)>;
    MODEL variable <(variable_options)> = <effects> < / options>;
    UNITS <independent1=list1> ... < / options>;
RUN;
```

Note: 1 observation was deleted due to missing values for the response or explanatory variables.

Backward Elimination Procedure

Class Level Information			
Class	Value	Design Variables	
Fireplaces	0	0	0
	1	1	0
	2	0	1
Lot_Shape_2	Irregular	1	
	Regular	0	

Step 0. The following effects were entered:

Intercept Basement_Area Fireplaces Basement_*Fireplaces Lot_Shape_2 Basement_*Lot_Shape_ Fireplace*Lot_Shape_

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

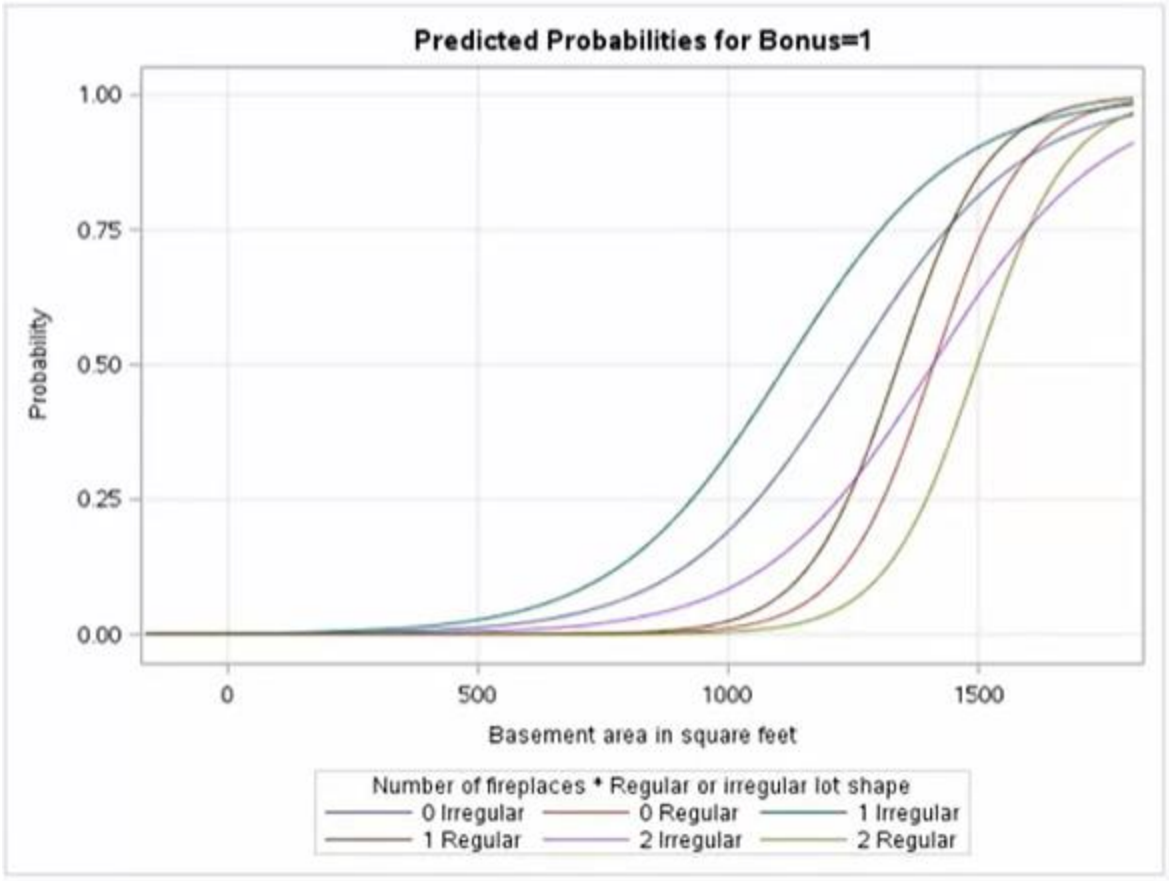
Summary of Backward Elimination						
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Variable Label
1	Fireplace*Lot_Shape_	2	5	3.2305	0.1988	
2	Basement_*Fireplaces	2	4	1.7237	0.4224	

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Basement_Area	1	18.2896	<.0001
Fireplaces	2	4.7171	0.0946
Lot_Shape_2	1	5.0247	0.0250
Basement_*Lot_Shape_	1	3.1127	0.0777

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-15.3017	3.2407	22.2952	<.0001
Basement_Area		1	0.0109	0.00254	18.2896	<.0001
Fireplaces	1	1	0.7671	0.4687	2.6781	0.1017
Fireplaces	2	1	-0.9405	0.9503	0.9795	0.3223
Lot_Shape_2	Irregular	1	8.0362	3.5850	5.0247	0.0250
Basement_*Lot_Shape_	Irregular	1	-0.00503	0.00285	3.1127	0.0777

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	93.8	Somers' D	0.876
Percent Discordant	6.2	Gamma	0.876
Percent Tied	0.1	Tau-a	0.221
Pairs	11220	c	0.938

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Fireplaces 1 vs 0	1.0000	2.153	0.865	5.500
Fireplaces 2 vs 0	1.0000	0.390	0.047	2.251




```
/*st107d06.sas*/  /*Part A*/
proc logistic data=STAT1.ameshousing3 plots(only)=(effect oddsratio);
   class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
   model Bonus(event='1')=Basement_Area|Fireplaces|Lot_Shape_2 @2 /
        selection=backward clodds=pl slstay=0.10;
   units Basement_Area=100;
   title 'LOGISTIC MODEL (3): Backward Elimination '
         'Bonus=Basement_Area|Fireplaces|Lot_Shape_2';
run;
```

Question 7.11

Complete the MODEL statement to add **Gender** as a main effect and specify the backward elimination method.

```
proc logistic data=stat1.titanic;
    class Age (param=ref ref='child')
          Gender (param=ref ref='female')
    model Survived(event='yes')=Age [COMPLETE THE MODEL STATEMENT];
run;
```

The completed MODEL statement is shown below:

```
model Survived(event='yes')=Age Gender / selection=backward;
```

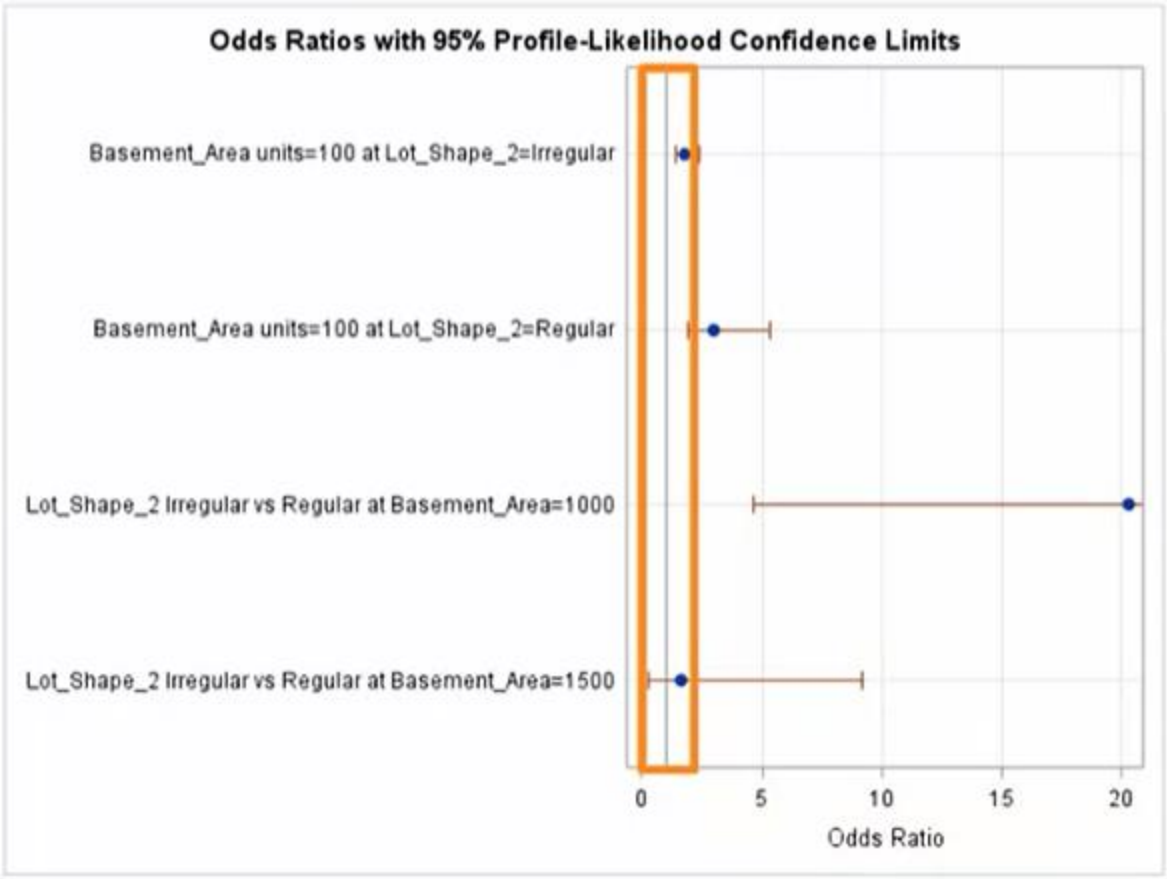
Demo Fitting a Multiple Logistic Regression Model with All Odds Ratios Using PROC LOGISTIC

```
/*st107d06.sas*/  /*Part B*/
proc logistic data=STAT1.ameshousing3
              plots(only)=oddsratio(range=clip);
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
    model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
    units Basement_Area=100;
    oddsratio Basement_Area / at (Lot_Shape_2=ALL) cl=pl;
    oddsratio Lot_Shape_2 / at (Basement_Area=1000 1500) cl=pl;
    title 'LOGISTIC MODEL (3.1): Bonus=Basement_Area|Lot_Shape_2 Fireplaces';
run;
```

```
PROC LOGISTIC DATA=SAS-data-set <options>;
    CLASS variable <(options)> ... <( / options)>;
    MODEL variable <(variable_options)> = <effects> < / options>;
    UNITS <independent1=list1> ... < / options>;
    ODDSRATIO <'label'> variable < / options>;
RUN;
```

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Odds Ratio	Estimate	95% Confidence Limits	
Basement_Area units=100 at Lot_Shape_2=Irregular	1.791	1.421	2.396
Basement_Area units=100 at Lot_Shape_2=Regular	2.960	1.932	5.315
Lot_Shape_2 Irregular vs Regular at Basement_Area=1000	20.278	4.623	146.987
Lot_Shape_2 Irregular vs Regular at Basement_Area=1500	1.643	0.283	9.145




```
/*st107d06.sas*/  /*Part B*/
proc logistic data=STAT1.ameshousing3
          plots(only)=oddsratio(range=clip);
   class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
   model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
   units Basement_Area=100;
   oddsratio Basement_Area / at (Lot_Shape_2=ALL) cl=pl;
   oddsratio Lot_Shape_2 / at (Basement_Area=1000 1500) cl=pl;
   title 'LOGISTIC MODEL (3.1): Bonus=Basement_Area|Lot_Shape_2 Fireplaces';
run;
```

LOGISTIC MODEL (3.1): Bonus=Basement_Area|Lot_Shape_2 Fireplaces

The LOGISTIC Procedure

Model Information		
Data Set	STAT1.AMESHOUSING3	
Response Variable	Bonus	Sale Price > $175,000
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	300
Number of Observations Used	299

Response Profile		
Ordered Value	Bonus	Total Frequency
1	0	255
2	1	44

Probability modeled is Bonus='1'.

Note: 1 observation was deleted due to missing values for the response or explanatory variables.

Class Level Information			
Class	Value	Design Variables	
Fireplaces	0	0	0
	1	1	0
	2	0	1
Lot_Shape_2	Irregular	1	
	Regular	0	

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	251.812	138.872
SC	255.513	161.074
-2 Log L	249.812	126.872

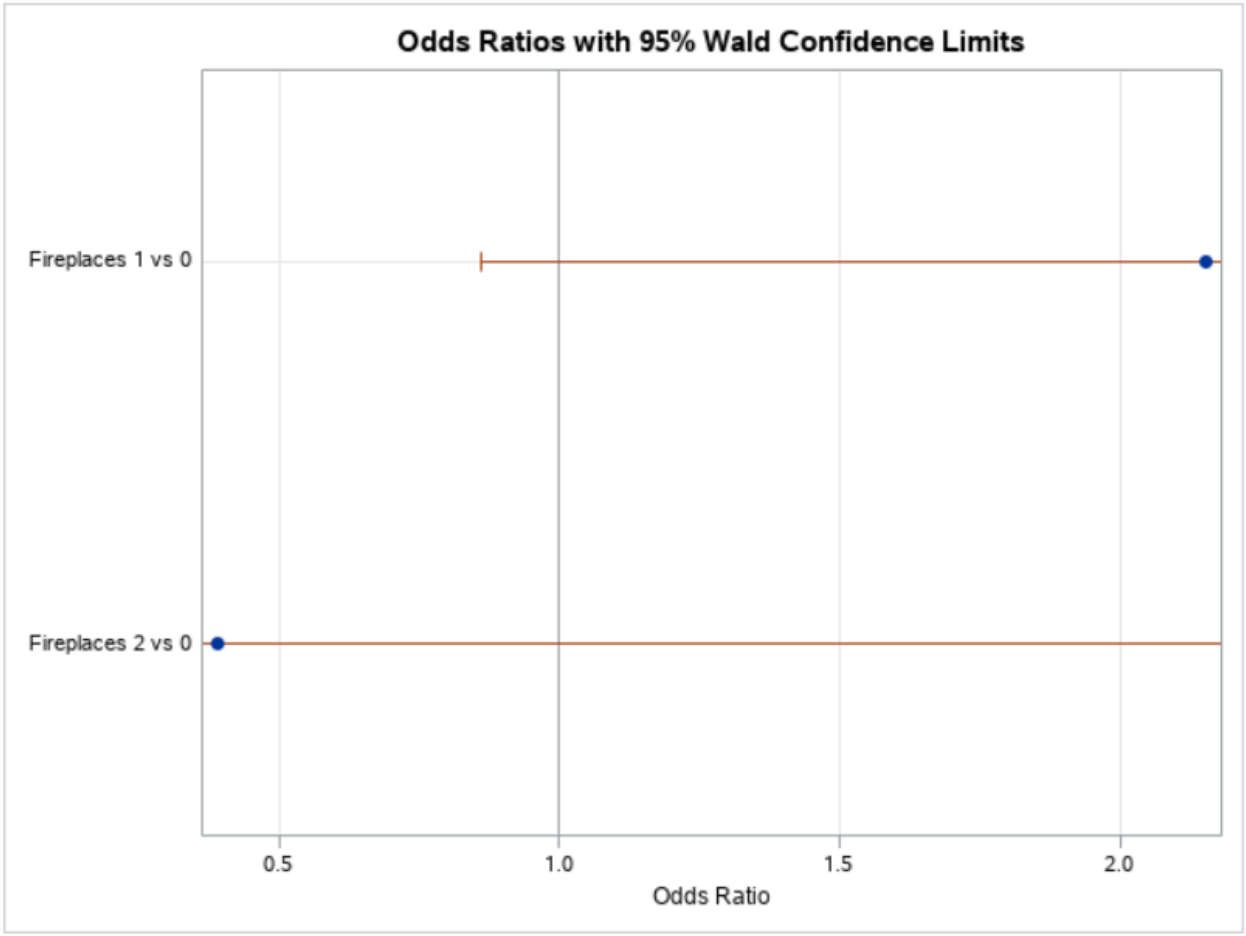
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	122.9405	5	<.0001
Score	102.6370	5	<.0001
Wald	42.9826	5	<.0001

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Basement_Area	1	18.2896	<.0001
Lot_Shape_2	1	5.0247	0.0250
Basement_*Lot_Shape_	1	3.1127	0.0777
Fireplaces	2	4.7171	0.0946

Note: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests. The joint test for an effect is a test that all the parameters associated with that effect are zero. Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

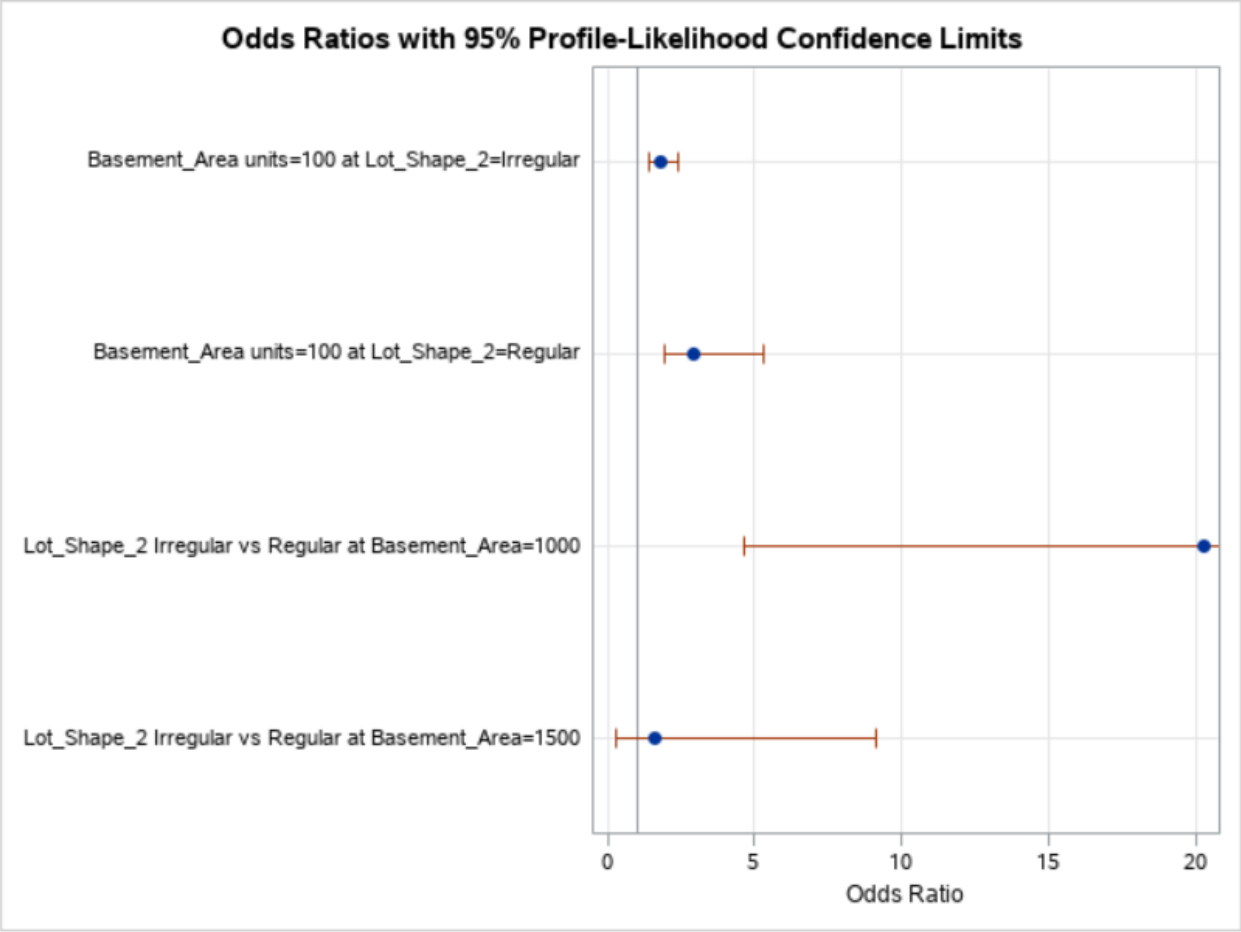
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-15.3017	3.2407	22.2952	<.0001
Basement_Area		1	0.0109	0.00254	18.2896	<.0001
Lot_Shape_2	Irregular	1	8.0362	3.5850	5.0247	0.0250
Basement_*Lot_Shape_	Irregular	1	-0.00503	0.00285	3.1127	0.0777
Fireplaces	1	1	0.7671	0.4687	2.6781	0.1017
Fireplaces	2	1	-0.9405	0.9503	0.9795	0.3223

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Fireplaces 1 vs 0	2.153	0.859	5.396
Fireplaces 2 vs 0	0.390	0.061	2.514



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	93.8	Somers' D	0.876
Percent Discordant	6.2	Gamma	0.876
Percent Tied	0.1	Tau-a	0.221
Pairs	11220	c	0.938

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
Basement_Area units=100 at Lot_Shape_2=Irregular	1.791	1.421	2.396
Basement_Area units=100 at Lot_Shape_2=Regular	2.960	1.932	5.315
Lot_Shape_2 Irregular vs Regular at Basement_Area=1000	20.278	4.623	146.987
Lot_Shape_2 Irregular vs Regular at Basement_Area=1500	1.643	0.283	9.145



Question 7.12

Complete the MODEL statement to specify the predictor variables **Age**, **Gender**, and **Class**. Indicate that you want to include an interaction between **Gender** and **Class** in addition to the three main effects.

```
proc logistic data=stat1.titanic;
    class Age (param=ref ref='child')
          Gender (param=ref ref='female')
          Class (param=ref ref='crew');
    model Survived(event='yes')=[COMPLETE THE MODEL STATEMENT]
                / selection=backward;
run;
```

The completed MODEL statement is shown below. You place a bar operator only between the variables that you want to include in an interaction.

```
model Survived(event='yes')=Age Gender | Class
      / selection=backward;
```

Alternate Solution: The bar operator saves typing, but is not required for identifying interactions.

```
model Survived(event='yes')=Age Gender Class Gender*Class
      / selection=backward;
```

Demo Generating Predictions Using PROC PLM

```
/*st107d07.sas*/

ods select none;
proc logistic data=STAT1.ameshousing3;
    class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
    model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
    units Basement_Area=100;
    store out=isbonus;
run;
ods select all;
```

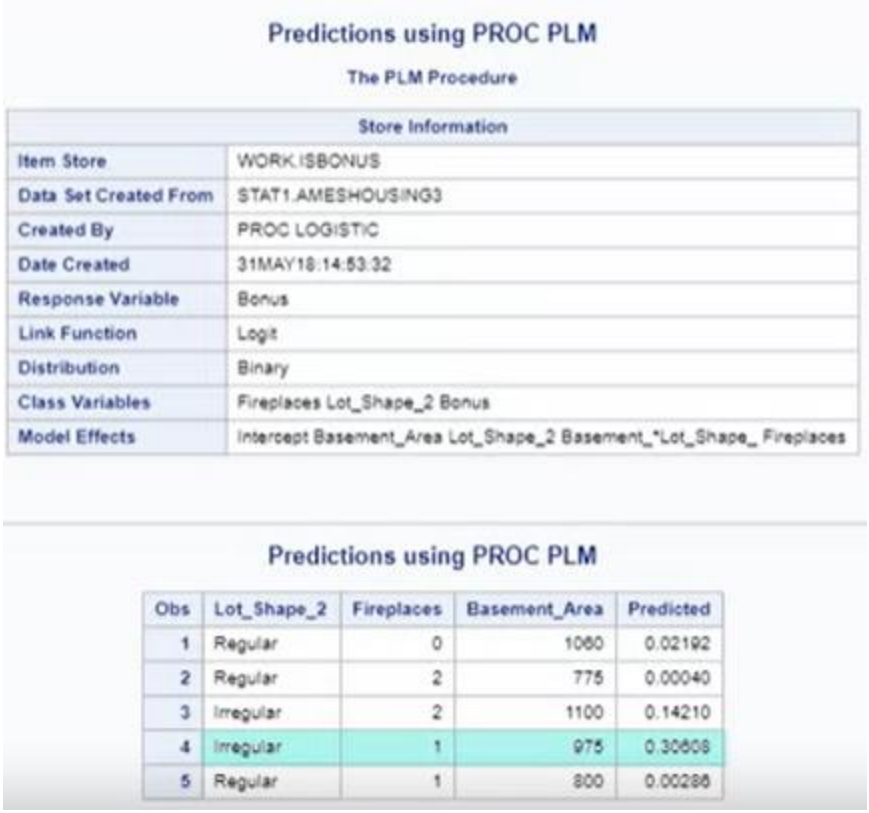
```
PROC LOGISTIC DATA=SAS-data-set <options>;
    CLASS variable <(options)> ... <( / options);
    MODEL variable <(variable_options)> = <effects> < / options>;
    UNITS <independent1=list1> ... < / options>;
    STORE <OUT=>item-store-name < / LABEL= 'label'>;
RUN;
```

```
data newhouses;
    length Lot_Shape_2 $9;
    input Fireplaces Lot_Shape_2 $ Basement_Area;
    datalines;
    0  Regular    1060
    2  Regular     775
    2  Irregular  1100
    1  Irregular   975
    1  Regular     800
    ;
run;
```

```
proc plm restore=isbonus;
    score data=newhouses out=scored_houses / ILINK;
    title 'Predictions using PROC PLM';
run;
```

```
PROC PLM RESTORE=item-store-specification <options>;
    SCORE DATA=SAS-data-set <OUT=SAS-data-set><keyword<=name>>...< / options>;
RUN;
```

```
proc print data=scored_houses;
run;
```



Predictions using PROC PLM

The PLM Procedure

Store Information	
Item Store	WORK.ISBONUS
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC LOGISTIC
Date Created	31MAY18:14:53:32
Response Variable	Bonus
Link Function	Logit
Distribution	Binary
Class Variables	Fireplaces Lot_Shape_2 Bonus
Model Effects	Intercept Basement_Area Lot_Shape_2 Basement_*Lot_Shape_ Fireplaces

Predictions using PROC PLM

Obs	Lot_Shape_2	Fireplaces	Basement_Area	Predicted
1	Regular	0	1060	0.02192
2	Regular	2	775	0.00040
3	Irregular	2	1100	0.14210
4	Irregular	1	975	0.30608
5	Regular	1	800	0.00286

/*st107d07.sas*/

```
ods select none;
proc logistic data=STAT1.ameshousing3;
   class Fireplaces(ref='0') Lot_Shape_2(ref='Regular') / param=ref;
   model Bonus(event='1')=Basement_Area|Lot_Shape_2 Fireplaces;
   units Basement_Area=100;
        store out=isbonus;
run;
ods select all;



data newhouses;
        length Lot_Shape_2 $9;
        input Fireplaces Lot_Shape_2 $ Basement_Area;
```

```
        datalines;
        0  Regular    1060
        2  Regular     775
        2  Irregular  1100
        1  Irregular   975
        1  Regular     800
        ;
run;

proc plm restore=isbonus;
        score data=newhouses out=scored_houses / ILINK;
        title 'Predictions using PROC PLM';
run;

proc print data=scored_houses;
run;
```

Predictions using PROC PLM

The PLM Procedure

Store Information	
Item Store	WORK.ISBONUS
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC LOGISTIC
Date Created	06SEP21:06:18:33
Response Variable	Bonus
Link Function	Logit
Distribution	Binary
Class Variables	Fireplaces Lot_Shape_2 Bonus
Model Effects	Intercept Basement_Area Lot_Shape_2 Basement_*Lot_Shape_ Fireplaces

Predictions using PROC PLM

Obs	Lot_Shape_2	Fireplaces	Basement_Area	Predicted
1	Regular	0	1060	0.02192
2	Regular	2	775	0.00040
3	Irregular	2	1100	0.14210
4	Irregular	1	975	0.30608
5	Regular	1	800	0.00286

```
/*st107s04.sas*/
ods graphics on;
proc logistic data=STAT1.safety plots(only)=(effect oddsratio);
   class Region (param=ref ref='Asia')
                Size (param=ref ref='Small');
         model Unsafe(event='1') = Weight Region Size
                  / clodds=pl selection=backward;
         units Weight = -1;
         store isSafe;
   format Size sizefmt.;
         title 'Logistic Model: Backwards Elimination';
run;

data checkSafety;
   length Region $9.;
         input Weight Size Region $ 5-13;
         datalines;
4 1 N America
3 1 Asia
5 3 Asia
5 2 N America
         ;
run;

proc plm restore=isSafe;
         score data=checkSafety out=scored_cars / ILINK;
         title 'Safety Predictions using PROC PLM';
run;
```

```
proc print data=scored_cars;
run;
```

Logistic Model: Backwards Elimination

The LOGISTIC Procedure

Model Information	
Data Set	STAT1.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.

Backward Elimination Procedure

Class Level Information			
Class	Value	Design Variables	
Region	Asia	0	
	N America	1	
Size	Large	1	0
	Medium	0	1
	Small	0	0

Step 0. The following effects were entered:

Intercept Weight Region Size

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

Step 1. Effect Region is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	121.249	92.455
SC	123.813	102.712
-2 Log L	119.249	84.455

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.7937	3	<.0001
Score	32.4658	3	<.0001
Wald	23.9471	3	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
0.4526	1	0.5011

Step 2. Effect Weight is removed:

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	121.249	92.629
SC	123.813	100.322
-2 Log L	119.249	86.629

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	32.6199	2	<.0001
Score	31.3081	2	<.0001
Wald	24.2875	2	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
2.5983	2	0.2728

Note: No (additional) effects met the 0.05 significance level for removal from the model.

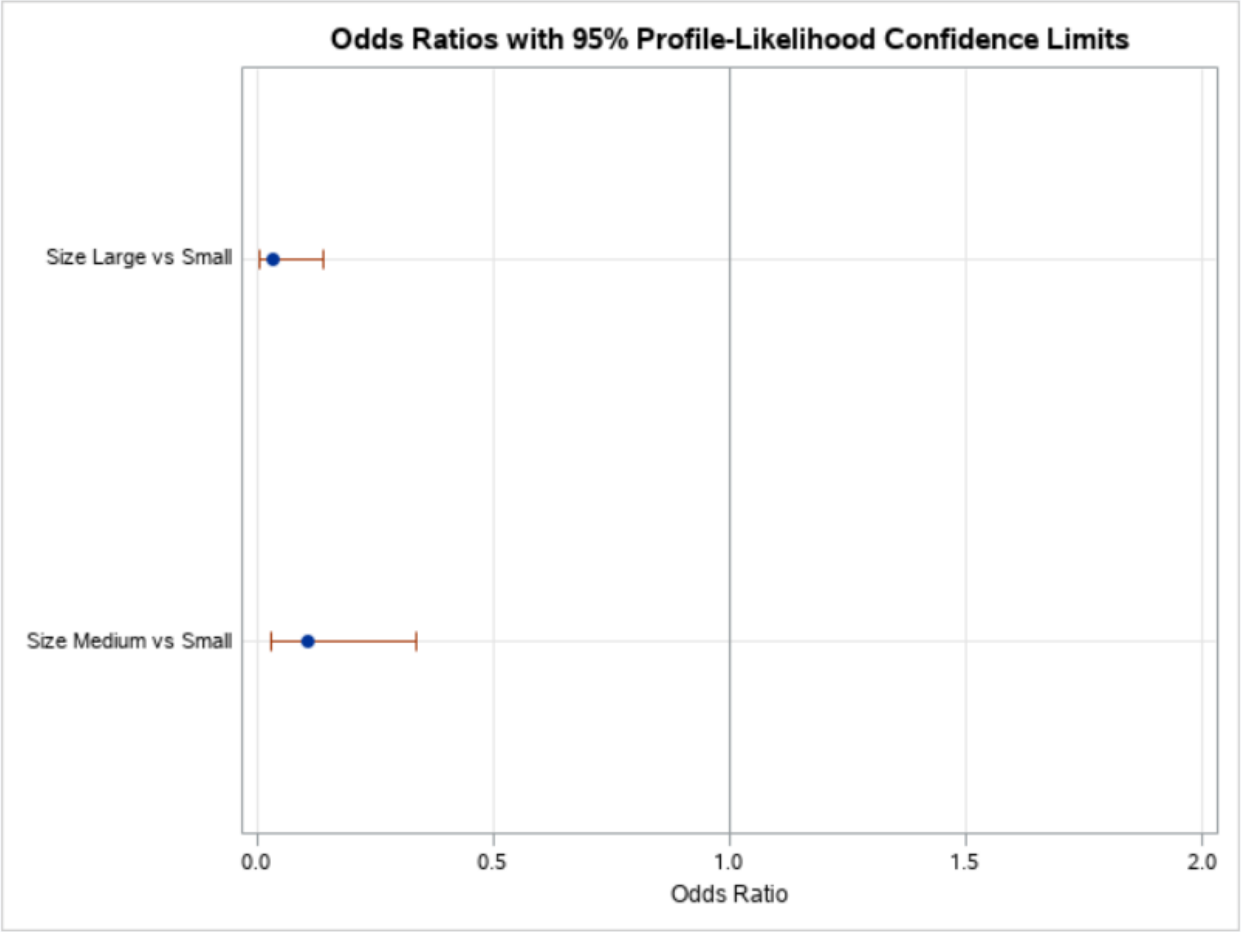
Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	Region	1	2	0.4506	0.5020
2	Weight	1	1	2.1565	0.1420

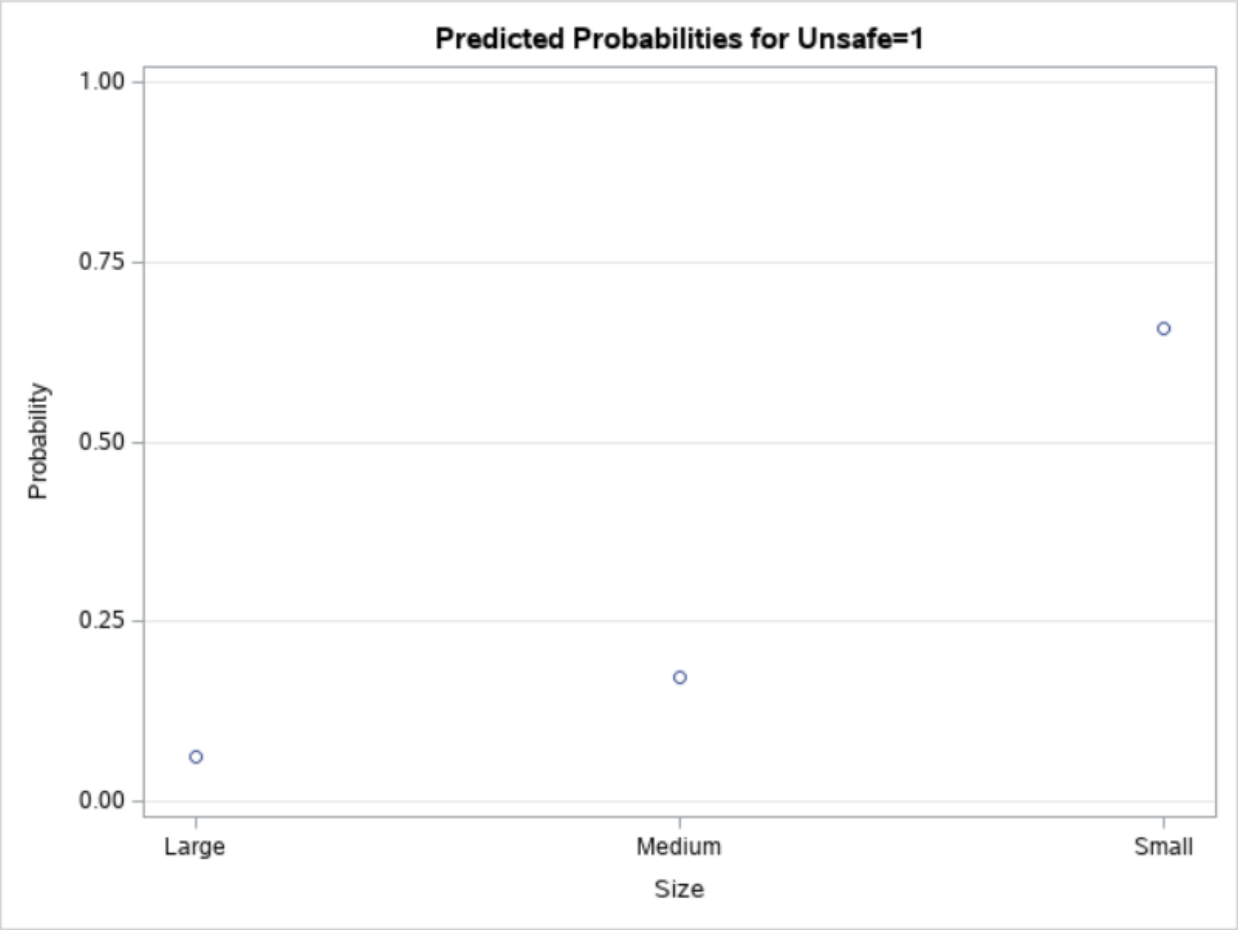
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Size	2	24.2875	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.6506	0.3561	3.3377	0.0677
Size	Large	1	-3.3585	0.8125	17.0880	<.0001
Size	Medium	1	-2.2192	0.6070	13.3654	0.0003

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	70.3	Somers' D	0.636
Percent Discordant	6.7	Gamma	0.827
Percent Tied	23.0	Tau-a	0.276
Pairs	1980	c	0.818

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Size Large vs Small	1.0000	0.035	0.005	0.141
Size Medium vs Small	1.0000	0.109	0.030	0.336


Odds Ratios with 95% Profile-Likelihood Confidence Limits
Size Large vs Small
Size Medium vs Small
0.0
0.5
1.0
1.5
2.0
Odds Ratio


Predicted Probabilities for Unsafe=1
1.00
0.75
0.50
0.25
0.00
Probability
Large
Medium
Small
Size

Safety Predictions using PROC PLM

The PLM Procedure

Store Information	
Item Store	WORK.ISSAFE
Data Set Created From	STAT1.SAFETY
Created By	PROC LOGISTIC
Date Created	06SEP21:06:22:02
Response Variable	Unsafe
Link Function	Logit
Distribution	Binary
Class Variables	Region Size Unsafe
Model Effects	Intercept Size

Safety Predictions using PROC PLM

Obs	Region	Weight	Size	Predicted
1	N America	4	Small	0.65714
2	Asia	3	Small	0.65714
3	Asia	5	Large	0.06251
4	N America	5	Medium	0.17241

Practice: Using PROC LOGISTIC to Perform Backward Elimination and PROC PLM to Generate Predictions

Question 1

The insurance company wants to model the relationship between three of a car's characteristics, weight, size, and region of manufacture, and its safety rating. Run PROC LOGISTIC and use backward elimination. Start with a model using only main effects. The **stat1.safety** data set contains the data about vehicle safety.

1. Use PROC LOGISTIC to fit a multiple logistic regression model with **Unsafe** as the response variable and **Weight**, **Size**, and **Region** as the predictor variables.
2. Use the EVENT= option to model the probability of Below Average safety scores.
3. Apply the SIZEFMT. format to the variable **Size**.
4. Specify **Region** and **Size** as classification variables and use reference cell coding. Specify *Asia* as the reference level for **Region**, and *1* (small cars) as the reference level for **Size**.
5. Add a UNITS statement with -1 as the unit for **Weight** so that you can see the odds ratio for lighter cars over heavier cars.
6. Add a STORE statement to save the analysis results as **isSafe**.
7. Request any relevant plots.
8. Submit the code and view the results.

Which terms appear in the final model?

Only **Size** appears in the final model.

```
/*st107s04.sas*/

ods graphics on;
proc logistic data=STAT1.safety plots(only)=(effect oddsratio);
   class Region (param=ref ref='Asia')
           Size (param=ref ref='Small');
   model Unsafe(event='1') = Weight Region Size
         / clodds=pl selection=backward;
   units Weight = -1;
   store isSafe;
   format Size sizefmt.;
   title 'Logistic Model: Backwards Elimination';
run;
```

Notice that the reference level for **Size** is set to 'Small' in the solution, rather than '1'. When a format is applied to a CLASS statement variable, the reference level option should refer to the formatted value and not the internal value.

Question 2

If you compare these results with those from the previous practice (a model fit with only one variable, **Region**), do you think that this is a better model?

Comparing the model fit statistics, you see that the AIC (92.629) and SC (100.322) are both smaller in the model fit by the backward elimination method, 119.854 and 124.982, respectively. This indicates that the **Size**-only model is doing better than the **Region**-only model. Using the c statistics, you can also see improvement beyond the **Region**-only model, that is, 0.818 in this model compared with 0.598 in the previous model.

Question 3

Using the final model that was chosen by backward elimination, and using the STORE statement, generate predictive probabilities for the cars in the following DATA step:

```
data checkSafety;
   length Region $9.;
   input Weight Size Region $ 5-13;
   datalines;
   4 1 N America
   3 1 Asia
   5 3 Asia
   5 2 N America
    ;
run;
```

size small has the highest probability (no matter what the region is)

```
proc plm restore=isSafe;
   score data=checkSafety out=scored_cars / ILINK;
   title 'Safety Predictions using PROC PLM';
run;

proc print data=scored_cars;
run;
```