Part 4
Lecture 1 Categorical Data

1

Who I am…

Pascal Tyrrell, PhD

Associate Professor

Department of Medical Imaging, Faculty of Medicine

Institute of Medical Science, Faculty of Medicine

Department of Statistical Sciences, Faculty of Arts and Science

Examining Categorical Variables

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

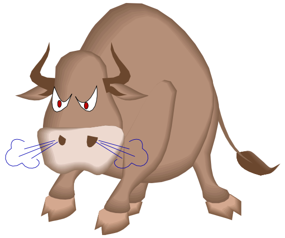
- determine the frequencies of data values.
- recognize possible associations among variables

Categorical Variables Association

- An association exists between two categorical variables if the distribution of one variable changes when the level (or value) of the other variable changes.

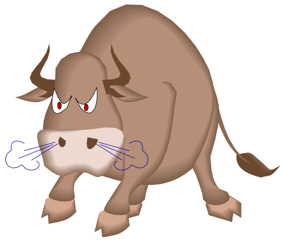
- If there is no association, the distribution of the first variable is the same regardless of the level of the other variable.

No Association

		
	72%	28%
	72%	28%

Is your manager's mood associated
with the weather?

Association

		
	82%	18%
	60%	40%

Is your manager's mood associated
with the weather?

Frequency Tables

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

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

Cross Tabulation Tables

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

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

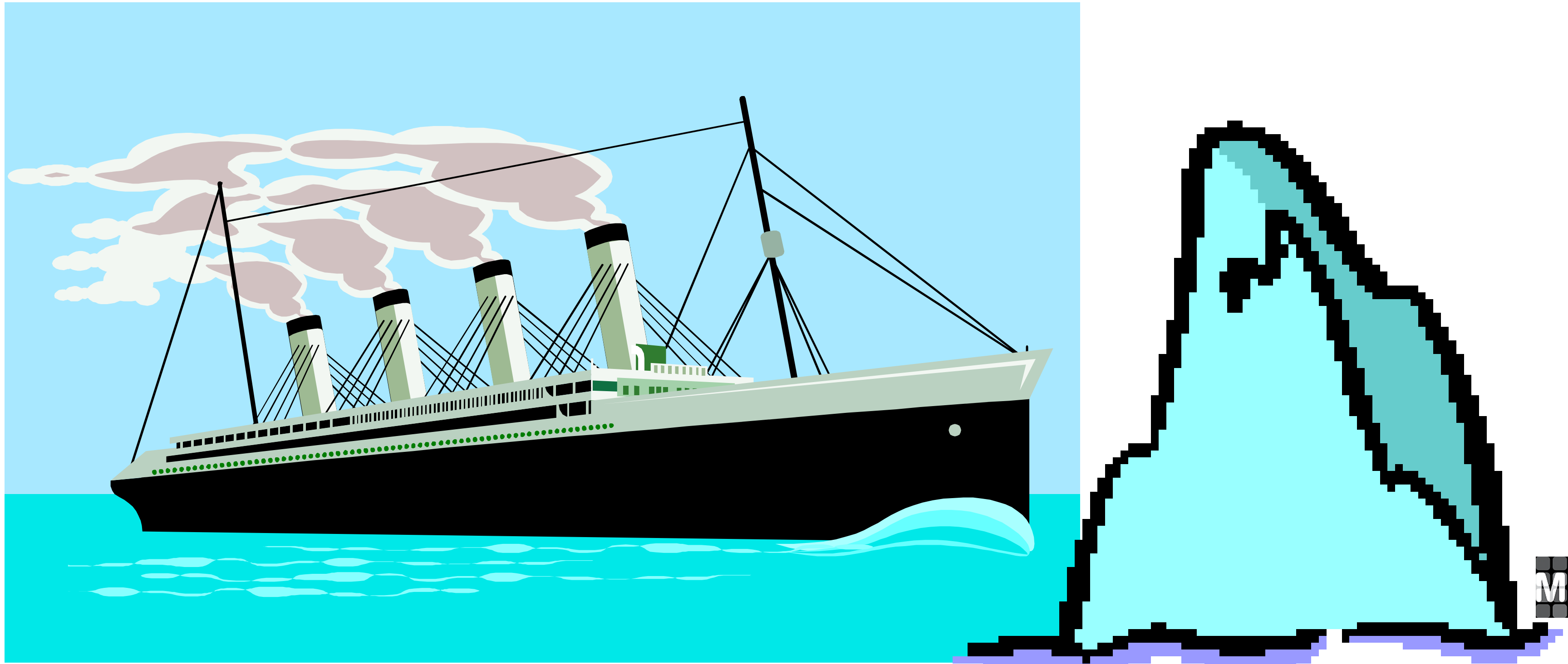
The FREQ Procedure

- General form of the FREQ procedure:

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

Titanic Example

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



Question

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

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

Correct Answer

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

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

Objectives

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

Overview

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

Introduction

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

Null Hypothesis

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

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

Chi-Square Test

NO ASSOCIATION

observed frequencies=expected frequencies

ASSOCIATION

observed frequencies≠expected frequencies

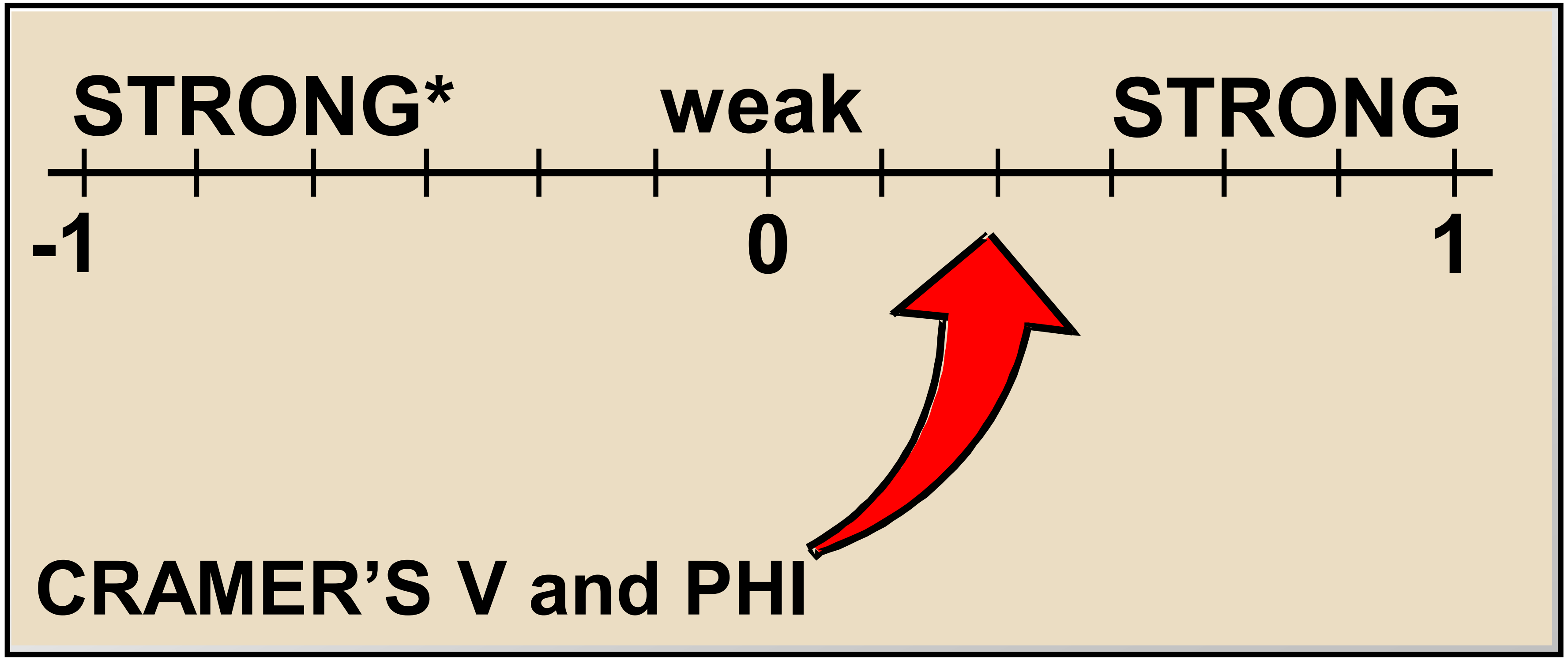
- The expected frequencies are calculated by the formula: (row total*column total) / sample size.

Chi-Square Tests

- Chi-square tests and the corresponding *p*-values
  - determine whether an association exists
  - do not measure the strength of an association
  - depend on and reflect the sample size.

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

Measures of Association



Cramer's V is always non negative for tables larger than 2*2.
Use Phi for 2*2 tables.

Odds Ratios

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

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

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

Probability versus Odds of an Outcome

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

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

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

Probability versus Odds of an Outcome

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

Probability of **Yes** in Group B=0.90

÷

Probability of **No** in Group B=0.10

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

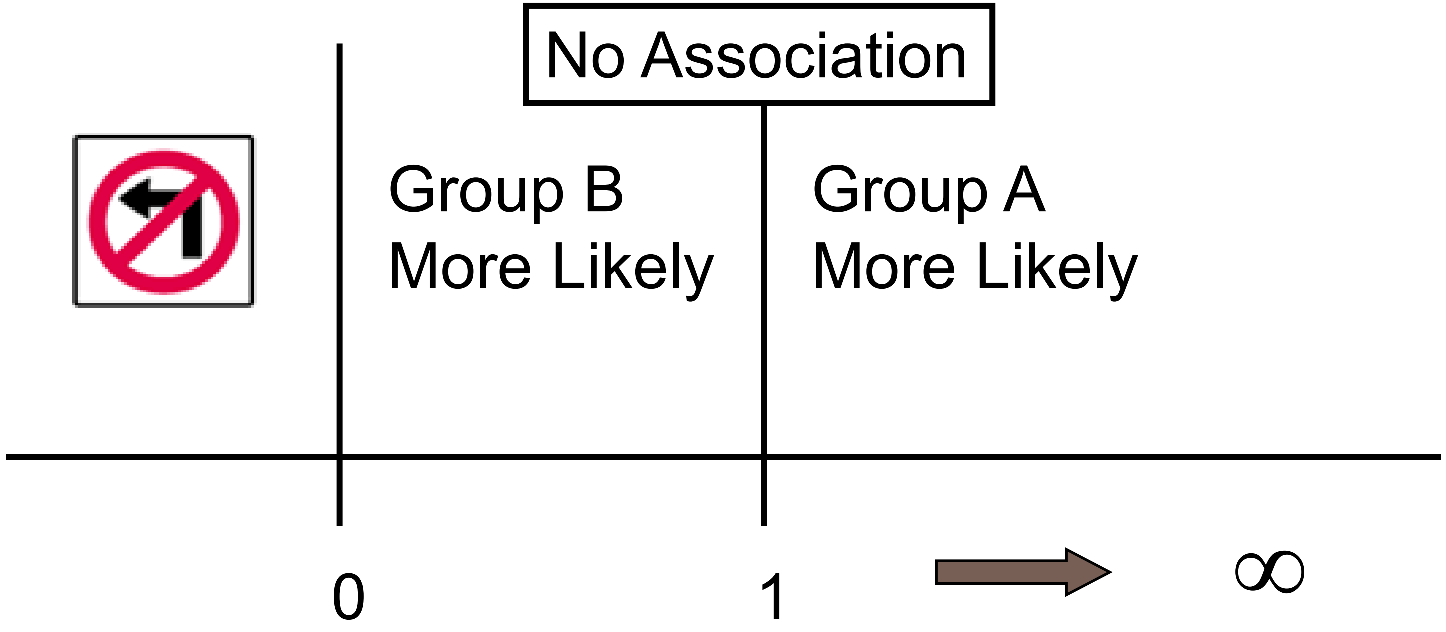
Odds Ratio

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

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

Odds Ratio, **A** to **B**=**3**÷**9**=**0.3333**

Properties of the Odds Ratio, A to B



Multiple Answer Poll

- What tends to happen when sample size decreases?

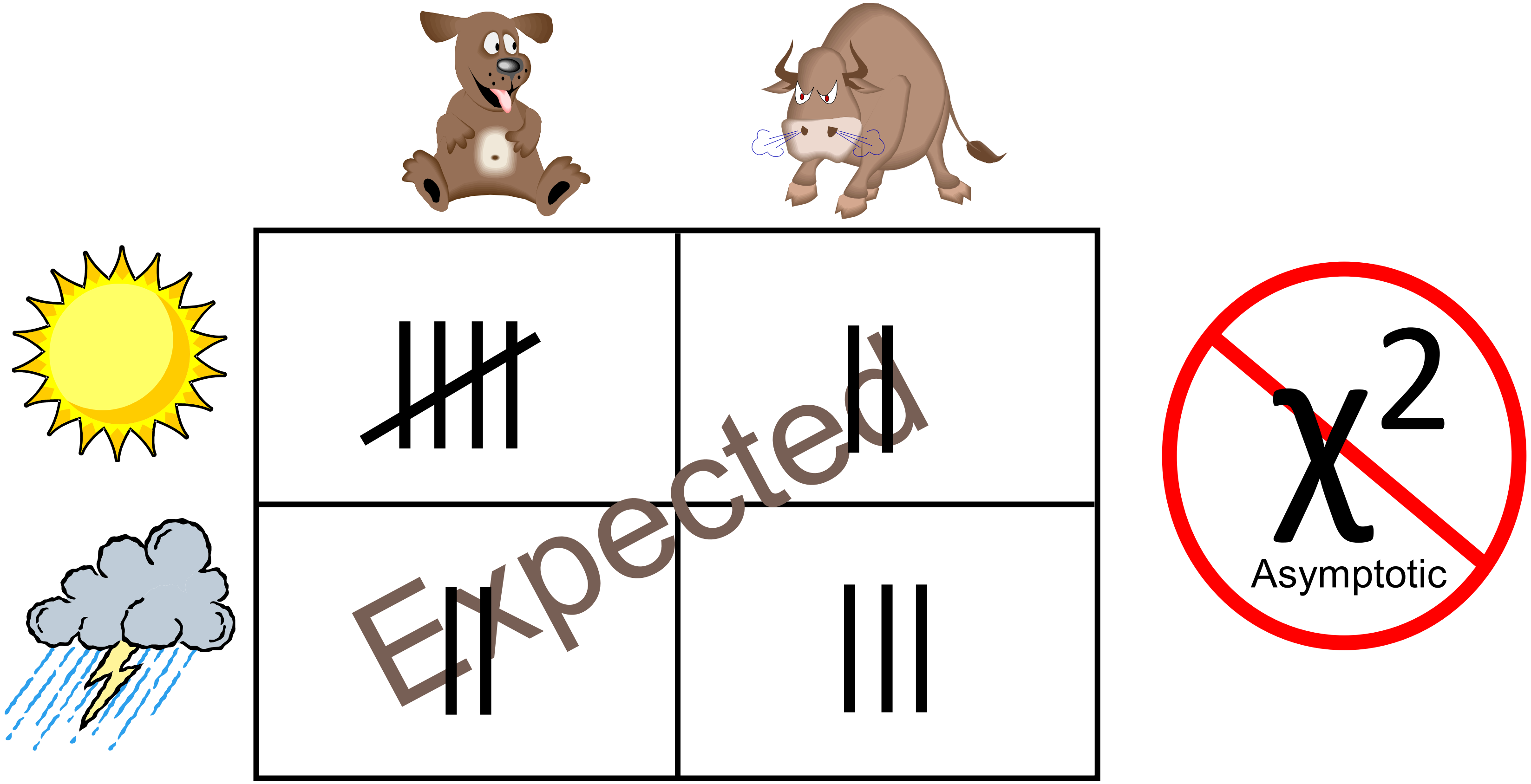
  a. The chi-square value increases.
  b. The *p*-value increases.
  c. Cramer's V increases.
  d. The Odds Ratio increases.
  e. The width of the CI for the Odds Ratio increases.

Multiple Answer Poll – Correct Answers

- What tends to happen when sample size decreases?

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

When Not to Use the Asymptotic χ^2



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

Observed versus Expected Values

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

Small Samples – Exact *p*-Values

Sample Size

Small

Large

→ **Exact *p*-values**

→ **Asymptotic *p*-values**

Exact *p*-Values for Pearson Chi-Square

Observed Table		
<u>0</u>	<u>3</u>	3
<u>2</u>	<u>2</u>	4
2	5	7

Expected Table		
.86	2.14	3
1.14	2.86	4
2	5	7

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

Could the <u>underlined</u> sample values occur just by chance?

Exact *p*-Values for Pearson Chi-Square

Observed Table

0	3	3
2	2	4
2	5	7

χ^2=2.100
prob=0.286

Possible Table 2

Most like expected table

1	2	3
1	3	4
2	5	7

χ^2=0.058
prob=0.571

Possible Table 3

2	1	3
0	4	4
2	5	7

χ^2=3.733
prob=0.143

Most likely, given **marginal** values

Exact p-Values for Pearson Chi-Square

Observed Table

0	3	3
2	2	4
2	5	7

χ^2=2.100
prob=0.286

Possible Table 2

1	2	3
1	3	4
2	5	7

χ^2=0.058
prob=0.571

Possible Table 3

2	1	3
0	4	4
2	5	7

χ^2=3.733
prob=0.143

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

p-value=0.286+0.143=0.429

End of Lecture 1

Next up in Part 4 Lecture 2: Logistic Regression