CATEGORICAL DATA ANALYSIS

Analytics Primer

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	Tests of Association	Logistic Regression	Logistic Regression

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	Tests of Association	Logistic Regression	Logistic Regression

DESCRIBING CATEGORICAL DATA

Qualitative Data Types

- **Qualitative**:
  - Data whose measurement scale is inherently categorical.
  - **Nominal** – categories with no logical ordering
  - **Ordinal** – categories with a logical order / only two ways to order the categories (binary IS ordinal)

Examining Categorical Variables

- By examining the distributions of categorical variables, you can do the following:
  1. Determine the frequencies of data values
  2. 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

		
	78%	22%
	78%	22%

Is your manager’s mood associated with the weather?

Association

		
	87%	13%
	40%	60%

Is your manager’s mood associated with the weather?

TESTS OF ASSOCIATION

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	Tests of Association	Logistic Regression	Logistic Regression

Association

		
	87%	13%
	40%	60%

How **much of a change** is required
to believe there actually is a difference?

Tests of Association - Hypotheses

- **Null Hypothesis**
  - There is no association between **Mood** and **Weather**.
  - The probability of being happy was the same whether it was sunny or rainy.

- **Alternative Hypothesis**
  - There ***is*** an association between **Mood** and **Weather**.
  - The probability of being happy was **not** the same whether it was sunny or rainy.

Chi-Square Tests

H_0: ***NO ASSOCIATION***

observed frequencies = expected frequencies

H_a: ***ASSOCIATION***

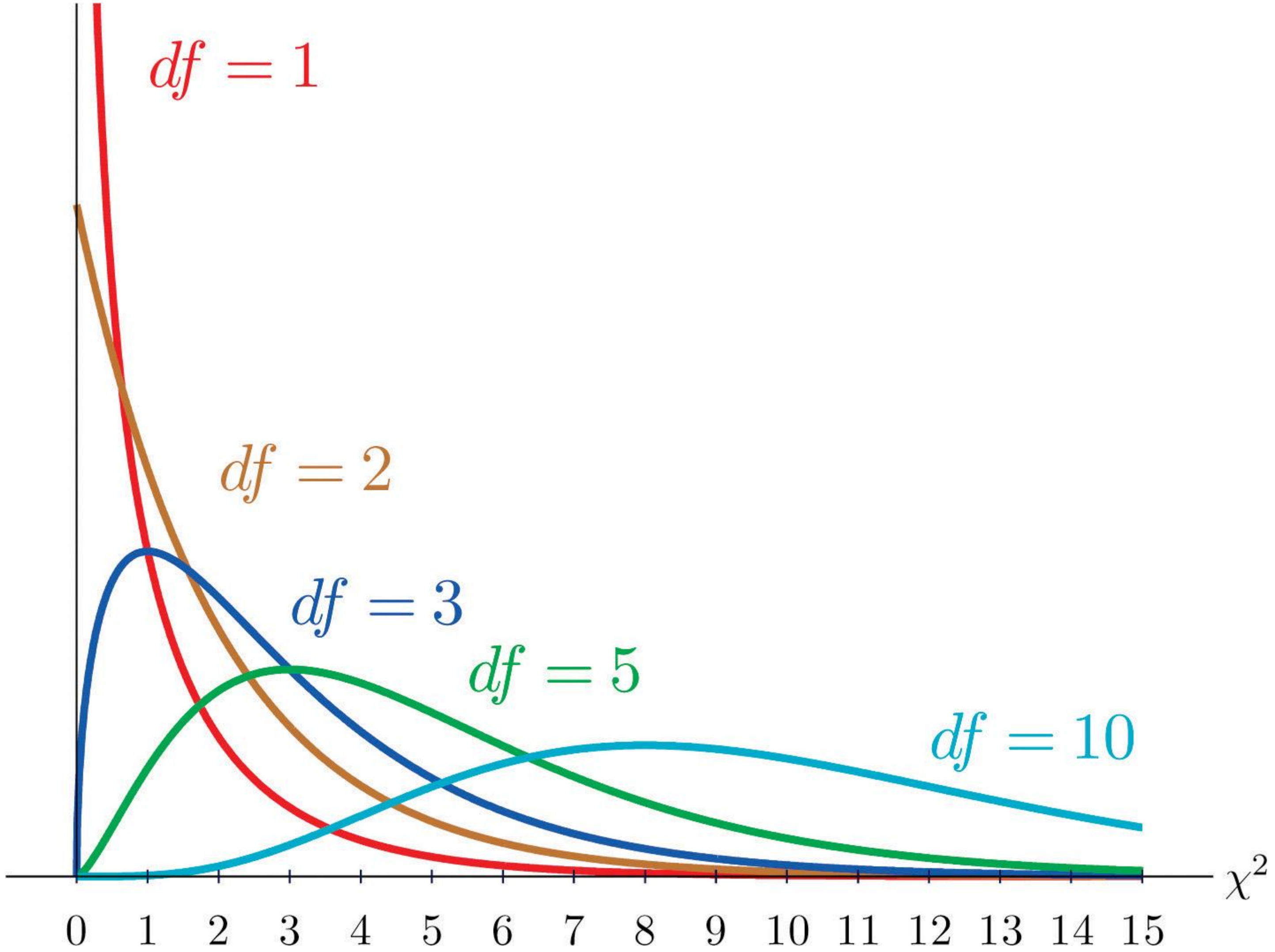
observed frequencies ≠ expected frequencies

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

χ^2-Distribution

- The Chi-Square test comes from the **χ^2-distribution**.
- Characteristics of the χ^2-distribution:
  1. Bounded Below By Zero
  2. Right Skewed
  3. One set of Degrees of Freedom

χ^2-Distribution



Pearson Chi-Square Test

$$Q_P = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(Obs_{i,j} - Exp_{i,j}\right)^2}{Exp_{i,j}}$$

Pearson Chi-Square Test

$$Q_P = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(Obs_{i,j} - Exp_{i,j}\right)^2}{Exp_{i,j}}$$

D.F. = (# Rows – 1)(# Columns – 1)

Likelihood Ratio Chi-Square Test

$$Q_{LR} = 2 \times \sum_{i=1}^{R} \sum_{j=1}^{C} Obs_{i,j} \times \log\left(\frac{Obs_{i,j}}{Exp_{i,j}}\right)$$

Likelihood Ratio Chi-Square Test

$$Q_{LR} = 2 \times \sum_{i=1}^{R} \sum_{j=1}^{C} Obs_{i,j} \times \log\left(\frac{Obs_{i,j}}{Exp_{i,j}}\right)$$

D.F. = (# Rows – 1)(# Columns – 1)

Example

- A manager of a major car dealership wants to determine if the membership of a client in the loyalty program is associated with the color of car that they buy. With this knowledge, it potentially could help the sales staff show different cars to different clients to help improve the likelihood of a sale. The manager pull information from the previous years sales.

Example

1. Calculate the expected counts in the right table:

Observed

Color	Yes	No	Total
Black	149	101	250
White	101	66	167
Blue	72	108	180
Red	96	161	257
Green	39	65	104
Total	457	501	958

Expected

Color	Yes	No	Total
Black			250
White			167
Blue			180
Red			257
Green			104
Total	457	501	958

Example

1. Calculate the expected counts in the right table:

Observed

Color	Yes	No	Total
Black	149	101	250
White	101	66	167
Blue	72	108	180
Red	96	161	257
Green	39	65	104
Total	457	501	958

Expected

Color	Yes	No	Total
Black			250
White			167
Blue			180
Red			257
Green			104
Total	457	501	958

$$\frac{457}{958} \times 250 = 119.26$$

Population % of Loyal Customers

Example

1. Calculate the expected counts in the right table:

Observed

Color	Yes	No	Total
Black	149	101	250
White	101	66	167
Blue	72	108	180
Red	96	161	257
Green	39	65	104
Total	457	501	958

Expected

Color	Yes	No	Total
Black			250
White			167
Blue			180
Red			257
Green			104
Total	457	501	958

$$\frac{457}{958} \times 250 = 119.26$$

Customers Bought Black Car

Example

1. Calculate the expected counts in the right table:

Observed

Color	Yes	No	Total
Black	149	101	250
White	101	66	167
Blue	72	108	180
Red	96	161	257
Green	39	65	104
Total	457	501	958

Expected

Color	Yes	No	Total
Black	119.26		250
White			167
Blue			180
Red			257
Green			104
Total	457	501	958

$$\frac{457}{958} \times 250 = 119.26$$

Expected # Loyal Buying Black Car

Example

1. Calculate the expected counts in the right table:

Observed

Color	Yes	No	Total
Black	149	101	250
White	101	66	167
Blue	72	108	180
Red	96	161	257
Green	39	65	104
Total	457	501	958

Expected

Color	Yes	No	Total
Black	119.26	130.74	250
White	79.66	87.34	167
Blue	85.87	94.13	180
Red	122.60	134.40	257
Green	49.61	54.39	104
Total	457	501	958

Example

2. Compute Q_P and Q_{LR} and summarize results.

Example

2. Compute Q_P and Q_{LR} and summarize results.

$$Q_P = \frac{(149 - 119.26)^2}{119.26} + \cdots + \frac{(65 - 54.39)^2}{54.39} = 44.77$$

$$Q_{LR} = 2 \times \left(149 \times \log\left(\frac{149}{119.26}\right) + \cdots + 65 \times \log\left(\frac{65}{54.39}\right)\right) = 45.09$$

Ordinal Compared to Nominal Tests

- Both the Pearson and Likelihood Ratio Chi-Square tests can handle any type of categorical variable – either ordinal, nominal, or both.
- However, ordinal variables provide us extra information since the order of the categories actually matters compared to nominal.
- We can test for even more with ordinal variables against other ordinal variables – whether two ordinal variables have a **linear relationship** as compared to just a general one.

Ordinal vs. Ordinal Chi-Square Tests

H_0: ***NO LINEAR ASSOCIATION***

H_a: ***LINEAR ASSOCIATION***

Mantel-Haenszel Chi-Square Test

$$Q_{MH} = (n-1)r^2$$

Pearson correlation
between row and
column variables.

Mantel-Haenszel Chi-Square Test

$$Q_{MH} = (\mathrm{n} - 1)\mathrm{r}^2$$

D.F. = 1

MEASURES OF ASSOCIATION

Chi-Square Tests

- Determines whether an association exists
- **DOES NOT** measure the strength of the association
- Depends on and reflects the sample size

Measures of Association

- **DOES NOT** determines whether an association exists
- Measures the **strength of the association**

- There are many different measures of association.
- Two common measures of association are the following:
  1. Odds Ratios (Only for 2x2 tables – binary vs. binary)
  2. Cramer’s V (Any size table)

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.
- The **odds** of an event occurring is NOT the same as the probability that an event occurs.

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.
- The **odds** of an event occurring is NOT the same as the probability that an event occurs.

$$Odds = \frac{p}{1-p}$$

Probability versus Odds of an Outcome

	Buy Product		
	Yes	No	Total
Loyal	20	60	80
Non-Loyal	10	90	100
Total	30	150	180

Total Yes outcomes in Non-Loyal ÷ Total outcomes in Non-Loyal

Probability of a **Yes** in Non-Loyal = **10**÷**100**=**0.1**

Probability versus Odds of an Outcome

	Buy Product		
	Yes	**No**	**Total**
Loyal	20	60	80
Non-Loyal	10	90	100
Total	30	150	180

Probability of **Yes** in Non-Loyal = 0.10 ÷ Probability of **No** in Non-Loyal = 0.90

Odds of **Yes** in Non-Loyal = **0.10**÷**0.90**=1/9

Odds Ratio

	Buy Product		
	Yes	**No**	**Total**
Loyal	20	60	80
Non-Loyal	10	90	100
Total	30	150	180

Odds of Yes in **Loyal** = **1/3** ÷ Odds of Yes in **Non-Loyal** = **1/9**

Odds Ratio, **Loyal** to **Non-Loyal** = **1/3**÷**1/9**=**3**

Odds Ratio

Odds of Yes in **Loyal** = **1/3**

÷

Odds of Yes in **Non-Loyal** = **1/9**

Odds Ratio, **Loyal** to **Non-Loyal**= **1/3**÷**1/9**=**3**

Loyal program customers have **3 times the odds** of buying the product as compared to cutomers not in the loyalty program.

Cramer's V

- Odds ratios provide value for binary vs. binary relationships, but when you have more than two categories in one or both variables use **Cramer's V**.

$$V = \sqrt{\frac{\left(\frac{Q_P}{n}\right)}{\min(\#\text{Rows} - 1, \#\text{Columns} - 1)}}$$

Cramer's V

- Odds ratios provide value for binary vs. binary relationships, but when you have more than two categories in one or both variables use **Cramer's V**.

$$V = \sqrt{\frac{\left(\frac{Q_P}{n}\right)}{\min(\#\text{Rows} - 1, \#\text{Columns} - 1)}}$$

- Bounded between 0 and 1 (-1 and 1 for 2x2 scenario) where closer to 0 the weaker the relationship.

Example

- The same manager as the previous example now wants to know the strength of the relationship between the color of car and loyalty program. Use the appropriate measure of association to calculate this.

Example

- The same manager as the previous example now wants to know the strength of the relationship between the color of car and loyalty program. Use the appropriate measure of association to calculate this.

$$V = \sqrt{\frac{\left(\frac{Q_P}{n}\right)}{\#\text{Columns} - 1}} = \sqrt{\frac{\left(\frac{44.77}{958}\right)}{2 - 1}} = 0.216$$