Introduction to Categorical Data Analysis

Overview for Today

Today we will be learning about:

- Categorical data
- Maximum likelihood estimation
- Statistical inference for proportions

What is a Categorical Variable?

Categorical variable is a variable that consists of a set of two or more categories:

- Customer churn: Remained or Left
- Political ideology: Democrat, Republican, or Independent
- Medical diagnosis: Positive or Negative
- Attitude measures: Satisfied or Not Satisfied

Categorical Variable as a Predictor

So far we have talked about categorical variables as predictors of some quantitative variable:

- How does an employees' work status (full-time or part-time) impact their job satisfaction?

Now we will start to talk about categorical variables as outcomes.

Common Ways to Model Categorical Data

We are going to cover three common ways to analyze categorical data:

- Comparing a single proportion to a null value
- Comparing two proportions to one another
- Comparing two or more proportions at once

Common Probability Distributions for Categorical Data

The two most common probability distributions used to model categorical data are the:

- **Binomial distribution** for binary categorical variables
- **Multinomial distribution** for multicategorical variables

Understanding the Binomial Distribution

The binomial distribution tells us the probability of seeing **k** successes in a sequence of **n** trials:

$$P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

- $\binom{n}{k}$: Tells us how many ways we can see k success in n trials
- π: The probability of a success
- n: The number of trials
- k: The number of successes

The Mean and Variance of a Binomial Variable

Much like we do with a quantitative variable, we will often want to describe a categorical variable with its mean and variance. For a binary variable (binomial distribution), we can calculate its mean and variance as:

$$\text{Mean} = \pi, \text{Variance} = \pi(1 - \pi)$$

Modeling Two (Fair) Coin Flips

With two coin flips, there are four possible outcomes:

1. H, H - 2 Successes
2. H, T - 1 Success
3. T, H - 1 Success
4. T, T - 0 Successes

The probability of 1 success:

$$P(X = 1) = \binom{2}{1} .50^1 (1 - .50)^{2-1} = 2 \times .50 \times .50 = .5$$

Probability Models, Parameters, and Estimates

A **probability model** is a function (equation) that tells us how the **probability** of an event changes as a function of the **observed data** and the **parameters** of the probability model. Often, we need to use the data to **estimate** the parameters of the probability model.

- The binomial distribution can be used as a probability model for categorical variables that take on two categories
- The multinomial distribution can be used as a probability model for categorical variables that take on more than two categories
- The normal distribution can be used as a probability model for quantitative variables

Using Data to Estimate Probability Model Parameters

We usually want to learn about the probability model by collecting data that could have been plausibly generated from our hypothesized probability model and then we use the data to estimate the unknown probability model parameters:

- In linear regression, we collect data in order to estimate and test the relationships (regression slops) between the outcome and predictor variables.
- In election years, pollsters collect data in order to estimate the chances of one candidate winning over another.

Maximum Likelihood Estimation

To estimate the parameters, we can use a method called **maximum likelihood estimation**.

You can think of ML estimation as answering the question: “What parameters of the probability model make my observed data most likely?”

The parameter that answers this question is called the **maximum likelihood estimate or MLE**.

Guessing the MLE

You have flipped a coin 100 times and heads came up 57 times (57 %). You believe the data was generated from a binomial distribution, what value does the π have to be to make your data most likely?

- π: .05, Likelihood = 0.46
- π: .25, Likelihood = 0.88
- π: .57, Likelihood = 1
- π: .80, Likelihood = 0.93

How Does the MLE Relate to Statistical Inference?

When estimated from “enough” data all MLE have some nice characteristics:

- Normally distributed sampling distribution
- Small standard errors
- Estimates are usually very close to the parameter they are estimating

When it comes time to make inferences about MLEs, these characteristics allow us to do the same thing we have been doing when we make inferences about linear regression coefficients – calculate a test statistic and see how extreme it is given a normally distributed null distribution!

The MLE of the Binomial Parameter: The Proportion

It turns out that the value of π that is always going to maximize the Likelihood function for a binomial probability model is:

$$\hat{\pi} = \frac{\#\,\text{Successes}}{\#\,\text{Trials}}$$

This is just the proportion of successes to trials!

Modeling Customer Churn from a Content Streaming Service

You are an analyst working at a content streaming provider who has been asked to make inferences about which customers are likely to cancel their subscriptions (churn):

```
# A tibble: 1,000 × 2
   generation_cat churn_cat
   <chr>          <chr>
 1 Millenials     Y
 2 Millenials     Y
 3 Baby Boomers   N
 4 Gen X          N
 5 Millenials     N
 6 Millenials     N
 7 Gen Z          N
 8 Gen Z          N
 9 Millenials     N
10 Millenials     Y
# ℹ 990 more rows
```

Statistical Inference for One Proportion

To start this question, maybe we want to know if the proportion of customers who cancel their subscription is less than the industry proportion of .40. We could set up a hypothesis test:

$$H_0 : \pi = .40$$

$$H_a : \pi \neq .40$$

Setting Up the Statistical Test

With hypothesis testing, we are asking the question: "How many standard errors is our estimate away from the null value, assuming that our null hypothesis is true?"

$$\frac{\hat{\pi} - .40}{SE_0}, \; SE_0 = \sqrt{\frac{.40 \times (1 - .40)}{1000}}$$

Typically, if our estimate is about 2 standard errors away from the null value (p-value a little less than .05), then we can reject our null hypothesis.

Using R to Test a Proportion

```r
cust_churn <- sum(data_churn$churn)
cust_total <- length(data_churn$churn)

prop.test(x = cust_churn, n = cust_total, p = .40, alternative = "two.sided",
```

```
    1-sample proportions test without continuity correction

data:  cust_churn out of cust_total, null probability 0.4
X-squared = 10.838, df = 1, p-value = 0.0009946
alternative hypothesis: true p is not equal to 0.4
95 percent confidence interval:
 0.3200860 0.3790697
sample estimates:
    p
0.349
```

Analyzing Relationships Between Variables with a Contigency Table

We are often interested in how the chances of success on an outcome variable change at different levels of a predictor variable:

- How do the chances of customer churn might change based on customer demographics?
- How does one's political party identification relate to their sex?
- How does the development of a disease relate to behaviors like smoking?

To answer these types of questions, it is helpful to build and analyze a contingency table (also known as a cross-tabulation or cross-tabs table).

Building a Contingency Table in R

```r
xtabs(~generation_cat + churn_cat, data = data_churn) |> addmargins()
```

```
               churn_cat
generation_cat    N    Y  Sum
  Baby Boomers   88   11   99
  Gen X         244   86  330
  Gen Z         110   62  172
  Millenials    209  190  399
  Sum           651  349 1000
```

Joint, Marginal, and Conditional Probabilites

From the contingency table, you can calculate joint, marginal, and conditional probabilities:

- **Joint Probability**: The probability a customer belongs to the Gen Z generation **and** cancelled their subscription.
- **Marginal Probability**: The probability a customer belongs to the Gen Z generation.
- **Conditional Probability**: The probability a customer cancels their subscription **given** they belong to the Gen Z generation.

Joint, Maringal, and Conditional Probabilities in R

```
1 xtabs(~generation_cat + churn_cat, data = data_churn) |> prop.table() |> round
```

```
              churn_cat
generation_cat    N    Y
  Baby Boomers 0.09 0.01
  Gen X        0.24 0.09
  Gen Z        0.11 0.06
  Millenials   0.21 0.19
```

```
1 xtabs(~generation_cat + churn_cat, data = data_churn) |> prop.table() |> rowSu
```

```
Baby Boomers        Gen X        Gen Z   Millenials
         0.10         0.33         0.17         0.40
```

```
1 xtabs(~generation_cat + churn_cat, data = data_churn) |> prop.table(1) |> roun
```

```
              churn_cat
generation_cat    N    Y
  Baby Boomers 0.89 0.11
  Gen X        0.74 0.26
  Gen Z        0.64 0.36
  Millenials   0.52 0.48
```

Statistical Inference for Two Proportions

Is the proportion of Gen Z customers who cancel their subscription different from the proportion of Millennial customers who cancel their subscription?

$$H_0 : \pi_{\text{Gen Z}} = \pi_{\text{Mill.}}$$

$$H_a : \pi_{\text{Gen Z}} \neq \pi_{Mill.}$$

Setting Up the Statistical Test

With hypothesis testing, we are asking the question: "How many standard errors is our estimate away from the null value, assuming that our null hypothesis is true?"

$$\frac{\hat{\pi}_{\text{Gen Z}} - \hat{\pi}_{\text{Mill.}}}{SE}$$

Typically, if our estimate is about 2 standard errors away from the null value (p-value a little less than .05), then we can reject our null hypothesis.

Using R to Compare Two Proportions

```r
churn_z <- sum(data_churn$churn[data_churn$generation_cat == "Gen Z"])
churn_mill <- sum(data_churn$churn[data_churn$generation_cat == "Millenials"])

cust_z <- sum(data_churn$generation_cat == "Gen Z")
cust_mill <- sum(data_churn$generation_cat == "Millenials")

prop.test(x = c(churn_z, churn_mill), n = c(cust_z, cust_mill), alternative =
```

```
    2-sample test for equality of proportions without continuity correction

data:  c(churn_z, churn_mill) out of c(cust_z, cust_mill)
X-squared = 6.5283, df = 1, p-value = 0.01062
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2026169 -0.0288338
sample estimates:
   prop 1    prop 2
0.3604651 0.4761905
```

Other Ways to Compare Two Proportions

There are multiple ways we can compare and communicate the differences between two proportions:

- Absolute Risk: The simple difference between two proportions (useful when both proportions are far away form 0 or 1).
- Relative Risk: The ratio of two proportions (useful when both proportions are close to 0 or 1).
- Odds Ratio: The ratio of the odds calculated from both proportions (used in logistic regression).

Statistical Inference for Relative Risk Using R

The relative risk tells us that the probability that a customer categorized as a Millennial cancels their subscription is 1.32 times greater (or 32% greater) than the probability that a customer categorized as Generation Z cancels their subscription.

```r
prop_mill <- mean(data_churn$churn[data_churn$generation_cat == "Millenials"])
prop_z <- mean(data_churn$churn[data_churn$generation_cat == "Gen Z"])

round(prop_mill / prop_z, 2)
```

```
[1] 1.32
```

```r
PropCIs::riskscoreci(x1 = churn_mill, n1 = cust_mill, x2 = churn_z, n2 = cust_
```

```
data:

95 percent confidence interval:
 1.064382 1.664729
```

What are Odds? What is an Odds Ratio?

The **odds** of success are defined as:

$$\text{odds} = \frac{\pi}{1 - \pi}$$

and the **odds ratio** is defined as:

$$\text{odds ratio} = \frac{\text{odds}_{\text{Mill.}}}{\text{odds}_{\text{Gen. Z}}}$$

Statistical Inference for the Odds Ratio Using R

The odds ratio tells us that the odds that a customer categorized as a Millennial cancels their subscription are 1.61 times greater than the odds that a customer categorized as Generation Z cancels their subscription.

```
1 (prop_mill / (1-prop_mill)) / (prop_z / (1-prop_z)) # Odds Ratio
```

```
[1] 1.612903
```

```
1 PropCIs::orscoreci(churn_mill, cust_mill, churn_z, cust_z, conf.level = .95)
```

```
data:

95 percent confidence interval:
 1.116124 2.330795
```

Chi-Squared Test: Statistical Inference for Two or More Proportions

A chi-squared test tests the extent to which the observed contingency table cells differ from what would be expected if the categorical variables were independent:

$$\chi^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

$$n_{ij} = \text{Obs. cell count}$$

$$\mu_{ij} = \text{Exp. cell count} = n\hat{\pi}_{i+}\hat{\pi}_{+j}$$

Calculating a Chi-Squared Test

The key thing to remember is that if two variables are independent, then their joint probability (cell proportion) is the product of their marginal probabilities:

```
# A tibble: 8 × 9
  churn_cat generation_cat prop_churn prop_gen prop_cell total_sample
  <chr>     <chr>                <dbl>    <dbl>     <dbl>        <dbl>
1 Y         Millenials           0.349    0.399    0.139          1000
2 Y         Baby Boomers         0.349    0.099    0.0346         1000
3 Y         Gen X                0.349    0.33     0.115          1000
4 Y         Gen Z                0.349    0.172    0.0600         1000
5 N         Millenials           0.651    0.399    0.260          1000
6 N         Baby Boomers         0.651    0.099    0.0644         1000
7 N         Gen X                0.651    0.33     0.215          1000
8 N         Gen Z                0.651    0.172    0.112          1000
  expected_cell obs_cell chi_squared
          <dbl>    <int>       <dbl>
1         139.       190     18.5
2          34.6       11     16.1
3         115.        86      7.39
4          60.0       62      0.0648
5         260         209     9.92
```

Statistical Inference with the Chi-Squared Test

The Chi-Squared tests the following hypothesis:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \text{ for all i and j}$$

$$H_a : \pi_{ij} \neq \pi_{i+}\pi_{+j} \text{ for at least one cell}$$

Chi-Squared Test in R

The code below will create a contingency table and store it in an object named `contigency_table`. Then, it will conduct a chi-squared test using the `chisq.test` function.

```
1 contingency_table <- xtabs(~ generation_cat + churn_cat, data_churn)
2
3 results_chisq <- chisq.test(contingency_table)
4
5 results_chisq
```

```
    Pearson's Chi-squared test

data:  contingency_table
X-squared = 64.518, df = 3, p-value = 6.361e-14
```

Chi-Squared Residuals

The `chisq.test` function also provides us the residuals of the chi-squared test, which shows us whether our observed counts exceeded or fell below their expected counts. Absolute values greater than 3 represent cells that do not fit the null hypothesis.

```
1 results_chisq$stdres
```

```
              churn_cat
generation_cat           N          Y
  Baby Boomers  5.2314991 -5.2314991
  Gen X         4.1156539 -4.1156539
  Gen Z        -0.3466764  0.3466764
  Millenials   -6.8754416  6.8754416
```