Categorical Data Analysis

Categorical Data Analysis

Categorical data refers to qualitative information organized into distinct categories or groups rather than represented by numerical measurements. There are different types of categorical data:

1. Binary Data: This special type has only two categories, such as yes/no, true/false, or success/failure.
2. Nominal Data: Categories represent distinct groups without any inherent order or ranking, like gender (male, female), eye color (blue, brown, green), or car brands (Toyota, Ford, Honda).
3. Ordinal Data: These categories have a meaningful order or ranking, even though they are distinct. Examples include educational levels (elementary, high school, college, graduate) or customer satisfaction ratings (poor, fair, good, excellent).

Probability distributions for categorical data are different from those for continuous numerical data. Common probability distributions used for categorical data include the binomial distribution and multinomial distribution.

Knee Injuries Dataset

- The Knee Injuries dataset contains observations on 127 patients with sport-related injuries who were treated with two different therapies. The variables in the dataset are:

N: Patient's number Th: Therapy (placebo = 1, treatment = 2)

Age: Age in years

Sex: Gender (male = 0, female = 1)

R1: Pain before treatment (no pain = 1, severe pain = 5)

R2: Pain after three days of treatment

R3: Pain after seven days of treatment

R4: Pain after ten days of treatment

Data Preparation To perform categorical data analysis, we first convert certain variables into factor variables to indicate their categorical nature:

```
library(catdata)

## Warning: package 'catdata' was built under R version 4.3.1
```

```
## Loading required package: MASS

data(knee)
head(knee)

##   N Th Age Sex R1 R2 R3 R4
## 1 1  1  28   1  4  4  4  4
## 2 2  1  32   1  4  4  4  4
## 3 3  1  41   1  3  3  3  3
## 4 4  2  21   1  4  3  3  2
## 5 5  2  34   1  4  3  3  2
## 6 6  1  24   1  3  3  3  2

str(knee)

## 'data.frame':    127 obs. of  8 variables:
##  $ N  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Th : int  1 1 1 2 2 1 2 2 2 1 ...
##  $ Age: int  28 32 41 21 34 24 28 40 24 39 ...
##  $ Sex: num  1 1 1 1 1 1 1 1 0 0 ...
##  $ R1 : int  4 4 3 4 4 3 4 3 4 4 ...
##  $ R2 : int  4 4 3 3 3 3 3 2 4 4 ...
##  $ R3 : int  4 4 3 3 3 3 3 2 4 4 ...
##  $ R4 : int  4 4 3 2 2 2 2 2 3 3 ...
```

- Convert Therapy and Sex variables to factor variables

```
knee$Th = as.factor(knee$Th)
knee$Sex = as.factor(knee$Sex)
```

- Change factor levels for Therapy and Sex variables

```
levels(knee$Th) = c("Placebo", "Treatment")
levels(knee$Sex) = c("Male", "Female")
head(knee)

##   N        Th Age    Sex R1 R2 R3 R4
## 1 1   Placebo  28 Female  4  4  4  4
## 2 2   Placebo  32 Female  4  4  4  4
## 3 3   Placebo  41 Female  3  3  3  3
## 4 4 Treatment  21 Female  4  3  3  2
## 5 5 Treatment  34 Female  4  3  3  2
## 6 6   Placebo  24 Female  3  3  3  2
```

- Tabulated Summaries We can create tabulated summaries to better understand the distribution of data in different categories:

- Tabulated summary for Therapy variable

```
T1 = table(knee$Th)
T1
```

```
##
##   Placebo Treatment
##        63        64
```

- Proportions for Therapy variable

```
prop.table(T1)

##
##   Placebo Treatment
##  0.496063  0.503937
```

- Cross-tabulated summary for Therapy and Sex variables

```
T2 = table(knee$Th, knee$Sex)
T2

##
##             Male Female
##   Placebo     17     46
##   Treatment   21     43
```

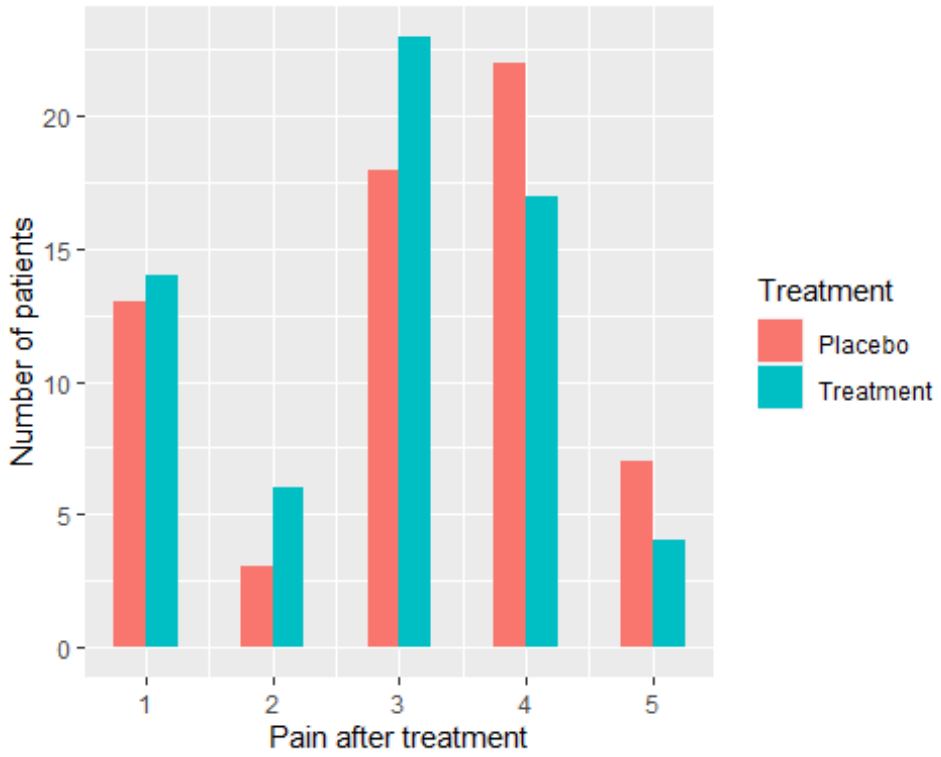
- Proportions for the cross-tabulated summary

```
prop.table(T2)

##
##                   Male    Female
##   Placebo   0.1338583 0.3622047
##   Treatment 0.1653543 0.3385827
```

Categorical Data Visualization

Visualization is useful to gain insights from categorical data. We can use ggplot2 to create histograms for categorical variables:

```
library(ggplot2)
ggplot(knee, aes(R2, fill = knee$Th)) +
  geom_histogram(position = "dodge", binwidth = 0.5) +
  labs(x = "Pain after treatment", y = "Number of patients", fill =
"Treatment")

## Warning: Use of `knee$Th` is discouraged.
## ℹ Use `Th` instead.
```



Chi-Square Goodness of Fit Test

The chi-square goodness of fit test allows us to determine whether a variable is likely to come from a specified distribution. For example, we can check if the patients were randomly allocated to the treatment and placebo groups:

```
library(lsr)

## Warning: package 'lsr' was built under R version 4.3.1

probabilities <- c(Treatment = 0.5, Placebo = 0.5)
probabilities

## Treatment   Placebo
##       0.5       0.5

library(lsr)
goodnessOfFitTest(x=knee$Th) # No need to input probabilities if they are
equal

##
##      Chi-square test against specified probabilities
##
## Data variable:   knee$Th
##
## Hypotheses:
```

```
##    null:        true probabilities are as specified
##    alternative: true probabilities differ from those specified
##
## Descriptives:
##           observed freq. expected freq. specified prob.
## Placebo               63           63.5             0.5
## Treatment             64           63.5             0.5
##
## Test results:
##    X-squared statistic:  0.008
##    degrees of freedom:  1
##    p-value:  0.929
```

Chi-Square Test of Independence

The chi-square test of independence helps us determine whether two categorical variables are related or not. For instance, we can check if Therapy (Th) and Pain after three days of treatment (R2) are independent:

```
library(lsr)
knee$R2= as.factor(knee$R2)
associationTest(formula = ~Th+as.factor(R2), data = knee )
```

```
## Warning in associationTest(formula = ~Th + as.factor(R2), data = knee):
## Expected frequencies too small: chi-squared approximation may be incorrect

##
##      Chi-square test of categorical association
##
## Variables:   Th, R2
##
## Hypotheses:
##    null:        variables are independent of one another
##    alternative: some contingency exists between variables
##
## Observed contingency table:
##            as.factor(R2)
## Th           1  2  3  4  5
##   Placebo   13  3 18 22  7
##   Treatment 14  6 23 17  4
##
## Expected contingency table under the null hypothesis:
##            as.factor(R2)
## Th             1    2    3    4    5
##   Placebo   13.4 4.46 20.3 19.3 5.46
##   Treatment 13.6 4.54 20.7 19.7 5.54
##
## Test results:
##    X-squared statistic:  3.098
```

```
##    degrees of freedom:  4
##    p-value:  0.542
##
## Other information:
##    estimated effect size (Cramer's v):  0.156
##    warning: expected frequencies too small, results may be inaccurate

associationTest(formula = ~Th + R2, data = knee)

## Warning in associationTest(formula = ~Th + R2, data = knee): Expected
## frequencies too small: chi-squared approximation may be incorrect

##
##      Chi-square test of categorical association
##
## Variables:   Th, R2
##
## Hypotheses:
##    null:        variables are independent of one another
##    alternative: some contingency exists between variables
##
## Observed contingency table:
##            R2
## Th           1  2  3  4  5
##   Placebo   13  3 18 22  7
##   Treatment 14  6 23 17  4
##
## Expected contingency table under the null hypothesis:
##            R2
## Th             1    2    3    4    5
##   Placebo   13.4 4.46 20.3 19.3 5.46
##   Treatment 13.6 4.54 20.7 19.7 5.54
##
## Test results:
##    X-squared statistic:  3.098
##    degrees of freedom:  4
##    p-value:  0.542
##
## Other information:
##    estimated effect size (Cramer's v):  0.156
##    warning: expected frequencies too small, results may be inaccurate
```

#Assumptions of Chi-Square Test There are assumptions for performing the chi-square test:

1. Expected frequencies should be sufficiently large.
2. Observations should be independent. If these assumptions are violated, alternative tests like Fisher exact test or McNemar test can be used.

Fisher Exact Test

The Fisher exact test is an alternative to the chi-square test when expected cell counts are small:

```
T3=table(knee$Th,knee$R2)
fisher.test(T3)

##
##  Fisher's Exact Test for Count Data
##
## data:  T3
## p-value = 0.5641
## alternative hypothesis: two.sided
```

McNemar Test

The McNemar test is used when the same set of observations is correlated, such as when comparing R2 and R3:

```
R2.merge = factor(ifelse(knee$R2 %in% c(1, 2), 1, 2))
R3.merge = ifelse(knee$R3 %in% c(1, 2), 1, 2)
T4 = table(R2.merge, R3.merge)
mcnemar.test(T4)

##
##  McNemar's Chi-squared test with continuity correction
##
## data:  T4
## McNemar's chi-squared = 9.0909, df = 1, p-value = 0.002569
```

Odds Ratio and 95% CI

We can compute the odds ratio and its confidence intervals to understand the relationship between two categorical variables:

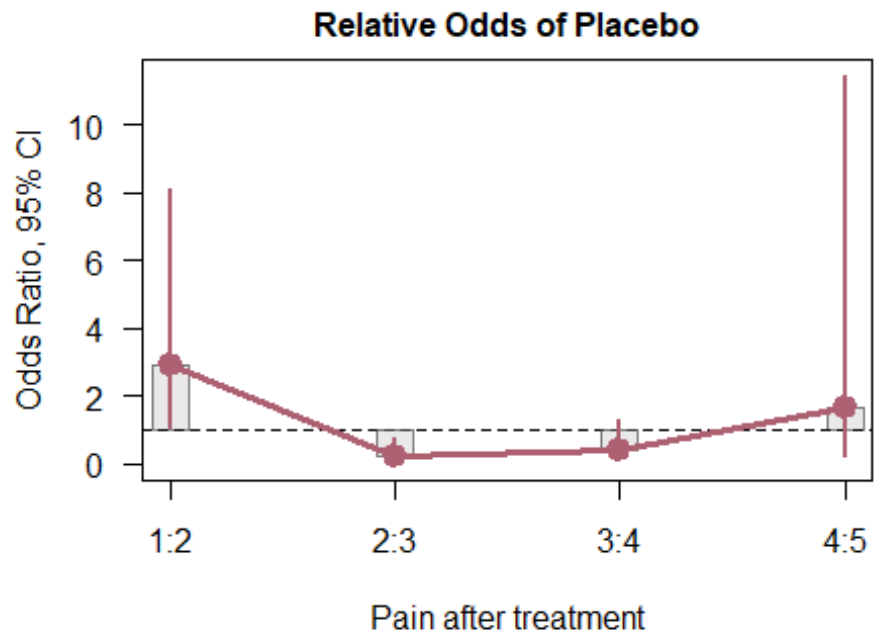
```
library(vcd)

## Warning: package 'vcd' was built under R version 4.3.1

## Loading required package: grid

T5 <- table(knee$R4, knee$Th)
odds.2cb <- oddsratio(T5, log = FALSE)
summary(odds.2cb)

##
## z test of coefficients:
##
```

```
##                         Estimate Std. Error z value Pr(>|z|)
## 1:2/Placebo:Treatment  2.90789    1.52468  1.9072  0.05649 .
## 2:3/Placebo:Treatment  0.24176    0.13799  1.7520  0.07978 .
## 3:4/Placebo:Treatment  0.38182    0.23506  1.6243  0.10430
## 4:5/Placebo:Treatment  1.66667    1.63865  1.0171  0.30911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(odds.2cb)
```

```
##                              2.5 %     97.5 %
## 1:2/Placebo:Treatment 1.04057283  8.1261509
## 2:3/Placebo:Treatment 0.07898152  0.7400092
## 3:4/Placebo:Treatment 0.11424274  1.2760997
## 4:5/Placebo:Treatment 0.24263538 11.4483623
```

```
plot(odds.2cb, main = "Relative Odds of Placebo", xlab = "Pain after
treatment", ylab = "Odds Ratio, 95% CI")
```



Kendall Rank Correlation

Kendall rank correlation is used to test the similarities in the ordering of data, especially useful when sample size is small and has many tied ranks:

```
cor.test(knee$R3, knee$R4, method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  knee$R3 and knee$R4
## z = 12.086, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.8869367
```

These are some of the essential techniques for analyzing categorical data and exploring relationships between categorical variables.