Homework Assignment #1

Dr. Wynne's Partial Solutions

2/17/2023

Key Terms (5 pts)

1. An output variable (response variable, dependent variable) is the (possibly unknown) response we are interested in modeling as a function of (typically known) one or more input variables (explanatory variables, independent variables, predictors, features).
2. Reducible error is error due to a "mismatch" between the true model $\hat{Y} = f(X)$ and the estimated model $\hat{Y} = \hat{f}(X)$. Irreducible error is further variation due to things that we cannot account for in the model, including pure random chance. For example, if I wanted to predict your grade on the final exam, I might include your two midterm scores, your Math 338 grade, and how much sleep you got the week before finals, but there is clearly some variation still that can only be explained by how well the specific problems I pick match up with what you know well/studied.
3. Generally, inference questions ask us to infer characteristics of the population (for example, the relationship between X and Y), while prediction questions ask us to predict an unknown value of the response variable. It is possible to use the same model for both inference and prediction; for example, linear regression does both.
4. In general, regression methods are used to predict values of a numerical response variable and classification methods are used to predict values of a categorical response variable.
5. The major advantage of nonparametric methods is that they can model a much wider variety of possible relationships between input and output variables without having to worry about a mismatch between the estimated form of f and the true form of f. However, the assumptions of parametric methods do a lot of the heavy lifting, meaning we typically need much more data for nonparametric methods to produce reasonably-accurate estimates and predictions.
6. The loss function the chapter describes for regression problems is MSE (mean squared error), while for classification problems it is misclassification rate (proportion of observations classified to the wrong group).
7. When we fit the model on the entire dataset, we typically will underestimate the true MSE/misclassification rate. Therefore, it is important to only fit the model on the training set so that we can more accurately evaluate the model's performance on a dataset the model has *not* seen before. We hold out the rest of the data and make predictions on that dataset, in order to get a more accurate estimate of the "true" MSE/misclassification rate on a new dataset.
8. Generally, more complex models have higher variance and lower bias. They are more flexible, so we expect the predicted values to be more-or-less correct on average, but the tradeoff is that they can be extremely sensitive to a couple of changes in the training set. Less complex models have lower variance and higher bias. They are less flexible, meaning we are likely to systematically overestimate or underestimate the true response, but they are much less sensitive to small changes in the training data.
9. Overfitting means that the model has stopped fitting the true functional form $\hat{Y} = f(X)$ and has started fitting the irreducible error ϵ. This results in the training MSE continuing to decrease but the test MSE skyrocketing, because the decrease in bias is no longer enough to offset the increase in variance.

10. A Bayes classifier assigns each observation to the class with the highest posterior probability, that is, the most likely class given the value(s) of the predictor variable(s).

Conceptual Problems

Conceptual Problem 1 (4 pts)

Your solution should include some (probably not all) of these points:

- When H_0 is true and the assumptions of the test are met, we should have a uniform distribution of p-values on [0,1].
- When H_a is true and the assumptions of the test are met, we should no longer have a uniform distribution of p-values. In fact, much smaller p-values are much more common. As the power increases, very small p-values become much more likely.
- When H_0 is true and the assumptions of the test are not met, we sometimes get a roughly uniform distribution of p-values and sometimes do not. At least for this particular test, whether we get a uniform distribution of p-values appears to depend on how badly the assumptions are violated and how big the sample is. This should perhaps not be surprising given the Central Limit Theorem.
- The distribution of p-values under H_0 doesn't really depend on the significance level. The distribution of p-values under H_a does depend on the significance level, but that's primarily because changing the significance level changes the critical value, which changes the distance between μ_0 and μ_a that we can detect with a given power.
- All we really learned about power in Math 338 was some mathematical procedures involving it; we are just now learning really what power means in general and why it's important.

Conceptual Problem 3 (3 pts)

Part (a)

The variable A_j takes the value 1 if the null hypothesis is rejected (which has probability α of happening) and the value 0 if the null hypothesis is not rejected (which has probability $1 - \alpha$ of happening). Therefore, each $A_j \sim Bernoulli(\alpha)$ (or equivalently, $B(1, \alpha)$).

Part (b)

The sum of n independent and identically distributed Bernoulli random variables is a binomial random variable with parameters n and p. Here we are summing m iid $Bernoulli(\alpha)$ random variables, so the total number of incorrectly rejected null hypotheses $V \sim B(m, \alpha)$.

Part (c)

The standard deviation of a $B(n, p)$ random variable is $\sqrt{np(1-p)}$. Plugging in our answers from part (b), we have that the standard deviation of the number of incorrectly rejected null hypotheses is $\sqrt{m\alpha(1-\alpha)}$.

Simulation Problems

Simulation Problem 1 (Code: 1.5 pts; Explanation: 3.5 pts)

A parametric method assumes a specific functional form for the model; for example, the two-sample t-test is a parametric method because it assumes that the variable is normally distributed in each population being compared. On the other hand, a nonparametric method does not assume a functional form; for example, the Mann-Whitney test is designed to work for any arbitrary continuous variable regardless of distribution.

Our simulation activity in class suggested that when the assumptions of the t-test are met, we did not see any clear difference in performance between the t-test and the Mann-Whitney test. Both tests achieved a 5%

Type I Error rate using a 5% significance level, and both had pretty terrible power at small sample sizes or small values of d.

When the assumptions of the t-test are violated due to outlier issues in each sample, the two-sample t-test is terrible. It was nearly impossible for us to reject H_0 when H_0 was true or when $d = 0.2$. However, the Mann-Whitney test performed less poorly when H_0 was true (about 2% Type I Error rate, closer to the expected 5%) and clearly had superior power to the t-test at all values of d.

When there are clear outlier issues, but the distributions in the two samples look similar enough that the assumptions of the Mann-Whitney test might be met, we should definitely use the Mann-Whitney test instead of a two-sample t-test (even at $n = 50$).

Applied Problems

```
library(ISLR2)
library(ggplot2)
library(dplyr)
```

Applied Problem 1 (Code: 6 pts; Explanation: 3 pts)

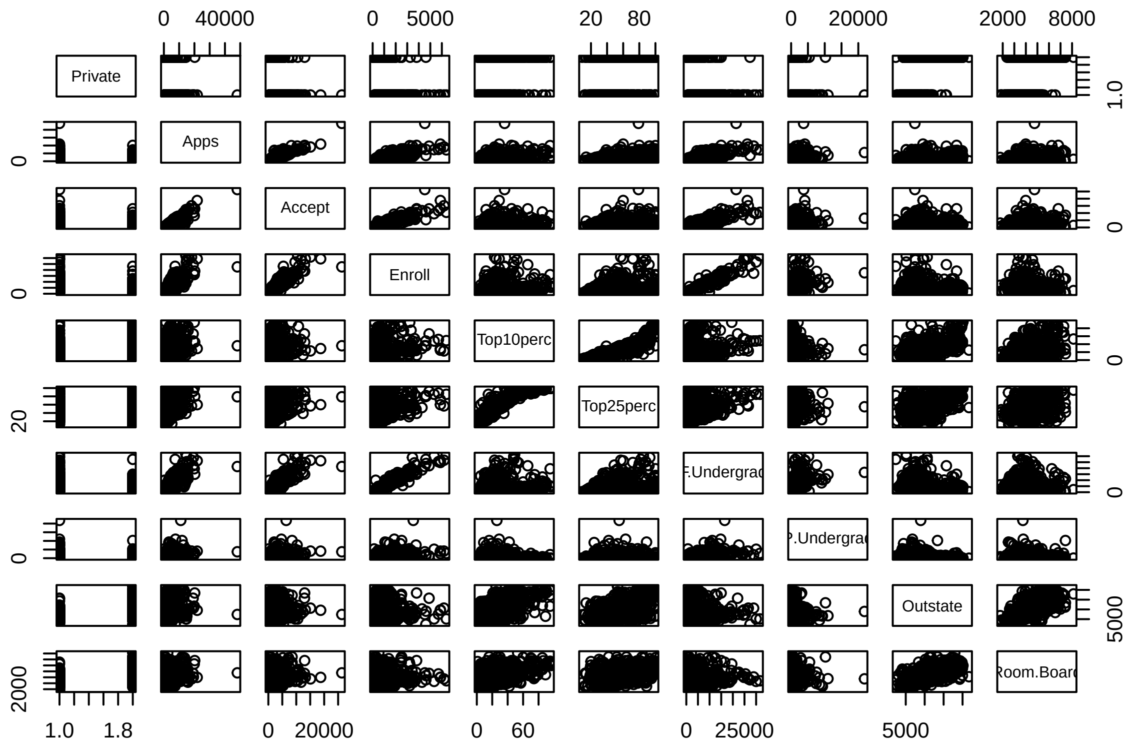
Part (c.i)

```
summary(College)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad          Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

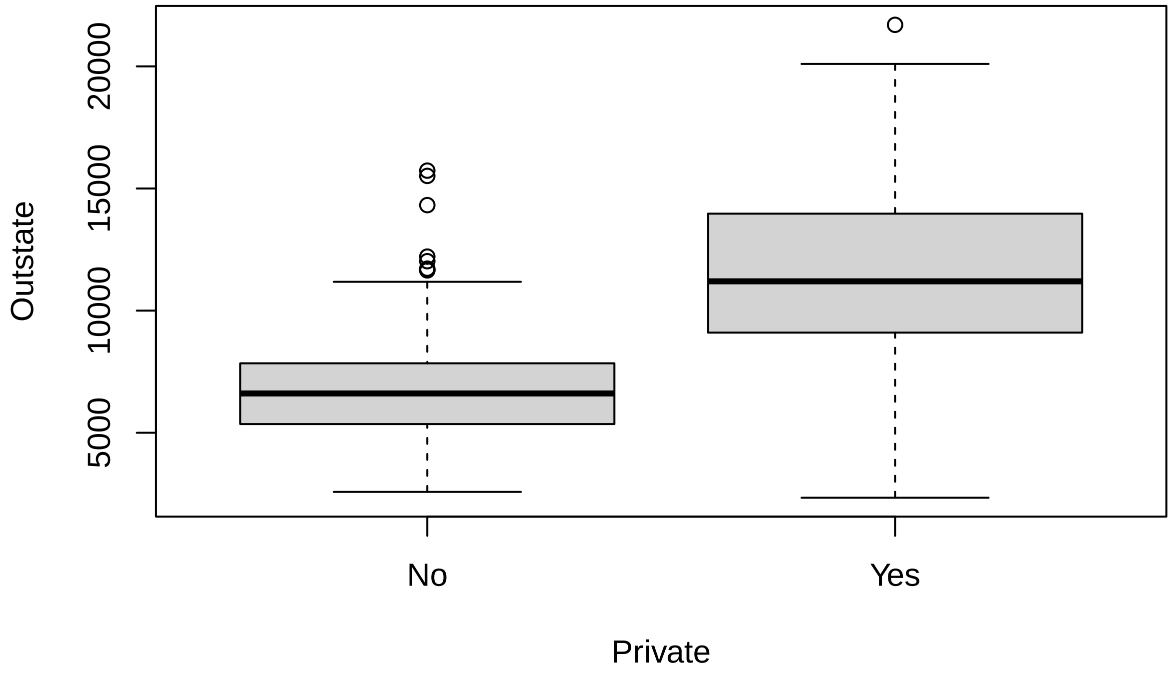
Part (c.ii)

```
pairs(College[,1:10])
```



Part (c.iii)

```
plot(Outstate ~ Private, data = College)
```



Part (c.iv)

```
library(dplyr)
College <- College %>% mutate(
  Elite = if_else(Top10perc > 50, "Yes", "No")
)
```

Part (c.vi)

For full credit, I expected you to follow the instructions to "continue exploring the data," especially by creating additional bivariate plots (blow up scatterplots from part ii and/or make more boxplots).

Here are all of the issues I found with this dataset, in rough order of how likely I think it is that you found them.

- Cazenovia College graduated 118% of its students.
- 103% of Texas A&M-Galveston's faculty have Ph.D.'s.
- The dataset was recorded in 1995, so we don't really have a good frame of reference to know whether the recorded prices (Outstate, Room.Board, etc.) are reasonable. Certainly none of us want to go to the Center for Creative Studies, which charged an estimated $2340 for books!
- Twelve other colleges have more faculty with Ph.D.'s than terminal degrees. A Ph.D. is a terminal degree.

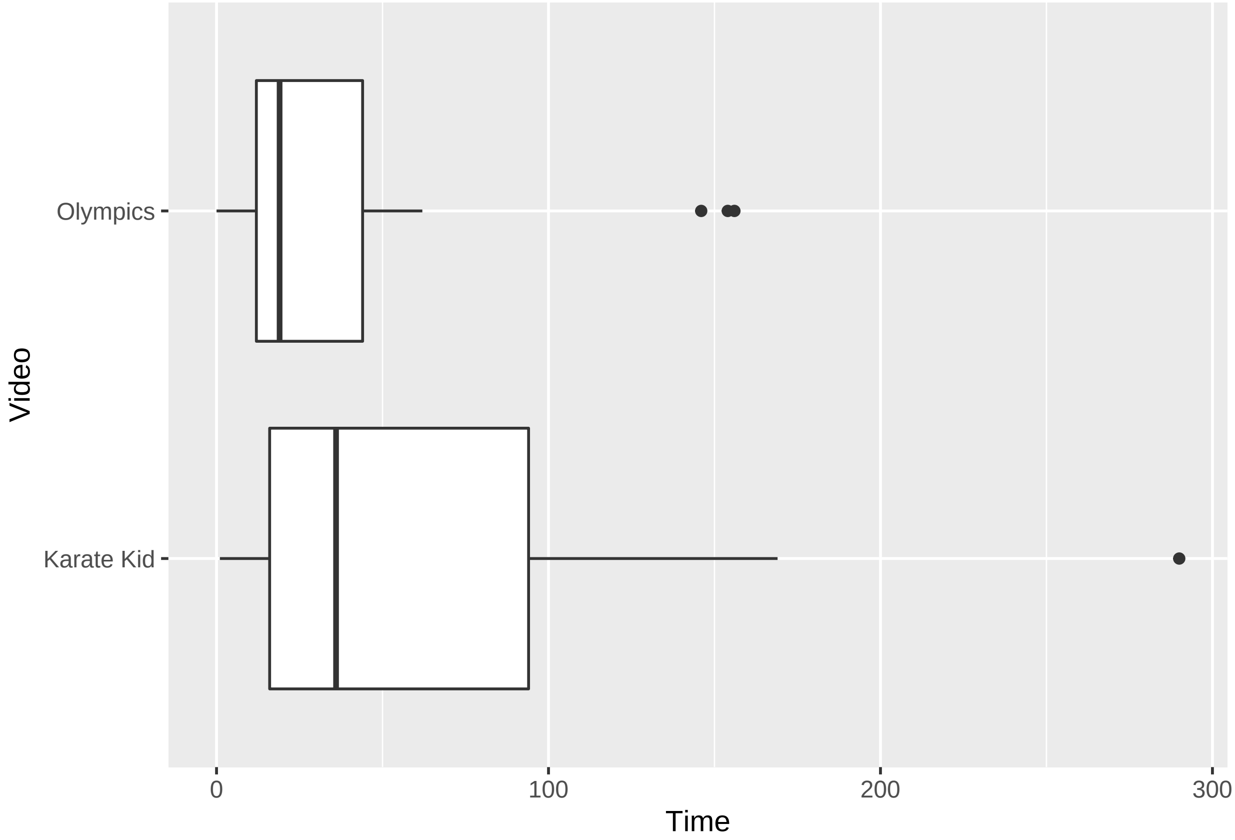
Applied Problem 2 (Code: 1 pt; Explanation: 2 pts)

```
violence <- read.csv("violence.csv", header = TRUE)
head(violence)
```

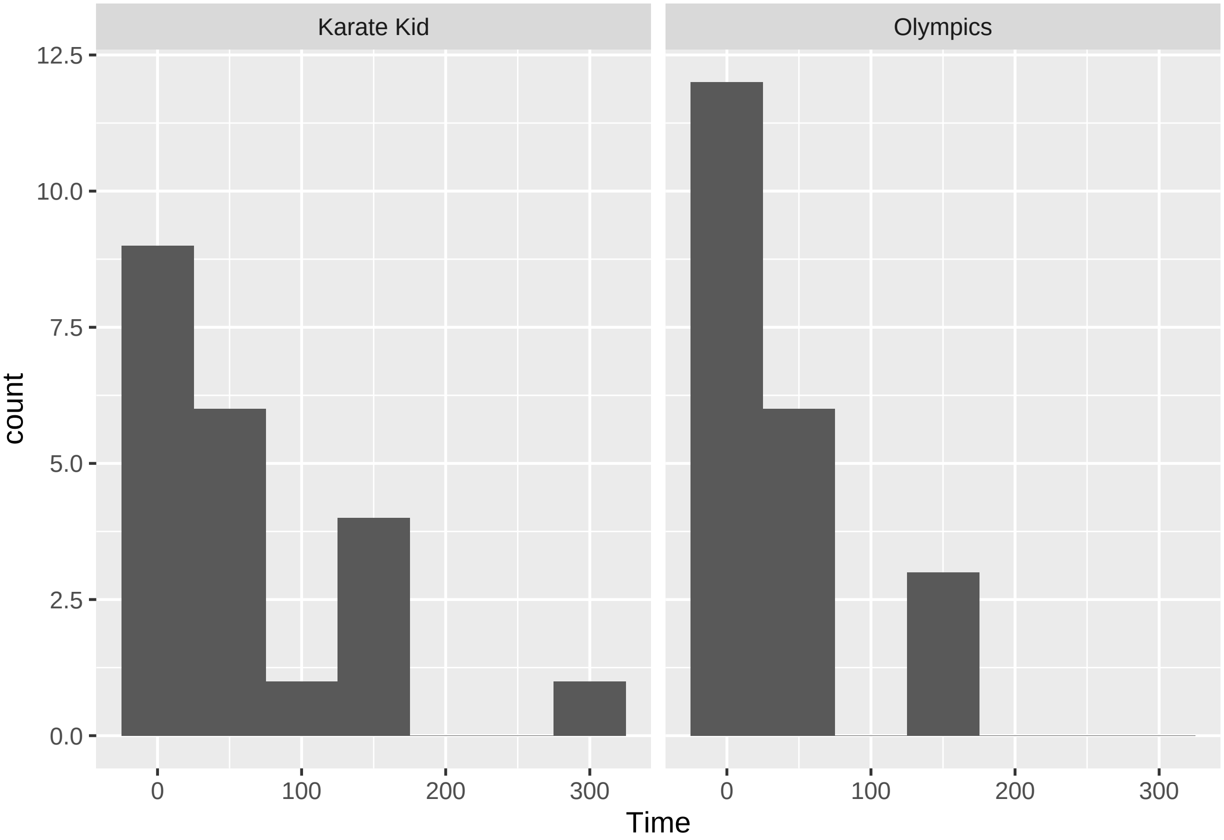
```
##      Video Time
## 1 Olympics   12
## 2 Olympics   44
## 3 Olympics   34
## 4 Olympics   14
## 5 Olympics    9
## 6 Olympics   19
```

You could do either a set of boxplots or a set of histograms to show the distribution of `Time` by `Video`.

```
ggplot(violence, aes(x = Video, y = Time)) +
  geom_boxplot() + coord_flip()
```



```
ggplot(violence, aes(x = Time)) +
  geom_histogram(center = 50, binwidth = 50) +
  facet_wrap(~Video)
```



It is apparent on both plots that the distribution of Time in each Video group is pretty heavily right-skewed with major outliers in both groups.

```
violence %>% group_by(Video) %>%
  summarize(n = n(),
            meanTime = mean(Time),
            sdTime = sd(Time))
```

```
## # A tibble: 2 x 4
##   Video          n meanTime sdTime
##   <chr>      <int>    <dbl>  <dbl>
## 1 Karate Kid    21     65.0   73.4
## 2 Olympics      21     40.2   49.2
```

With only 21 in each sample group I'd be worried about doing a t-test here. Depending on the book you *might* be able to get away with it. The standard deviations are not quite problematic.

Applied Problem 3 (Code: 1 pt; Explanation: 2 pts)

Use the permutation test function you wrote in the lab section to determine whether the research hypothesis in the previous question was supported. Be sure to follow all steps of hypothesis testing, up to and including writing a conclusion that answers the research question in context.

H_0: watching violent videos has no effect on children's tolerance of violent behavior. H_a: children who watch violent videos are more tolerant of violent behavior.

We have a numerical variable being recorded in two groups, so we should use our `permutation_t_test` function.

```r
permutation_t_test <- function(formula, data, alternative = "t", B = 9999, seed = 100){
  # formula: a formula of the form response ~ explanatory
  # data: a data frame containing the explanatory and response variables
  # alternative: the sign in the alternative hypothesis
  # B: the number of permutation resamples
  # seed: the seed to use for the permutation resampling

  t.obs <- t.test(formula = formula, data = data, var.equal = TRUE)$statistic

  t.perm <- numeric(B)

  permutation.df <- model.frame(formula = formula, data = data)

  set.seed(seed)
  for(i in 1:B){


  permutation.df[[1]] <- sample(permutation.df[[1]])

  t.perm[i] <- t.test(formula = formula,
                      data = permutation.df,
                      var.equal = TRUE)$statistic
}

  t.all <- c(t.obs, t.perm) # Finish this line
  p.left <- sum(t.all <= t.obs)/(B+1) # compute the p-value for a left-sided test
  p.right <- sum(t.all >= t.obs)/(B+1) # compute the p-value for a right-sided test

  p.value <- dplyr::case_when(alternative == "g" ~ p.right,
                              alternative == "l" ~ p.left,
                              alternative == "t" ~ 2*min(p.left, p.right),
                              TRUE ~ NaN # output NaN if alternative is anything else
)

  results <- list(
    statistic.observed = t.obs,
    statistic.simulated = t.perm,
    p.value = p.value,
    n.resamples = B,
    seed = seed
  ) # let's pack up everything we want to return to the user in one list

  invisible(results)

}
```

```r
violence_test <- permutation_t_test(Time ~ Video, data = violence,
                                    alternative = "g")
violence_test$statistic.observed
```

```
##        t
## 1.286887
```

```
violence_test$p.value
```

```
## [1] 0.1074
```

I obtained a t-statistic of 1.29 and a p-value of 0.107. Therefore, I fail to reject my null hypothesis. I do not have strong enough evidence to claim that children who watch violent videos are more tolerant of violence.