Pair Programming Module 06 Classic Hypothesis Tests

HDS

2024-08-30

Pair Programming Exercise 6 DSE5001

Hypothesis tests based on Algebraic Tests

These are “classic” tests of various hypotheses. We can’t cover them all in depth this semester, but we can survey them. These ideas crop up repeatedly throughout data science, so we need to be aware of them.

Most of these tests are about comparisons between groups.

We will use the mtcars data set

HD Sheets, August 30, 2024 checked 01/03/2025

Student Info

Your name:Ryan Waterman Your team-mates names: Trinity Tobin

Libraries

```
library("ggplot2")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

The mtcars data set

This is a nice data set as there are a variety of types of cars in it.

This is a set set from Motor Trend (MT) magazine from sometime in the mid 1970s, so it is more or less historical data

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Cyl as a factor

We want to think of cyl as factor, a group membership, so we will force it to be a factor

When calling functions, R will always treat variable as indicating group membership if they are of type factor, so it is worth converting all group memberships as factors

```
mtcars$cyl=factor(mtcars$cyl)
```

We will break the cars up into three categories based on the number of cylinders, 4,6 or 8

Use dplyrs to do this, create separate data frames for 4,6,8 cylinder cars

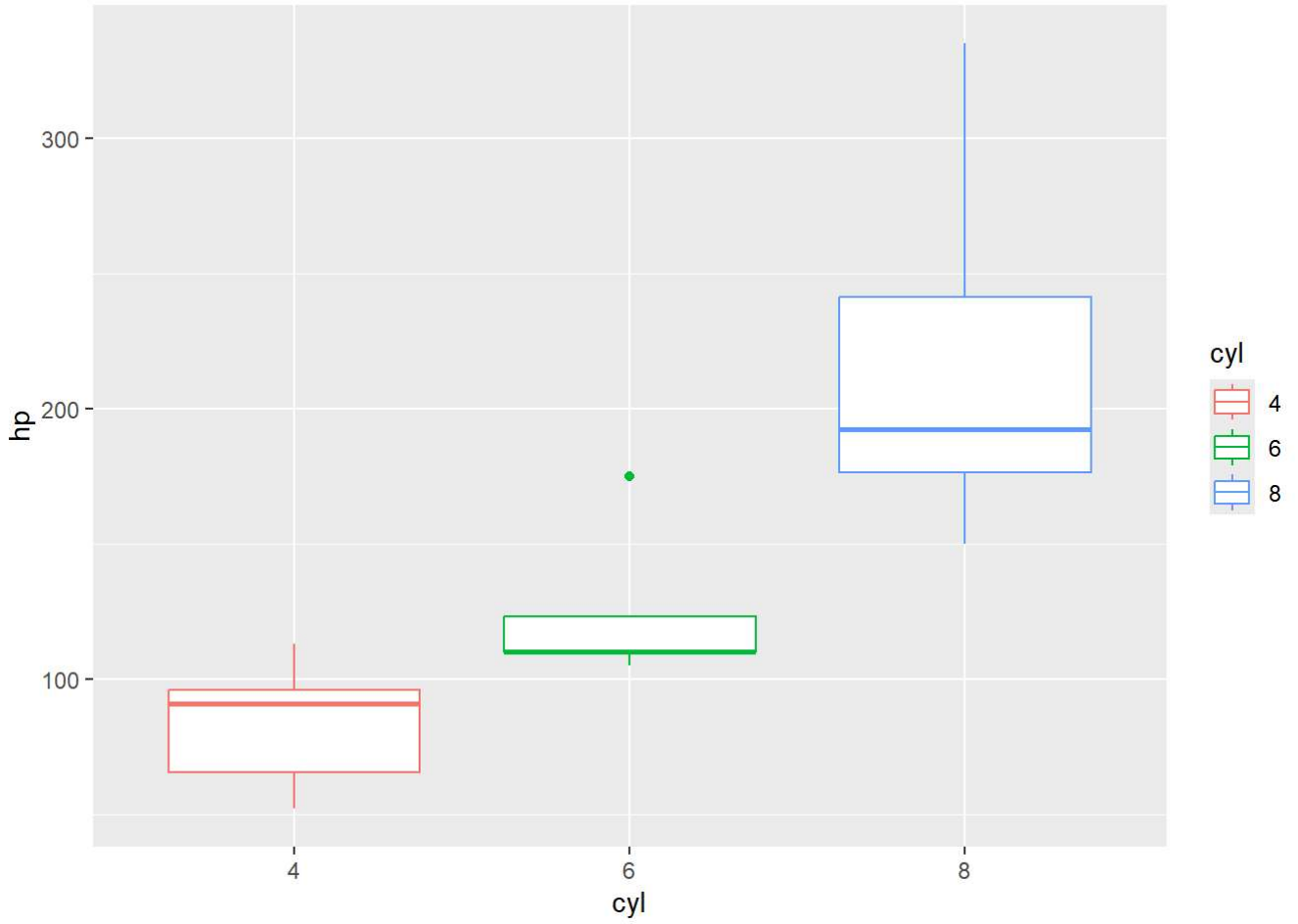
```
mtcars4<-mtcars %>% filter(cyl ==4)
mtcars6<-mtcars %>% filter(cyl==6)
mtcars8<-mtcars %>% filter(cyl==8)
```

t-test for the Difference in the means of two groups

Suppose we state the hypothesis that 6 cylinder cars on average have more power than 4 cylinder cars. This seems a fairly common sense idea

First, lets plot this, using a boxplot with color by cylinder using the whole data set

```
ggplot(mtcars,aes(y=hp,x=cyl,color=cyl))+geom_boxplot()
```



Well, the plot looks pretty convincing, there is little or now overlap in the horsepower of 4 and 6 cylinder engines

Running a formal t-test

```
# we are sending in lists of hp values for 4 and 6 cylinders
# we specify that the alternative hypothesis is "greater", meaning the first
# variable value is hypothesized to be greater than the second

t.test(x=mtcars6$hp,y=mtcars4$hp,alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars6$hp and mtcars4$hp
## t = 3.5617, df = 11.486, p-value = 0.002088
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  19.73455      Inf
## sample estimates:
## mean of x mean of y
## 122.28571  82.63636
```

The results show us the mean values of x (6 cylinder hp) and of y (4 cylinder hp)

the observed t value was 3.5617, or about 3.5 times the pooled standard deviation

with a degree of freedom of 11.49, the value is 0.002088, or about a 0.2% chance the observed difference is due to chance

Question/Action

Test the hypothesis that the mean weight of 8 cylinder cars is greater than the mean weight of 6 cylinder cars, using a t-test

```
t.test(x=mtcars8$wt,y=mtcars6$wt,alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars8$wt and mtcars6$wt
## t = 3.6212, df = 18.991, p-value = 0.0009098
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4608734       Inf
## sample estimates:
## mean of x mean of y
##  3.999214  3.117143
```

It appears that there is a 0.09% chance that the difference in mean weight by cylinder count is due to chance.

ANOVA, differences among the means of more than two groups

Let's look at the mean hp of the 3 cyl categories

```
mtcars %>% group_by(cyl) %>% summarize(mean_hp=mean(hp))
```

```
## # A tibble: 3 × 2
##   cyl   mean_hp
##   <fct>   <dbl>
## 1 4        82.6
## 2 6       122.
## 3 8       209.
```

Looks like there are meaningful difference among all three We can test this hypothesis using an ANOVA test

We are specifying a formula here, that hp is predicted by cyl

this is written as hp~cyl

while this called an ANOVA, it is actually a form of generalized linear model (GLM)

```
results_anova= aov(hp~cyl,data=mtcars)
summary(results_anova)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2 104031   52015   36.18 1.32e-08 ***
## Residuals   29  41696    1438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This test is giving us an F-statistic

The F is (variance explained by grouping by cylinder)/(average error variance)

F values are always ratios of variances, high F means the grouping explains a lot of differences in the data relative to the random fluctuation in the data

The p-value of an F ratio depends on the degrees of Freedom (DF), the DF for a factor is one less than the number of groups, so we have 3 groups and the DF is 2

Typically F values over about 2 are meaningful, at reasonable sample sizes.

Here the p-value is very low, indicating no meaningful chance the variation in hp is unrelated to the number of cylinders

Question/Action

Determine if the differences in mean weights among 4,6 and 8 cylinder carries is related to the cylinder category, using an ANOVA as above

```
results_anova= aov(wt~cyl,data=mtcars)
summary(results_anova)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  18.18   9.088   22.91 1.07e-06 ***
## Residuals   29  11.50   0.397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tests of Counts in Categories

Contingency tables and Contingency tests (Chi-square)

The mtcar data set has a category for automatic or manual transmission. the variable am = 0 for automatics and am=1 for manual

We can force am to be a factor and also set the labels names as “auto” and “manual”

```
mtcars$am=factor(mtcars$am,labels=c("auto","manual"))
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs     am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0 manual    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0 manual    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1 manual    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1   auto    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0   auto    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1   auto    3    1
```

Contingency table

Let's look at the counts of transmission type per cylinder, this is contingency table

```
mtcars %>% group_by(cyl,am) %>% summarize(n=n())
```

```
## `summarise()` has grouped output by 'cyl'. You can override using the `.groups`
## argument.
```

```
## # A tibble: 6 × 3
## # Groups:   cyl [3]
##   cyl   am         n
##   <fct> <fct>  <int>
## 1 4     auto       3
## 2 4     manual     8
## 3 6     auto       4
## 4 6     manual     3
## 5 8     auto      12
## 6 8     manual     2
```

We could also use the R function table to show the counts in each category

This is sometimes called a contingency table

```
con_table=table(mtcars$cyl,mtcars$am)
con_table
```

```
##
##     auto manual
##   4    3      8
##   6    4      3
##   8   12      2
```

We can get the sums over rows and columns

```
library("MASS")
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
addmargins(con_table)
```

```
##
##       auto manual Sum
##   4      3      8  11
##   6      4      3   7
##   8     12      2  14
##   Sum   19     13  32
```

The Null hypothesis we might state is that the number of manual vs automatic transmissions is independent of the number of cylinders. As we can see in the table, that doesn't look to be true, but the predominance of manual transmissions in 4 cylinder cares might just be chance

The *Chi-squared test* is a test against the hypothesis of equal ratios in each cell of the contingency table,

Notice that we have to feed data into the test as a contingency table

```
chisq.test(con_table)
```

```
## Warning in chisq.test(con_table): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  con_table
## X-squared = 8.7407, df = 2, p-value = 0.01265
```

The Chi-squared value is the sum of the squared departures from the expected count in each cell, if the cells are randomly distributed among auto and manual

A large Chi-square is unlikely, the p value is determined by the df in the system

The df is (number of rows-1)*(number of columns-1)

This p value is low, so the chance of transmission type not being related to the number of cylinders is very low

The *Fisher exact test* is an alternative to the Chi-Square

We got a warning that the Chi-Square might not be a correct approximation, the Chi-Square can fail at small sample sizes.

The Fisher exact test is more reliable,but hard to calculate at large sample sizes, so it is best suited to small sample sizes where the Chi-square is likely to fail

```
fisher.test(con_table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  con_table
## p-value = 0.009105
## alternative hypothesis: two.sided
```

The Fisher exact test returned a lower p-value than the Chi-squared, but the implication is the same.

Question/Action

Create a contingency table for cylinders and gears.

You will need to convert gears to a factor, you don't need labels for the gears though, the numbers are fine.

Show the table and then run the Chi-square and Fisher exact test.

Explain what it means.

```
mtcar_factor<- mtcars %>% mutate(gear_factor = as.factor(gear))

con_table_cg=table(mtcar_factor$cyl,mtcar_factor$gear_factor)
con_table_cg
```

```
##
##      3  4  5
##   4  1  8  2
##   6  2  4  1
##   8 12  0  2
```

```
chisq.test(con_table)
```

```
## Warning in chisq.test(con_table): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  con_table
## X-squared = 8.7407, df = 2, p-value = 0.01265
```

```
fisher.test(con_table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  con_table
## p-value = 0.009105
## alternative hypothesis: two.sided
```

The chi-squared test indicates a relationship between the number of factors and gears, as does the fisher test. Each result in a low p-value.

F test for equality of variance

In some tests, there is an assumption of equal variance among groups. We may need to test this, or differences in variance may be our hypothesis

High variance in manufactured parts is a bad thing, so we may be looking to see if one set of parts had lower variance than another, as an example.

We can test to see if our three groups of cars by cylinder have equal variance in mpg

```
mtcars %>% group_by(cyl) %>% summarise("Variance"=var(mpg))
```

```
## # A tibble: 3 × 2
##   cyl   Variance
##   <fct>    <dbl>
## 1 4        20.3
## 2 6         2.11
## 3 8         6.55
```

Wow, sure looks like a no on that one!

In the 1970s, there were some really poor 4 cylinder engines

Here's the test

We can only compare two groups, so we'll compare 4 and 6, omitting 8

```
mtcars46=mtcars[mtcars$cyl!=8,]
```

we have the formula mpg predicted by cyl, but we are testing variances, not means

In the var test we can only have two groups, so i used cyl==4, to compare 4 cylinders vs 6s an d8s

```
var.test(mpg~cyl, data=mtcars46)
```

```
##
##  F test to compare two variances
##
## data:  mpg by cyl
## F = 9.6261, num df = 10, denom df = 6, p-value = 0.01182
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   1.762592 39.198688
## sample estimates:
## ratio of variances
##           9.626086
```

F is a ratio of variances, so high F means inequality

the p value depends on both df values, here p=0.01182 for the null, so we can reject the null

Question/Action

Run an F-test for equality of variance on 6 cylinder vs 8 cylinder cars

```
mtcars %>% group_by(cyl) %>% summarise("Variance"=var(mpg))
```

```
## # A tibble: 3 × 2
##   cyl   Variance
##   <fct>    <dbl>
## 1 4        20.3
## 2 6         2.11
## 3 8         6.55
```

```
mtcars68=mtcars[mtcars$cyl!=4,]
```

```
var.test(mpg~cyl, data=mtcars68)
```

```
##
##  F test to compare two variances
##
## data:  mpg by cyl
## F = 0.32238, num df = 6, denom df = 13, p-value = 0.1728
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.08944544 1.71799232
## sample estimates:
## ratio of variances
##          0.3223843
```