Abdulaziz Alqumayzi G200007615 Module6_CT3

Rudra S Bandhu

11/13/2021

```
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

library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(ggplot2)
library(multcomp)

## Loading required package: mvtnorm

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:caret':
##
##     cluster

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'MASS'

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

```
##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser

library(Metrics)

##
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':
##
##     precision, recall

library(stats)
library(Metrics)
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

library(jmv)
library(readxl)
```

Import datasets

```
problem_1 <- read_excel("problems.xlsx", sheet = "problem 1")
problem_3 <- read_excel("problems.xlsx", sheet = "problem 3")
problem_4 <- read_excel("problems.xlsx", sheet = "problem 4")
```

Problem 1

```
# Code for problem 1

# Apply Anova test
model_p1 <- aov(test ~ temp, data = problem_1)
summary(model_p1)

##             Df Sum Sq Mean Sq F value Pr(>F)
## temp         2   50.4   25.19   1.332  0.289
## Residuals   18  340.3   18.91

print(model_p1)

## Call:
##    aov(formula = test ~ temp, data = problem_1)
```

```
##
## Terms:
##                      temp Residuals
## Sum of Squares   50.3810  340.2857
## Deg. of Freedom        2        18
##
## Residual standard error: 4.347961
## Estimated effects may be unbalanced
```

Problem 1 Questions and Answers

a Specify an appropriate null hypothesis

H0: The null hypothesis is that the mean of test at different temperatures is equal

b Test the hypothesis that the polymer performs equally well at all three temperatures at 5 percent level of significance

At 5 percent level of significance. The p-value (0.289) is greater than 0.05 so we fail to reject the null hypothesis.

c Test the hypothesis that the polymer performs equally well at all three temperatures at 1 percent level of significance

At 1 percent level of significance. The p-value (0.289) is greater than 0.1 so we fail to reject the null hypothesis.

d State your conclusion from the analysis.

We can conclude that the ability of the polymer to remove toxic wastes from water tests has no different at three different temperatures.

Problem 2

```
# Code for problem 2

gmean <- (32+40+30)/3 # grand mean
SSTR<-(12*((32-gmean)^2+(40-gmean)^2+(30-gmean)^2)) # Sum of Squares
treatment
MSTR<-SSTR/(3-1)
SSe<-(12-1)*(33+44+40) # Sum of Squares error
MSE<-SSe/(36-3)
F<-MSTR/MSE # f-test
P <- 1-pf(8.6154,2,33) # p-value
q<-3.48
MOE<-3.48*sqrt(MSE/12)

cat(paste("F-test = ",round(F, digits = 5), " p-value = ", round(P, digits =
5),'\n\n',
          'CI of A-B: (',round(32-40-MOE, digits = 5),round(32-40+MOE,digits
= 5),')\n',
      'CI of A-C: (',round(32-30-MOE, digits = 5),round(32-30+MOE,digits =
```

```
5),')\n',
      'CI of B-C: (',round(40-30-MOE, digits = 5),round(40-30+MOE,digits =
5),')'))

## F-test =  8.61538  p-value =  0.00098
##
##  CI of A-B: ( -14.27366 -1.72634 )
##  CI of A-C: ( -4.27366 8.27366 )
##  CI of B-C: ( 3.72634 16.27366 )
```

Problem 2 Questions and Answers

a Test the hypothesis that the mean time to clear a mild asthmatic attack is the same for all three steroids. Use the 5 percent level of significance.

At 5 percent level of significance. The p-value (0.00098) is less than 0.05 so we have significant evidence to reject the null hypothesis and accept the alternative.

b Find confidence intervals for all differences of means ( ) that, with 95 percent confidence, are valid.

At 95% confidence:

Confidence interval of A-B: (-14.27366 -1.72634) Confidence interval of A-C: (-4.27366 8.27366) Confidence interval of B-C: (3.72634 16.27366)

Problem 3

```
# Code for problem 3

#Apply Anova test
model_p3 <- aov(Yield ~ Area, data = problem_3)
summary(model_p3)

##             Df Sum Sq Mean Sq F value Pr(>F)
## Area         2  16.20   8.100    5.12  0.013 *
## Residuals   27  42.72   1.582
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

print(model_p3)

## Call:
##    aov(formula = Yield ~ Area, data = problem_3)
##
## Terms:
##                     Area Residuals
## Sum of Squares  16.20067  42.71800
## Deg. of Freedom        2        27
##
## Residual standard error: 1.257835
## Estimated effects may be unbalanced
```

Problem 3 Questions and Answers

a State your null hypothesis.

H0: The null hypothesis is that the mean of planting snow peas in different regions is equal.

b What analysis strategy

One-way independent ANOVA Test to compare several means when those means have come from different groups.

c Calculate the test statistic and p-value.

F-statistic value = 5.12 and P-value = 0.013

d Calculate the residual error.

Residual standard error = 1.257835

e State your conclusion.

The p-value (0.013) is less than 0.05 at a 5 percent level of significance. So we have sufficient evidence to reject the null.

We can conclude that planting snow peas in different regions are different in different regions.

f What approach can you take to reduce the residual error. Discuss at least 2 ideas and justify

1- Reduce variability: The less that your data varies, the more precisely you can estimate a population parameter.That's because reducing the variability of your data decreases the standard deviation and, thus, the margin of error for the estimate.

2- Elimination of confounds: Unmeasured variables confound the results. If any variables are known to impact the dependent variable being assessed, ANCOVA is an excellent way to eliminate their bias. Once a potential confounding variable is identified, it may be tested and included as a covariate in the analysis.

3- Increase the sample size: Often, the most practical way to decrease the margin of error is to increase the sample size. Usually, the more observations that you have, the narrower the interval around the sample statistic is. Thus, you can often collect more data to obtain a more precise estimate of a population parameter.

Problem 4

```
# Code for problem 4


problem_4$Area <- as.factor(problem_4$Area)

model_p4_1 <- ancova(data = problem_4, dep =  Yield, factors = Area, covs =
```

```
'Height(inches)',postHoc = ~ Area, modelTest = T )
print(model_p4_1)

##
##  ANCOVA
##
##  ANCOVA - Yield
##  -------------------------------------------------------------------------
----------
##                        Sum of Squares    df    Mean Square    F            p
##  -------------------------------------------------------------------------
----------
##    Overall model           4.092615     3      1.3642050    4.0455540
0.0173929
##    Area                    1.542927     2      0.7714634    0.4993501
0.6126274
##    Height(inches)          2.549688     1      2.5496883    1.6503530
0.2102391
##    Residuals              40.168312    26      1.5449351
##  -------------------------------------------------------------------------
----------
##
##
##  POST HOC TESTS
##
##  Post Hoc Comparisons - Area
##  -------------------------------------------------------------------------
-------------------
##    Area         Area    Mean Difference    SE           df           t
p-tukey
##  -------------------------------------------------------------------------
-------------------
##    FS      -    PS           -0.5848511    0.6044905    26.00000    -
0.9675109    0.6034443
##            -    SH           -0.1780560    1.0130785    26.00000    -
0.1757574    0.9831254
##    PS      -    SH            0.4067951    1.2186513    26.00000
0.3338076    0.9405663
##  -------------------------------------------------------------------------
-------------------
##    Note. Comparisons are based on estimated marginal means

cat(paste("New Residual error = ",round(sqrt(40.168312/26), digits = 5),
      "\nOld Residual error = ", round(1.257835, digits = 5)))

## New Residual error =  1.24295
## Old Residual error =  1.25784
```

a What analysis strategy would you use to take advantage of the additional data you have

The Analysis of Covariance (ANCOVA) is used to compare means of an outcome variable between two or more groups taking into accountvariability of other variables, called covariates.

b What is the dependence of the yield on the height for each of the areas? Does it differ significantly between the 3 areas. What can you conclude (max 10 sentences)

We can see the p-value of dependency between Yield on height is 0.2102391. Which indicates that there is a weak dependency.

Code Results:

Area Area Mean Difference SE df t p-tukey

--- FS - PS -0.5848511 0.6044905 26.00000 -0.9675109 0.6034443
- SH -0.1780560 1.0130785 26.00000 -0.1757574 0.9831254
PS - SH 0.4067951 1.2186513 26.00000 0.3338076 0.9405663

We can see the all p-values of the test are less than 0.05 from the code results. Which indicate that there are nor deffirence between the dependences for each area.

c Calculate the test statistic and p-value.

```
                    Sum of Squares    df    Mean Square    F            p
```

--- Overall model 4.092615 3 1.3642050 4.0455540 0.0173929

We can see from the Overall model results that:

F-statistic value = 4.0455540 and P-value = 0.0173929

d Calculate the residual error and compare that to what you found in problem 2.

In general, the smaller the residual standard deviation/error, the better the model fits the data.

New model residual error = 1.24295 Old model residual error = 1.25784

We can see that the residuals are slightly near to each other. But in this new model, there is an indication that the new model does improve slightly which gives us more confidence in our decision in the previous model.

e Specify your final model.

We used ancova() function for analysis of covariance to test the model that contains one dependent variable, one factor variable, and one covariate variable.

The code used is the following:

ancova(data = problem_4, dep = Yield, factors = Area, covs = ‘Height(inches)’,postHoc = ~ Area, modelTest = T)

modelTest parameter used to show the overall model result.

The code result is the following:

ANCOVA

ANCOVA - Yield

	Sum of Squares	df	Mean Square	F	p
Overall model	4.092615	3	1.3642050	4.0455540	0.0173929
Area	1.542927	2	0.7714634	0.4993501	0.6126274
Height(inches)	2.549688	1	2.5496883	1.6503530	0.2102391
Residuals	40.168312	26	1.5449351		

POST HOC TESTS

Post Hoc Comparisons - Area

Area		Area	Mean Difference	SE	df	t	p-tukey
FS	-	PS	-0.5848511	0.6044905	26.00000	-0.9675109	0.6034443
	-	SH	-0.1780560	1.0130785	26.00000	-0.1757574	0.9831254
PS	-	SH	0.4067951	1.2186513	26.00000	0.3338076	0.9405663

The residual error was calculated using the following code:

sqrt(40.168312/26)

ƒ State your conclusion

As we tested in the previous problem 3. The p-value of model 3 is 0.013 and the p-value of model 4 is 0.017. Both models indicate that we have sufficient evidence to reject the null and accept the alternative. Which was we can conclude that planting snow peas in different regions is different in different regions. Also, the decreased residual gives us more confidence in our previous decision.

References:

Field, A. P., Miles, J., & Field Zoë. (2017). Discovering statistics using R. W. Ross MacDonald School Resource Services Library.

Choueiry, G.(2021)Residual Standard Deviation/Error: Guide for Beginners. Retrieved November 13, 2021, from https://quantifyinghealth.com/residual-standard-deviation-error/

finnstats (2021) How to perform ANCOVA in R. Retrieved November 13, 2021, from https://www.r-bloggers.com/2021/07/how-to-perform-ancova-in-r/