Chapter 12: Analysis of Variance

These notes reflect material from our text, Exploring the Practice of Statistics, by Moore, McCabe, and Craig, published by Freeman, 2014.

ANOVA

One-way ANOVA

Which means differ, and by how much? Fisher method

Multiple comparisons : Bonferroni method, Tukey method

ANOVA table for comparing means and ANOVA table for regression

Two-way ANOVA

Main effects, interaction

Outline for inference : theoretical, simulation, conclusion in context

One-way ANOVA : Being put on hold

Agresti and Franklin (AFS, 3e, section 14.1) describe an experiment to evaluate customer’s tolerance of being put on hold when they called in to a busy airline reservation system. While the customer waited on the line, either an advertisement, or Muzak, or classical music was played in the hope of keeping the customer on the line until an attendant became available. Which type of recording is the most successful in keeping waiting customers on the line?

The data records the number of minutes that the customer stayed on the line under one of the three conditions. Import this data into R.

```
# data : on hold
advertisement <- c(5, 1, 11, 2, 8)
muzak <- c(0, 1, 4, 6, 3)
classical <- c(13, 9, 8, 15, 7)
time <- c(advertisement, muzak, classical)
tag <- c(rep("advertisement", 5), rep("muzak", 5), rep("classical", 5))
hold <- data.frame(time, tag)
hold
```
Time on hold

Summarize the data.

```r
# stats
hold.level <- c("advertisement", "muzak", "classical")
hold.mean <- sapply(list(advertisement, muzak, classical), mean)
hold.sd <- sapply(list(advertisement, muzak, classical), sd)

hold.stats <- data.frame(hold.level, hold.mean, hold.sd)
colnames(hold.stats) <- c("level", "mean", "s")
hold.stats
#  level mean        s
# 1 advertisement 5.4 4.159327
# 2         muzak  2.8 2.387467
# 3     classical 10.4 3.435113
```

```
advertisement: ● ● ● ● ●
Muzak: ● ● ● ● ●
classical: ● ● ● ● ●
```

Linear model

Construct a linear model.

```r
hold.lm <- lm(time ~ tag, data=hold)
summary(hold.lm)

# Call:
# lm(formula = time ~ tag, data = hold)

# Residuals:
#   Min 1Q Median 3Q Max
#  -4.4 -2.6  -0.4  2.6  5.6

# Coefficients:
# Estimate  Std. Error   t value  Pr(>|t|)
# (Intercept)  5.400    1.523     3.545  0.00403 **
# tagclassical 5.000    2.154     2.321  0.03868 *
# tagmuzak    -2.600    2.154    -1.207  0.25068
```
Analysis of variance

Use an $F$ test to explore whether the evidence in this data supports the null hypothesis that there is no difference in the mean times that clients will remain on the line, based on the recordings they hear.

```r
anova(hold.lm)
```

The test statistic is the $F$ value in the row labelled \texttt{tag}. The associated $p$-value is 0.01264, which is lower than our customary significance level of 0.05, so we reject the null hypothesis that all of the means are the same and conclude that at least one type of recording is more effective than others in determining the time that a customer remains on the line.

$F$ distribution

The $F$ distribution that we just used has 2 and 12 degrees of freedom, so $F \sim F(df_1 = 2, df_2 = 12)$, and $P(F \geq 6.431) = 0.01264$. 

![F(df1=2, df2=12)](image.png)
Estimating differences of means

Once the $F$ test has established that the group means are not all the same, the next question becomes “Which group means are different?”

The strip chart suggests a difference in the times that customers would stay on the line while listening to classical music or to an advertisement. Let $\mu_1$ denote the mean population time that customers would stay on the line while listening to classical music, and let $\mu_2$ denote the mean population time that customers would stay on the line while listening to an advertisement. Our point estimate for the difference $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$, which we calculate from the sample. The formula for a 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE,$$

where $t^*$ can be calculated with the R command

$$\text{qt}(0.975, df = 12)$$

and

$$SE = RSE \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $RSE$ is the residual standard error reported in the summary of the linear model, and $n_1$ and $n_2$ are the sizes of the two groups. Therefore, the required confidence interval is

$$\bar{x}_1 - \bar{x}_2 \pm t^* \times SE = (10.4 - 5.4) \pm 2.179(3.406)\sqrt{1/5 + 1/5} \rightarrow [0.306524, 9.693476].$$

We are 95% confident that the difference in the times that customers will remain on the line while listening to classical music is between 0.3 and 9.7 minutes longer than the time that they will stay on the line listening to an advertisement. The fact that this confidence interval comes so close to including 0 among the plausible values should make us hesitant to rely too strongly on this result.

```r
# boxplot
boxplot(time ~ tag, data=hold,
  col=c("lightseagreen", "lightslateblue", "cadetblue"),
  ylab="Time (min)", main="Holding Time")
```

```
Holding Time

advertisement classical muzak
0 5 10 15
Time (min)
```
A function for computing a confidence interval for the difference of two means

We can write a function for computing a confidence interval for the difference of two means. This simplifies and automates the discussion in the previous section. Here we are using the same 95% significance level, regardless of the (possibly large) number of pairs of means that we might be working with, so this is known as the Fisher method.

\[
\bar{y}_i - \bar{y}_j \pm t_{0.975} \times RSE \times \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}
\]

```r
# CI for difference of means (Fisher Method)
# note : rse and its df are reported in the summary of the linear model
ci.difference.of.means <- function(y.bar.i, n.i, y.bar.j, n.j, rse, df){
t975 <- qt(0.975, df=df)
ci <- (y.bar.i - y.bar.j) + t975 * rse * sqrt(1/n.i + 1/n.j) * c(-1, 1)
return(ci)
}
```

Agresti and Franklin consider a survey which relates happiness levels to the number of close friends reported by the participant (AFS, 3e, section 14.2, pp.692-693). There are three levels of happiness, so there are three pairs of means (of number of close friends) to consider. We can use our function to compute all three confidence intervals. How should we interpret the results?

```r
# AFS 3e, example 5, p.692
y.bar.1 <- 10.4
y.bar.2 <- 7.4
y.bar.3 <- 8.3
n.1 <- 276
n.2 <- 468
n.3 <- 87
rse <- sqrt(234.2)
df <- 828
# compare very happy with pretty happy
ci.difference.of.means(y.bar.1, n.1, y.bar.2, n.2, rse, df)
# 0.7202573 5.2797427
# compare very happy with not too happy
.ci.difference.of.means(y.bar.1, n.1, y.bar.3, n.3, rse, df)
# -1.593312 5.793312
# compare pretty happy with not too happy
ci.difference.of.means(y.bar.2, n.2, y.bar.3, n.3, rse, df)
# -4.40704 2.60704
```
Multiple Comparisons - Fisher method

In the Fisher method, used for constructing the above confidence intervals, a fixed significance level applies to each particular confidence interval, regardless of how many other confidence intervals we might be constructing. But as the number of comparisons increases, the likelihood that at least one of the intervals in that increasingly large set does not contain its true difference of means also increases. This is the problem of multiple comparisons.

Think of the following analogous problem: The probability that each individual person will make a mistake is 5%, but you have 100 people working on your problem. What is the probability that at least one person in the group of 100 makes a mistake? Answer: $1 - 0.95^{100} = 0.994$. So the likelihood that at least one member of your group makes a mistake is 99.4%.

The solution to the problem of multiple comparisons is to design a comparison method for which the level of significance applies to the entire set of comparisons being made, not just to each individual comparison. One such solution is called the Tukey Honest Significant Differences method, and it is implemented in R by the procedure TukeyHSD. Tukey’s confidence intervals will be wider than the intervals obtained by the Fisher method, because the 95% confidence level must apply to the entire set of comparisons, not just to each individual comparison of means.

Let’s apply both methods to the study of callers being put on hold. First, use the Fisher method by applying the above confidence interval function to each pair of differences. Interpret these three Fisher confidence intervals.

```r
# CIs for difference of means (on hold : Fisher method)

y.bar.1 <- 2.8 # muzak
y.bar.2 <- 10.4 # classical
y.bar.3 <- 5.4 # advertisement

n.1 <- 5
n.2 <- 5
n.3 <- 5

rse <- sqrt(3.406)
df <- 12

# muzak - classical

CI.difference.of.means(y.bar.1, n.1, y.bar.2, n.2, rse, df)
# -10.143152 -5.056848

# muzak - advertisement

CI.difference.of.means(y.bar.1, n.1, y.bar.3, n.3, rse, df)
# -5.14315155 -0.05684845

# classical - advertisement

CI.difference.of.means(y.bar.2, n.2, y.bar.3, n.3, rse, df)
# 2.456848 7.543152
```
Multiple Comparisons - TUKEY HSD method

Now construct the three Tukey HSD confidence intervals. They are certainly wider than the corresponding Fisher intervals. Do you reach the same conclusions with the Tukey confidence intervals as you did with the Fisher confidence intervals?

```r
# CIs for difference of means (on hold : Tukey HSD)

hold.aov <- aov(time ~ tag, data=hold)

TukeyHSD(hold.aov)
# Tukey multiple comparisons of means
# 95% family-wise confidence level

# Fit: aov(formula = time ~ tag, data = hold)

# $tag
# # diff lwr upr p adj
# classical-advertisement 5.0 -0.7467542 10.746754 0.0910409
# muzak-advertisement -2.6 -8.3467542 3.146754 0.4715973
# muzak-classical -7.6 -13.3467542 -1.853246 0.0107397

plot(TukeyHSD(hold.aov),
     col="springgreen4")
```

Differences in mean levels of tag

95% family-wise confidence level
Analyzing Association

Associations involve explanatory variables and response variables. Order them like this:

explanatory × response.

categorical × categorical  (EPS, chapter 9)

r × c contingency table, test for independence
1 × c contingency table, goodness of fit

Test for independence or goodness of fit with a \( \chi^2 \) test statistic

quantitative × quantitative  (EPS, chapters 10, 11)

Linear model for the population

\[
\mu_y = \beta_0 + \beta_1 x + \beta_2 x
\]

Linear model describing the sample

\[
\hat{y} = b_0 + b_1 x + b_2 x
\]

Test for relevance of the model with an \( F \) test statistic.

\[
H_0 : \text{all } \beta_i \text{'s are 0}
\]

Estimate the parameters \( \beta_i \) with \( t \) statistics and confidence intervals.

(quantitative and categorical) × quantitative

Subsume this case into the previous one with indicator variables.

categorical × quantitative  (EPS, chapter 12)

The categorical variable divides quantitative measurements into groups, and the question becomes one of comparing the mean responses of the groups.

Test that all of the means are the same with an \( F \) test (ANOVA)

\[
H_0 : \beta_1 = \cdots = \beta_g
\]

Find which means are different with \( t \) tests and confidence intervals for \( \beta_i - \beta_j \)

Control the significance level for multiple comparisons with Tukey HSD

quantitative × categorical  (EPS, chapter 14)

Use quantitative variables to predict a categorical variable with logistic regression
Exercises

We will attempt to solve some of the following exercises as a community project in class today. Finish these solutions as homework exercises, write them up carefully and clearly, and hand them in at the beginning of class next Friday.

Homework 12a – ANOVA

*Exercises from Chapter 12 exercises:*
  12.9 (example), 12.10 (p.value), 12.11 (p.value), 12.13 (ANOVA models)

Homework 12b – the ANOVA table

*Exercises from Chapter 12 exercises:*
  12.15 (ANOVA models), 12.38 (poets), 2.43 (bones), 12.44 (bones),

Homework 12c – multiple comparisons

*Exercises from Chapter 12 exercises:*
  12.45 (cooking pots), 12.46 (cooking pots), 12.47 (scaffolds), 12.48 (scaffolds)