PSTAT 120C HW 3

TJ Sipin

2022-08-26

Reading

Define the ANOVA procedure in your own words. How does the model detect a difference in means by comparing variances?

The ANOVA procedure aims to explain and assign the amount of variation of responses to each independent variable in order to identify which ones are important and how they affect the response.

Given two treatments/populations and $n_1 = n_2$, we can find the SST:

$$SST = n_1 \sum_{i=1}^{2} (\bar{Y}_i - \bar{Y})^2 = \frac{n_1}{2}(\bar{Y}_1 - \bar{Y}_2)^2.$$

If SST is large, then we know that $|\bar{Y}_1 - \bar{Y}_2|$ is large, giving evidence that a difference in means is large just by comparing variances.

Write equations for the following in a general one-way ANOVA:

The total sum of squares:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$$

The sum of squares for treatment

$$SS_{Treatment} = \sum_{i=1}^{k} n_k (\bar{Y}_i - \bar{Y})^2$$

The sum of squares for error

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2$$

What are the null and alternative hypotheses for a one-way ANOVA?

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \iff H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$$
$$H_\alpha : \mu_i \neq \mu_j \text{ for some } i \neq j \iff \tau_i \neq 0 \text{ for some } i \in [1, k]$$

What assumptions should be met when we conduct an ANOVA F-test?

- Independent random samples to be selected from the k populations.
- The k populations are normally distributed with equal variances σ^2 and means $\mu_1, \mu_2, \ldots, \mu_k$.

Write the general statistical model for a one-way ANOVA.

For $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$,

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

where

- Y_{ij} is the jth observation from population (treatment) i
- μ is the overall mean
- τ_i is the nonrandom effect of treatment i, where $\sum_{i=1}^{k} \tau_i = 0$
- ϵ_{ij} are the random error terms such that ϵ_{ij} are independent normally distributed random variables with $\mathbb{E}(\epsilon_{ij}) = 0$ and $\text{var}(\epsilon_{ij}) = \sigma^2$.

Write the general statistical model for a two-way ANOVA.

For $i = 1, \ldots, k$ and $j = 1, \ldots, b$,

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij},$$

- Y_{ij} is the observation on treatment i in block j
- μ is the overall mean
- τ_i is the nonrandom effect of treatment i, where $\sum_{i=1}^{k} \tau_i = 0$

- β_j is the nonrandom effect of block j, where $\sum_{j=1}^{b} \beta_j = 0$
- ϵ_{ij} are the random error terms such that ϵ_{ij} are independent normally distributed random variables with $\mathbb{E}(\epsilon_{ij}) = 0$ and $\text{var}(\epsilon_{ij}) = \sigma^2$.

Practice

1. The Florida Game and Fish Commission desires to compare the amounts of residue from three chemicals found in the brain tissue of brown pelicans. Independent random samples of ten pelicans each yielded the accompanying results (measurements in parts per million). Is there evidence of sufficient differences among the mean residue amounts at the 5% level of significance?

		Chemical	
Statistic	DDE	DDD	DDT
Mean	0.032	0.022	0.041
Standard deviation	0.014	0.008	0.017

```r
statistics = data_frame(S = c(0.014, 0.008, 0.017),
                        mu = c(0.032, 0.022, 0.041))

SSE = 0
for(i in 1:3){
  SSE = (10 - 1) * as.numeric(statistics[i, 'S']^2)
}
SST = 0
for(i in 1:3){
  SST = SST + (10 * as.numeric(statistics[i, 'mu']))
}
SST = SST / 30
MSE = SSE / (30 - 3)
MST = SST / (3 - 1)


F = MST / MSE
F; qf(0.05, 2, 27, lower.tail = FALSE)
```

```
## [1] 164.3599
```

```
## [1] 3.354131
```

Since $F = \frac{MST}{MSE} = \frac{0.0158}{0.0000963} = 164.3599 > F_\alpha = 3.354$, we reject the null hypothesis and conclude that there is sufficient evidence of significant difference among the mean residue amounts at the 5% level of significance.

2. It has been hypothesized that treatments (after casting) of a plastic used in optic lenses will improve wear. Four different treatments are to be tested. To determine whether any differences in mean wear exist among treatments, 28 casting from a single formulation of the plastic were made and 7 castings were randomly assigned to each of the treatments. Wear was determined by measuring the increase in "haze" after 200 cycles of abrasion (better wear being indicated by smaller increases). The data collected are reported in the accompanying table.

```r
optic = data_frame(A = c(9.16, 13.29, 12.07, 11.97,
                         13.31, 12.32, 13.06),
                   B = c(11.95, 15.15, 14.75, 14.79,
                         15.48, 13.47, 13.06),
                   C = c(11.47, 9.54, 11.26, 13.66,
                         11.18, 15.03, 14.86),
                   D = c(11.35, 8.73, 10.00, 9.75,
                         11.71, 12.45, 12.38))

optic %>%
  kbl() %>%
  kable_classic(full_width = F, html_font = 'Cambria') %>%
  add_header_above(c("Treatment" = 4))
```

Treatment			
A	B	C	D
9.16	11.95	11.47	11.35
13.29	15.15	9.54	8.73
12.07	14.75	11.26	10.00
11.97	14.79	13.66	9.75
13.31	15.48	11.18	11.71
12.32	13.47	15.03	12.45
13.06	13.06	14.86	12.38

a. Is there evidence of a difference in mean wear among the four treatments? Use a $\alpha = 0.05$.

```r
means = sapply(optic, mean)
optic_mean = optic %>% as.matrix() %>% mean()
CM = 28 * optic_mean^2
SST = 0
for(i in colnames(optic)) {
  SST = SST + sum(optic[i])^2 / 7
}
SST = SST - CM
SS_Total = 0
for (i in optic) {
  for (j in i) {
    SS_Total = SS_Total + j^2
  }
}
SS_Total = SS_Total - CM
SSE = SS_Total - SST

MST = SST / (4  - 1)
MSE = SSE / (28 - 4)

F = MST / MSE; F
```

```
## [1] 4.706571
```

```r
qf(0.05, 3, 24, lower.tail = FALSE)
```

```
## [1] 3.008787
```

Since $F = 4.71 > F_\alpha = F_{0.05} = 3.01$, we reject the null hypothesis and conclude that there is sufficient evidence of a difference in mean wear among the four treatments.

b. Estimate the mean difference in haze increase between treatments B and C using a 99% confidence interval.

We use the formula below to estimate the mean difference in haze increase between treatments B and C using 99% confidence interval:

$$(14.09 - 12.43) \pm t_{\alpha/2, \nu=(n-k)} S \sqrt{\frac{1}{7} + \frac{1}{7}},$$

where $S = \sqrt{MSE} = 1.60$ and $t_{\alpha/2, \nu=(n-k)} = t_{0.005, \nu=(28-4)} = 2.80$

```r
sqrt(MSE)
```

```
## [1] 1.596349
```

```r
qt(0.005, 24, lower.tail = FALSE)
```

```
## [1] 2.79694
```

So our confidence interval is obtained from the following:

```r
(14.09 - 12.43) - qt(0.005, 24, lower.tail = FALSE) * sqrt(MSE) * sqrt(2/7)
```

```
## [1] -0.7265848
```

```r
(14.09 - 12.43) + qt(0.005, 24, lower.tail = FALSE) * sqrt(MSE) * sqrt(2/7)
```

```
## [1] 4.046585
```

So with 99% confidence, we estimate the mean difference in haze increase between treatments B and C to be $(-0.7265848, 4.046585)$.

c. Find a 90% confidence interval for the mean wear for lenses receiving treatment A.

We use the formula below to estimate the mean wear for lenses receiving treatment A:

$$12.17 \pm t_{0.05} \frac{1.60}{\sqrt{7}}$$

```r
12.16857 - qt(0.05, 24, lower.tail = FALSE) * (sqrt(MSE))/(sqrt(7))
```

```
## [1] 11.13629
```

```
12.16857 + qt(0.05, 24, lower.tail = FALSE) * (sqrt(MSE))/(sqrt(7))
```

```
## [1] 13.20085
```

A 90% confidence interval for the mean wear for lenses receiving treatment A is $(11.13629, 13.20085)$.

3. Fill in the blanks in the following two-way ANOVA table, using the information provided:

Source	SS	df	MS	F
Block		7		14.5
Treatment	5797.5		1932.5	
Interaction	11363.1			
Error	14841.6		154.6	
Total		127		

$$F_{Block} = 14.5 = \frac{MSB}{MSE} = \frac{MSB}{154.6} \implies MSB = 2241.7$$
$$MSB = \frac{SSB}{7} \implies SSB = 15691.9$$
$$MS_{Treatments} = 1932.5 = \frac{SS_{Treatments}}{Treatments_{df}} = \frac{5797.5}{Treatments_{df}} \implies Treatments_{df} = 3$$
$$F_{Treatment} = \frac{MS_{Treatment}}{MSE} \implies F_{Treatment} = 12.5$$
$$MSE = 154.6 = \frac{SSE}{Error_{df}} = \frac{14841.6}{Error_{df}} \implies Error_{df} = 96$$
$$Interaction_{df} = Total_{df} - (Block_{df} + Treatment_{df} + Error_{df}) = 127 - (7 + 3 + 96) \implies Interaction_{df} = 21$$
$$MS_{Interaction} = \frac{SS_{Interaction}}{Interaction_{df}} \implies MS_{Interaction} = 541.1$$
$$F_{Interaction} = \frac{MS_{Interaction}}{MSE} \implies F_{Interaction} = 3.5$$
$$SS_{Total} = SSB + SS_{Treatment} + SS_{Interaction} + SSE \implies SS_{Total} = 47694.1$$

Source	SS	df	MS	F
Block	15691.9	7	2241.7	14.5
Treatment	5797.5	3	1932.5	12.5
Interaction	11363.1	21	541.1	3.5
Error	14841.6	96	154.6	
Total	47694.1	127		

```r
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE, eval = T)

library(dplyr)
library(knitr)
library(kableExtra)
statistics = data_frame(S = c(0.014, 0.008, 0.017),
                        mu = c(0.032, 0.022, 0.041))

SSE = 0
for(i in 1:3){
  SSE = (10 - 1) * as.numeric(statistics[i, 'S']^2)
}
SST = 0
for(i in 1:3){
  SST = SST + (10 * as.numeric(statistics[i, 'mu']))
}
SST = SST / 30
MSE = SSE / (30 - 3)
MST = SST / (3 - 1)


F = MST / MSE
F; qf(0.05, 2, 27, lower.tail = FALSE)
optic = data_frame(A = c(9.16, 13.29, 12.07, 11.97,
                         13.31, 12.32, 13.06),
                   B = c(11.95, 15.15, 14.75, 14.79,
                         15.48, 13.47, 13.06),
                   C = c(11.47, 9.54, 11.26, 13.66,
                         11.18, 15.03, 14.86),
                   D = c(11.35, 8.73, 10.00, 9.75,
                         11.71, 12.45, 12.38))

optic %>%
  kbl() %>%
  kable_classic(full_width = F, html_font = 'Cambria') %>%
  add_header_above(c("Treatment" = 4))
means = sapply(optic, mean)
optic_mean = optic %>% as.matrix() %>% mean()
CM = 28 * optic_mean^2
SST = 0
for(i in colnames(optic)) {
  SST = SST + sum(optic[i])^2 / 7
}
SST = SST - CM
SS_Total = 0
for (i in optic) {
  for (j in i) {
    SS_Total = SS_Total + j^2
  }
}
SS_Total = SS_Total - CM
SSE = SS_Total - SST

MST = SST / (4  - 1)
MSE = SSE / (28 - 4)

F = MST / MSE; F
qf(0.05, 3, 24, lower.tail = FALSE)
sqrt(MSE)
qt(0.005, 24, lower.tail = FALSE)
(14.09 - 12.43) - qt(0.005, 24, lower.tail = FALSE) * sqrt(MSE) * sqrt(2/7)
(14.09 - 12.43) + qt(0.005, 24, lower.tail = FALSE) * sqrt(MSE) * sqrt(2/7)
12.16857 - qt(0.05, 24, lower.tail = FALSE) * (sqrt(MSE))/(sqrt(7))
12.16857 + qt(0.05, 24, lower.tail = FALSE) * (sqrt(MSE))/(sqrt(7))
```