Testing for Linear Combinations in Spock Example

David Gerard

2018-12-07

Objectives

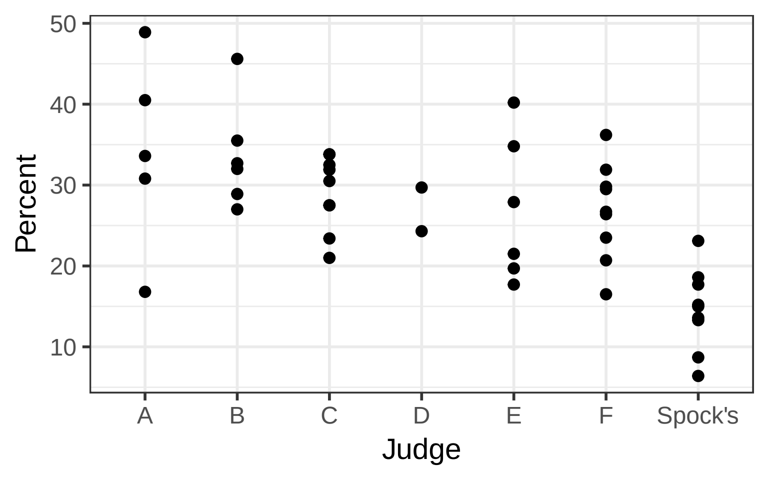
- Demonstrate how to interact with linear combinations of means in R.
- Analyze the Spock trial data in R.

Load in Data

```r
library(Sleuth3)
library(ggplot2)
data("case0502")
```

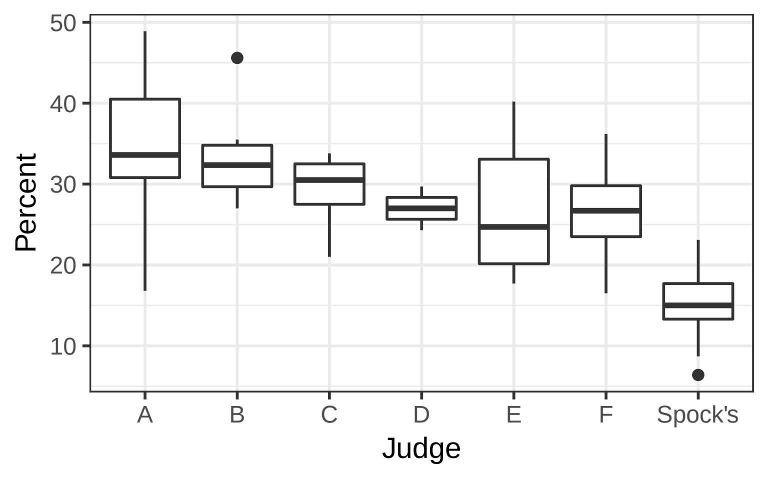
Spock EDA

```
qplot(Judge, Percent, data = case0502)
```



Spock EDA

```
qplot(Judge, Percent, data = case0502, geom = "boxplot")
```



Before Fitting

- Always make sure the grouping variable (the explanatory variable) is a "`factor`" with the `class()` function.

```
class(case0502$Judge)
```

```
## [1] "factor"
```

- Things will go wrong if this is any other type (even `logical` or `character` in the case of linear hypotheses)
- You can force a variable to be a `factor` with the `as.factor()` function:

```
case0502$Judge <- as.factor(case0502$Judge)
```

Fit the full model

- Use `aov()` function (for **A**nalysis **O**f **V**ariance) to fit the model that assumes μ_1, μ_2, ..., μ_I are all *different*.
- Always save this output.
- The **response** variable goes on the left of the tilde (~) and the **explanatory** variable goes to the right of the tilde.

Fit the full model

```
aout_alldiff <- aov(Percent ~ Judge, data = case0502)
aout_alldiff
```

```
## Call:
##    aov(formula = Percent ~ Judge, data = case0502)
##
## Terms:
##                  Judge Residuals
## Sum of Squares    1927      1864
## Deg. of Freedom      6        39
##
## Residual standard error: 6.914
## Estimated effects may be unbalanced
```

Wrong Parameterization

```r
coef(aout_alldiff)
```

```
## (Intercept)      JudgeB      JudgeC      JudgeD
##     34.1200     -0.5033     -5.0200     -7.1200
##      JudgeF JudgeSpock's
##     -7.3200    -19.4978
```

- Returns estimates of μ, δ_2, δ_3, δ_4, δ_5, δ_6, δ_7 in model where group i has mean $\mu + \delta_i$.

Do this to get correct parameterization

- **Subtracting 1** is notation for removing an "intercept column" and gives you the parameterization you expect.

```
aout_alldiff <- aov(Percent ~ Judge - 1, data = case0502)
coef(aout_alldiff)
```

```
##         JudgeA        JudgeB        JudgeC        JudgeD
##          34.12         33.62         29.10         27.00
##         JudgeF JudgeSpock's
##          26.80         14.62
```

- Returns estimates of μ_1, μ_2, μ_3, μ_4, μ_5, μ_6, μ_7 in model where group i has mean μ_i.
- **Never do this unless you are testing for linear combinations**
- The above code will **ruin** an anova table (when you call `anova()` or `summary()`).

Linear Combinations

- The book says "A computer can produce the averages and the pooled estimate of variability, but hand calculations are usually required from there."
- But they don't know that there is *almost always* an R package that can do what you want (not true for any other statistical software).
- We will use the `linearHypothesis()` function from the `car` packages to run a test for a general linear combination of means.

```r
install.packages("car")
```

```r
library(car)
```

Look at order of factors

- Use the `levels()` command to see what the order of the factors is.

```
levels(case0502$Judge)
```

```
## [1] "A"         "B"         "C"         "D"         "E"         "F
```

Set up coefficient vector

- The order is `("A", "B", "C", "D", "E", "F", "Spock's")`
- The following coefficient vector will test against

$$H_0: \text{ Spock's} - \frac{1}{6}A - \frac{1}{6}B - \frac{1}{6}C - \frac{1}{6}D - \frac{1}{6}E - \frac{1}{6}F = 0$$

```
combo_vec1 <- c(-1/6, -1/6, -1/6, -1/6, -1/6, -1/6, 1)
```

- The following will test against

$$H_0: \frac{1}{3}A + \frac{1}{3}B + \frac{1}{3}C - \frac{1}{3}D - \frac{1}{3}E - \frac{1}{3}F = 0$$

```
combo_vec2 <- c(1/3, 1/3, 1/3, -1/3, -1/3, -1/3, 0)
```

Run the hypothesis test

```
lhout <- linearHypothesis(model = aout_alldiff,
                          hypothesis.matrix = combo_vec1)
lhout
```

```
## Linear hypothesis test
##
## Hypothesis:
## - 0.166666666666667 JudgeA - 0.166666666666667 JudgeB -
##
## Model 1: restricted model
## Model 2: Percent ~ Judge - 1
##
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1     40 3401
## 2     39 1864  1      1537 32.1 1.5e-06
```

Show equivalent to t-test

- $\hat{\gamma}$ turns out to be -14.9783
- $SE(\hat{\gamma})$ turns out to be 2.6418
- So the t-statistic is $\hat{\gamma}/SE(\hat{\gamma})$ = -5.6697
- We compare this to a $t_{n-l} = t_{46-7} = t_{39}$

```
2 * pt(-5.67, df = 39)
```

```
## [1] 1.488e-06
```

```
lhout$`Pr(>F)`
```

```
## [1]          NA 1.489e-06
```

Correspondance

- The t-statistic is the square root of the F-statistic
- t-stat is -5.67

```r
sqrt(lhout$F)
```

```
## [1]   NA 5.67
```