

Tutorial 7

Swarag

2023-10-09

Problem

- ▶ What if we have more than two groups/ samples? How will we compare?
- ▶ We can do a pairwise comparison, but, it will increase the probability of a type I error.

Analysis Of Variance(ANOVA)

- ▶ We can compare the means of more than two groups.
- ▶ Testing for variation among groups is equivalent to asking whether the means differ.
- ▶ We will discuss one-way or single-factor analysis of variance, which investigates the means of several groups differing by one explanatory variable or factor.
- ▶ It works by assessing whether samples chosen from different groups are, on average, more different than individuals chosen from the same group.

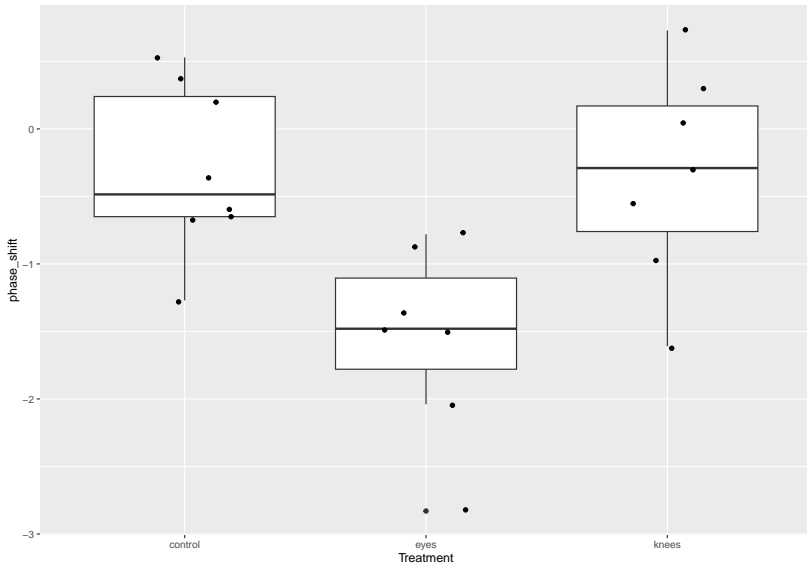
The knees who say night

Traveling to a different time zone can cause jet lag, but people adjust as the schedule of light to their eyes in the new time zone gradually resets their internal, circadian clock. This change in their internal clock is called a phase shift. Campbell and Murphy (1998) reported that the human circadian clock can also be reset by exposing the back of the knee to light, a finding met with skepticism by some, but hailed as a major discovery by others. Aspects of the experimental design were subsequently challenged. Wright and Czeisler (2002) re-examined the phenomenon.

Continued ...

The new experiment measured circadian rhythm by the daily cycle of melatonin production in 22 people randomly assigned to one of three light treatments. Participants were awakened from sleep and subjected to a single three-hour episode of bright lights applied to the eyes only, to the knees only, or to neither (the control group). Effects of treatment on the circadian rhythm were measured two days later by the magnitude of phase shift in each participant's daily cycle of melatonin production. A negative measurement indicates a delay in melatonin production, which is the predicted effect of light treatment; a positive number indicates an advance. Does light treatment affect phase shift?

Data



Hypotheses

- ▶ The null hypothesis is that the mean(μ_i) is same for all the groups.
- ▶ Here, $\mu_{control} = \mu_{knees} = \mu_{eyes}$
- ▶ Under the null hypothesis, the sample mean, \bar{Y}_i will differ only because of random sampling error.
- ▶ The alternative hypothesis is that the mean phase shift is not the same in all three light-treatment populations.

Hypotheses

- ▶ $H_0 : \mu_{control} = \mu_{knees} = \mu_{eyes}$
- ▶ $H_A : \text{At least one } \mu_i \text{ is different from the others}$

Important Note

- ▶ The alternative hypothesis does not state that every mean is different from all the others, but only that at least one mean stands apart.
- ▶ Rejecting H_0 in ANOVA is evidence that the mean of at least one group is different from the others.

How ANOVA deals with sampling error?

- ▶ Even if all the groups in a study had the same true mean, the data would likely show a different sample mean for each group - sampling error.
- ▶ The key insight of ANOVA is that we can estimate how much variation among group means ought to be present from sampling error alone if the null hypothesis is true.
- ▶ ANOVA lets us determine whether there is more variance among the sample means than we would expect by chance alone.
- ▶ If so, then we can infer that there are real differences among the population means.

Group Mean Square(MS_{Group})

- ▶ Proportional to the observed amount of variation among the group sample means.
- ▶ Representing the variation among the sampled subjects that belong to different groups.

Error Mean Square(MS_{error})

- ▶ Estimates the variance among subjects that belong to the same group.
- ▶ It is analogous to the pooled sample variance in two-sample comparisons.

What to expect under H_0

- ▶ Individuals belonging to different groups will on average be no more different from one another than individuals belonging to the same groups.
- ▶ The group mean square(MS_{Group}) and the error mean square(MS_{error}) should be equal (except by chance).
- ▶ But if the null hypothesis is false, we expect the group mean square to exceed the error mean square.
- ▶ In this case the variation among individuals belonging to different groups is expected to be greater than the variation among subjects belonging to the same group.

The F ratio

$$F = \frac{\text{group mean square}}{\text{error mean square}} = \frac{MS_{Group}}{MS_{error}}$$

The F ratio

- ▶ Let $Y_{i,j}$ be the j^{th} element in the i^{th} group
- ▶ Mean of the i^{th} group be \bar{Y}_i
- ▶ The grand mean across all mean be \bar{Y}
- ▶ Now, we can split the deviation of each observation from grand mean as following
- ▶ $Y_{i,j} - \bar{Y} = (Y_{i,j} - \bar{Y}_i) + (\bar{Y}_i - \bar{Y})$

The F ratio

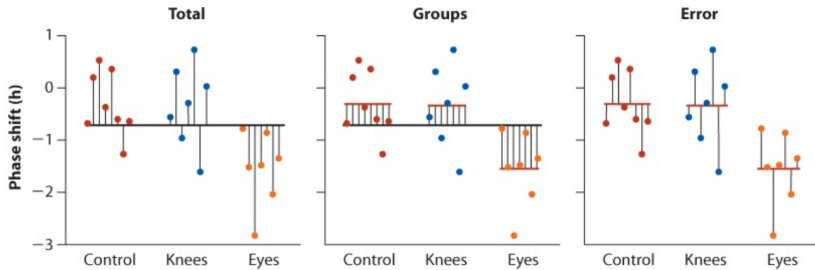


Figure 1: Picture showing MS_{total} , MS_{Group} and, MS_{error}

Computing Sum of Squares

- ▶ $SS_{total} = SS_{error} + SS_{group}$
- ▶ $SS_{group} = \sum_i n_i (\bar{Y}_i - \bar{Y})^2$
- ▶ $SS_{error} = \sum_i \sum_j (Y_{i,j} - \bar{Y}_i)^2$
- ▶ $SS_{total} = \sum_i \sum_j (Y_{i,j} - \bar{Y})^2$
- ▶ $\bar{Y} = \frac{\sum_i n_i \bar{Y}_i}{\sum_i n_i}$

Group Mean Square(MS_{group})

- ▶ $MS_{group} = \frac{SS_{group}}{df_{group}}$
- ▶ $df_{group} = k - 1$, where k is the number of groups

The **group mean square** of ANOVA represents variation among the sampled individuals belonging to different groups. It will on average be similar to the error mean square if population means are equal.

Error Mean Square(MS_{error})

- ▶ $MS_{error} = \frac{SS_{error}}{df_{error}}$
- ▶ $df_{error} = \sum_i (n_i - 1) = N - k$, where k is the number of groups

The **error mean square** of ANOVA is the pooled sample variance, a measure of the variation among individuals within the same groups.

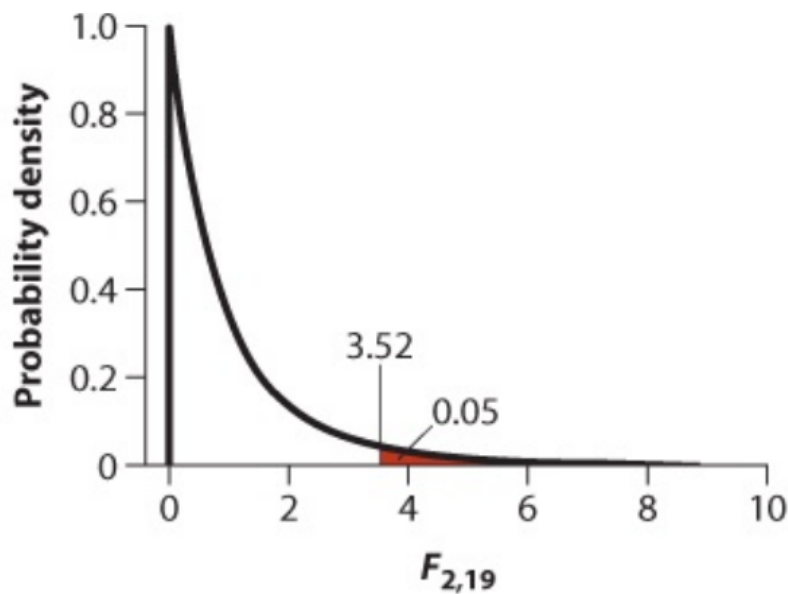
The Variance Ratio(F)

- ▶ $F = \frac{MS_{Group}}{MS_{error}}$
- ▶ Under the null hypothesis, F will on average lie close to one, differing from it only because of sampling variation in the numerator and denominator.
- ▶ If the null hypothesis is false, however, and the alternative hypothesis is correct, then MS_{groups} should exceed MS_{error} and we expect F to be greater than one.

Calculating P value

- ▶ To calculate the P-value, we need the sampling distribution for the F-statistic under H_0 .
- ▶ This null distribution for the F-statistic is called the F-distribution.
- ▶ The F-distribution has a pair of degrees of freedom, one for the numerator (top) of the F-ratio and a second for the denominator (bottom).
- ▶ We need F distribution with $k - 1$ and $N - k$ degree of freedoms - $F_{k-1, N-k}$

F distribution and P value



Assumptions of ANOVA

The assumptions of analysis of variance are the same as those of the two-sample t-test, but they must hold for all k groups.

- ▶ The measurements in every group represent a random sample from the corresponding population.
- ▶ The variable is normally distributed in each of the k populations.
- ▶ The variance is the same in all k populations.

Robustness of ANOVA

- ▶ The ANOVA is surprisingly robust to deviations from the assumption of normality, particularly when the sample sizes are large.
- ▶ This robustness stems from a property of sample means described by the central limit theorem.
- ▶ ANOVA is also robust to departures from the assumption of equal variance in the k populations, but only if the samples are all large, about the same size, and if there is no more than about a tenfold difference among the variances.

aov function in R, in stats package

First, let's apply aov

```
mdl <- aov(phase_shift ~ Treatment , data)
```

Let's see the results

```
summary mdl)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Treatment      2  7.224    3.612    7.289 0.00447 **
## Residuals     19  9.415    0.496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```