Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA)

How would we analyze the results of a clinical trial that randomly assigns not two but *three* treatments to a sample of patients?

Two of the treatments might be different medications and the third a placebo control.

For example, are both medications better than the placebo?

If they are, then by how much?

Is one medication superior to the other?

If so, how much better is it?

How do we compare the means of the three groups?

The best solution is the **analysis of variance,** or **ANOVA,** which compares means of multiple groups simultaneously in a single analysis.

Analysis of *variance* might seem like a misnomer, given our intention to compare *means,* but testing for variation among groups is equivalent to asking whether the means differ.

Analysis of Variance (ANOVA)

The analysis of variance
Analysis of variance is the most powerful approach known for simultaneously testing whether the means of k groups are equal.

It works by assessing whether individuals chosen from different groups are, on average, more different than individuals chosen from the same group.

Hypotheses
H0: $\mu_1 = \mu_2 = \mu_3$.
HA: At least one μ_i is different from the others.

Rejecting H0 in ANOVA is evidence that the mean of at least one group is different from the others.

Analysis of Variance (ANOVA)

ANOVA in a nutshell

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.

This is because of sampling error—the chance difference between a sample estimate and the true value of a population parameter caused by random sampling.

Thus, we expect to see variation among sample means taken from different groups even when the null hypothesis is true and the all groups have the same mean.

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.

Analysis of Variance (ANOVA)

The two measures of variation that are calculated from the data and compared in a test of the null hypothesis.

The *group mean square* (MS_{groups}) is proportional to the observed amount of variation among the group sample means. This quantity represents the variation among the sampled subjects that belong to different groups.

The *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.

Under the null hypothesis that the true means of groups do not differ, 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 and the error mean square 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.

Analysis of Variance (ANOVA)

The comparison of mean squares is done with an *F*-ratio.

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

If the null hypothesis is true, and the means do not differ, the group and error mean squares on average will be similar and *F* should be close to 1.

If the null hypothesis is false, the real differences among group means should inflate the group mean square and *F* is expected to exceed 1.

Analysis of Variance (ANOVA)

Partitioning the sum of squares
First one needs to separate the two sources of variation in the data: within groups and among groups.

Let *Yij:* be jth individual of the *ith* group. Then, the deviation between each observation *Yij* and the grand mean $\bar{Y}$ (the mean of all the observations) can be split into its two parts,

$$Y_{ij} - \bar{Y} = \left(Y_{ij} - \bar{Y}_i\right) + (\bar{Y}_i - \bar{Y})$$

The first part of the split, $\left(Y_{ij} - \bar{Y}_i\right)$, is the deviation between each observation and its group mean. The second part of the split, $(\bar{Y}_i - \bar{Y})$, is the deviation between the mean of the group to which the observation belongs and the grand mean.

Analysis of Variance (ANOVA)

Repeating this split for all observations, and then squaring the deviations and taking their sum, partitions the total variation in the data set into its within- and among-group components,

$$ss_{total} = \sum_i \sum_j \left(Y_{ij} - \bar{Y} \right)^2 = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_i \right)^2 + \sum_i n_i (\bar{Y}_i - \bar{Y})^2$$

where n_i is the number of observations in group i.

The term on the left of the equation is the *total sum of squares,* symbolized as SS_{total}.

The two parts on the right of the equal sign are the *error sum of squares,* SS_{error}, and the *group sum of squares,* SS_{groups}.

Thus,

$$SS_{total} = SS_{error} + Ss_{groups}$$

Analysis of Variance (ANOVA)

The grand mean $\bar{Y}$ (the mean of all the data from all groups combined):

$$\bar{Y} = \frac{\sum_i n_i \bar{Y}_i}{N}$$

where N is the total sample size, $N = \Sigma n_i$. Note: The grand mean is not the same as the average of the group sample means if the sample size is not the same in every group.

The group sum of squares is then calculated as

$$SS_{groups} = \sum_i n_i (\bar{Y}_i - \bar{Y})^2$$

The error sum of squares can be calculated as

$$SS_{error} = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i s_i^2 (n_i - 1)$$

where s_i is the sample standard deviations from group i.

Analysis of Variance (ANOVA)

Calculating the mean squares

The *group mean square,* MS_{groups}, is calculated from the deviations of group sample means ($\bar{Y_i}$) around the grand mean of all the measurements ($\bar{Y}$) :

$$MS_{groups} = \frac{SS_{groups}}{df_{groups}}$$

df_{groups}=k−1, where *k* is the number of groups.

Note: 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.

Analysis of Variance (ANOVA)

The *error mean square,* MS_{error}, measures variance within groups. ANOVA assumes that σ^2 (i.e., the variance of Y) is the same in every population.

In other words, we assume that $\sigma^2 = \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ for all k groups. The best estimate of this variance within groups is the pooled sample variance, just as in the two-sample t-test.

In ANOVA, the pooled sample variance is called the error mean square, and it is calculated as

$$MS_{error} = \frac{SS_{error}}{df_{error}}$$

df_{error} is just the sum of the degrees of freedom for the different groups:

$$df_{error} = \sum_i (n_i - 1) = N - k$$

Analysis of Variance (ANOVA)

The variance ratio, *F*
Under the null hypothesis that the population means of all groups are the same, the variation among individuals belonging to different groups (MS_{groups}) will on average be the same as the variation among individuals belonging to the same group (MS_{error}). In ANOVA, therefore, we test for a difference by calculating the ratio of Msgroups over MS_{error}:

$$F = \frac{MS_{groups}}{MS_{error}}$$

This *F*-ratio is the test statistic in analysis of variance. 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.

To calculate the *P*-value, we need the sampling distribution for the *F*-statistic under H0.

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). The numerator (MS_{groups}) has $k - 1$ degrees of freedom, and the denominator (MS_{error}) has $N - k$ degrees of freedom.

Analysis of Variance (ANOVA)

Variation explained: R^2
The R^2 **value** is used in ANOVA to summarize the contribution of group differences to the total variation in the data. R^2 measures the fraction of the variation in Y that is explained by group differences.

$$R^2 = \frac{SS_{\text{groups}}}{SS_{\text{total}}}$$

It is a reflection of how much narrower the scatter of measurements is around the group means compared with the scatter around the grand mean.

R^2 takes on values between zero and one. When R^2 is close to zero, the group means are all very similar and most of the variability is within groups—that is, the explanatory variable defining the groups explains *very little* of the variation in Y.

Conversely, an R^2 close to one indicates that little variation in Y is left over after the different group means are taken into account—that is, the explanatory variable explains *most* of the variation in Y.

EXAMPLE

Analysis of Variance (ANOVA)

EXAMPLE: The knees who say night

Traveling to a different time zone can cause jet lag, but people adjust as the schedule Trof 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 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.

The 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?

Analysis of Variance (ANOVA)

Raw data and descriptive statistics of phase shift, in hours, for the circadian rhythm experiment.

| Treatment	Data (h)	$\bar{Y}$	s	n
Control	0.53, 0.36, 0.20, -0.37, -0.60, -0.64, -0.68, -1.27	-0.3088	0.6176	8
Knees	0.73, 0.31, 0.03, -0.29, -0.56, -0.96, -1.61	-0.3357	0.7908	7
Eyes	-0.78, -0.86, -1.35, -1.48, -1.52, -2.04, -2.83	-1.5514	0.7063	7
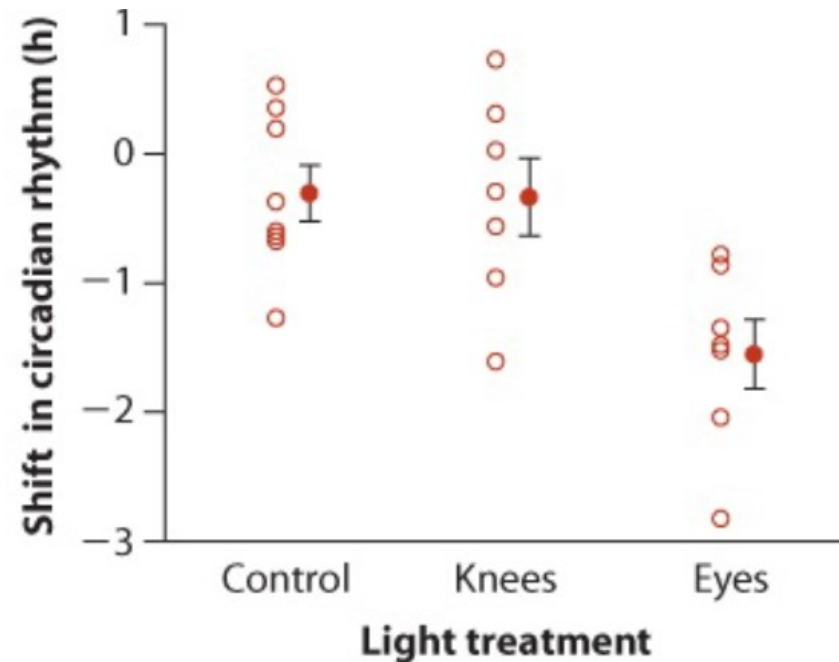
Analysis of Variance (ANOVA)

ANOVA tables

ANOVA table for the results of the circadian rhythm experiment

Source of variation	Sum of squares	df	Mean squares	F-ratio	P
Groups (treatment)	7.224	2	3.6122	7.29	0.004
Error	9.415	19	0.4955		
Total	16.639	21			

Analysis of Variance (ANOVA)

Partitioning the sum of squares

$$\bar{Y} = \frac{8(-0.3087) + 7(-0.3357) + 7(-1.5514)}{22} = -0.7127$$
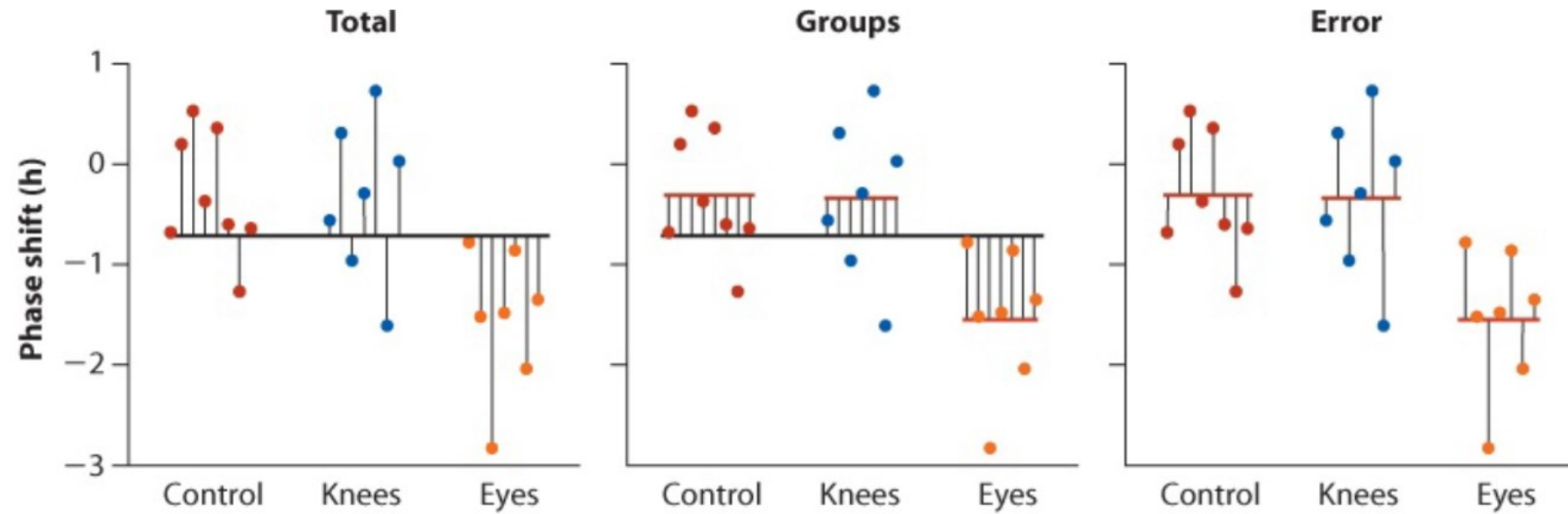
Analysis of Variance (ANOVA)

The group sum of squares is

$$SS_{groups} = 8[-0.3087 - (-0.7127)]^2 + 7[-0.3357 - (-0.7127)]^2 + 7[-1.5514 - (-0.7127)]^2 = 7.224.$$

And error sum of squares is

$$SS_{error} = (0.6176)^2(8-1) + (0.7908)^2(7-1) + (0.7063)_2(7-1) = 9.415.$$

Thus, total sum of squares is

$$SS_{total} = SS_{error} + SS_{groups} = 9.415 + 7.224 = 16.639.$$

Analysis of Variance (ANOVA)

Calculating the mean squares

Thus, the group mean squares is

$$MS_{groups} = \frac{SS_{groups}}{df_{groups}} = \frac{7.224}{3-1} = 3.6122$$

And, the error mean of squares is

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{9.415}{22-3} = 0.4955$$

## The variance ratio, *F*

Thus, the variance ratio F is

$$F = \frac{3.6122}{0.4955} = 7.29$$

Analysis of Variance (ANOVA)

Here, $N = 22$ and $k = 3$, so there are $k - 1 = 3 - 1 = 2$ degrees of freedom for the numerator and $N - k = 22 - 3 = 19$ degrees of freedom for the denominator.

The F-distribution with 2 and 19 df is therefore the appropriate null distribution for F. The number of degrees of freedom for the numerator is always presented first when specifying the F-distribution. This is important because the F-distribution with 2 and 19 degrees of freedom is *not* the same as the F-distribution having 19 and 2 degrees of freedom.
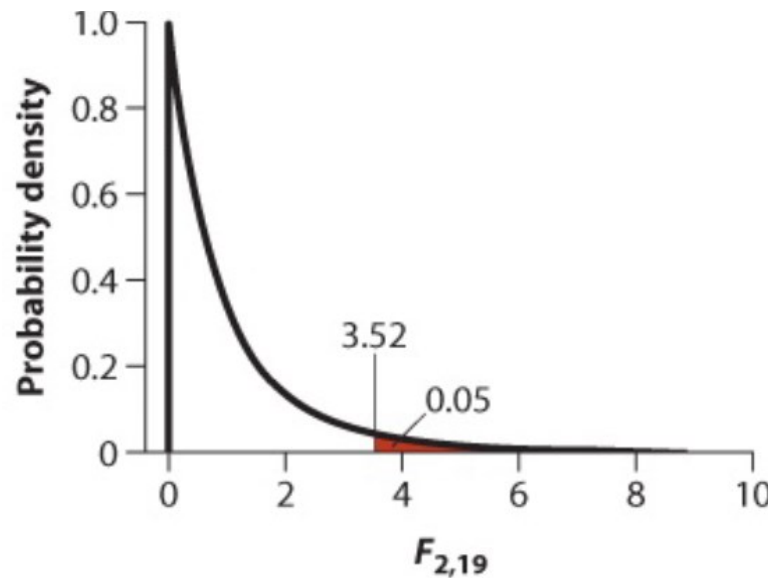
If H0 is false, then MS_{groups} should *exceed* MS_{error} and lead to an F-ratio in the right tail of the F-distribution. F-ratios less than one might occur, but only by chance.

Analysis of Variance (ANOVA)

Here,

$$F_{0.05(1),2,19} = 3.52$$

Because our observed value of F (i.e., 7.29) is larger than 3.52, it lies farther out in the right tail of the F-distribution, so P must be less than 0.05. Therefore, we reject the null hypothesis.

The exact P-value is the area under the curve of the F-distribution to the *right* of the observed F-value: That is

P-vale = P[F>7.29] = 0.040 < 0.05.

Rejecting H0 indicates only that at least one of the population means μ_i is different from the others, not that every μ_i is different from all of the others.

The R^2 value is

$$R^2 = 7.224 / 16.639 = 0.43$$

In other words, 43% of the total sum of squares among subjects in the magnitude of phase shift is explained by differences among them in light treatment. The remaining 57% of the variability among subjects is "error"—variance unexplained by the explanatory variable, light treatment.

Thank You