ANOVA

STA 198: Introduction to Health Data Science

Yue Jiang

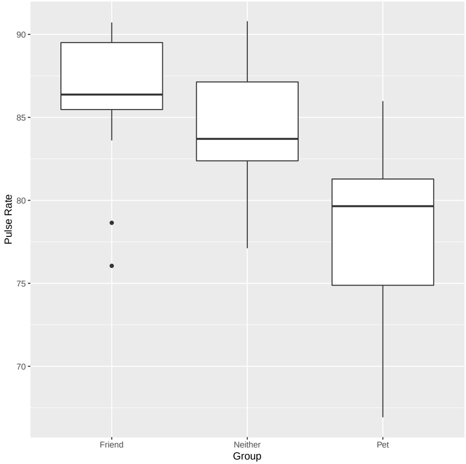
June 19, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Motivating example: pets and stress

CIs for group means



How do we compare across three groups?
Which groups are different?

Example: pets and stress

We are interested in testing

$$H_0 : \mu_P = \mu_F = \mu_N$$

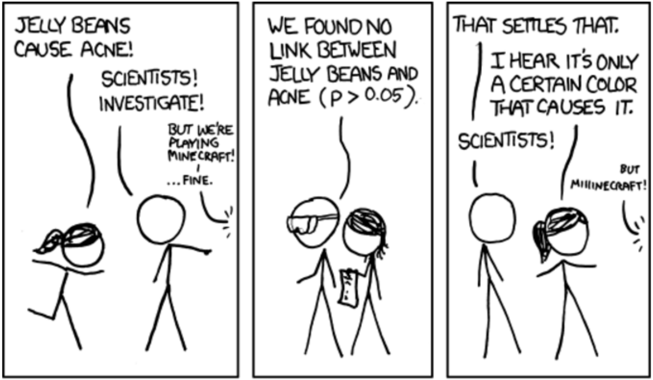
against the alternative that at least one mean is different from the others.

One way to do this would be to use t-tests on all possible pairs of tests (here there are just three). However, if we have more groups, this becomes quite complicated. For example, with 10 groups we need to do $\binom{10}{2} = 45$ tests!

Multiple comparisons

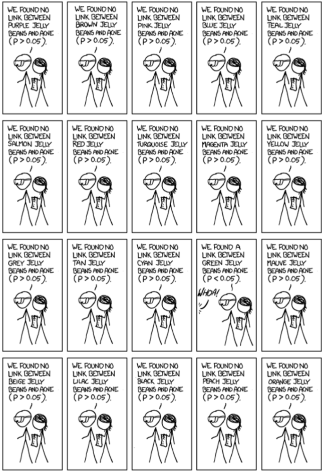
However, in addition to being time-consuming, carrying out multiple tests can lead to an inflated Type I error rates which puts into question the validity of a given study if these multiple comparisons are not accounted for.

Multiple comparisons in action: overall test



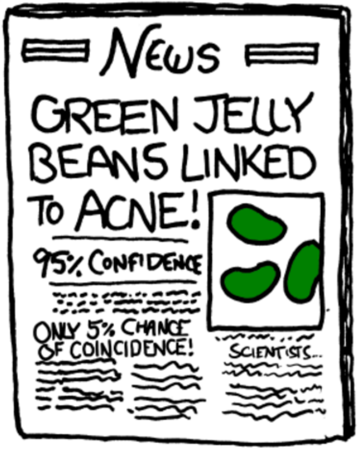
Randall Munroe, xkcd #882

Extra tests...whoa!



Randall Munroe, xkcd #882

Green jelly beans!



Randall Munroe, xkcd #882

So what went wrong?

Multiple comparisons

Let's revisit the pets / stress example, where we had three pairwise comparisons.

- Suppose all means are truly equal (H_0 is true), and we conduct all three pairwise tests
- Suppose also the tests are independent and done at a 0.05 significance level
- Then the probability we fail to reject all three tests (the correct decision) is $(1 - 0.05)^3 = 0.95^3 = 0.857$, and so the probability of rejecting at least one of the three null hypotheses, called the family-wise error rate, is $1 - 0.857 = 0.143 > 0.05$
- With 45 tests, the probability of rejecting at least 1 of them (incorrectly!) is almost 1

Multiple comparisons

- In reality, this is a little more complicated (tests are not independent), but we still have the problem of an inflated Type I error rate
- ANOVA extends the t-test and is one way to control the overall Type I error rate at a fixed level α, if we only test pairwise differences when the overall test is rejected

ANOVA

ANOVA stands for analysis of variance. We use ANOVA when we want to compare more than two groups.

For the two-sample t-test comparing the means of two groups, we could consider

$$H_0 : \mu_P = \mu_F$$

$$H_1 : \mu_P \neq \mu_F.$$

What if we have three groups? Our null might be that the groups are all the same: $H_0 : \mu_P = \mu_F = \mu_N$, and our alternative would be the complement of H_0, or that there is some type of difference between the groups.

ANOVA null hypothesis

In ANOVA, we typically follow this testing procedure. First, we conduct an overall test of the null hypothesis that the means of all of the groups are equal.

1. If you reject this null hypothesis, then we usually step down to see which means are different from each other. A multiple comparisons correction is sometimes used for these pairwise comparisons of means.
2. If you fail to reject the null hypothesis, this means that we have insufficient evidence to conclude that there is a difference among the groups. Generally, no further testing should be done.

ANOVA alternative hypothesis

For ANOVA with three groups, our null hypothesis is $H_0 : \mu_P = \mu_F = \mu_N$. What could happen under the alternative?

- $\mu_P \neq \mu_F \neq \mu_N$
- $\mu_P = \mu_F$, but μ_N is different
- $\mu_P = \mu_N$, but μ_F is different
- $\mu_F = \mu_N$, but μ_P is different

If we reject the null hypothesis, any of these situations could be true, and we may wish to conduct further tests to discover what setting we are in. Conducting further tests without rejecting the overall test of H_0 will lead to an inflated Type I error rate unless we use another method to adjust for multiple comparisons.

Why analyze variance when we're talking about means?

Remember, ANOVA stands for analysis of variance.

What does variance have to do with our null hypothesis, which is about equality of means (say, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$)?

General ANOVA setup

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$
- H_1 : at least one of the means is different
  - Not only $H_1 : \mu_1 \neq \mu_2 \neq \cdots \neq \mu_K$, since this would force all means to be different
- Let the response of subject j in group i be given by y_{ij}, $i = 1, 2, ..., K$, $j = 1, 2, ..., n_i$
- K is the number of groups
- n_i is the number of subjects in group i

General ANOVA setup

- ANOVA model is

$$y_{ij} = \mu_i + \epsilon_{ij}$$

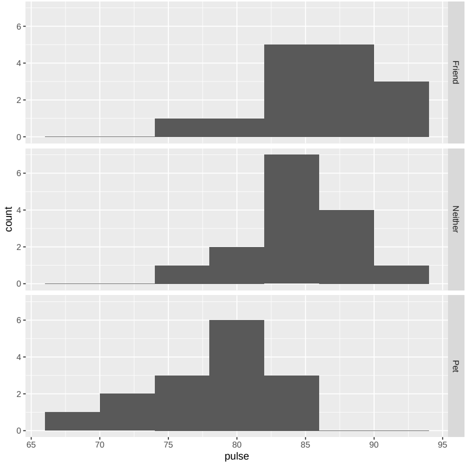
  where we assume the ϵ_{ij} are independent and normally distributed with mean 0 and variance σ^2
- ϵ_{ij} represent random error about each group mean, μ_i
- Thus, the model assumes the y_{ij} are independent and normally distributed with mean μ_i and variance σ^2

Assumptions of ANOVA

1. Outcomes within groups are normally distributed
2. Homoscedastic variance (the within-group variance is the same for all groups)
3. Samples are independent

If these assumptions are violated, then results from ANOVA may not be valid. We will discuss some alternatives later in the course.

Validity of ANOVA for pet data



Variances appear to be similar, but normality looks questionable!
For now, let's proceed despite this problem.

ANOVA for pet data

- Let the pulse rate of subject j in group i be given by y_{ij}, $i = 1, 2, 3$, $j = 1, 2, ..., 15$
- We have three groups (pets, friends, neither) and $n_1 = n_2 = n_3 = 15$ subjects in each group for a total of 45 people with measurements
- For each group, calculate the group mean and variance, so now we have $\bar{y}_1$ and s_1^2 (pet group mean and variance), $\bar{y}_2$ and s_1^2 (friend group mean and variance), and $\bar{y}_3$ and s_3^2 (mean and variance for the neither group)
- ANOVA is concerned with two different variances:
  - Within-groups variance: variance of the individual observations around their respective group means
  - Between-groups variance: variance of the group means around the overall mean of all observations, $\bar{y}$

Reminder: variance

The total variance of y is estimated as

$$s^2 = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}{n - 1}$$

where K is the number of groups, n_i is the sample size for group i, and n is the total sample size across all groups.

Within-groups variance

Each group's variance, $s_i^2 = \frac{\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i \right)^2}{n_i - 1}$ is a measure of the variance of the individuals around that population group mean. To get a pooled estimate of the common variance of individuals around their group means, we can calculate

$$s_W^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)^2 + \cdots + (n_K - 1)s_K^2}{n - K},$$

where K is the number of groups and $n = n_1 + n_2 + \cdots + n_K$. We can think of the within-groups variance as the inherent variability in the population.

(Don't worry about this exact expression)

Between-groups variance

The between groups variance is estimated by

$$s_B^2 = \frac{n_1 (\bar{y}_1 - \bar{y}.)^2 + n_2 (\bar{y}_2 - \bar{y}.)^2 + \cdots + n_K (\bar{y}_K - \bar{y}.)^2}{K - 1}.$$

We can think of the between-groups variance as the sum of inherent variability *and* any kind of systematic variability due to the group effect.

(Don't worry about this exact expression)

F-test

If the sample means vary around the overall mean (measured by s_B^2) more than the individual observations vary around the sample means (measured by s_W^2), we have evidence that the corresponding population means are in fact different (so that all K means are not the same).

(Examples to be drawn during live session)

F-test

How do we compare formally compare these variances? Consider the F statistic given by

$$F = \frac{s_B^2}{s_W^2},$$

which if H_0 is true, has an F distribution with $K - 1$ and $n - K$ degrees of freedom. The df associated with s_B^2 are called the numerator degrees of freedom and correspond to the total number of groups minus 1. The df associated with s_W^2 are called the denominator degrees of freedom and equal to the total sample size minus the number of groups.

F-test

The F-test for ANOVA is inherently one-tailed, rejecting H_0 only if F is considerably larger than one.

If there are only two groups, then the F-test gives the same result as the t-test.

Importantly: This does not mean we have a one-sided alternative; we just look at one tail of the F distribution to get the p-value.

Output

Source	Sum Sq.	df	MS	F	p-value
Between	2387.7	2	1193.8	14.1	<0.0001
Within	3561.0	42	84.8		
Total	5948.7	44	135.2		

You may also see "Between" denoted by the grouping variable and "Within" by "residuals." (Why?)

F-test for pet data

Note that $F = s_B^2/s_W^2 = 1193.8/84.8 = 14.1$, with ndf $=$ $3 - 1 = 2$ and ddf $= 45 - 3 = 42$. This corresponds to a p-value $<$0.0001. At $\alpha = 0.05$, we reject the null hypothesis. There is sufficient evidence that at least one of the three groups comes from a population with a different mean from the others.

Next: which groups are different?

Bonferroni correction

As we showed earlier, conducting multiple tests on a data set increases the *family-wise error rate*. One very conservative way to ensure this is not the case is to simply divide α by the number of tests to be done and to use that as the significance level.

This procedure is called the Bonferroni correction.

Bonferroni correction

For example for two tests, to preserve an overall 0.05 type I error rate, the Bonferroni correction would use $\alpha/2 = 0.025$ as the significance level for each individual test instead of 0.05.

Bonferroni is a conservative correction, making it harder to reject the null hypothesis, but it is a safe bet in controlling the Type I error rate.

Pets and stress: group differences

We can compare the groups using a Bonferroni correction (here we have three tests, so the significance level for each test is $\alpha/3$). `R` handles this by multiplying each p-value by 3 before showing it to you, so you can still use $\alpha = 0.05$ to assess significance of these pairwise comparisons.

The raw (uncorrected) p-values for the t-test comparing friend vs. pet was < 0.0001; for friend vs. neither was 0.021, and for pet vs. neither was 0.009.

What do we conclude?