ANOVA

STA 102: Introduction to Biostatistics

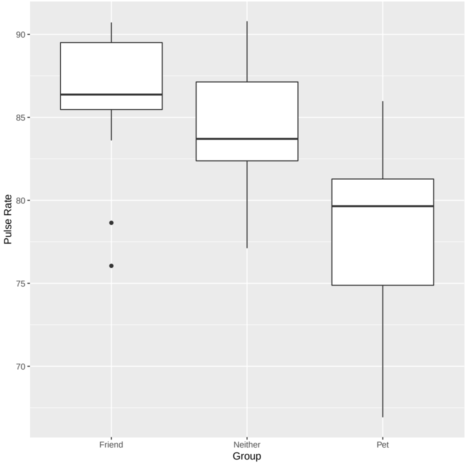
Yue Jiang

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

Distribution of heart rate data



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

Example: pets and stress

We are interested in testing

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

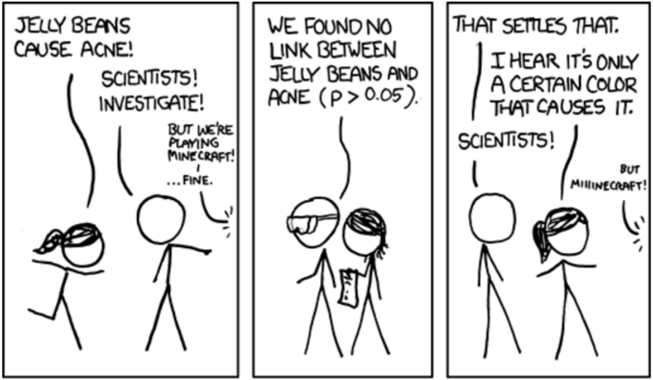
against the alternative that at least one mean is different from the others.

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

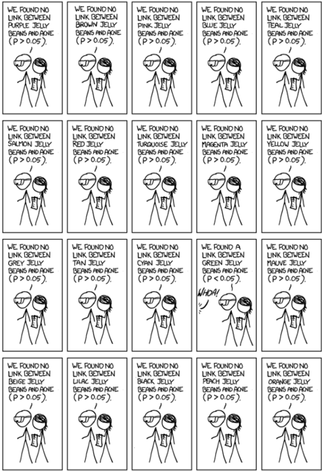
However, in addition to being time-consuming, carrying out multiple tests can lead to an inflated Type I error rates which puts into question the validity of a given study if these multiple comparisons are not accounted for.

Multiple comparisons in action: overall test


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

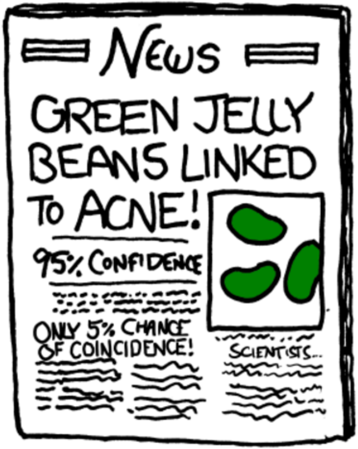
Randall Munroe, xkcd #882

Extra tests...whoa!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Randall Munroe, xkcd #882

Green jelly beans!


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

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 over 90%!

Multiple comparisons

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

ANOVA null hypothesis

In ANOVA, we typically follow this testing procedure:

1. First, we conduct an overall test of the null hypothesis that the means of all of the groups are equal.
2. If this is rejected, then we step down to see which means are different from each other. A multiple comparisons correction is sometimes used for these pairwise comparisons of means.
3. If we fail to reject the null hypothesis, then 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

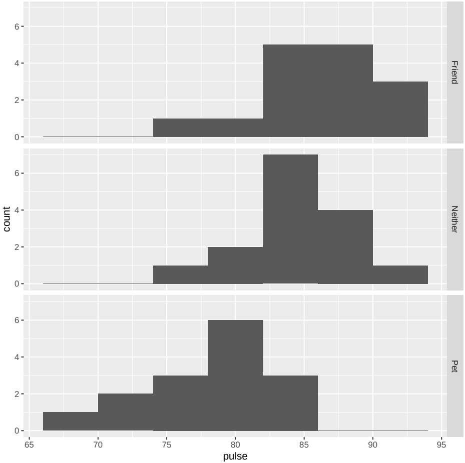
The alternative hypothesis for ANOVA is that at least one of the means is different from the others.

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.

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, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$)?

F-test

If the group-specific means vary around the overall grand mean more than the individual observations vary around their group-specific sample means, then we have evidence that the corresponding population means are in fact different.

(Examples to be visualized during live session)

F-test

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

$$F = \frac{s^2_{between}}{s^2_{within}},$$

which if H_0 is true, has an F distribution with $K - 1$ numerator degrees of freedom and $n - K$ denominator degrees of freedom, where K is the number of groups and n is the total number of observations.

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?