

# Unit 5: Inference for Numerical Data

Statistics 102 Teaching Team

March 09, 2020

## Comparing many means with ANOVA

# ANALYSIS OF VARIANCE (ANOVA)

Suppose we are interested in comparing means across more than two groups. Why not conduct several two-sample  $t$ -tests?

- If there are  $k$  groups, then  $\binom{k}{2} = \frac{k(k-1)}{2}$   $t$ -tests are needed.
- Conducting multiple tests on the same data increases the overall rate of Type I error.

ANOVA uses a single hypothesis test to assess whether means across many groups are equal:

- $H_0$ : mean outcome is same across all groups ( $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ )
- $H_A$ : at least one mean is different from the others (i.e., means are not all equal)

# IDEA BEHIND ANOVA

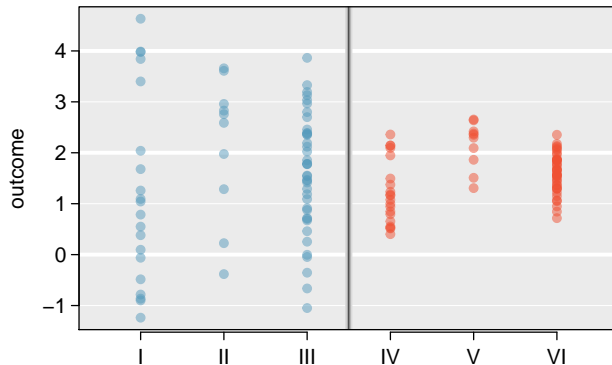
Is the variability in the sample means large enough that it seems unlikely to be from chance alone?

Compare two quantities:

- Variability between groups ( $MSG$ ): how different are the group means from each other, i.e., how much does each group mean vary from the overall mean?
- Variability within groups ( $MSE$ ): how variable are the data within each group?

$MSG$  denotes mean square between groups, while  $MSE$  denotes mean square error. Refer to *OI Biostat* Section 5.5.1 for details.

## IDEA BEHIND ANOVA...



- I, II, and III: difficult to discern differences in means, variability within each group is high
- IV, V, and VI: appears to be differences in means, these differences are large relative to variance within each group

## IDEA BEHIND ANOVA...

Under the null hypothesis, there is no real difference between the groups; thus, any observed variation in group means is due to chance.

- Think of all observations as belonging to a single group.
- Variability between group means should equal variability within groups

The *F-statistic* is the test statistic for ANOVA.

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} = \frac{MSG}{MSE}$$

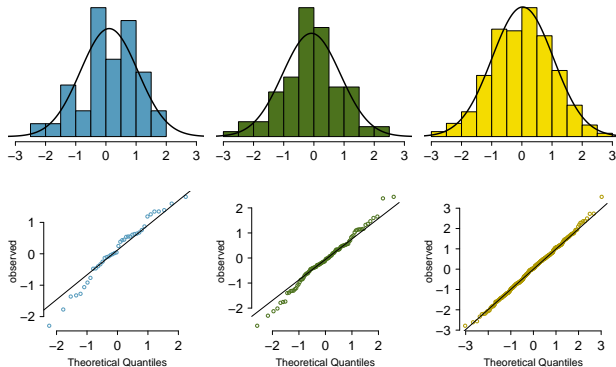
- When the population means are equal, the *F*-statistic is approximately 1.
- When the population means differ, *F* will be larger than 1. Larger values of *F* represent stronger evidence against the null.
- The *F* statistic follows an *F* distribution, with two degrees of freedom,  $df_1$  and  $df_2$ ;  $df_1 = n_{groups} - 1$ ,  $df_2 = n_{obs} - n_{groups}$ .
- The *p*-value for the *F*-statistic is the probability *F* is larger than the *F*-statistic.

# ASSUMPTIONS FOR ANOVA

It is important to check whether the assumptions for conducting ANOVA are reasonably satisfied:

1. Observations independent within and across groups
  - Think about study design/context
2. Data within each group are nearly normal
  - Look at the data graphically, such as with a histogram
  - Normal Q-Q plots can help. . .
3. Variability across groups is about equal
  - Look at the data graphically
  - Numerical rule of thumb: ratio of largest variance to smallest variance  $< 3$  is considered “about equal”

# NORMAL PROBABILITY PLOTS (Q-Q PLOTS)



If points fall on or near the line, data closely follow a normal distribution.

- Difficult to evaluate in small datasets
- Plots show three simulated normal datasets: from L to R,  $n = 40$ ,  $n = 100$ ,  $n = 400$

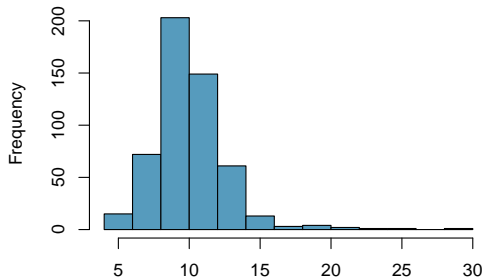


# NORMAL PROBABILITY PLOTS (Q-Q PLOTS)...

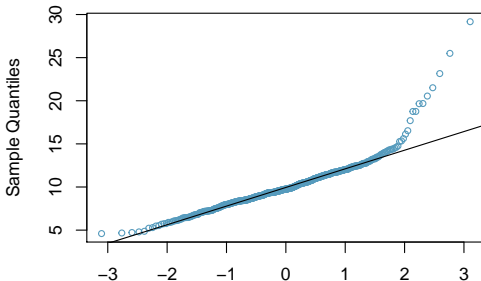
```
#simulate right-skewed distribution  
set.seed(2019)  
sim.data <- c(rnorm(500, 10, 2), rnorm(25, 15, 5))
```

```
#plots  
par(mfrow = c(1, 2))  
hist(sim.data, col = COL[1])  
qqnorm(sim.data, cex = 0.75, col = COL[1]); qqline(sim.data)
```

Histogram of sim.data



Normal Q-Q Plot



## PAIRWISE COMPARISONS

If the  $F$ -test indicates there is sufficient evidence that the group means are not all equal, proceed with pairwise comparisons to identify which group means are different.

Pairwise comparisons are made using the two-sample  $t$ -test for independent groups.

- To maintain the overall Type I error rate at  $\alpha$ , each pairwise comparison is conducted at an adjusted significance level referred to as  $\alpha^*$ .
- The Bonferroni correction is one method for adjusting  $\alpha$ .

$$\alpha^* = \alpha/K, \text{ where } K = \frac{k(k-1)}{2} \text{ for } k \text{ groups}$$

- Note that the Bonferroni correction is a very stringent (i.e., conservative) correction, made under the assumption that all tests are independent.

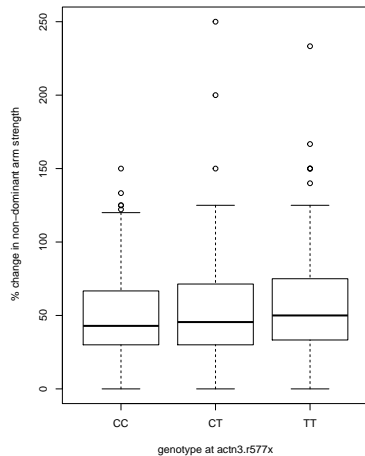
## FAMuSS: COMPARING NDRM.CH BY GENOTYPE

The main question of interest in the FAMuSS study can be approached with ANOVA.

*Is change in non-dominant arm strength after resistance training associated with genotype?*

Questions 1 - 3 in Lab 3 step through this analysis.

## FAMuSS: COMPARING NDRM.CH BY GENOTYPE...



## FAMUSS: COMPARING NDRM.CH BY GENOTYPE...

The null and alternative hypotheses are

- $H_0 : \mu_{CC} = \mu_{CT} = \mu_{TT}$ , the mean percent change in non-dominant arm strength is equal across the three genotypes
- $H_A$ : At least one group has mean percent change in non-dominant arm strength that is different from the other groups

Let  $\alpha = 0.05$ .

## LETTING R DO THE WORK

Formulas for hand calculations shown in *Ol Biostat* Section 5.5.1.

```
#use summary(aov())  
summary(aov(famuss$ndrm.ch ~ famuss$actn3.r577x))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)  
## famuss$actn3.r577x    2    7043     3522   3.231 0.0402 *  
## Residuals          592 645293     1090  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:  $p < \alpha$ , sufficient evidence to reject  $H_0$  in favor of  $H_A$ . There is at least one group with a mean different from the other groups.

- But which groups have different means?

## CONTROLLING TYPE I ERROR RATE

If the ANOVA  $F$ -test is significant, then it is appropriate to proceed to conducting pairwise comparisons; i.e., using two-sample  $t$ -tests to compare each possible pairing of the groups.<sup>2</sup>

- Each test should be conducted at the  $\alpha^*$  significance level so that the overall Type I error rate remains at  $\alpha$ .
- These tests are still conducted under the assumption that the variance between groups is equal; thus, the test statistics are calculated using the pooled estimate of standard deviation between groups. Details are in *Ol Biostat* Section 5.5.3.
- We will use `pairwise.t.test( )` to perform these *post hoc* two-sample  $t$ -tests. Refer to Unit 5, Lab 2 for an example; this function is also discussed in the Unit 5 Lab Notes.

---

<sup>2</sup>These  $t$ -tests are typically referred to as *post hoc* tests.

## CONTROLLING TYPE I ERROR RATE...

Pairwise comparisons using two-sample  $t$ -tests (CC to CT, CC to TT, CT to TT) can now be done if the Type I error rate is controlled.

- Apply the Bonferroni correction.
- In this setting,  $\alpha^* = 0.05/3 = 0.0167$ .



## LETTING R DO THE WORK

Only CC versus TT resulted in a  $p$ -value less than  $\alpha^*$  of 0.0167.

- Mean strength change in non-dominant arm for CT individuals not distinguishable from strength change for CC and TT.
- However, evidence at  $\alpha = 0.05$  level that mean strength change for individuals of genotype CC and TT are different.

```
pairwise.t.test(famuss$ndrm.ch, famuss$actn3.r577x, p.adj = "none")
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data: famuss$ndrm.ch and famuss$actn3.r577x  
##  
##      CC      CT  
## CT 0.179 -  
## TT 0.011 0.144  
##  
## P value adjustment method: none
```

## LETTING R DO THE WORK...

Alternatively, set `p.adj` to "bonf"; this instructs R to rescale the  $p$ -values (by multiplying by  $K$ ) so they can be compared to the original  $\alpha$  level of 0.05.

```
pairwise.t.test(famuss$ndrm.ch, famuss$actn3.r577x, p.adj = "bonf")
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: famuss$ndrm.ch and famuss$actn3.r577x
##
##      CC      CT
## CT 0.537 -
## TT 0.034 0.433
##
## P value adjustment method: bonferroni
```