Analysis of Variance (ANOVA)

Review: Two-Sample T-Test

- Comparison between two groups
  - ✓(two groups in a categorical variable like Female/Male or Hispanic/non-Hispanic)
- Observed samples from each population
- Assume underlying normality in each population
- Use rank-based methods when not normal

Limitations of T-Test

- Single classification variable with only two groups
  - ✓What if there are more than two groups?
  - ✓What about the case of more than one categorical variable?

When do we use ANOVA model?

- Setting: **Continuous response** **&** **Categorical (grouping) variables**

- **Goal**: Analyze the difference among groups and study the behaviors of response variable depending on grouping variable

(E.g.) we are interested in blood sugar (continuous);

- Variable1: treatment (placebo/ treatment1/ treatment2)
- Variable2: diet (vegetarian/ vegan/ else)
- Variable3: exercise (<1 hr/ between 1 and 3 hrs/ >3 hrs)

Want to answer:

- Does type of treatment (or diet or exercise) affect blood sugar?
- If so, which treatment is the most efficient?
- Does diet help to decrease blood sugar?

ANOVA model

Kind of extension of two-sample t-test

- Compare means of two groups
- T-test can be applied only when **both** groups follow **normal**

  (parametric test)
- Two types of t-test under equal variance or unequal variance assumption

Similar in ANOVA test

- Can compare means from more than three groups
- Assumptions for classic ANOVA (again, parametric test):

  Normality for all groups, **equal variances, **iid** sample**

  More specific statement at slide 11
- Modified test when groups have different variances (**welch**'s ANOVA)

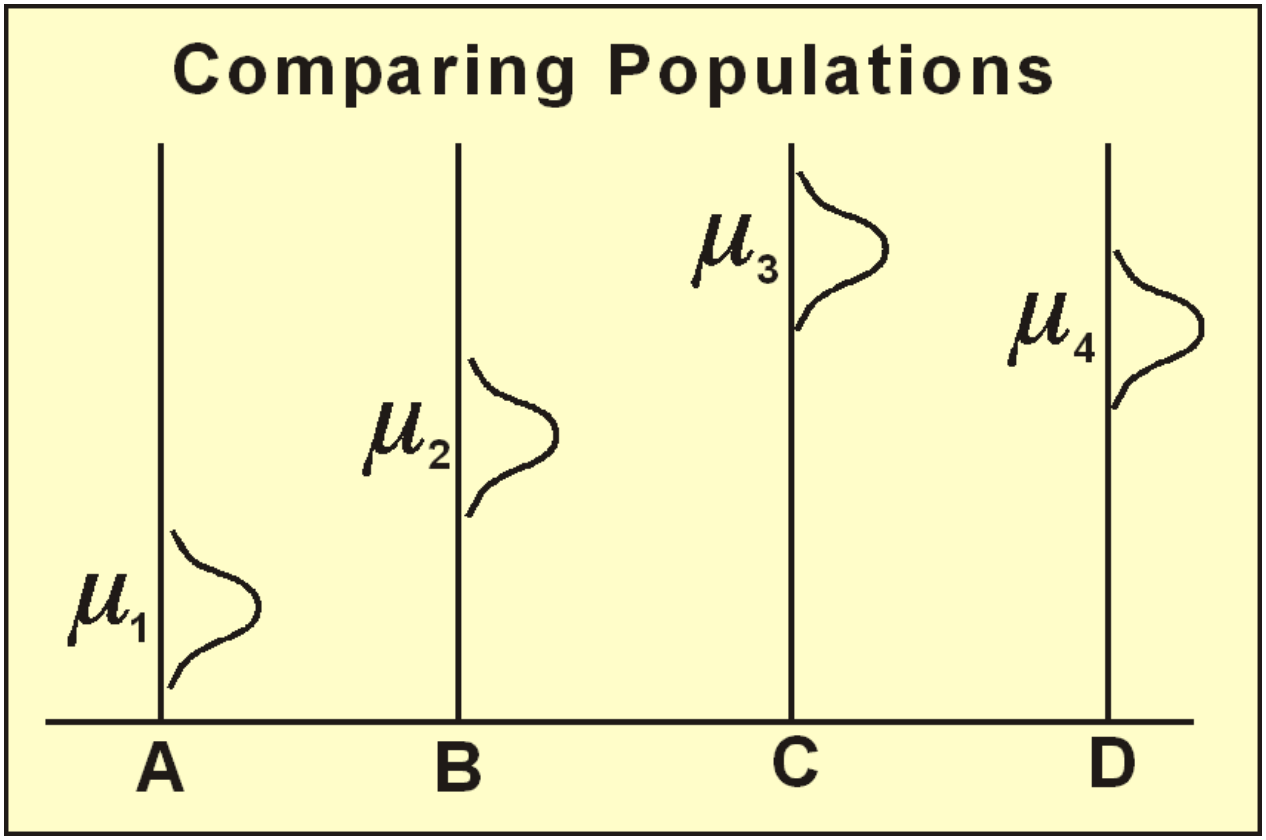
Definitions

- **One-way analysis of variance**: ANOVA based on a single categorical predictor variable
- **Two-way analysis of variance**: based on 2 independent categorical predictor variables
- **N-way analysis of variance**: based on *n* independent categorical variables

Definitions

- **Main effect**: effect of single categorical predictor
- **Interaction**: the combined effect of combination of categorical predictors -> for example, synergy effect
- **First-order interaction**: an interaction between two categorical predictors
- **N-th order interaction**: interaction of a categorical predictor with *n* other categorical predictors
  - hard interpretation or potential overfitting issue. In practice, include them only when needed
- **Balanced data:** data with an equal number of observations in each cell
- **Unbalanced data:** at least one cell has different number of observations

ANOVA assumption overview



ANOVA Hypotheses

- Null Hypothesis: There are no mean differences between the groups on response
  - **H0: $\mu_1 = \mu_2 = \ldots = \mu_g$,**
    where g is the number of groups.
    - E.g., means of salary are same regardless of different education levels

- Alternate Hypothesis:
  - **H1: At least ONE of the group means is significantly different from the others in the population**

    - NOTE: But we do not know which group has larger or smaller mean

ANOVA Hypotheses

- For **the interactions:**
  (for multi-way ANOVA like 2-way, 3-way ….)

We can also test the null hypothesis for interactions.

- Null hypothesis:

  H0: There is no interaction between independent variables in the population.

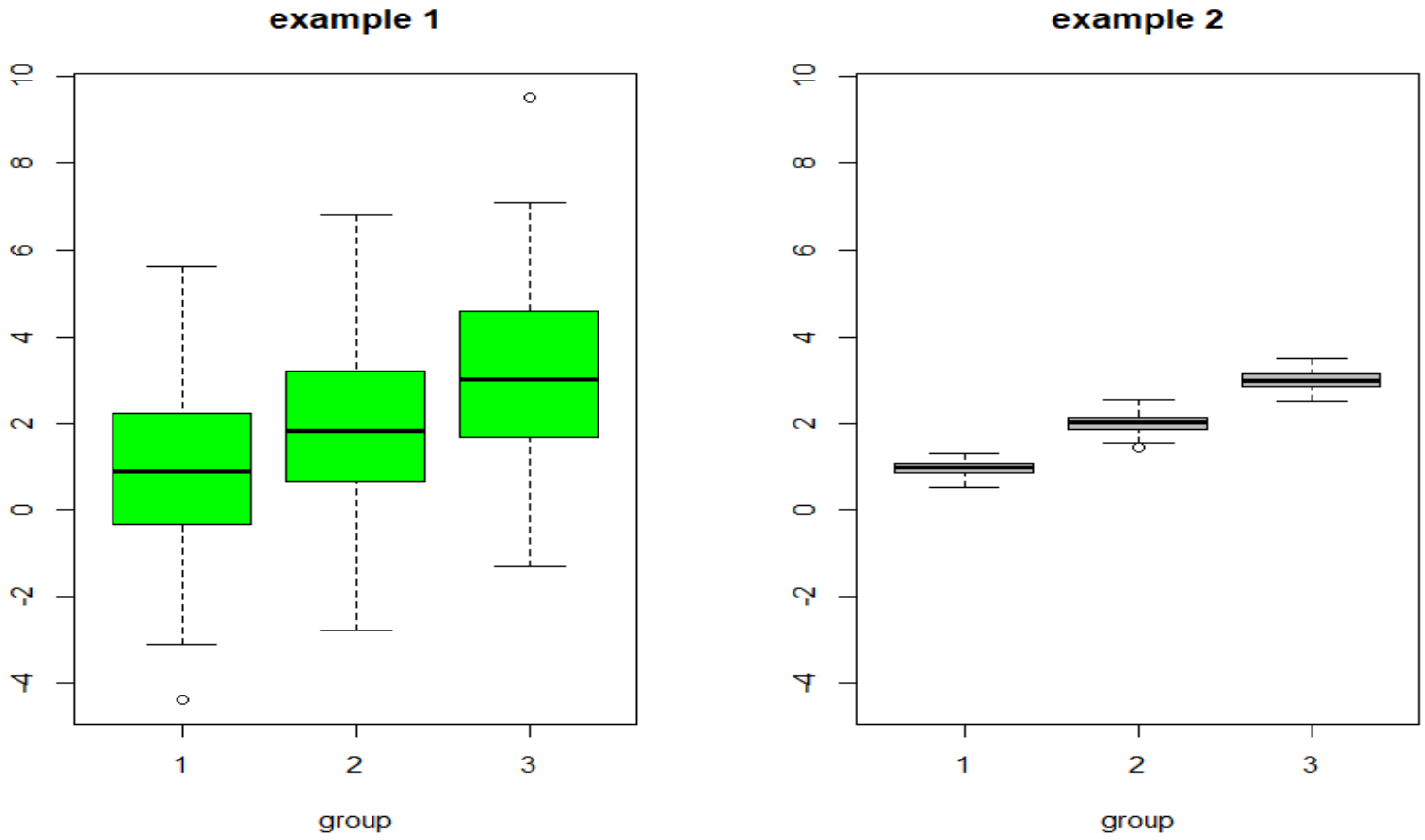
- Alternative hypothesis:

  H1: There is an interaction between independent variables in the population.

Assumptions of ANOVA

1) The response (dependent) variable is **continuous**
2) Populations from which samples were drawn **follow normal distribution**
   - i.e., Each group should be normally distributed
   - ✓ Note: ANOVA relatively *robust* to violations of normality
3) Populations from which samples were drawn must have **equal variances** (Homogeneity of Variance)
   - ✓ Need to perform equal variance test before applying ANOVA
4) Observations must be **independent of one another**

The F test



- **Between group mean variation (differences)** are same for both examples
- How about **within group variation**?

The F test

- Use **F-test** when all assumptions are satisfied
- The F test uses the F *statistic* to determine if there are any significant main effects or interactions
- Formula and Intuition:

  F = Between groups variation/ Within group variation

- Make a conclusion based on p-values of F-test

The F test

- If the F-statistic is NOT statistically significant, then you are done and there is no reason to conduct additional analyses. No difference among groups is found.

- If the F-statistic is statistically significant:
  - All you know now is that there is **at least one** mean that differs from the another.
    - ✓To determine which mean(s) differ, you need to conduct post-hoc test
    - ✓Able to get the information which group has significantly larger of smaller mean value

Example: ToothGrowth

- Response: Tooth length (continuous variable)
- **Supplement**: **VC** or **OJ**
- **Dose: 0.5, 1** or **2**
  - ✓Should be coded as a **factor** not as a numeric

Example: One-way ANOVA

- Install package “car”
- Perform analysis of variance of **Toothlength** as a function of **Dose**

I. Check balanced of unbalanced
II. Run one-way ANOVA with **aov()** or possibly, **lm()**
III. Check equal variance assumption – levene’s test
   - H0: all groups have the same variances vs.
     Ha: at least one group has different variance
   - Could use Welch adjustment if equal variance assumption is not valid
IV. Check Normality assumption - check diagnostics plot
   - qq plot and residual plot
V. What is conclusion?

Example: One-way ANOVA

- ANOVA table interpretation
- R-Square value for predictive power of the model
- Significance of **Dose** as a predictive variable
- Conclusion about impact of **Dose** on tooth growth

One-way ANOVA example:

```
tooth$Dose= as.factor(tooth$Dose)
str(tooth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ Toothlength: num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.
2 7 ...
##  $ Supplement : Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2
2 2 2 2 ...
##  $ Dose       : Factor w/ 3 levels "0.5","1","2": 1 1 1 1
1 1 1 1 1 1 ...
```

- Balance or Unbalanced?

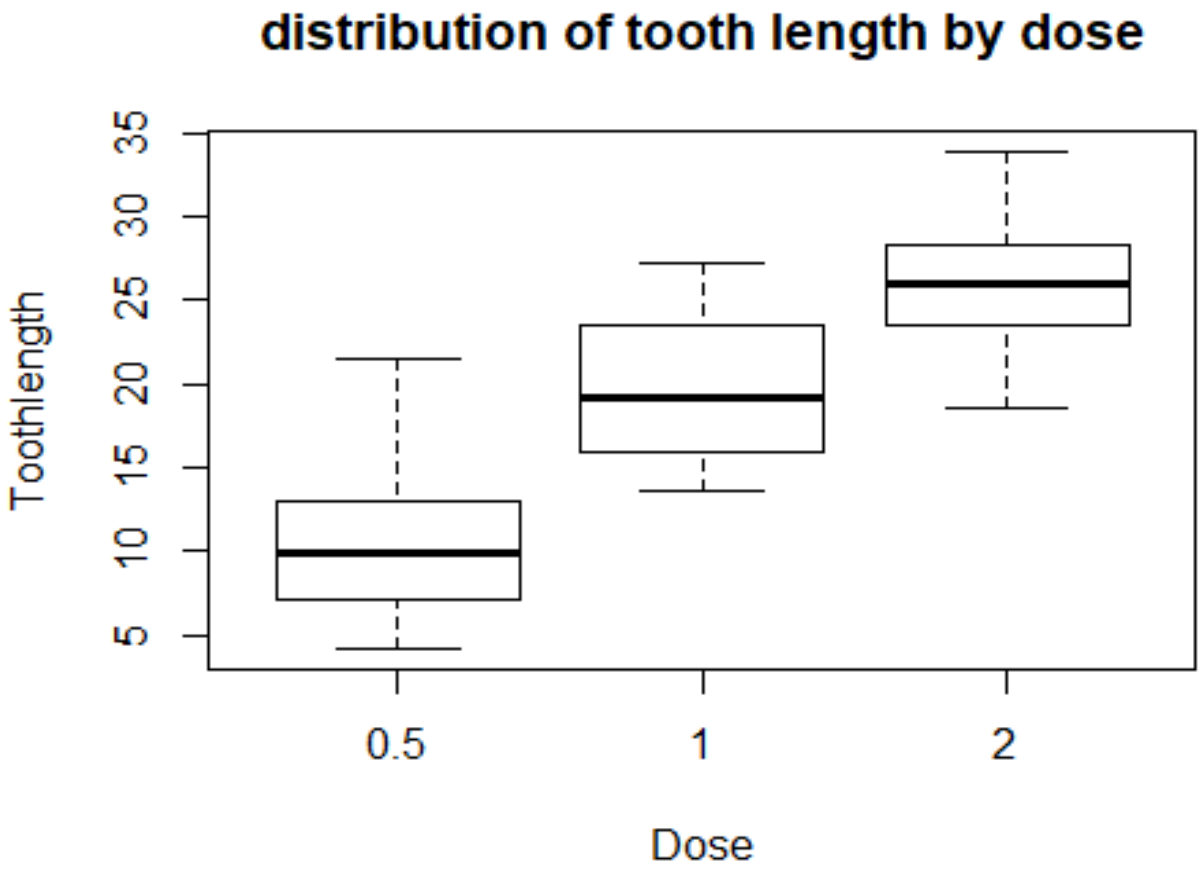
```
table(tooth$Dose); table(tooth$Supplement)
```

```
##
## 0.5   1   2
##  20  20  20
```

```
##
## OJ VC
## 30 30
```

One-way ANOVA example:

```
boxplot(Toothlength ~ Dose, data=tooth, main="distributio
n of tooth length by dose")
```



One-way ANOVA example:

```
aov.res= aov(Toothlength~Dose, data=tooth)

summary(aov.res)

##             Df Sum Sq Mean Sq F value   Pr(>F)
## Dose         2   2426    1213   67.42 9.53e-16 ***
## Residuals   57   1026      18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

Total sum of squares (3452)
(variation of Toothlength)
= sum of squares by Dose (2426)
(variation of Toothlength explained by Dose)
+ sum of squares by Error (1026)
(variation of Toothlength not be explained by the model)

H0: Dose has no effect on tooth growth
$(\mu_{0.5} = \mu_1 = \mu_2)$
Ha: Does has an effect on tooth growth
(at least one group in Dose has different mean of tooth length

One-way ANOVA example:

- To calculate **R-square**, need to run anova with **lm()**
- Results from **lm()** and **aov()** are exactly identical

```
lm.res= lm(Toothlength ~ Dose, data=tooth)
anova(lm.res)
```

```
## Analysis of Variance Table
##
## Response: Toothlength
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Dose       2 2426.4  1213.2  67.416 9.533e-16 ***
## Residuals 57 1025.8    18.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

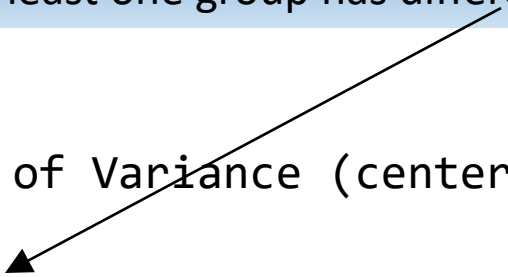
```
summary(lm.res)$r.squared
```

```
## [1] 0.7028642
```

R-square: percentage of variation in a response variable that is explained by the model (Dose)

One-way ANOVA example:

H0: all groups in Dose have the same variance
Ha: at least one group has different variance



```
leveneTest(aov.res)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.6457 0.5281
##       57
```

Welch's ANOVA – when homogeneity assumption is violated

```
oneway.test(Toothlength ~ Dose, data=tooth, var.equal=FALSE)
```

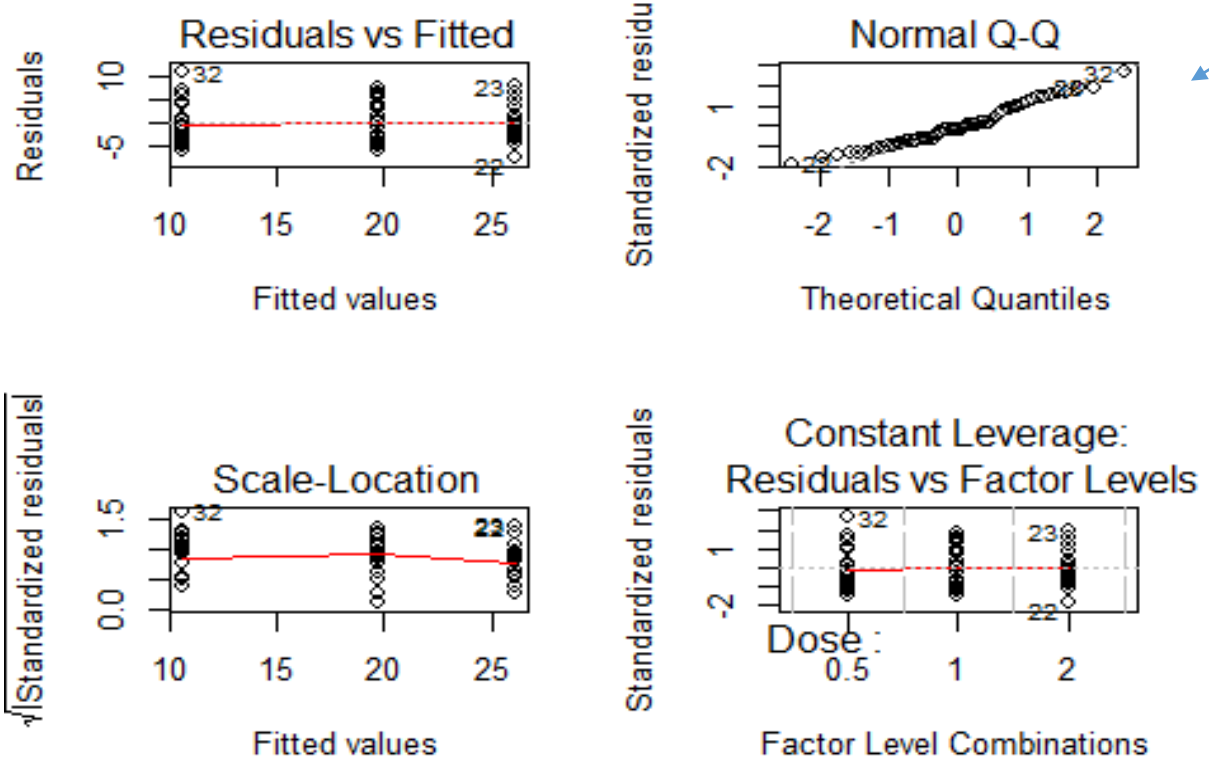
```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Toothlength and Dose
## F = 68.401, num df = 2.000, denom df = 37.743, p-value = 2.81
2e-13
```

One-way ANOVA example:

- Normality check – use diagnostics plot instead of rigorous shapiro test by group

```
par(mfrow=c(2,2))
plot(aov.res)
```

If normal qq plot shows almost straight line,
It supports normality assumption.



Multiple Comparisons

- Pairwise comparison with two-sample t-test
  - If there are 3 groups, there will be total 3 comparisons
  - (group1 vs. group2), (group1 vs. group3) and (group2 vs. group3)
- Making many comparisons at once!!
- Need to account for increased probability of making wrong decision
- Need correction in calculating p-value from t-test
  - **Scheffe method, Tukey's Method, etc.**
- **Should know how to interpret the result. What is null hypothesis and what kind of conclusion can we make?**

One-way ANOVA example:

Pairwise t-test with modified p-values:
H0: $\mu_1 = \mu_{0.5}$ vs. H1: $\mu_1 \neq \mu_{0.5}$

```
SheffeTest(aov.res)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $Dose
##         diff    lwr.ci    upr.ci    pval
## 1-0.5  9.130  5.758155 12.501845 4.3e-08 ***
## 2-0.5 15.495 12.123155 18.866845 1.2e-15 ***
## 2-1    6.365  2.993155  9.736845 7.6e-05 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Able to get information which pairs are significantly different
Final conclusion:
All three different Dose have different effect on tooth length and specifically, Dose 2 > Dose 1 > Dose 0.5

One-way ANOVA example:

```
TukeyHSD(aov.res)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Toothlength ~ Dose, data = tooth)
##
## $Dose
##         diff       lwr       upr    p adj
## 1-0.5  9.130  5.901805 12.358195 0.00e+00
## 2-0.5 15.495 12.266805 18.723195 0.00e+00
## 2-1    6.365  3.136805  9.593195 4.25e-05
```

- Different method but we can interpret the output in the same way
- In practice, Scheffe and Tukey are popular

Example: Two-way ANOVA

- Two main effects (**Dose**, **Supplement**) and their interaction
- Interpret significance of model, terms, etc.
- Model validity check (check assumptions)
- Interpretation of Post-hoc test result

Two-way ANOVA example:

```
aov.res2 <- aov(Toothlength ~ Dose * Supplement , data = too
th)
```

H0: Supplement has no effect on tooth growth
$(\mu_{OJ} = \mu_{VC})$
Ha: Supplement has an effect on tooth growth
(at least one group in supplement has different mean of tooth length; $\mu_{OJ} \neq \mu_{VC}$)

```
summary(aov.res2)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Dose            2 2426.4  1213.2  92.000  < 2e-16 ***
## Supplement      1  205.4   205.4  15.572 0.000231 ***
## Dose:Supplement 2  108.3    54.2   4.107 0.021860 *
## Residuals      54  712.1    13.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

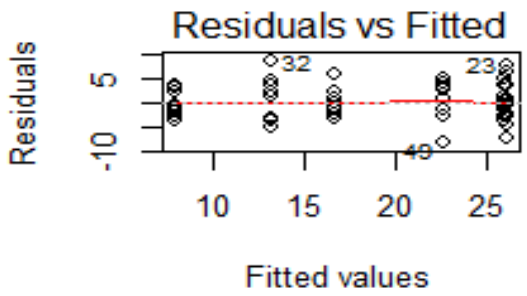
H0: no interaction between type and supplement
Ha: exist an interaction between type and supplement

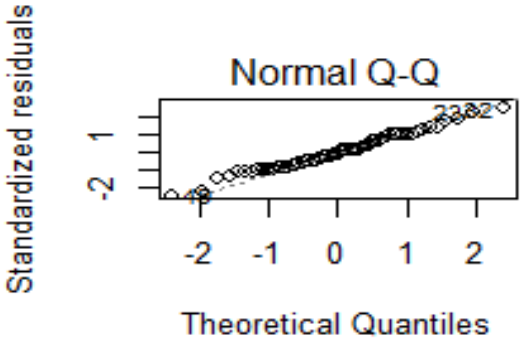
Two-way ANOVA example:

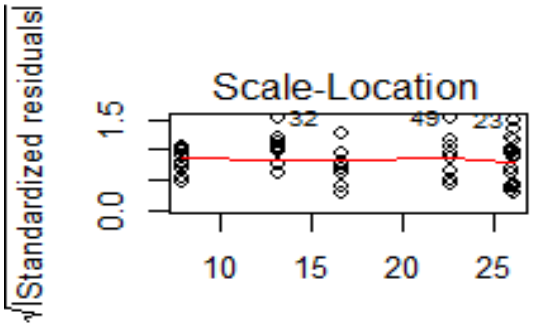
```
leveneTest(aov.res2)
```

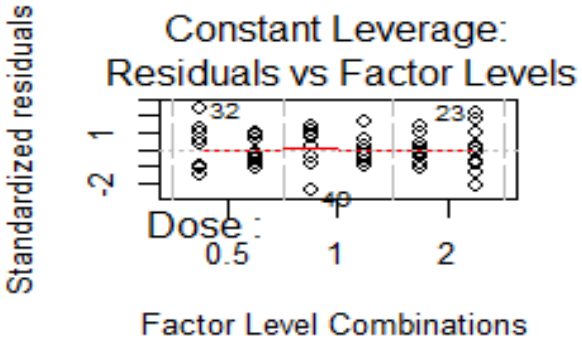
```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  5  1.7086 0.1484
##       54
```

```
par(mfrow=c(2,2))
plot(aov.res2)
```









Model validity check

- Homogeneity of variance?
- Normality?

Two-way ANOVA example:

```
lm.res2= lm(Toothlength ~ Dose * Supplement, data=tooth)
summary(lm.res2)$r.squared  # R-square
```

```
## [1] 0.7937246
```

- Compare R-square from one-way ANOVA model (Dose)
- R-square always increases as the model gets bigger (larger number of independent variables)

Two-way ANOVA example:

```
ScheffeTest(aov.res2)

##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $Dose
##         diff   lwr.ci   upr.ci    pval
## 1-0.5  9.130  5.16355 13.09645 3.8e-08 *
## 2-0.5 15.495 11.52855 19.46145 3.9e-16 *
## 2-1    6.365  2.39855 10.33145 0.00014 *
##
## $Supplement
##       diff    lwr.ci     upr.ci   pval
## VC-OJ -3.7 -6.938593 -0.4614069 0.0153 *
##
## $`Dose:Supplement`
##                 diff     lwr.ci      upr.ci    pval
## 1:OJ-0.5:OJ     9.47   3.860592  15.0794079 5.5e-05 ***
```

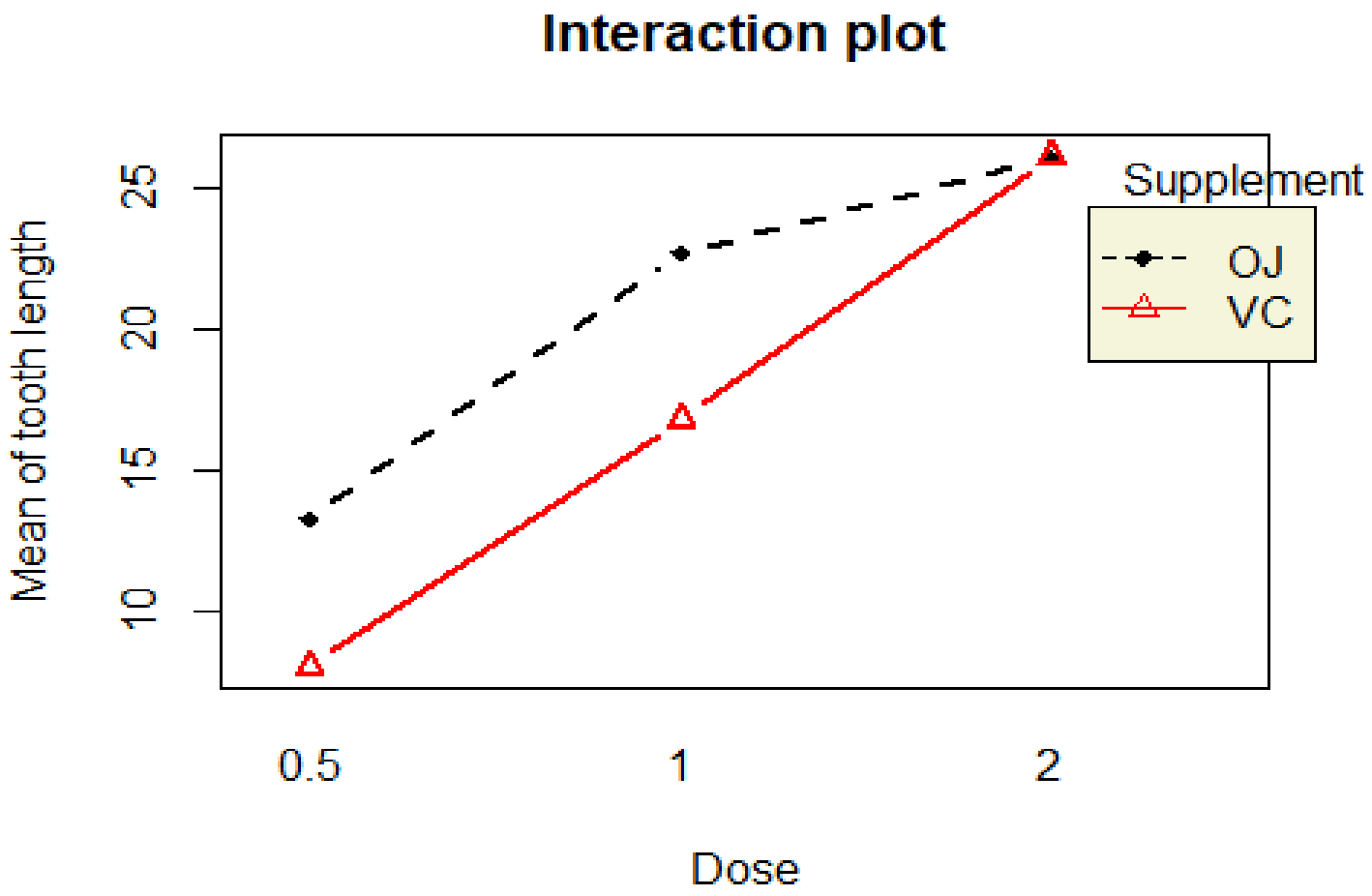
….. (omitted)

Focus on post-hoc analysis for main effects

Dose effect: Dose2>Dose1>Dose0.5

Supplement effect: VC < OJ

Two-way ANOVA example:



Some notes:

- Significance of model <-> R-square
  (0 or not) (prediction power)

- Model with higher R-square is always better?
  - What about the model with R-square = 100%?
  - What is the goal of the analysis?

- In post-hoc analysis, it can happen e.g.,
  $\mu_{0.5} = \mu_1$ and $\mu_{0.5} = \mu_2$ but $\mu_1 \neq \mu_2$

  - Why it happens and how can we make a conclusion?

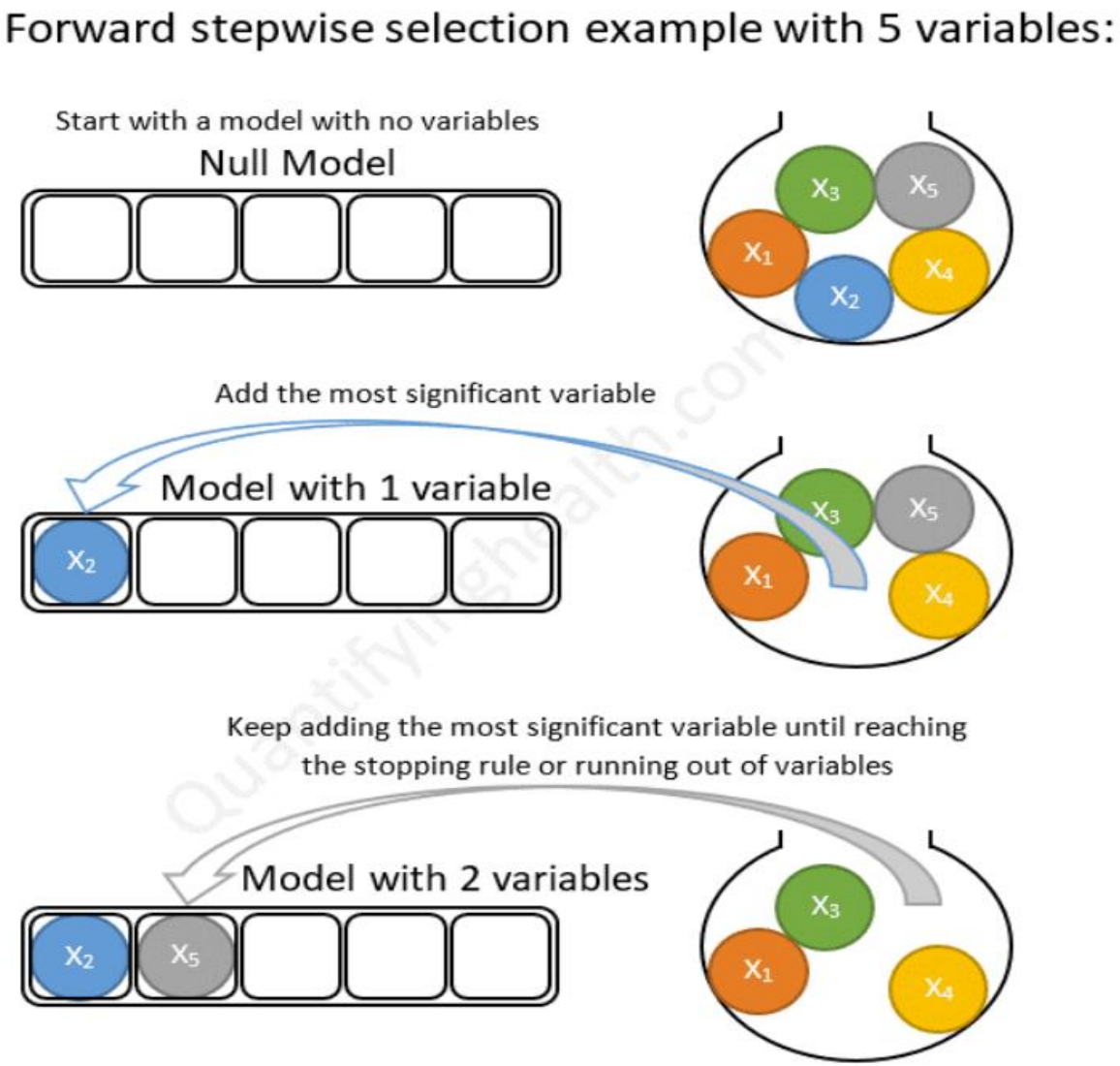
Model Selection

- For the case of n-way ANOVA, the largest model with all possible interactions has (2^n-1) terms
- How to choose the best model?

➢ Forward selection/ Backward elimination

➢ Stepwise selection (Backward + Forward)

Forward selection



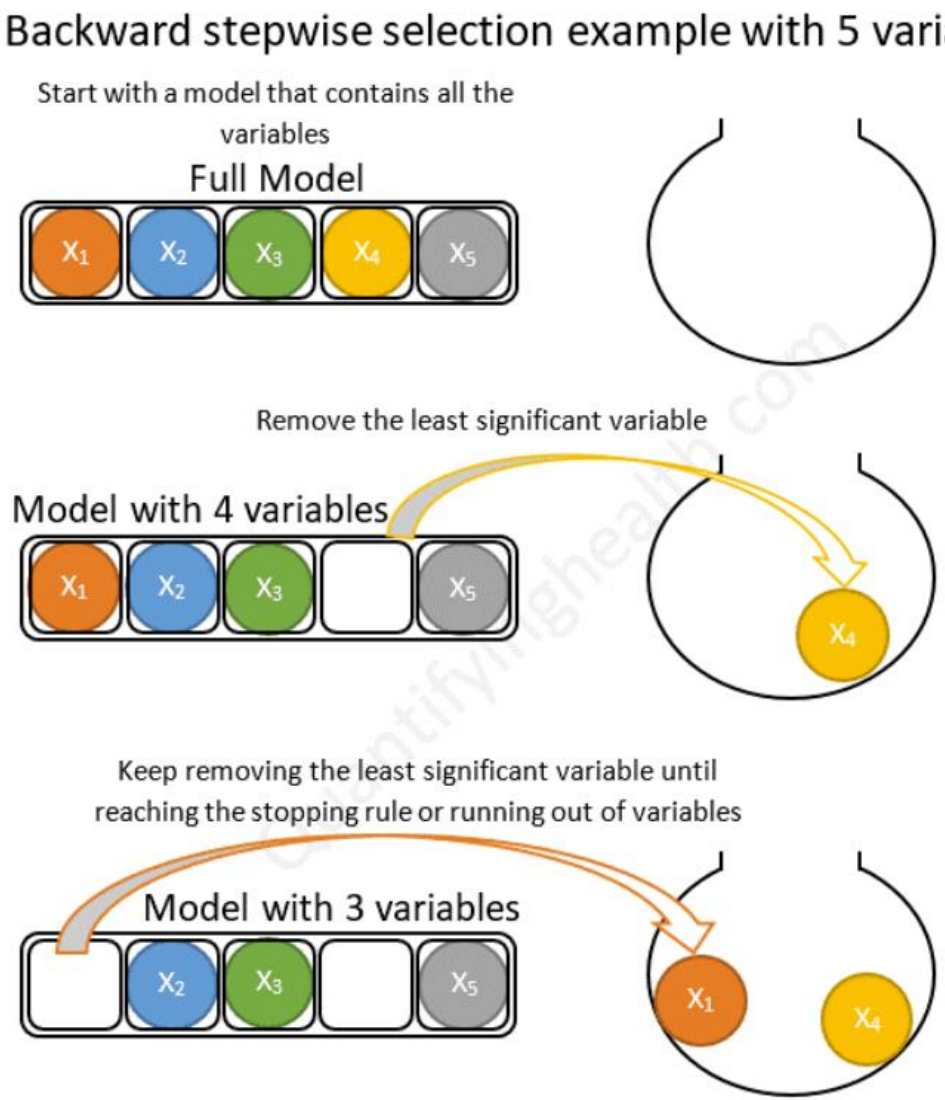
source: https://quantifyinghealth.com/stepwise-selection/

Forward selection

1. Begins with a model that contains no variables (called the Null Model)
2. Then starts adding the most significant variables one after the other
3. Until a pre-specified stopping rule is reached; Specifically, until there is no more variable which has p-value smaller than significance level (in general 0.05, but not necessarily)

- Once a variable is entered, there is no chance to be out
- Final model may include insignificant variables

Backward elimination



source: https://quantifyinghealth.com/stepwise-selection/

Backward elimination

1. Begins with a model that contains all variables under consideration (called the *Full Model*)
2. Then starts removing the least significant variables one after the other
3. Until a pre-specified stopping rule is reached – no more variable with p-value greater than significance level (0.05 but not necessarily)

- Once a variable is eliminated, there is no chance to be in
- All variables in the final model are always significant

Model Selection

- What should we do if interaction term (X1*X2) is significant but main effect (X1 or X2) is not?
  - In practice, if main effects are not significant, we do not include interaction between them even if it is significant
  1. Forward/ backward/ stepwise selection on main effect model first
  2. Test interaction among significant main effects

- Use package “MASS” in R
  - Use AIC criteria instead of p-value, but idea is the same
  - AIC will be covered in linear regression

Practice

- Using the **grass.csv** , let's start with a model that includes **Method**, **Variety**, and **Group** as independent variables and **Yield** as the response variable.

- Perform model selection
  - Backward elimination manually
  - Forward selection manually
  - Stepwise selection using stepAIC() in package "MASS"
- Find the final model from each approach