Lesson 3
Analysis of Variance

BIS 505b

Yale University
Department of Biostatistics

Date Modified: 2/23/2021

Goals for this Lesson

Addressing a Research Question

1. Comparing a quantitative variable among three or more independent groups
2. Identifying specific differences when an overall difference is found

More than Two Samples

- So far, have focused on one- and two-sample problems
  - One-sample: Comparing a single mean to a hypothesized value

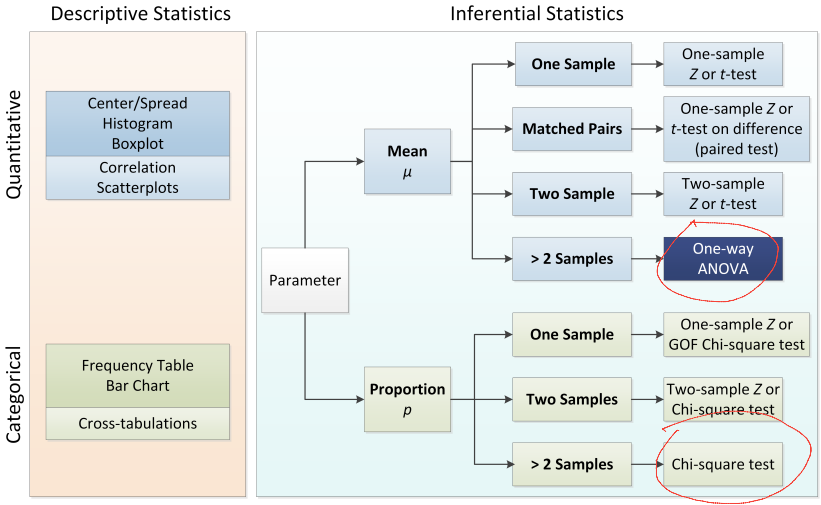
$$H_0 : \mu = \mu_0$$

  - Two-sample: Determining if a difference exists between two independent populations

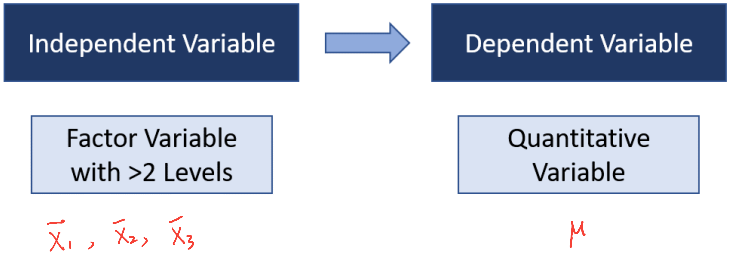
$$H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0$$

- The extension of the two-sample t-test to > 2 samples is known as Analysis of Variance (ANOVA)

More than Two Samples



One-Way ANOVA



Link to article 1
Link to article 2

Contents

1. One-Way ANOVA
   - Introduction
   - Sources of Variation
   - Types of ANOVA
2. Multiple Comparison Procedures
   - Introduction
   - Bonferroni Procedure
   - Bonus Material: Other MCPs

Progress this Unit

1. One-Way ANOVA
   - Introduction
   - Sources of Variation
   - Types of ANOVA
2. Multiple Comparison Procedures
   - Introduction
   - Bonferroni Procedure
   - Bonus Material: Other MCPs

ANOVA

- Analysis of Variance (ANOVA) is used to test the equality of $k > 2$ population means

Analysis of Variance Hypotheses a global test

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad \text{vs.}$$

H_1 : Means not all equal, or for at least one pair, $\mu_i \neq \mu_j$

- Assumptions:
  - k independent populations
  - Random samples from each of the k populations
  - Large samples ($n_i \geq 30$ for $i = 1, \ldots, k$) or Normal populations $X_i \sim N$
  - Equal population variances ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$) for central limit theorem apply

Exercise

Poll

- A recent study published in the *American Journal of Pharmaceutical Education* evaluated pharmacy students' knowledge of black box warnings for prescription drugs. Black box warnings are used to highlight potentially fatal, life-threatening, or disabling adverse effects.
- A cross-sectional survey instrument was administered to pharmacy students in their first (P1), second (P2), and third (P3) professional years at the end of the spring 2007 semester. The survey instrument assessed students' awareness of medications possessing a black box warning and familiarity with the warning content for 20 medications (15 with and 5 without).
- Mean ($\pm$ SD) number of correct responses identifying the presence or absence of a black box warning was 5.8 ± 3.3, 9.6 ± 4.0, and 14.8 ± 2.8 for the P1, P2, and P3 students, respectively ($p < 0.05$).

Link to article

Connecting Concepts: t-Test to ANOVA

- Comparing the mean of *two* samples, t-test:
  - **t-test statistic**: Function of the distance the means are apart from each other and the variability of each sample

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

$S_p^2 \rightarrow \sigma_1^2 = \sigma_2^2$

- Comparing the mean of *three or more* samples, ANOVA:
  - **ANOVA F-test statistic**:

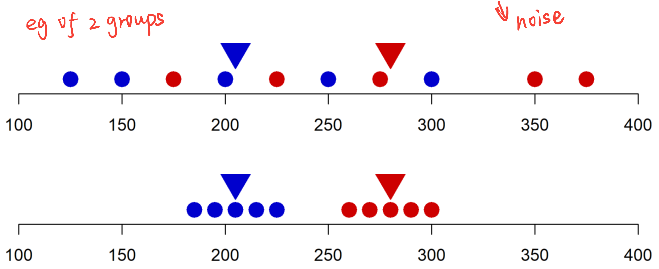
$$\text{F} = \frac{\text{Distance/variability between means}}{\text{Variability within each sample}}$$

$\bar{x}_1, \sim, \bar{x}_k$

$\sigma_1^2 = \sim = \sigma_k^2$

Why Analyze *Variance*?

- If our ultimate goal is to ascertain whether any of the group means are different, then why are we analyzing variance?
- Two types of variability:
  1. Between-groups variability (difference between the treatment means)
  2. Within-groups variability (sampling variability)

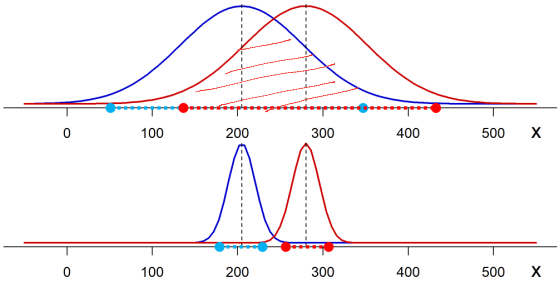


1. Relative to the sampling variability, the difference between means is **small**

2. Relative to the sampling variability, the difference between means is **big**

Variability

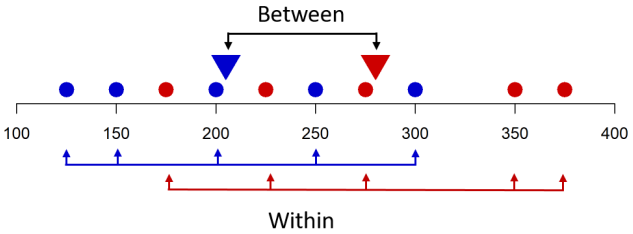
- The variability of individual samples impacts the relative difference in means



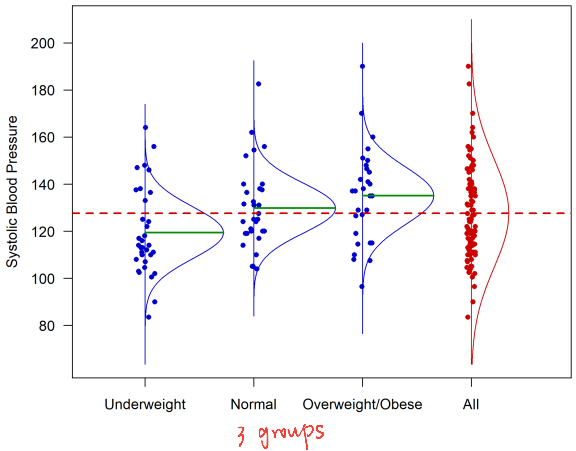
- The greater the variability of the individual samples, the less likely the population means will differ significantly more overlap
  - Extends to 3+ samples

Variability

- To summarize, we are not only concerned with
  - How far apart the sample means are from the overall mean, but
  - How far apart they are relative to the variability of individual observations
- Key: Compare the difference between treatment means (between-groups variability) to the sampling variability (within-groups variability)



Decompose Variation



- Partitioning the total variation (SS_T) in the response into variation between groups (SS_B) (treatment effect) and variation within groups (SS_W) (random variation/noise)

F-Test

- **Test statistic** to test

$$H_0 : \mu_1 = \ldots = \mu_k \quad \text{vs.} \quad H_1 : \text{At least one } \mu_i \neq \mu_j$$

is the ratio of estimates of these two measures of variability:

ANOVA F-Test Statistic

$$\text{F} = \frac{\text{variation between groups}}{\text{variation within groups}} = \frac{s_B^2}{s_W^2}$$

1. Between-groups variability (s_B^2 or MS_B): Variation of the population means about the overall mean
2. Within-groups variability (s_W^2 or MS_W): Variation of the individual values around their population means (pooled estimate of common variance, σ^2)

F-Test

ANOVA F-Test Statistic

$$\text{F} = \frac{\text{variation between groups}}{\text{variation within groups}} = \frac{s_B^2}{s_W^2}$$

- If the variability within the k populations is small relative to the variability among their means, this suggests the population means are different
  - F-statistic will be large and lead us to reject $H_0 : \mu_1 = \ldots = \mu_k$

k-Sample Problem: Notation

Population	Population Mean	Population Variance	Sample Size	Sample Mean	Sample Variance
1	μ_1	σ_1^2	n_1	$\bar{x}_1$	s_1^2
2	μ_2	σ_2^2	n_2	$\bar{x}_2$	s_2^2
$\vdots$					
k	μ_k	σ_k^2	n_k	$\bar{x}_k$	s_k^2

(standard deviation) s

- $\bar{x}_1 = \dfrac{\sum_{i=1}^{n_1} x_{1i}}{n_1}$
- $s_1^2 = \dfrac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1}$
- $N = n_1 + n_2 + \ldots + n_k$, total sample size
- In the 2-sample pooled t-test, recall $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

Estimates of Variability, s_W^2

- Estimate of **within-groups variance**, s_W^2 (MS_W) $F_{df_2}^{df_1} = \frac{S_B^2}{S_W^2} = \frac{SS_B/df_1}{SS_W/df_2}$
  - ANOVA assumes a common variance in each group, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$
  - Pooled estimate of common variance σ^2:

$$s_W^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2}{n_1 + n_2 + \ldots + n_k - k} = \frac{SS_W}{N - k} = \frac{SS_W}{df_2}$$

  where $N = n_1 + n_2 + \ldots + n_k$, total sample size

- Extension of the formula for s_p^2 in pooled t-test
- If $n_1 = n_2 = \ldots = n_k$, $s_W^2 = \frac{s_1^2 + s_2^2 + \ldots + s_k^2}{k}$

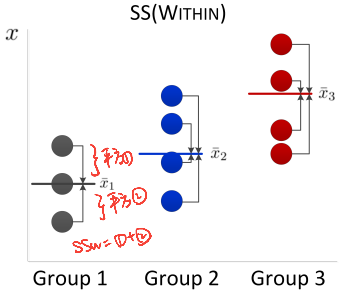
all sample size are the same

Within-Groups Sum of Squares

- **Within-groups sum of squares**:

$$SS_W = \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2 + \ldots + \sum_{i=1}^{n_k}(x_{ki} - \bar{x}_k)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_j)^2$$

  - Where $\bar{x}_j$ = sample mean in j^{th} group $(j = 1, \ldots, k)$; x_{ji}: i^{th} observation in group j



- SS_W is also known as "residual" sum of squares
- s_W^2 is also known as MSE or mean squared error

Comment on Equality of Variances

- There are formal hypothesis tests that check equality of variance assumption (Levene's or Bartlett's tests)
- Bartlett's test is sensitive to deviations from normality; Levene's more robust to deviations from normality. Both tests may be driven by sample size.

if sample is not normal

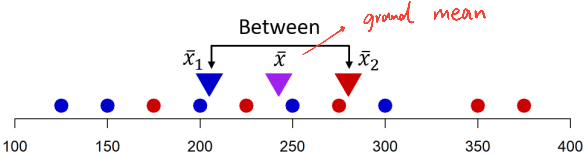
Rule of Thumb Concerning Variability for One-Way ANOVA $\frac{S_{max}}{S_{min}} < 2$

The ANOVA F-test is approximately correct when the *largest* sample standard deviation is no more than **twice as large** as the *smallest* sample standard deviation.

- When homogeneity of variance is violated, there is a greater probability of falsely rejecting the ANOVA null hypothesis
- Alternatives: Welch's ANOVA or non-parametric Kruskal-Wallis test

→ in lab3

Estimates of Variability, s_B^2



$F = \frac{SB^2}{Sw^2}$

- Estimate of **between-groups variance**, s_B^2 (MS_B)

$$s_B^2 = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots + n_k(\bar{x}_k - \bar{x})^2}{k-1} = \frac{SS_B}{k-1} = \frac{SS_B}{df_1}$$

- Where $\bar{x}$ is the **grand mean** over all observations

if know $\bar{x}_k$ use weighted mean

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots + n_k} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

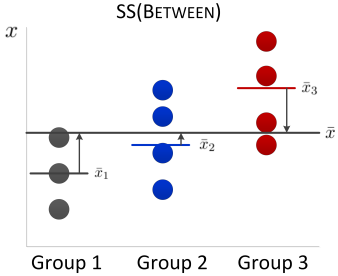
- The larger the differences in the sample means, the larger s_B^2

Between-Groups Sum of Squares

- **Between-groups sum of squares**:

$$SS_B = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_j - \bar{x})^2 = \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2 = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots + n_k(\bar{x}_k - \bar{x})^2$$

  - Where $\bar{x}_j$ = sample mean in the j^{th} group

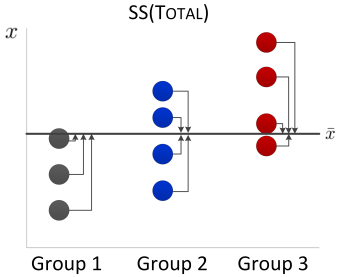


- SS_B is also known as SS(treatment) or SS(model) or "group" sum of squares

Total Variation

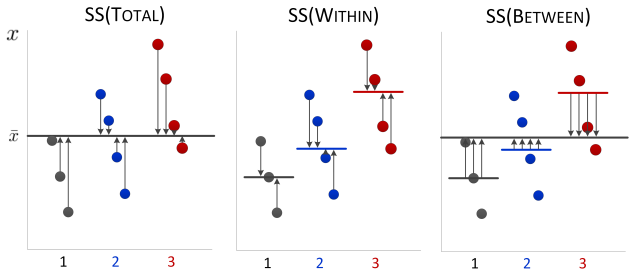
- **Total sum of squares**: $SS_T = SS_W + SS_B$

$$SS_T = \sum_{ij}(x_{ji} - \bar{x})^2 = \sum_{ij}(x_{ji} - \bar{x}_j)^2 + \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2$$



- SS_T: sum of the squared deviations of each observation from the overall mean, $\bar{x}$
- Numerator of the overall sample variance of $\{x_1, x_2, \ldots, x_N\}$

Partitioning Total Sum of Squares



$$x_{ji} - \bar{x} = (x_{ji} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

- Partition variability in the response (X) into the variability between groups and the variability within groups
- Within group variability is leftover variability in the outcome that cannot be explained by group membership

F-Test

- The two sum of squares are comparable by dividing by their degrees of freedom

ANOVA F-Test Statistic

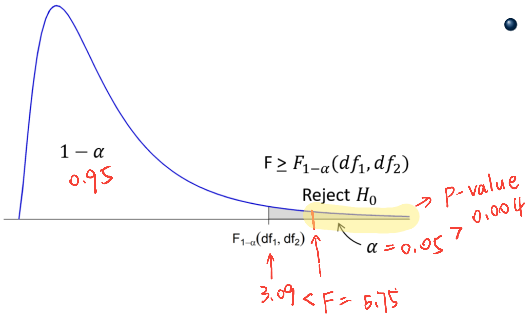
$$\mathsf{F} = \frac{s_B^2}{s_W^2} = \frac{SS_B \Big/ k-1}{SS_W \Big/ N-k} = \frac{MS_B}{MS_W}$$

- Under H_0, $\mathsf{F} \sim F(k-1, N-k)$, an F-distribution with
  - numerator degrees of freedom $df_1 = k-1$ and
  - denominator degrees of freedom $df_2 = N-k$
  - Where $N = n_1 + n_2 + \ldots + n_k$, the total sample size
    k is the number of groups

F-Distribution

- Reject $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ if the test statistic F falls in the rejection region

Figure: $F_{k-1,N-k}$ distribution



- $F = \dfrac{s_B^2}{s_W^2} \geq F_{1-\alpha}(df_1, df_2)$; Rejection region in upper tail if F is big reject H_0
- If variation between samples is large relative to variation within samples (F large), reject H_0

R Code, F-distribution

```
# Critical value, qf(1-alpha, df1, df2)
> qf(.95, df1 = 2, df2 = 91)
[1] 3.096553

# P-value, 1-pf(test stat, df1, df2)
> 1 - pf(5.75, df1 = 2, df2 = 91)
[1] 0.004450896
```

ANOVA Table

present by software

Table: ANOVA Table

Source of Variation	Sum of Squares (SS)		Degrees of Freedom (df)	Mean Squares (MS)	F
Between	SS_B	df1	$k-1$	$MS_B = s_B^2 = \frac{SS_B}{k-1}$	$\frac{MS_B}{MS_W}$
Within	SS_W	df2	$N-k$	$MS_W = s_W^2 = \frac{SS_W}{N-k}$	
Total	SS_T		$N-1$		
	additive		additive		

- Where $\bar{x}_j$ = sample mean in the j^{th} group
- N: Total sample size
- Grand mean: $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{N}$
- F $\sim F(k-1, N-k)$ under H_0

Effect Size Measure

- Effect sizes are often reported in the literature; with ANOVA, eta-squared (η^2) is used
  - Allow researchers to present the magnitude of reported effect in a standardized metric independent of scale used to measure dependent variable

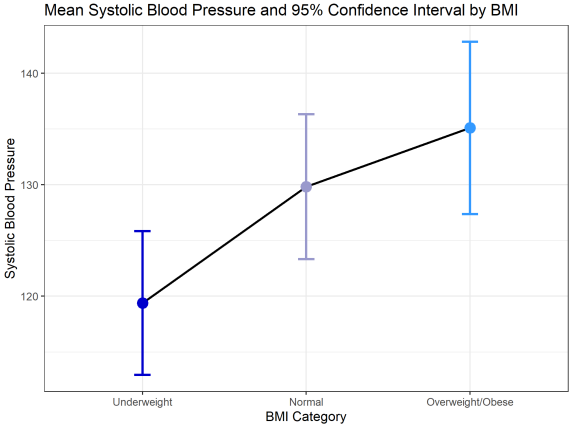
Eta-squared

$$\eta^2 = \frac{SS_B}{SS_T}$$

- 0 - .1 is a weak effect
- .1 - .3 is a modest effect
- .3 - .5 is a moderate effect
- $>$.5 is a strong effect

- η^2 - Proportion of total variation that is due to between-group differences (a.k.a. explained variation); between 0 and 1
- $\eta^2 \times 100\%$ of the total variability in X is explained by the group effect
- The other $(1 - \eta^2) \times 100\%$ remains unexplained (due to error or within-group differences)

ANOVA: Example



- Example: Goal is to determine if mean systolic blood pressure differs by BMI category (underweight, normal weight, overweight/obese), μ_1, μ_2, and μ_3. Assume systolic blood pressure is normally distributed.

ANOVA: Example

R Code, Summary Statistics

```
# Creating a function that will print several summary stats
> summze <- function(x) c(n = sum(!is.na(x)),
                          mn = mean(x, na.rm = TRUE),
                          sdev = sd(x, na.rm = TRUE),
                          varn = var(x, na.rm = TRUE))

# Summary statistics by group using aggregate() function
> aggregate(x = list(sysbp = fhs_anova$SYSBP), by = list(bmi = fhs_anova$BMIGRP2_factor),
            FUN = summze)
              bmi   sysbp.n  sysbp.mn sysbp.sdev sysbp.varn
1      Underweight  34.00000 119.38235   18.45431  340.56150
2           Normal  31.00000 129.82258   17.71513  313.82581
3 Overweight/Obese  29.00000 135.08621   20.31980  412.89409

# Summary statistics full sample
> summze(fhs_anova$SYSBP)
        n         mn       sdev       varn
 94.00000 127.67021  19.75339 390.19652
```

ANOVA: Example

3 groups

- Step 1: State the hypotheses
  - $H_0 : \mu_1 = \mu_2 = \mu_3$
  - H_1 : At least one $\mu_i \neq \mu_j$ (at least one of the population means differs from the others)
- Step 2: Specify the significance level $\alpha = 0.05$

ANOVA: Example

- Step 2.5: Check assumptions for ANOVA
  1. Random samples from each population
  2. Systolic BP approximately normally distributed or large samples ($n \geq 30$)
  3. Three populations are independent
  4. Population variances equal
     - Using the rule of thumb
       - $\frac{\text{Largest } s}{\text{Smallest } s} = \frac{s_o}{s_n} = \frac{20.32}{17.72} = 1.15 < 2$

ANOVA: Example

- Step 3: Compute the appropriate test statistic $\mathrm{F} = \frac{s_B^2}{s_W^2}; F \sim F(2, 91)$

BMI Category (j)	n_j	$\bar{x}_j$	s_j	s_j^2
Underweight (1)	34	119.38	18.45	340.56
Normal weight (2)	31	129.82	17.72	313.83
Overweight/Obese (3)	29	135.09	20.32	412.89
$N = \sum = 94$				

- Grand mean for use in calculating s_B^2: $\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3}$
  - $\bar{x} = \frac{34(119.38) + 31(129.82) + 29(135.09)}{94} = 127.67$

ANOVA: Example, s_B^2

BMI Category (j)	n_j	$\bar{x}_j$	s_j	s_j^2	$(\bar{x}_j - \bar{x})^2$
Underweight (1)	34	119.38	18.45	340.56	68.69
Normal weight (2)	31	129.82	17.72	313.83	4.63
Overweight/Obese (3)	29	135.09	20.32	412.89	55.00
	$\sum = 94$	$\bar{x} = 127.67$			

- $s_B^2 = \dfrac{SS_B}{k-1} = \dfrac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2}{k-1}$
$= \dfrac{34(68.69) + 31(4.63) + 29(55.00)}{2}$
$= \dfrac{4073.94}{2} = 2036.97$
- $SS_B = 4073.94$

ANOVA: Example, s_W^2

BMI Category (j)	n_j	$\bar{x}_j$	s_j	s_j^2
Underweight (1)	34	119.38	18.45	340.56
Normal weight (2)	31	129.82	17.72	313.83
Overweight/Obese (3)	29	135.09	20.32	412.89
	$\sum = 94$	$\bar{x} = 127.67$		

- $s_W^2 = \dfrac{SS_W}{N-k} = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2}{N-k}$
$= \dfrac{33(340.56) + 30(313.83) + 28(412.89)}{94-3}$
$= \dfrac{32,214.34}{91} = 354.00$
- $SS_W = 32214.34$

ANOVA: Example

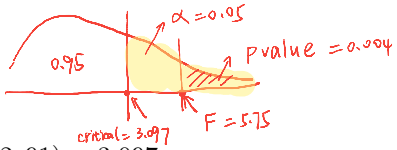
Table: ANOVA Table, $k = 3, N = 94$

Source of Variation	SS	df	MS	F
Between	$SS_B = 4073.94$	2	$MS_B = s_B^2 = \frac{4073.94}{2} = 2036.97$	5.75
Within	$SS_W = 32,214.34$	91	$MS_W = s_W^2 = \frac{32,214.34}{91} = 354$	
Total	$SS_T = 36,288.28$	93		

- Step 3: Compute the appropriate test statistic $\mathsf{F} = \frac{s_B^2}{s_W^2}; F \sim F(2, 91)$
  - $\mathsf{F} = \frac{2036.97}{354} = 5.75$

ANOVA: Example

- Step 4: Decision Role
  - Given $\alpha = 0.05$,
  - Reject H_0 if F $\geq F_{1-\alpha}(df_1, df_2) = F_{.95}(2, 91) = 3.097$ (critical value)




```
qf(.95, df1=2, df2=91)
```

- Step 5: Draw a conclusion about H_0

```
1-pf(5.75, df1=2, df2=91)
```

  - F $= 5.75$ (test statistic)
  - $F \geq 3.097 \rightarrow$ Reject H_0
  - $p = P(F \geq 5.75) = 0.0044$
  - $p \leq 0.05 \rightarrow$ Reject H_0

- Conclusion: There is evidence to reject H_0 and conclude that the mean systolic blood pressure is not equal in the three BMI categories ($p = 0.0044$)

F-Distribution Critical Values

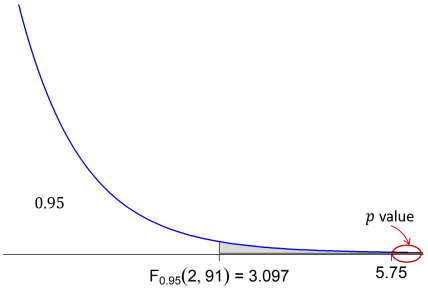
R Code, F-distribution

```
> alpha = .05    # type I error
> k = 3          # number of groups
> N = 94         # total sample size
> df.1 = k - 1
> df.2 = N - k
> teststat = 5.75

# Critical value
> qf(1 - alpha, df1 = df.1, df2 = df.2)
[1] 3.096553
> qf(.95, df1 = 2, df2 = 91)
[1] 3.096553

# P-value
> 1 - pf(teststat, df1 = df.1, df2 = df.2)
[1] 0.004450896
> 1 - pf(5.75, df1 = 2, df2 = 91)
[1] 0.004450896
```

Figure: $F(2, 91)$ Distribution



ANOVA: Example

R Code, ANOVA

```
> options(show.signif.stars = FALSE)
# ANOVA: response variable ~ *factor* group variable
> anovasys <- aov(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
> summary(anovasys)
               Df Sum Sq Mean Sq F value  Pr(>F)
BMIGRP2_factor  2   4074    2037   5.754 0.00443
Residuals      91  32214     354
```

- $s_B^2 = 2037, \quad k - 1 = 2$
- $s_W^2 = 354, \quad N - k = 91$

Pooled t-test and ANOVA

Same for 2 groups

- One-way ANOVA performed on 2 groups is equivalent to the pooled two-sample t-test assuming equal variance

One-way ANOVA	t-test assuming equal variance
$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$	$H_1 : \mu_1 \neq \mu_2$
F-statistic	t-statistic

reason: $F(1, N-2) = (t_{N-2})^2$

- Numerator df: $k - 1 = 2 - 1 = 1$ ($k-1=2-1$)
- $N = n_1 + n_2$
- Will yield identical p-values

different test statistics and p-value

Pooled t-test and ANOVA

- The t-test is more flexible
  - Can choose a one-sided alternative, can assume unequal variance (Welch's t-test)

R Code, ANOVA (1) vs. (3)

```
> summary(aov(SYSBP ~ BMIGRP2_factor,
          data = subset(fhs_anova,
           BMIGRP2_factor == "Underweight" |
           BMIGRP2_factor == "Overweight/Obese")))
                Df Sum Sq Mean Sq F value Pr(>F)
BMIGRP2_factor   1   3860    3860   10.33 0.0021
Residuals       61  22800     374
```

Note: $n_1 = 34, n_3 = 29$

$$F = 10.33 = t^2 = (-3.2135)^2$$

R Code, t-test (1) vs. (3)

```
> t.test(SYSBP ~ BMIGRP2_factor,
         data = subset(fhs_anova,
          BMIGRP2_factor == "Underweight" |
          BMIGRP2_factor == "Overweight/Obese"),
          var.equal = TRUE)

     Two Sample t-test

data:  SYSBP by BMIGRP2_factor
t = -3.2135, df = 61, p-value = 0.002096
alternative hypothesis: true difference in means
  is not equal to 0
95 percent confidence interval:
 -25.475741  -5.931967
 ...
```

One-Way ANOVA

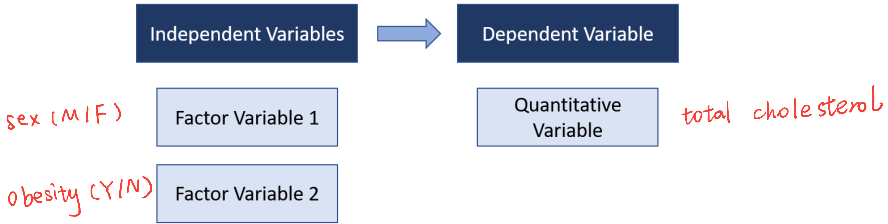
- All ANOVA methods study variables that explain the variability in the quantitative response (dependent) variable
- We will perform these analyses in a linear regression setting
- One-Way ANOVA compares the mean of a quantitative dependent variable across > 2 levels of one independent variable (1 factor)

BMI (underweight, normal, overweight)

Total cholesterol

Two-Way ANOVA

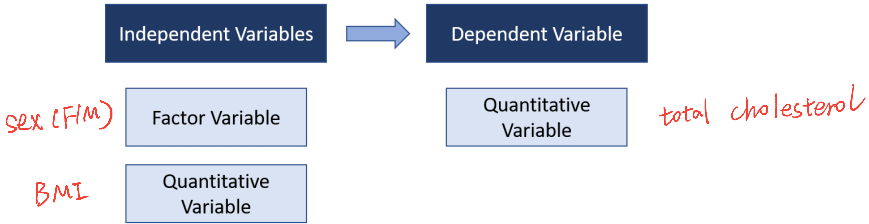
- Two-Way ANOVA compares the mean of a quantitative dependent variable across two independent variables (2 factors). Primary goal is to determine if there is an interaction between the two independent variables on the dependent variable (e.g., interaction between sex (M/F) and obesity (Y/N) on total cholesterol).



ANCOVA

do this in Regression

- Analysis of Covariance (ANCOVA) compares a quantitative response variable by both a factor and a quantitative independent variable (e.g. comparing total cholesterol (dependent variable) by both sex (M/F) and quantitative BMI)
- Generally use ANCOVA to compare a quantitative dependent variable (e.g., total cholesterol) by levels of a factor variable (e.g., sex), controlling for a quantitative covariate (e.g., BMI)



Progress this Unit

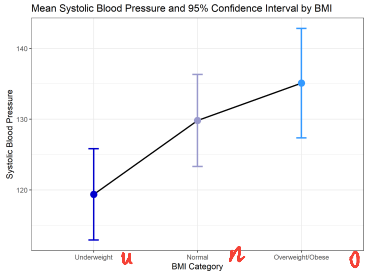
1. One-Way ANOVA
   - Introduction
   - Sources of Variation
   - Types of ANOVA
2. Multiple Comparison Procedures
   - Introduction
   - Bonferroni Procedure
   - Bonus Material: Other MCPs

Introduction

- When we reject $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$, we say there is evidence that the means are not all equal (or at least one pair of means are not equal)
- ANOVA lets us detect when at least two groups have different underlying means, but does not let us determine which of the groups have means that differ from each other
- The usual practice is to:
  1. Perform the overall F-test
  2. If H_0 is rejected, then specific groups are compared (referred to as *post hoc* tests)

ANOVA: Example

- In our example, we saw that mean systolic blood pressure was not equal in the three BMI categories
- *Which* pairs are significantly different?
  - $H_0 : \mu_u = \mu_n$ vs. $H_1 : \mu_u \neq \mu_n$
  - $H_0 : \mu_u = \mu_o$ vs. $H_1 : \mu_u \neq \mu_o$
  - $H_0 : \mu_n = \mu_o$ vs. $H_1 : \mu_n \neq \mu_o$



Pairwise Tests

- $\bar{X}_j \sim N\left(\mu_j, \frac{\sigma^2}{n_j}\right)$ and $\bar{X}_{j'} \sim N\left(\mu_{j'}, \frac{\sigma^2}{n_{j'}}\right)$
  - Because the samples are independent, $\bar{X}_j - \bar{X}_{j'} \sim N\left[\mu_j - \mu_{j'}, \sigma^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)\right]$ where $\sqrt{\sigma^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$ is the standard error (SE) of the difference in means

Under H_0, $\mu_j = \mu_{j'}$

$$\bar{X}_j - \bar{X}_{j'} \sim N\left[0, \sigma^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)\right]$$

Pairwise Tests

- If σ^2 were known, could divide by the standard error to obtain the test statistic,

Under H_0,

$$Z = \frac{\bar{X}_j - \bar{X}_{j'}}{\sqrt{\sigma^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}} \sim N(0,1)$$

Pairwise Tests

- How should σ^2 be estimated?
  - In a pooled t-test, $s^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
- Recall, the ANOVA assumption that underlying variance of each group is the same
  - In one-way ANOVA, there are k sample variances. Similar approach used to estimate σ^2.
  - Extending weighted average of k individual sample variances, gives s_w^2

use s_w^2 in one-way ANOVA $\rightarrow$ σ^2

Pooled Estimate of Variance for One-Way ANOVA

$$s_w^2 = \frac{\sum_{j=1}^{k}(n_j-1)s_j^2}{\sum_{j=1}^{k}(n_j-1)} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{n_1+n_2+\ldots+n_k-k} = \frac{SS_W}{N-k} = \text{MSE}$$

Pairwise Tests

BMI Category	Underweight	Normal weight	Overweight/Obese
$\bar{x}$	119.38	129.82	135.09
n	34	31	29

$s_w^2 = 354$

Standard Error for Pairwise Comparisons

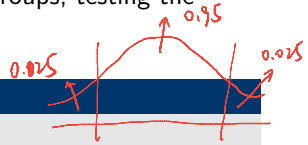
$$\hat{\text{SE}} = \sqrt{s_w^2 \left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)} = s_w \sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}$$

- $\hat{\text{SE}}(\bar{X}_u - \bar{X}_n) = \sqrt{354\left(\frac{1}{34} + \frac{1}{31}\right)} = 4.67$

Comparison	$\bar{x}_j - \bar{x}_{j'}$	$\text{SE}(\bar{X}_j - \bar{X}_{j'})$
U vs. N	-10.44	4.67
U vs. O	-15.70	4.76
N vs. O	-5.26	4.86

LSD Procedure: Test Statistic and CI

- Procedure below that performs pairwise t-test following 1-way ANOVA is referred to as Fisher's least significant difference (LSD) method
- To compare two specific groups (e.g., group j and j') among k groups, testing the hypothesis $H_0 : \mu_j = \mu_{j'}$ vs. $H_1 : \mu_j \neq \mu_{j'}$



Pairwise Tests: LSD Procedure

- Compute pooled estimate of variance $s^2 = s_w^2$ (i.e., MSE)
- Compute test statistic, $t = \dfrac{\bar{x}_j - \bar{x}_{j'}}{\sqrt{s_w^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}}$ with decision rules based on t_{N-k}

consistent with S_w^2 ($df = N-k$)

- $100(1-\alpha)\%$ CI for $\mu_j - \mu_{j'}$ is given by $\bar{x}_j - \bar{x}_{j'} \pm t_{N-k,1-\frac{\alpha}{2}}\sqrt{s_w^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$

$1-\frac{\alpha}{2} = 0.975$

LSD Procedure: Least Significant Difference

- Rearranging the test statistic gives the smallest difference in sample means that will be statistically significant, known as the least significant difference

Least Significant Difference: LSD Procedure

$$\text{if} \quad |\bar{x}_j - \bar{x}_{j'}| \geq t_{N-k,1-\frac{\alpha}{2}} \sqrt{s_w^2 \left(\frac{1}{n_j} + \frac{1}{n_{j'}} \right)}$$

reject H_0

LSD Procedure: Example

same t distribution t_{91} — Critical value $t_{91,.975}$

	Comparison	$\bar{x}_j - \bar{x}_{j'}$	$\hat{SE}$	t-Statistic	p (vs. α)	LSD
1	U vs. N	-10.44	4.67	$t = \frac{119.38 - 129.82}{\sqrt{354\left(\frac{1}{34} + \frac{1}{31}\right)}} = -2.23$	0.028	9.28
2	U vs. O	-15.70	4.76	$t = \frac{119.38 - 135.09}{\sqrt{354\left(\frac{1}{34} + \frac{1}{29}\right)}} = -3.30$	0.0014	9.45
3	N vs. O	-5.26	4.86	$t = \frac{129.82 - 135.09}{\sqrt{354\left(\frac{1}{31} + \frac{1}{29}\right)}} = -1.08$	0.28	9.66

$t_{91,.975} = 1.986$

$s_w^2 = 354$, $N - k = 91$, $t_{91,.975} = 1.986$

- LSD comparing U to N: $t_{91,.975} \times \hat{SE} = 1.986 \times 4.67 = 9.28$; $|-10.44| > 9.28$, supporting a significant difference

LSD Procedure: Example

R Code, LSD

```
# LSD test p-values (specify "none")
> pairwise.t.test(fhs_anova$SYSBP,
                  fhs_anova$BMIGRP2_factor,
                  p.adjust.method = "none")

  Pairwise comparisons using t tests with
    pooled SD

data:  fhs_anova$SYSBP and
    fhs_anova$BMIGRP2_factor

                  Underweight Normal
Normal            0.0279      -
Overweight/Obese  0.0014      0.2817

P value adjustment method: none
```

outcome
factor
no adjustment
p value
*
*

- On average, underweight and normal weight individuals and underweight and overweight/obese individuals have significantly different systolic BP

LSD Procedure: Example

R Code, LSD

```
# Install and load required package
> library(lsmeans)
# Requires model formulation of dependent variable ~ factor variable
> modelsys <- lm(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
# Pairwise tests and p-values
> lsdresults <- lsmeans(modelsys,
                        pairwise ~ BMIGRP2_factor,
                        adjust = "none")   # specify "none" for LSD test
> lsdresults$contrasts
 contrast                          estimate   SE df t.ratio p.value
 Underweight - Normal               -10.44 4.67 91 -2.234  0.0279
 Underweight - (Overweight/Obese)   -15.70 4.76 91 -3.302  0.0014
 Normal - (Overweight/Obese)         -5.26 4.86 91 -1.083  0.2817
```

$\bar{x}_j - \bar{x}_{j'}$ t-test statistic

Problem of Multiple Comparisons

- You could perform all $c = \binom{k}{2}$ or $\frac{k(k-1)}{2}$ t-tests, but performing multiple hypothesis tests increases the chance of making a Type I error (falsely rejecting H_0)
- In particular, the familywise error rate, or the probability of making at least 1 Type I error when performing c tests is equal to:

Familywise Error

$$\text{FW ER} = 1 - (1 - \alpha)^c$$

  - Where each of the c tests is performed at the α-level

Familywise Error

Table: Familywise error for c tests, each conducted at the $\alpha = 0.05$-level

c ↑	FW ER ↑
1	0.050
2	0.098
3	0.143
4	0.185
5	0.226
10	0.401
15	0.537
20	0.642
25	0.723
50	0.923
100	0.994

- Perform 10 hypothesis tests: probability of committing at least one type I error = 40.1%
- Would like a testing procedure in which the overall probability of making a type I error = α 5%

Familywise Error

- Goal: To maintain overall α-level, or chance of making a type I error
- Multiple comparison procedures are used to control the familywise error rate
- Ensure the overall probability of declaring any significant differences between pairs of groups is maintained at some fixed significance level (e.g., α)
- Simplest and most widely-used multiple comparison procedure is the Bonferroni procedure

Bonferroni Procedure

- Idea behind Bonferroni adjustment method is simple: By reducing the significance level used for each of the individual tests appropriately, can control the familywise error

Bonferroni Adjustment

If c post-hoc tests are conducted, perform each test at the level:

$$\alpha^* = \frac{\alpha}{c}$$

- A limitation of Bonferroni procedure is that it is very conservative, resulting in low power when the number of tests performed is large

Table: Bonferroni adjustment (α^*) and familywise error for c tests, each conducted at the $\alpha = 0.05$-level

c	$\alpha^* = \frac{0.05}{c}$	FW ER
1	0.0500	0.05000
2	0.0250	0.04938
3	0.0167	0.04917
4	0.0125	0.04907
5	0.0100	0.04901
10	0.0050	0.04889
15	0.0033	0.04885
20	0.0025	0.04883
25	0.0020	0.04882
50	0.0010	0.04879
100	0.0005	0.04878

Bonferroni Procedure

Same as LSD

- Just as in the LSD procedure, the Bonferroni procedure performs two-sample t-tests for each pairwise comparison of interest, using the pooled sample variance (s_w^2) as an estimate of σ^2
- Difference: α-level assumed for the test (α^*)

Pairwise Tests: Bonferroni Procedure

- Compute pooled estimate of variance $s^2 = s_w^2$ (i.e., MSE)
- Compute test statistic, $t = \dfrac{\bar{x}_j - \bar{x}_{j'}}{\sqrt{s_w^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}}$ with decision rules based on $\boxed{t_{N-k}}$

↑ ANOVA

- $100(1-\alpha)\%$ simultaneous CIs for $\mu_j - \mu_{j'}$: $\bar{x}_j - \bar{x}_{j'} \pm t_{N-k,1-\frac{\alpha^*}{2}}\sqrt{s_w^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$

Bonferroni Procedure

- Mean response is significantly different at a FW ER $= \alpha$ in groups j and j' if:

Minimum Difference for Significance: Bonferroni Procedure

$$|\bar{x}_j - \bar{x}_{j'}| \geq t_{N-k,1-\frac{\alpha^*}{2}} \sqrt{s_w^2 \left(\frac{1}{n_j} + \frac{1}{n_{j'}} \right)}$$

Decisions using the Bonferroni Procedure

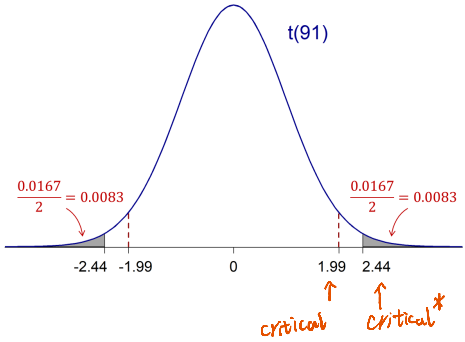
2 options

- **Option 1**:
  - Reject each individual hypothesis if p-value $< \alpha^*$
  - $\alpha^* = \dfrac{\alpha}{c}$
  - For example, if performing $c = 3$ pairwise hypothesis tests, $p < \dfrac{0.05}{3} = 0.0167$
- **Option 2**: often use — Reject H_0 if $p^* < \alpha$
  - Calculate a Bonferroni-adjusted p-value (p^*) and compare to α
  - $p^* = \min(p \times c, 1)$
  - For example, observe $p = 0.01$, then $p^* = 0.01 \times 3 = 0.03 < 0.05$ (reject H_0)

Bonferroni Critical Value: Example

R Code, $t_{df,1-\frac{\alpha^*}{2}}$ vs. $t_{df,1-\frac{\alpha}{2}}$

```
> alpha = 0.05          # type I error
> k = 3                 # number of groups
> N = 94                # total sample size
> c = k*(k-1)/2         # pairwise comparisons
> alphastar = 0.05/c    # Bonf-adjusted alpha
# With Bonferroni
> tcritstar = qt(1 - alphastar/2, df = N - k)
# Without Bonferroni
> tcrit = qt(1 - alpha/2, df = N - k)
> c(tcritstar, tcrit)
[1] 2.439040 1.986377
```



- At the 2-sided $\alpha^* = 0.0167$-level, for t_{91}, critical value $= t_{91,.992} = 2.44$
- At the 2-sided $\alpha = 0.05$-level, for t_{91}, critical value $= t_{91,.975} = 1.99$

Bonferroni Procedure: Example

- Assume all $k(k-1)/2 = 3(2)/2 = 3$ pairwise tests among $c = 3$ groups will be performed
- Significance level for each test: $\alpha^* = 0.05/3 = 0.0167$

Comparison	$\bar{x}_j - \bar{x}_{j'}$	$\hat{\text{SE}}$	t-Statistic	p (vs. α^*)	p^* (vs. α)	Min. Diff
U vs. N	-10.44	4.67	-2.23	0.028	0.084	11.40
U vs. O	-15.70	4.76	-3.30	0.0014	0.0041	11.60
N vs. O	-5.26	4.86	-1.08	0.28	0.85	11.86

- Minimum significant difference comparing U to N: $t_{91,.992} \times \hat{\text{SE}} = 2.439 \times 4.67 = 11.40$; $|-10.44| < 11.40$, supporting no significant difference

```
2*(1-pt(abs(-2.23), df=91)) = 0.028
```

- $t_{91,.998} = 2.439$

```
qt(1-(.05/3)/2, df=91) = 2.439
```

Bonferroni Procedure: Example

R Code, Bonferroni

```
# Bonferroni-adjusted p-values (p*)
> pairwise.t.test(fhs_anova$SYSBP,
            fhs_anova$BMIGRP2_factor,
          p.adjust.method = "bonferroni")

  Pairwise comparisons using t tests with
    pooled SD

data:  fhs_anova$SYSBP and
    fhs_anova$BMIGRP2_factor

                Underweight Normal
Normal          0.0837      -
Overweight/Obese 0.0041 *    0.8452

P value adjustment method: bonferroni
```

p* vs α = 0.05

- On average, ~~normal weight~~ underweight and overweight/obese individuals have significantly different systolic BP (Bonferroni-adjusted p-value $= 0.0041 < 0.05$)

Bonferroni Procedure: Example

R Code, Bonferroni

```
# Install and load required package
> library(lsmeans)

# Pairwise tests and Bonferroni-adjusted p-values
> modelsys <- lm(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
> bonresults <- lsmeans(modelsys,
                        pairwise ~ BMIGRP2_factor,
                        adjust = "bonferroni")

> bonresults$contrasts
 contrast                        estimate   SE df t.ratio p.value
 Underweight - Normal             -10.44 4.67 91 -2.234  0.0837
 Underweight - (Overweight/Obese) -15.70 4.76 91 -3.302  0.0041*
 Normal - (Overweight/Obese)       -5.26 4.86 91 -1.083  0.8452

P value adjustment: bonferroni method for 3 tests
```

same as pairwise.t.test

$\bar{x}_i - \bar{x}_{i^*}$ p^*

Bonferroni Procedure: Example

R Code, Bonferroni

```
# Install and load required package
> library(DescTools)

# Bonferroni simultaneous CIs and adjusted p-values
> anovasys <- aov(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
> PostHocTest(anovasys, method = "bonferroni")

  Posthoc multiple comparisons of means : Bonferroni
    95% family-wise confidence level

$BMIGRP2_factor
                                 diff     lwr.ci    upr.ci   pval
Normal-Underweight           10.440228 -0.9559496 21.83641 0.0837 .
Overweight/Obese-Underweight 15.703854  4.1039433 27.30376 0.0041 **
Overweight/Obese-Normal       5.263626 -6.5918391 17.11909 0.8452
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CI

Bonferroni Procedure: Example (U vs. N)

- Working through one pairwise test assuming all $c = 3$ tests will be performed
- Step 1: State the hypotheses
  - $H_0 : \mu_U = \mu_N$
  - $H_1 : \mu_U \neq \mu_N$
- Step 2: Specify the significance level $\alpha^* = \frac{0.05}{3} = 0.0167$
- Step 3: Compute the appropriate test statistic $T \sim t_{91}$
  - U vs. N: $t = \frac{\bar{x}_U - \bar{x}_N}{\sqrt{s_w^2\left(\frac{1}{n_U} + \frac{1}{n_N}\right)}} = \frac{119.38 - 129.82}{\sqrt{354\left(\frac{1}{34} + \frac{1}{31}\right)}} = \frac{-10.44}{4.67} = -2.23$

Bonferroni Procedure: Example (U vs. N)

- Step 4: Generate the decision rule
  - Critical value that $|t|$ must exceed to reject $H_0 = t_{N-k,1-\frac{\alpha^*}{2}}$
  - Reject H_0 if $|t| \geq t_{N-k,1-\frac{\alpha^*}{2}} = t_{94-3,1-\frac{0.0167}{2}} = t_{91,0.992} = 2.439$ `qt(1-(.05/3)/2, df=91)`

critical value

- Step 5: Draw a conclusion about H_0 `2*(1-pt(abs(-2.23), df=91))`
  - $t = -2.23$
  - $|t|$ not $\geq 2.439 \rightarrow$ Fail to reject H_0
  - $p = 2 \times P(T \geq 2.23) = 0.028$ (P value)
  - p not $\leq 0.0167 \rightarrow$ Fail to reject H_0 (α^*)
- Conclusion: Fail to reject the null hypothesis that the mean systolic BP is the same in the underweight and normal weight groups

Bonferroni Procedure: Example (U vs. N)

- Difference in means: $\bar{x}_U - \bar{x}_N = 119.38 - 129.82 = -10.44$
- Critical value: $t_{94-3,1-\frac{0.0167}{2}} = t_{91,0.992} = 2.439$
- Standard error: $\sqrt{s_w^2\left(\frac{1}{n_U} + \frac{1}{n_N}\right)} = \sqrt{354\left(\frac{1}{34} + \frac{1}{31}\right)} = 4.67$
- 95% simultaneous CI that controls family-wise error:

$$\bar{x}_U - \bar{x}_N \pm t_{N-k,1-\frac{\alpha^*}{2}}\sqrt{s_w^2\left(\frac{1}{n_U} + \frac{1}{n_N}\right)} = -10.44 \pm 2.439(4.67) = (-21.84, 0.96)$$

include 0

- *Note*: $10.44 < 2.439 \times 4.67 = 11.40$ (minimum significant difference in this comparison)

Bonferroni Procedure

- Can apply Bonferroni procedure for a fixed number of pre-planned tests (i.e., not required to perform all $c = k(k-1)/2$ pairwise tests)
- When performing r tests, $\alpha^* = \alpha/r$
- Reducing number of comparisons gives greater power while maintaining FW ER $< \alpha$

Exercise

Poll

- A study was conducted in patients with rheumatoid arthritis (RA) with functional disability. Outcomes of multidisciplinary rehabilitation for RA patients conducted in one Dutch rheumatology clinic were studied in 4 patient cohorts (1, 1992; 2a, 2001; 2b, 2003 and 3, 2008). The time periods correspond with the methotrexate era (1991-2000) and the biologic era (after 2000).
- Clinical assessment included HAQ score, which measures functional ability (range from 0–3.0, in 0.1 increments where higher scores indicate worse function and greater disability). HAQ scores at admission, discharge, and change scores were compared among the cohorts using one-way ANOVA with post hoc multiple comparisons using Bonferroni correction.

Study Year	1 1992–1993	2a 2001	2b 2003	3 2008–2009	*P*-value
HAQ admission (0–3), mean (s.d.)[d]	1.94 (0.74)	1.40 (0.74)*	1.39 (0.66)*	1.49 (0.59)*	0.00***
HAQ discharge (0–3), mean (s.d.)[d]	1.71 (0.78)	1.21 (0.62)*	1.22 (0.62)*	1.27 (0.69)*	0.00***
HAQ change scores[d]	0.21 (0.50)**	0.17 (0.49)**	0.15 (0.37)**	0.25 (0.46)**	0.69

[a]One-way ANOVA with post hoc Bonferroni correction between studies 2a, 2b, and 3. [b]Kruskall–Wallis test between studies 1, 2a, 2b, and 3. [c]Including paracetamol. [d]One-way ANOVA with post hoc Bonferroni correction between studies 1, 2a, 2b and 3. *Significant difference ($P < 0.05$) with study 1, one-way ANOVA after post hoc Bonferroni. **Significant change ($P < 0.05$) in HAQ score within the study; *t*-test for paired samples. ***Statistically significant difference ($P < 0.05$) between the studies 1, 2a, 2b and 3.

Tukey-Kramer Procedure

Bonus Material

- Another popular multiple comparison procedure for performing all pairwise comparisons is the Tukey-Kramer procedure
- Controls FW ER exactly at α for balanced design (i.e., $n_1 = n_2 = \ldots = n_k = n$) and approximately at α for unbalanced design
- More powerful than Bonferroni for many pairwise comparisons (Bonferroni usually better than Tukey for a small number of planned (i.e., pre-specified) comparisons)
- Most acceptable general method for all pairwise comparisons

Tukey-Kramer Procedure

- For a balanced design, we see

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{2s_w^2}{n}}} \leq \frac{\bar{x}_{\max} - \bar{x}_{\min}}{\sqrt{\frac{2s_w^2}{n}}} = \frac{q}{\sqrt{2}}$$

- where $q = \dfrac{\bar{x}_{\max} - \bar{x}_{\min}}{\sqrt{s_w^2/n}}$ follows a studentized range distribution, $q_{k,N-k}$
  - $k =$ number of groups
  - $N - k = df$ for s_w^2

Studentized Range Distribution

- Studentized range distribution is a continuous probability distribution over a non-negative range
- All of the α is in the upper tail of the distribution, thus, $1-\alpha$ level instead of $1-\alpha/2$ is used to determine critical values
- Critical value $= \frac{q_{k,N-k,1-\alpha}}{\sqrt{2}}$; p-value $= P(|t|\sqrt{2} \geq q_{k,N-k})$

R Code: Studentized Range Distribution

```
# Critical value, alpha = 0.05, 3 groups, 91 df for sw^2
> qtukey(0.95, 3, 91)/sqrt(2)
[1] 2.382662
# p-value if observe test statistic |t|=2.234, 3 groups, 91 df for sw^2
> 1 - ptukey(2.234*sqrt(2), 3, 91)
[1] 0.07087636
```

Tukey-Kramer Procedure

- Tukey-Kramer procedure generalizes Tukey procedure for unequal sample sizes

Pairwise Tests: Tukey-Kramer Procedure

- Compute pooled estimate of variance $s^2 = s_w^2$ (i.e., MSE)
- Compute test statistic, $t = \dfrac{\bar{x}_j - \bar{x}_{j'}}{\sqrt{s_w^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}}$ with decision rules based on $q_{k,N-k}$

  Same t test — different q
- Reject H_0 of $\mu_a = \mu_b$ vs. $\mu_j \neq \mu_{j'}$ if $|t| \geq \dfrac{q_{k,N-k,1-\alpha}}{\sqrt{2}}$
- $100(1-\alpha)\%$ simultaneous CIs for $\mu_j - \mu_{j'}$: $\bar{x}_j - \bar{x}_{j'} \pm \dfrac{q_{k,N-k,1-\alpha}}{\sqrt{2}}\sqrt{s_w^2\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$

Tukey-Kramer Procedure: HSD

- Mean response is significantly different at a FW ER $= \alpha$ in groups j and j' if:

Honestly Significant Difference (HSD): Tukey-Kramer Procedure

$$|\bar{x}_j - \bar{x}_{j'}| \geq \frac{q_{k,N-k,1-\alpha}}{\sqrt{2}} \sqrt{s_w^2 \left(\frac{1}{n_j} + \frac{1}{n_{j'}} \right)}$$

vs. LSD: $t_{N-k,\, 1-\frac{\alpha^*}{2}}$

- This difference is referred to as Tukey's Honestly Significant Difference (Tukey's HSD)

Tukey-Kramer Procedure: Example

Comparison	$\bar{x}_j - \bar{x}_{j'}$	$\hat{\text{SE}}$	t-Statistic	p (vs. α)	HSD
U vs. N	-10.44	4.67	-2.23	0.071	11.13
U vs. O	-15.70	4.76	-3.30	0.0039	11.33
N vs. O	-5.26	4.86	-1.08	0.53	11.58

- HSD comparing U to N: $\frac{q_{3,91,.95}}{\sqrt{2}} \times \hat{\text{SE}} = 2.383 \times 4.67 = 11.13$; $|-10.44| < 11.13$, supporting no significant difference

```
1-ptukey(2.23*sqrt(2),3,91) = 0.071
```

- $q_{3,91,0.95}/\sqrt{2} = 2.383$

```
qtukey(0.95,3,91)/sqrt(2) = 2.383
```

Tukey-Kramer Procedure: Example

R Code, Tukey-Kramer

```
# Requires lsmeans package

# Pairwise tests and Tukey-Kramer-adjusted p-values
> modelsys <- lm(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
> tukeyresults <- lsmeans(modelsys,
                          pairwise ~ BMIGRP2_factor,
                          adjust = "tukey")

> tukeyresults$contrasts
 contrast                          estimate   SE df t.ratio p.value
 Underweight - Normal                -10.44 4.67 91 -2.234  0.0708
 Underweight - (Overweight/Obese)    -15.70 4.76 91 -3.302  0.0039
 Normal - (Overweight/Obese)          -5.26 4.86 91 -1.083  0.5271

P value adjustment: tukey method for comparing a family of 3 estimates
```

Tukey-Kramer Procedure: Example

R Code, Tukey-Kramer

```
# Tukey-Kramer simultaneous CIs and adjusted p-values
> anovasys <- aov(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
> TukeyHSD(anovasys)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SYSBP ~ BMIGRP2_factor, data = fhs_anova)

$BMIGRP2_factor
                                diff        lwr      upr     p adj
Normal-Underweight            10.440228 -0.6925308 21.57299 0.0708044
Overweight/Obese-Underweight  15.703854  4.3720714 27.03564 0.0038899
Overweight/Obese-Normal        5.263626 -6.3178039 16.84506 0.5270810
```

Contrasts

- Post hoc comparisons are often pairwise comparisons between the factor level means; however, we can also draw inferences for a linear combination of the means
- A linear combination (a.k.a., contrast) is anything of the form,

Linear Contrast

$$L = \sum_j c_j \mu_j$$

where c_j are constants that sum to zero, $\sum_j c_j = 0$

- A pairwise comparison is also a contrast

Contrasts

- General contrast: $L = \sum_j c_j \mu_j$

	Comparison	Re-written	Constants
Pairwise comparison	$H_0 : \mu_1 = \mu_2$	$\mu_1 - \mu_2 = 0$	$c_1 = 1,\ c_2 = -1,\ (c_3 = 0)$
Contrast	$H_0 : \frac{\mu_1 + \mu_2}{2} = \mu_3$	$0.5\mu_1 + 0.5\mu_2 - \mu_3 = 0$	$c_j = \{0.5, 0.5, -1\}$

weight

- Contrast above: Average of the means in groups 1 and 2 vs. the mean in group 3

Contrasts: $L = \sum_j c_j \mu_j$

Unbiased Estimate of L

$$\hat{L} = \sum_j c_j \bar{X}_j$$

SE of $\hat{L}$

$$\hat{\mathsf{SE}}(\hat{L}) = \sqrt{s_w^2 \sum_j \frac{c_j^2}{n_j}}$$

- Contrast $\frac{\mu_1 + \mu_2}{2} - \mu_3 = 0$ is estimated by $0.5\bar{X}_1 + 0.5\bar{X}_2 - \bar{X}_3$
- $\hat{\mathsf{SE}}\left(0.5\bar{X}_1 + 0.5\bar{X}_2 - \bar{X}_3\right) = \sqrt{\frac{0.5^2 s_w^2}{n_1} + \frac{0.5^2 s_w^2}{n_2} + \frac{(-1)^2 s_w^2}{n_3}}$

Scheffé Procedure

- Scheffé method controls FW ER at α for all possible linear contrasts, not just pairwise comparisons
- Less sensitive (more conservative) than Tukey for pairwise comparisons but more sensitive for complex comparisons
- Should not be used to perform solely pairwise comparisons and should not be used to perform a priori comparisons (specifically designed as a post hoc test): too conservative

Scheffé Procedure

- Scheffé procedure tests general contrasts of the form $L = \sum_j c_j \mu_j$

Linear Contrasts: Scheffé Procedure

- Compute pooled estimate of variance $s^2 = s_w^2$ (i.e., MSE)
- Compute test statistic, $t = \dfrac{\hat{L}}{\widehat{\mathrm{SE}}(\hat{L})}$ with decision rules based on $F_{k-1,N-k,1-\alpha}$
- Reject H_0 of $L = 0$ vs. $L \neq 0$ if $|t| \geq \sqrt{(k-1)F_{k-1,N-k,1-\alpha}}$
- $100(1-\alpha)\%$ simultaneous CIs for $\mu_j - \mu_{j'}$: $\bar{x}_j - \bar{x}_{j'} \pm \sqrt{(k-1)F_{k-1,N-k,1-\alpha}}\ \widehat{\mathrm{SE}}(\hat{L})$

- Critical value $= \sqrt{(k-1)F_{k-1,N-k,1-\alpha}}$; p-value $= P\left(t^2/(k-1) \geq F_{k-1,N-k}\right)$

Scheffé Procedure: Example

- Scheffé procedure applied to our pairwise comparisons

Comparison	$\bar{x}_j - \bar{x}_{j'}$	$\hat{\text{SE}}$	t-Statistic	p (vs. α)
U vs. N	-10.44	4.67	-2.23	0.088
U vs. O	-15.70	4.76	-3.30	0.0058
N vs. O	-5.26	4.86	-1.08	0.56

- $\sqrt{2 \times F_{2,91,0.95}} = 2.489$

```
sqrt((3 - 1) * qf(.95, df1 = 3 - 1, df2 = 91)) = 2.489
1-pf((-2.23)^2/(3-1), df1=3-1, df2=91) = 0.088
```

Scheffé Procedure: Example

R Code, Scheffe

```
# Install and load required package
> library(lsmeans)

# Code below performs pairwise tests and gives Scheffe-adjusted p-values
> modelsys <- lm(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
> schefferesults <- lsmeans(modelsys,
                            pairwise ~ BMIGRP2_factor,
                            adjust = "scheffe")

> schefferesults$contrasts
 contrast                          estimate   SE df t.ratio p.value
 Underweight - Normal                -10.44 4.67 91 -2.234  0.0880
 Underweight - (Overweight/Obese)    -15.70 4.76 91 -3.302  0.0058
 Normal - (Overweight/Obese)          -5.26 4.86 91 -1.083  0.5585

P value adjustment: scheffe method with rank 2
```

Scheffé Procedure: Example

R Code, Scheffe

```
# Install and load required package
> library(DescTools)

# Scheffe simultaneous CIs and adjusted p-values for pairwise comparisons
> anovasys <- aov(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
> ScheffeTest(anovasys)

  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$BMIGRP2_factor
                                diff    lwr.ci   upr.ci   pval
Normal-Underweight            10.440228 -1.187491 22.06795 0.0880 .
Overweight/Obese-Underweight 15.703854  3.868263 27.53945 0.0058 **
Overweight/Obese-Normal       5.263626 -6.832712 17.35996 0.5585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Scheffé Procedure: Example

R Code, Scheffe

```
# Install and load required package
> library(DescTools)

# Scheffe simultaneous CIs and adjusted p-values: Can specify contrast
> anovasys <- aov(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
> ScheffeTest(anovasys, contrasts = c(1, -1, 0))  # allows you to specify contrast

  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$BMIGRP2_factor
                       diff    lwr.ci   upr.ci   pval
Underweight-Normal -10.44023 -22.06795 1.187491 0.0880 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Scheffé Procedure: Example

R Code, Scheffe

```
# Install and load required package
> library(DescTools)

# Scheffe simultaneous CIs and adjusted p-values: Can specify contrast
> anovasys <- aov(SYSBP ~ BMIGRP2_factor, data = fhs_anova)
> ScheffeTest(anovasys, contrasts = c(0.5, 0.5, -1))  # allows you to specify contrast

  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$BMIGRP2_factor
                                        diff   lwr.ci      upr.ci   pval
Underweight,Normal-Overweight/Obese -10.48374 -20.9432 -0.02427787 0.0493 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparison of MCPs: Example

R Code, Critical Values for $\alpha = 0.05$

```
> alpha = 0.05
> k = 3                  # Number of groups
> c = k*(k - 1)/2        # Number of pairwise comparisons
> N = 94                 # Total sample size
> qt(1 - alpha/2, df = N - k)                          # Fisher's LSD
[1] 1.986377
> qt(1 - (alpha/c)/2, df = N - k)                      # Bonferroni
[1] 2.43904
> qtukey(1 - alpha, k, N - k)/sqrt(2)                  # Tukey's HSD
[1] 2.382662
> sqrt((k - 1)*qf(1 - alpha, df1 = k - 1, df2 = N - k))   # Scheffe
[1] 2.488595
```

Comparison of MCPs

Comparison	$\bar{x}_j - \bar{x}_{j'}$	t-Statistic	LSD p	Bonferroni p	Tukey-Kramer p	Scheffe p
U vs. N	-10.44	-2.23	0.028	0.084	0.071	0.088
U vs. O	-15.70	-3.30	0.0014	0.0041	0.0039	0.0058
N vs. O	-5.26	-1.08	0.28	0.85	0.53	0.56

- Fisher's LSD: Most powerful, but does not control FW ER for > 2 post hoc tests
- Tukey-Kramer: Useful for unplanned pairwise comparison of means, more powerful than Bonferroni and Scheffé for pairwise comparisons
- Bonferroni: Easy to apply, good for small number of planned contrasts or pairwise comparisons
- Scheffé: Controls appropriately for unplanned contrasts, too conservative for pairwise comparisons more conservative than Bonferroni

Lesson Summary

- ANOVA is used to test equality of population means when there are more than 2 populations (global test) = overall F test
  - Function of between-group variability and within-group variability
- Post hoc tests used to pinpoint specific differences after significant ANOVA
  - Important to consider impact of multiple tests on type I error