# Prepping for the Final

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

# Prepping for the Final

1. Introduction
2. The Generalized $t$ Statistic for Independent Samples
   - Practice Exam Problem 19
   - Practice Exam Problem 20
3. ANOVA
   - Introduction
   - Simplified ANOVA with Equal $n$
   - ANOVA with Unequal $n$
   - ANOVA without Raw Data Revisited

# Introduction

I'm sorry to have had to miss class. I was unconscious on the operating table during class time yesterday. Talking is still a bit painful.

Some of you may have some concerns about exam questions on materials we did not discuss in class.

The purpose of these brief notes is to help "set you up" for these problems.

Many of you will not need these notes, but the package includes some routines you may find useful.

# The Generalized $t$ Statistic for Independent Samples

The generalized $t$ statistic can be used to test hypotheses about linear combinations of means from $J$ populations.

It includes the 1-Sample $t$, 2-Sample independent sample $t$, and many other tests as special cases.

Let's review how it works, and then write a very brief R function to execute it.

# The Generalized $t$ Statistic for Independent Samples

One way of thinking about the family of $t$-tests is as follows.

There is a linear combination of means, and you are interested in its value.

For example, if you have two means for two distinct populations, and you are interested in whether they are the same, you are interested in whether the linear combination $\mu_1 - \mu_2$ is equal to zero.

# The Generalized $t$ Statistic for Independent Samples

Any linear combination of variables is defined, essentially, by

- The list of items being linearly combined, and
- The list of linear weights used to do the combining.

If we are combining two means, we have $\mu_1, \mu_2$ as our list of items being combined, and we can use the notation $c_1, c_2$ to stand for the linear weights.

So to define the linear combination $\mu_1 - \mu_2$, we use $c_1 = +1, c_2 = -1$.

# The Generalized $t$ Statistic for Independent Samples

More generally, we can define the generalized $t$ statistic for testing the null hypothesis $\kappa = \sum_{j=1}^{J} c_j \mu_j = \kappa_0$ as

$$t_\nu = \frac{K - \kappa_0}{\sqrt{W \hat{\sigma}^2}} \tag{1}$$

where

$$W = \sum_j c_j^2 / n_j \tag{2}$$

$$K = \sum_j c_j \bar{X}_{\bullet j} \tag{3}$$

the degrees of freedom are

$$\nu = \sum_j n_j - J = n_\bullet - J \tag{4}$$

and the pooled unbiased variance estimator is

$$\hat{\sigma}^2 = \frac{\sum_j (n_j - 1) S_j^2}{\nu} \tag{5}$$

# The Generalized *t* Statistic for Independent Samples

Note that we can compute the generalized *t* statistic as long as we know the entries in 4 lists, i.e., the $n_j$, the $c_j$, the $\bar{X}_j$, and the $S_j$, and the one numerical value of $\kappa_0$ (which is often zero).

The routine shown below computes the generalized *t* from the 3 lists and $\kappa_0$. Note that the function requests a vector of standard deviations as one of its inputs. If you are given variances, you'll need to take the square root!

```
> GeneralizedT <- function(means, sds, ns, wts, k0 = 0) {
+     J <- length(means)
+     df <- sum(ns) - J
+     VarEstimate <- sum((ns - 1) * sds^2)/df
+     num <- sum(wts * means) - k0
+     den <- sqrt(sum(wts^2/ns) * VarEstimate)
+     t <- num/den
+     return(c(t, df))
+ }
```

# The Generalized $t$ Statistic for Independent Samples

Confidence intervals are constructed using the same building blocks as the test statistic.

The general form of the confidence interval for the linear combination $\kappa$ is

$$K \pm t^* \sqrt{W \hat{\sigma}^2} \tag{6}$$

with $K$, $W$ and $\hat{\sigma}^2$ as defined in Equations 1–5.

# The Generalized $t$ Statistic for Independent Samples

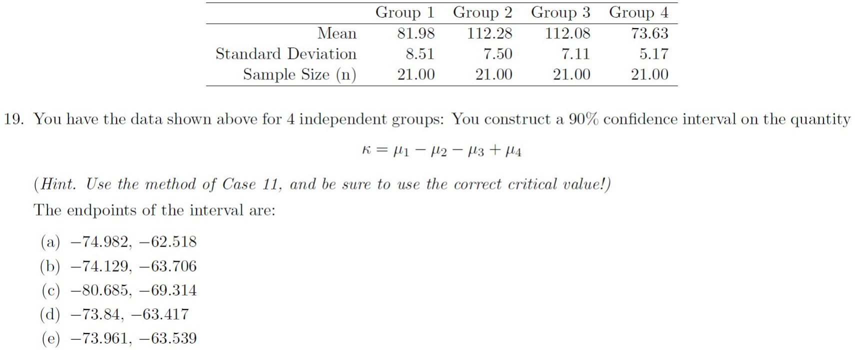
Here is a general routine for confidence intervals on *kappa*.

```
> IndependentMeanCI <- function(means, sds, ns, wts, conf) {
+     alpha <- 1 - conf
+     J <- length(ns)
+     df <- sum(ns) - J
+     t <- qt(1 - alpha/2, df)
+     k <- sum(wts * means)
+     mse <- sum((ns - 1) * sds^2)/df
+     serr <- sqrt(sum(wts^2/ns) * mse)
+     dist <- t * serr
+     lower <- k - dist
+     upper <- k + dist
+     return(c(lower, upper))
+ }
```

Let's use our two routines to solve problems 19 and 20 from the practice exam.

# The Generalized $t$ Statistic for Independent Samples

## Practice Exam Problem 19



| | Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|---|
| Mean | 81.98 | 112.28 | 112.08 | 73.63 |
| Standard Deviation | 8.51 | 7.50 | 7.11 | 5.17 |
| Sample Size (n) | 21.00 | 21.00 | 21.00 | 21.00 |

19. You have the data shown above for 4 independent groups: You construct a 90% confidence interval on the quantity

$$\kappa = \mu_1 - \mu_2 - \mu_3 + \mu_4$$

*(Hint. Use the method of Case 11, and be sure to use the correct critical value!)*

The endpoints of the interval are:

(a) $-74.982, -62.518$

(b) $-74.129, -63.706$

(c) $-80.685, -69.314$

(d) $-73.84, -63.417$

(e) $-73.961, -63.539$

We enter the data as

```
> means <- c(81.98, 112.28, 112.08, 73.63)
> sds <- c(8.51, 7.5, 7.11, 5.17)
> ns <- c(21, 21, 21, 21)
> wts <- c(1, -1, -1, 1)
```

# The Generalized $t$ Statistic for Independent Samples

## Practice Exam Problem 19

The solution is then computed as

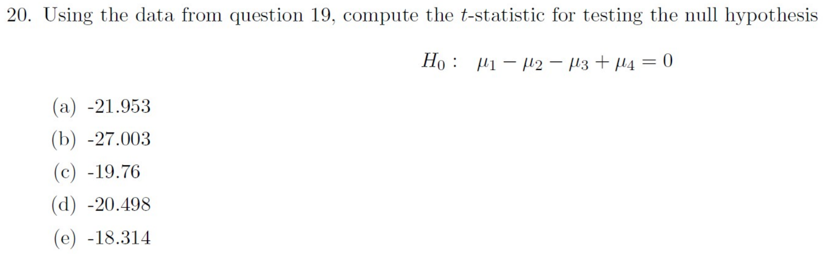
```
> options(digits = 5)
> IndependentMeanCI(means, sds, ns, wts, 0.9)

[1] -73.961 -63.539
```

Alternative (e) is the correct answer.

Next, we solve problem 20.

# The Generalized $t$ Statistic for Independent Samples

## Practice Exam Problem 20



20. Using the data from question 19, compute the $t$-statistic for testing the null hypothesis

$$H_0: \quad \mu_1 - \mu_2 - \mu_3 + \mu_4 = 0$$

(a) -21.953

(b) -27.003

(c) -19.76

(d) -20.498

(e) -18.314

The data are already loaded! Since the linear combination of interest is the same as in problem 19, and $\kappa_0$ is the default value of zero, we simply enter

```
> GeneralizedT(means, sds, ns, wts)

[1] -21.953  80.000
```

The $t$ statistic is $-21.953$ with 80 degrees of freedom.

Alternative (a) is the correct one.

# ANOVA

## Introduction

As I explain in the lecture slides on ANOVA, and explore in much more detail in Psychology 311, there are numerous ways of viewing ANOVA.

In this section, we simply concentrate on performing calculation of the ANOVA $F$-statistic and associated $p$-value in two basic situations:

- In the equal $n$ case, when the $J$ independent samples are of equal size $n$.
- In the more general, unequal $n$ case, when the $J$ independent samples have possibly unequal sample sizes $n_j$.

# ANOVA

## Introduction

Note that the method shown for unequal *n* can also be used for the equal *n* case when raw data are available. Moreover, even if raw data are not available, one can perform the calculations in numerous ways from summary statistics, as we shall see.

One-way fixed-effects ANOVA posits several groups, randomly taken from "populations" that are either selected groups or created by experimental manipulation.

In the overall test of significance, one examines the null hypothesis whether all groups have the same mean.

# ANOVA

## Simplified ANOVA with Equal $n$

In class notes, I demonstrate that when sample sizes are all the same, and are all equal to $n$, then the ANOVA $F$ test can be computed as

$$F_{J-1,J(n-1)} = \frac{nS^2_{\overline{X}}}{\hat{\sigma}^2} = \frac{S^2_{\overline{X}}}{\hat{\sigma}^2/n} \tag{7}$$

Equation 7 can be used if you have a table that lists group means and standard deviations (or variances). $S^2_{\overline{X}}$ is the variance of the group means, and $\hat{\sigma}^2$ is the mean of the sample variances.

At its foundation, the $F$-test in ANOVA compares the variance of the sample means with an estimate of $\sigma^2/n$, which would be the long run variance of the sample means *if all sample means came from the same population*.

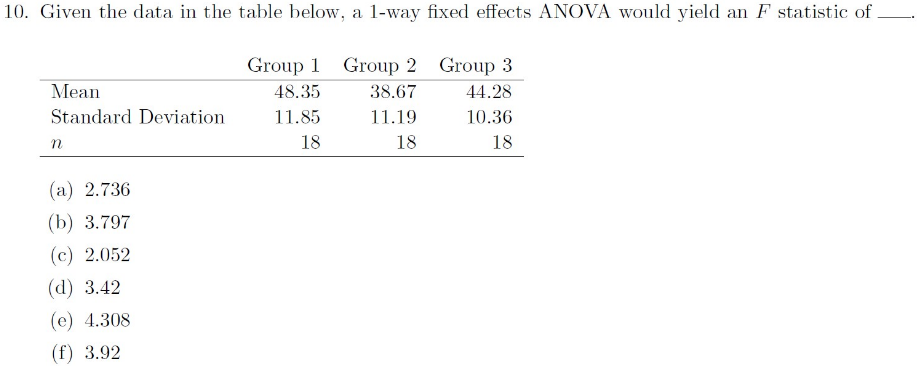
If all sample means did not come from populations with the same mean, then the variance of the sample means will, in the long run, be larger than $\sigma^2/n$.

So the ANOVA $F$-test is a one-tailed test.

# ANOVA

## Simplified ANOVA with Equal *n*

Let's try the example from the practice exam.



10. Given the data in the table below, a 1-way fixed effects ANOVA would yield an $F$ statistic of ___.

| | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Mean | 48.35 | 38.67 | 44.28 |
| Standard Deviation | 11.85 | 11.19 | 10.36 |
| $n$ | 18 | 18 | 18 |

(a) 2.736

(b) 3.797

(c) 2.052

(d) 3.42

(e) 4.308

(f) 3.92

# ANOVA

## Simplified ANOVA with Equal *n*

We'll enter the data directly from the table. For our purposes, the standard deviations will need to be squared to convert them to variances. The calculations show an $F$ of 3.42, with 2 and 51 degrees of freedom.

```
> means <- c(48.35, 38.67, 44.28)
> sds <- c(11.85, 11.19, 10.36)
> n <- 18
> vars <- sds^2
> F <- n * var(means)/mean(vars)
> J <- length(means)
> df1 <- J - 1
> df2 <- J * (n - 1)
> F

[1] 3.4203

> df1

[1] 2

> df2

[1] 51
```

# ANOVA

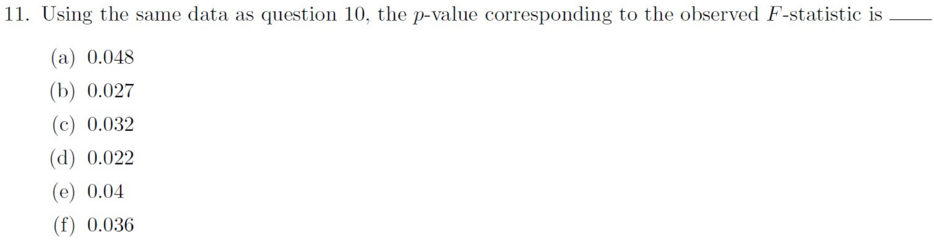
## Simplified ANOVA with Equal $n$

Problem 11 on the practice exam asks you to provided a $p$-value corresponding to the $F$-statistic obtained in problem 10.

As mentioned before, the $F$ test in ANOVA is a *one-sided* test, and so the $p$-value needs to be computed on that basis.

# ANOVA

## Simplified ANOVA with Equal $n$



11. Using the same data as question 10, the $p$-value corresponding to the observed $F$-statistic is ____

(a) 0.048

(b) 0.027

(c) 0.032

(d) 0.022

(e) 0.04

(f) 0.036

Since the $F$ test is one sided, the $p$ value is simply the probability of obtaining an $F$ larger than the observed value. This is easily computed in R from our previously computed values as

```
> 1 - pf(F, df1, df2)

[1] 0.040375
```

Alternative (e) is best.

# ANOVA

## ANOVA with Unequal $n$

When sample sizes are unequal, and you have raw data available, the easiest way to compute the ANOVA $F$-test is to employ R and process the ANOVA as a regression analysis.

The method is straightforward, and the hardest part is the data entry. Data are entered in two variables. There is a group code, which can be entered into a grouping variable either using numbers (1,2,3 if there are 3 groups), or using characters, like `'Experimental1', 'Experimental2', 'Control'`.

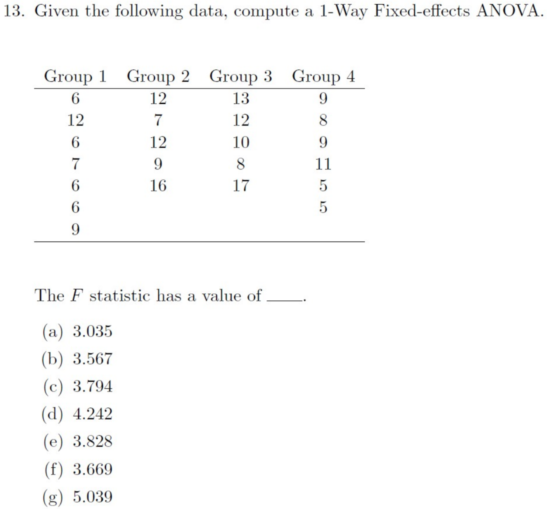
It is important that the grouping variable be explicitly typed as a `factor` variable, for reasons that are explained in detail in Psychology 311. If you enter group codes as text strings, R will automatically type the variable as factor, but if you enter the group codes as numbers, it will not.

Each subject's score is entered alongside his/her group code in the dependent variable, which is a standard R `numeric` variable.

# ANOVA

## ANOVA with Unequal *n*

We'll illustrate with problem 13 from the practice exam.



13. Given the following data, compute a 1-Way Fixed-effects ANOVA.

| Group 1 | Group 2 | Group 3 | Group 4 |
|---|---|---|---|
| 6 | 12 | 13 | 9 |
| 12 | 7 | 12 | 8 |
| 6 | 12 | 10 | 9 |
| 7 | 9 | 8 | 11 |
| 6 | 16 | 17 | 5 |
| 6 | | | 5 |
| 9 | | | |

The $F$ statistic has a value of ____.

(a) 3.035

(b) 3.567

(c) 3.794

(d) 4.242

(e) 3.828

(f) 3.669

(g) 5.039

# ANOVA

## ANOVA with Unequal $n$

Note that there are 7 observations in Group 1, 5 in Group 2, 5 in Group 3, and 6 in Group 4. So we can enter the Group codes quickly as

```
> Group <- c(rep(1, 7), rep(2, 5), rep(3, 5), rep(4, 6))
> Group <- factor(Group)
> Group

 [1] 1 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 4
Levels: 1 2 3 4
```

Note how I converted the `Group` variable to a factor.

Next, I enter the scores, in order. To aid in proofreading, I enter each group's scores separately, then cocatenate them into one large vector called `Score`.

```
> Group1 <- c(6, 12, 6, 7, 6, 6, 9)
> Group2 <- c(12, 7, 12, 9, 16)
> Group3 <- c(13, 12, 10, 8, 17)
> Group4 <- c(9, 8, 9, 11, 5, 5)
> Score <- c(Group1, Group2, Group3, Group4)
```

Now that the data are entered, there are several ways I can proceed to compute the ANOVA.

# ANOVA

## ANOVA with Unequal $n$

Perhaps the simplest way is to use the general linear model to fit a model where `Score` is predicted from `Group`, then apply the `anova` command to the fit object.

```
> fit <- lm(Score ~ Group)
> anova(fit)

Analysis of Variance Table


Response: Score
          Df Sum Sq Mean Sq F value Pr(>F)
Group      3   91.9   30.62    3.79  0.027 *
Residuals 19  153.3    8.07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the $F$ statistic is 3.7943, and the $p$-value is 0.02749.

Alternative (c) is the correct answer.

# ANOVA

## ANOVA without Raw Data Revisited

In the equal $n$ case, it is ridiculously simple to compute an ANOVA from a summary table of means and standard deviations for the $J$ groups.

In the unequal $n$ case, if we do not have raw data, the formulas are substantially more complicated.

This doesn't mean that the calculations have to be complicated!

We'll revisit a trick we learned early in the semester, and reprocess problems 10 and 11 using this trick. Regardless of whether the sample sizes are equal or unequal, this trick can be used to compute an ANOVA directly from the table of group means and standard deviations (or variances).

What is the trick? You don't have the data, so you just `make.exact.data` using the R function we learned in an early course lecture. Any set of data satisfying the summary statistics in the table will do, so we just create our own!

# ANOVA

## ANOVA without Raw Data Revisited

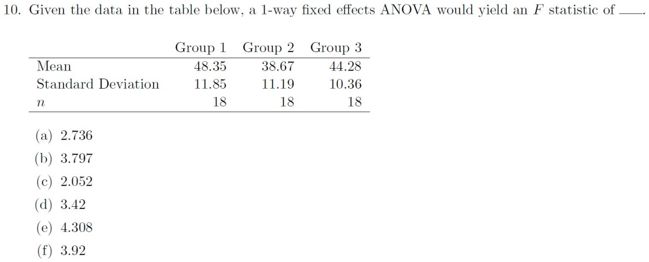
We'll pull the code from the lecture notes on *Using R to Answer Theoretical Questions*

```
> z.score <- function(x) {
+     (x - mean(x))/sd(x)
+ }
>
> rescale.numbers <- function(x, mean, sd) {
+     z <- z.score(x)
+     return(z * sd + mean)
+ }
>
> make.exact.data <- function(n, mean, sd) {
+     x <- rnorm(n)
+     return(rescale.numbers(x, mean, sd))
+ }
```

# ANOVA

## ANOVA without Raw Data Revisited

Let's revisit Problem 10.



10. Given the data in the table below, a 1-way fixed effects ANOVA would yield an $F$ statistic of ____.

| | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Mean | 48.35 | 38.67 | 44.28 |
| Standard Deviation | 11.85 | 11.19 | 10.36 |
| $n$ | 18 | 18 | 18 |

(a) 2.736

(b) 3.797

(c) 2.052

(d) 3.42

(e) 4.308

(f) 3.92

To create fake data mimicking the results in the table, we first create the grouping variable with $n = 18$ per group.

```
> Group <- c(rep(1, 18), rep(2, 18), rep(3, 18))
> Group <- factor(Group)
```

Next, we create the actual data

```
> Group1 <- make.exact.data(18, 48.35, 11.85)
> Group2 <- make.exact.data(18, 38.67, 11.19)
> Group3 <- make.exact.data(18, 44.28, 10.36)
> Score <- c(Group1, Group2, Group3)
```

# ANOVA

## ANOVA without Raw Data Revisited

Next we perform the computations. In one pass, we generate the answers to problems 10 and 11 that we generated earlier from formulas.

However, remember that this approach can be used with either equal or unequal *n* when only summary stats are available. Just make sure to distinguish between variances and standard deviations.

```
> options(digits = 7)
> fit <- lm(Score ~ Group)
> anova(fit)

Analysis of Variance Table

Response: Score
          Df Sum Sq Mean Sq F value  Pr(>F)
Group      2  850.4  425.22  3.4203 0.04037 *
Residuals 51 6340.5  124.32
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```