Chapter 5 Section 5.1

Review of two-sample t-test
Analysis of Variance = ANOVA or AOV

In both cases:

- The response variable is quantitative.
- The explanatory variable is categorical
  - For a two-sample t-test, it has 2 categories.
  - For ANOVA, it has 2 or more categories.
  - However, when k = 2, ANOVA is equivalent to a two-sided two-sample t-test.

Some basic definitions

- A factor is a categorical explanatory variable.
- A level of a factor is one category.
- Categories are sometimes called groups.

Example

Does average time spent studying per week differ by type of major? Take random sample from each type of major, or one random sample and divide into the 3 majors.

- Y = time spent studying per week (hours) [response var.]
- Factor = Category of major (sciences, social sciences, humanities) [explanatory variable]
- The 3 <u>levels</u> of the factor (the 3 groups) are sciences, social sciences, humanities.

Two-sample t-test (Review)

Data: Independent samples from two groups

Summary statistics:

$$n_1, \bar{Y}_1, s_1$$
$$n_2, \bar{Y}_2, s_2$$

Conditions:

1. Normal populations (or large *n*'s)
2. Equal variances (sometimes)

Hypotheses:

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Write as $Y_{ik} \sim N(\mu_k, \sigma)$, where

k = group (1 or 2)

i = individual within group = 1, 2, …, n_k

Pooled Two-sample t-test (Review?)

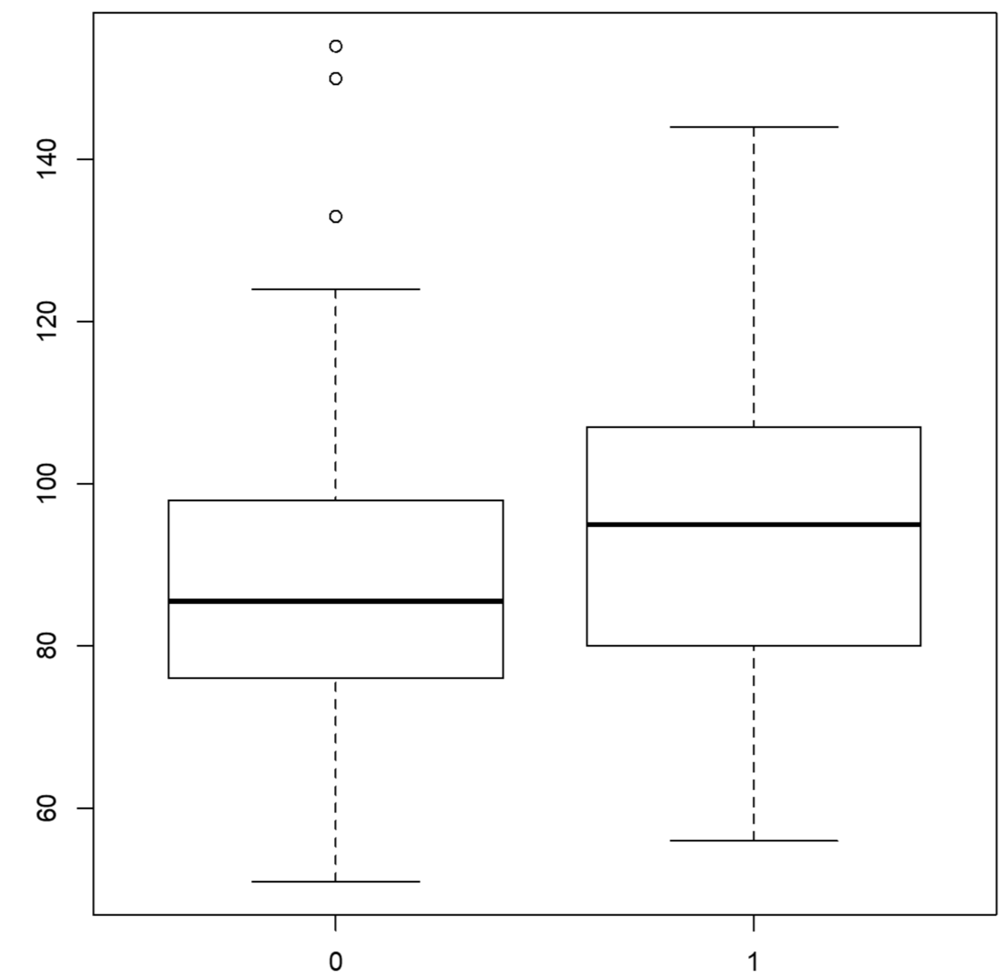
Pooled variance: $$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Test statistic: $$t.s. = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Explain why on white board.

Reference distribution: $$t_{n_1 + n_2 - 2}$$

Does Active Pulse Depend on Gender?



Two-sample t-test (*R*)

```
> t.test(Active~Gender,var.equal=TRUE)
      Two Sample t-test

data:  Active by Gender
t = -2.7436, df = 230, p-value = 0.006556
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.503416  -1.887046
sample estimates:
mean in group 0 mean in group 1
       88.12295        94.81818
```

Two-sample t-test

```
> t.test(Active~Gender,var.equal=TRUE)
        Two Sample t-test

data:  Active by Gender
t = -2.7436, df = 230, p-value = 0.006556
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.503416  -1.887046
sample estimates:
mean in group 0 mean in group 1
       88.12295        94.81818
```

ANOVA for Means

```
> summary(aov(Active~Gender))
             Df Sum Sq Mean Sq F value   Pr(>F)
Gender        1   2593 2592.96  7.5274 0.006556 **
Residuals   230  79228  344.47
---
> oneway.test(Active~Gender,var.equal=TRUE)
        One-way analysis of means

data:  Active and Gender
F = 7.5274, num df = 1, denom df = 230, p-value = 0.006556
```

ANOVA: Test for Difference in K Population Means

Data: Samples from K different groups

Summary statistics:

n_1	$\bar{Y}_1$	s_1
n_2	$\bar{Y}_2$	s_2
$\vdots$	$\vdots$	$\vdots$
n_K	$\bar{Y}_K$	s_K

For each group

Combine all

$n \quad \bar{Y} \quad S_Y$

Test: $H_0: \mu_1 = \mu_2 = \ldots = \mu_K$

H_1: Some $\mu_k \neq \mu_j$

Conditions and assumptions

1. Normal populations (or large n for each group)
2. Equal variances for all observations
3. All observations are independent, within and between groups.

Write as $Y_{ik} \sim N(\mu_k, \sigma)$, all independent, where

i = individual within each group = 1, 2, …, n_k

k = group, with k = 1, 2, …, K

See picture on white board.

Some possible ways to get independent data

1. K separate populations, take random sample from each.

   Ex: Groups = 4 regions of the US

   Y_{ik} = time spent commuting to work

2. Take one random sample and measure response variable Y, and categorical explanatory variable X.

   Ex: Groups = type of major (Science, SocSci, Humanities)

   Y_{ik}= time spent studying per week

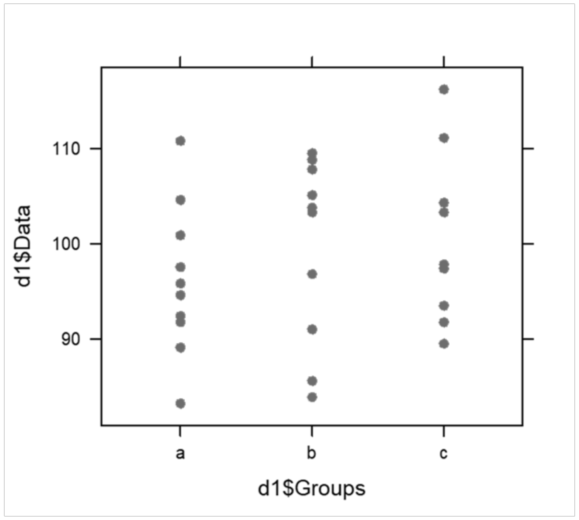
3. Randomized experiment with K treatments

   Ex: 30 cities available for experiment with 3 roadside billboards

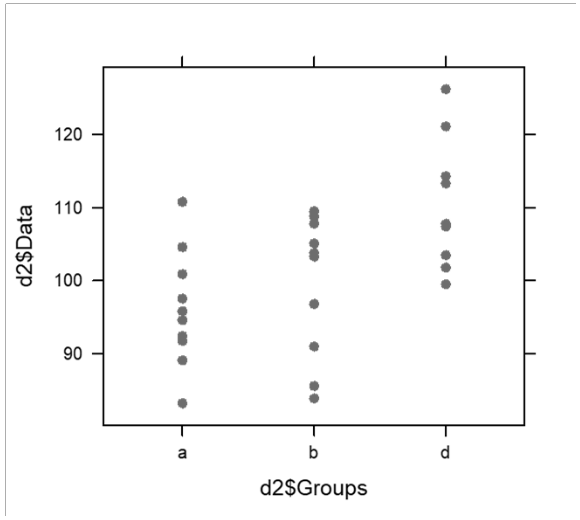
   Randomly assign 10 cities to each type of billboard

   Y_{ik} = Sales of product after 6 months in City *i,* with billboard *k.*

Test: Are Group Means Equal (in the Population)?



p-value = 0.39



p-value = 0.0015

What's different?

Same *n* and SDs but a shift in the third group

Effect size $= \frac{|\mu_1 - \mu_2|}{\sigma}$

Summary of Data
For categories in Groups
No Selector

Count	Mean	StdDev
10	96.0820	7.90629
10	99.5640	9.63299
10	101.601	9.09347

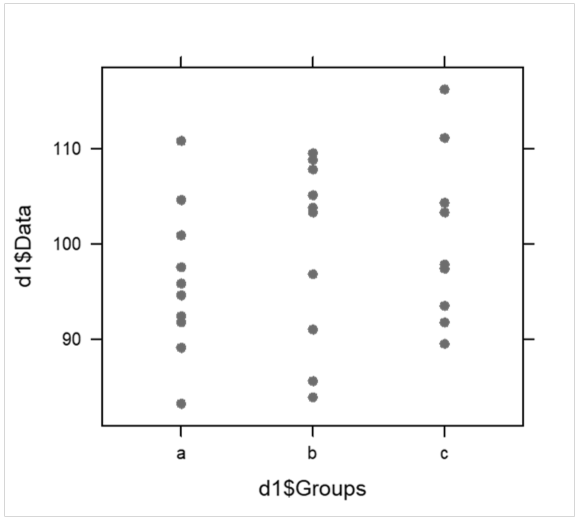
Effect size = 0.6

Summary of Data
For categories in Groups
No Selector

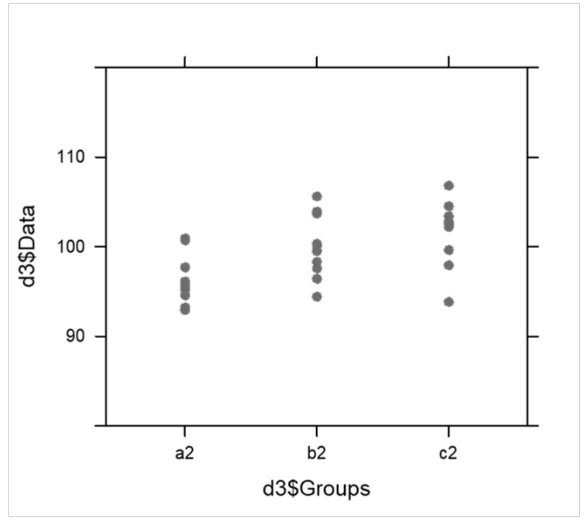
Count	Mean	StdDev
10	96.0820	7.90629
10	99.5640	9.63299
10	111.601	9.09052

Effect size = 1.6

Test: Are Group Means Equal (in the Population)?



p-value = 0.39



p-value = 0.0036

What's different?

Same n and means but smaller SDs

Summary of Data
For categories in Groups
No Selector

Count	Mean	StdDev
10	96.0820	7.90629
10	99.5640	9.63299
10	101.601	9.09347

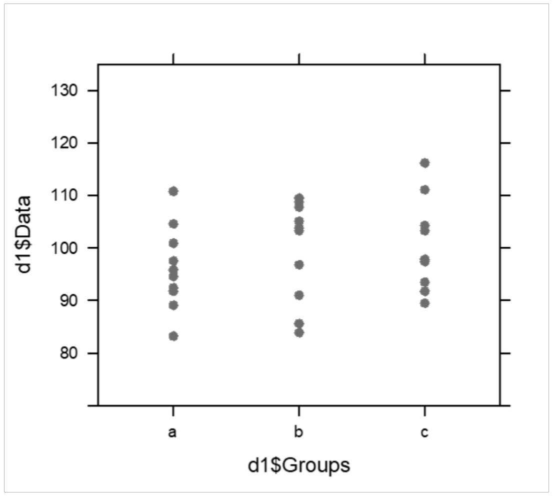
Effect size = 0.6

Summary of Data
For categories in Groups
No Selector

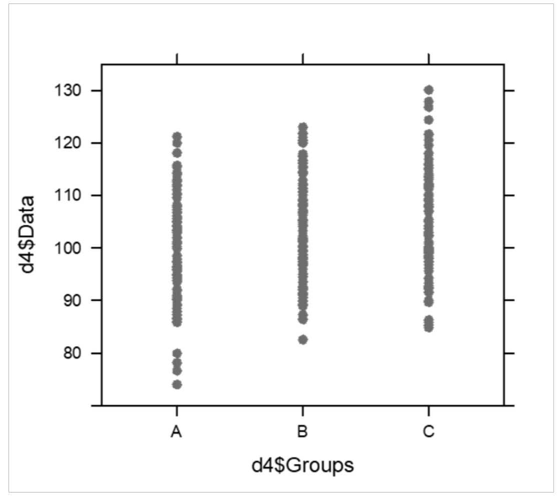
Count	Mean	StdDev
10	96.2640	2.75993
10	99.9780	3.55353
10	101.806	3.75886

Effect size = 1.5

Test: Are Group Means Equal (in the Population)?



p-value = 0.39



p-value = 0.0002

What's different?

Same (approx.) range among the means but larger *n*

Summary of Data
For categories in Groups
No Selector

Count	Mean	StdDev
10	96.0820	7.90629
10	99.5640	9.63299
10	101.601	9.09347

Effect size = 0.57 to two decimal places

Summary of Data
For categories in Groups
No Selector

Count	Mean	StdDev
100	99.8757	10.3175
100	103.405	9.34201
100	105.702	10.1690

Effect size = 0.56

Summary of what decreases *p*-value and increases power of the test (easier to reject null hypothesis)

- Bigger difference between the means
  - Increased effect size
- Smaller standard deviations
  - Increased effect size
- Larger sample sizes
  - Not an increase in effect size

Example: Random sample of $n_k = 5$ scores (Ys) from each of K = 4 exams (there are 4 levels)

Exam #1:	**62,**	**94,**	**68,**	**86,**	**50**
Exam #2:	**87,**	**95,**	**93,**	**97,**	**63**
Exam #3:	**74,**	**86,**	**82,**	**70,**	**28**
Exam #4:	**77,**	**89,**	**73,**	**79,**	**47**

	n_1	Mean	S_i
	5	72.0	17.89
	5	87.0	13.93
	5	68.0	23.24
	5	73.0	15.68
Overall	20	75.0	18.11

Is there a difference in population mean score among the four exams?

Test: $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: Some $\mu_k \neq \mu_j$

Helpful *R* Command

```
> means=tapply(X=Grade,INDEX=Exam,FUN=mean) #FUNction = mean
> means
 1  2  3  4
72 87 68 73

> sds=tapply(Grade,Exam,sd) #we don't have to state "X=", etc.
> Sds                       #standard deviations
 1        2        3        4
17.88854 13.92839 23.23790 15.68439

> ns=tapply(Grade,Exam,length) #length = sample size
> ns
1 2 3 4
5 5 5 5
```

ANOVA (Means) Model

$$Y = \mu_k + \varepsilon$$

Mean for group #*k*

$N(0,\sigma_\varepsilon)$ random error

Under H_0 (μ_k's all equal) $\rightarrow$ $\hat{\mu}_k = \bar{\bar{Y}}$

Under H_1 (μ_k's differ) $\rightarrow$ $\hat{\mu}_k = \bar{Y}_k$

These are the least squares estimates for μ_k for the two hypotheses.

"Predicting" in ANOVA Model

If the group means are the same (H_0):

$\hat{Y} = \overline{Y}$ for all groups → $residual = Y - \overline{Y}$

If the group means can be different (H_1):

$\hat{Y} = \overline{Y}_k$ for k^{th} group → $residual = Y - \overline{Y}_k$

Do we do "significantly" better with separate means?

Compare sums of squared residuals…

$SSTotal = \sum\left(Y - \overline{Y}\right)^2$ vs. $SSE = \sum\left(Y - \overline{Y}_k\right)^2$

Partitioning Variability

Data	=	Model	+	Error
Y	=	μ_k	+	ε
TOTAL variation in response, Y	=	Variation explained by MODEL	+	Unexplained variation in RESIDUALS

Key question: Does the MODEL explain a "significant" amount of the TOTAL variability?

Partitioning Variability ANOVA for Group Means

$$Y = \mu_k + \varepsilon$$

$$(y - \bar{y}) = (\bar{y}_k - \bar{y}) + (y - \bar{y}_k)$$

$$\sum (y - \bar{y})^2 = \sum (\bar{y}_k - \bar{y})^2 + \sum (y - \bar{y}_k)^2$$

$$SSTotal = SSGroups + SSE$$

Using familiar regression terminology

$$\sum (y-\bar{y})^2 = \sum (\bar{y}_k - \bar{y})^2 + \sum (y - \bar{y}_k)^2$$

Residuals if H_0 is true (same mean) = *"Explained" by model with separate means* + *Still unexplained with separate means*

SSTotal = *SSGroups = SSModel* + *SSE*

Example: Four Exams

						n_k	Mean	S_k
Exam #1:	**62,**	**94,**	**68,**	**86,**	**50**	5	72.0	17.89
Exam #2:	**87,**	**95,**	**93,**	**97,**	**63**	5	87.0	13.93
Exam #3:	**74,**	**86,**	**82,**	**70,**	**28**	5	68.0	23.24
Exam #4:	**77,**	**89,**	**73,**	**79,**	**47**	5	73.0	15.68
					Overall	20	75.0	18.11

$$SSGroups = 5(72-75)^2 + 5(87-75)^2 + 5(68-75)^2 + 5(73-75)^2 = 1030$$

$$SSE = (62-72)^2 + (94-72)^2 + \cdots + (47-73)^2 = \underline{5200}$$

$$SSTotal = (62-75)^2 + (94-75)^2 + \cdots + (47-75)^2 = 6230$$

Decomposition: Four Exams

Exam #1: **62, 94, 68, 86, 50**
Exam #2: **87, 95, 93, 97, 63**

Group Mean
72.0
87.0

Overall (Grand Mean) = 75.0

	Observed value		Grand mean		Group effect		Residual
Exam #1:	**62**	**=**	**75.0**	**+**	**-3**	**+**	**-10**
Exam #1:	**94**	**=**	**75.0**	**+**	**-3**	**+**	**22**
Exam #2:	**87**	**=**	**75.0**	**+**	**12**	**+**	**0**
Exam #2:	**95**	**=**	**75.0**	**+**	**12**	**+**	**8**

Etc.

ANOVA Table (for K Group Means)

H_0: $\mu_1 = \mu_2 = \ldots = \mu_K$
H_1: Some $\mu_k \neq \mu_j$

Note: n = total sample size

Source	d.f.	S.S.	M.S.	t.s.	p-value
Groups	$K-1$	$SSGroups$	$\frac{SSGroups}{K-1}$	$\frac{MSGroups}{MSE}$	use $F_{K-1,n-K}$
Error	$n-K$	SSE	$\frac{SSE}{n-K}$		
Total	$n-1$	$SSTotal$			

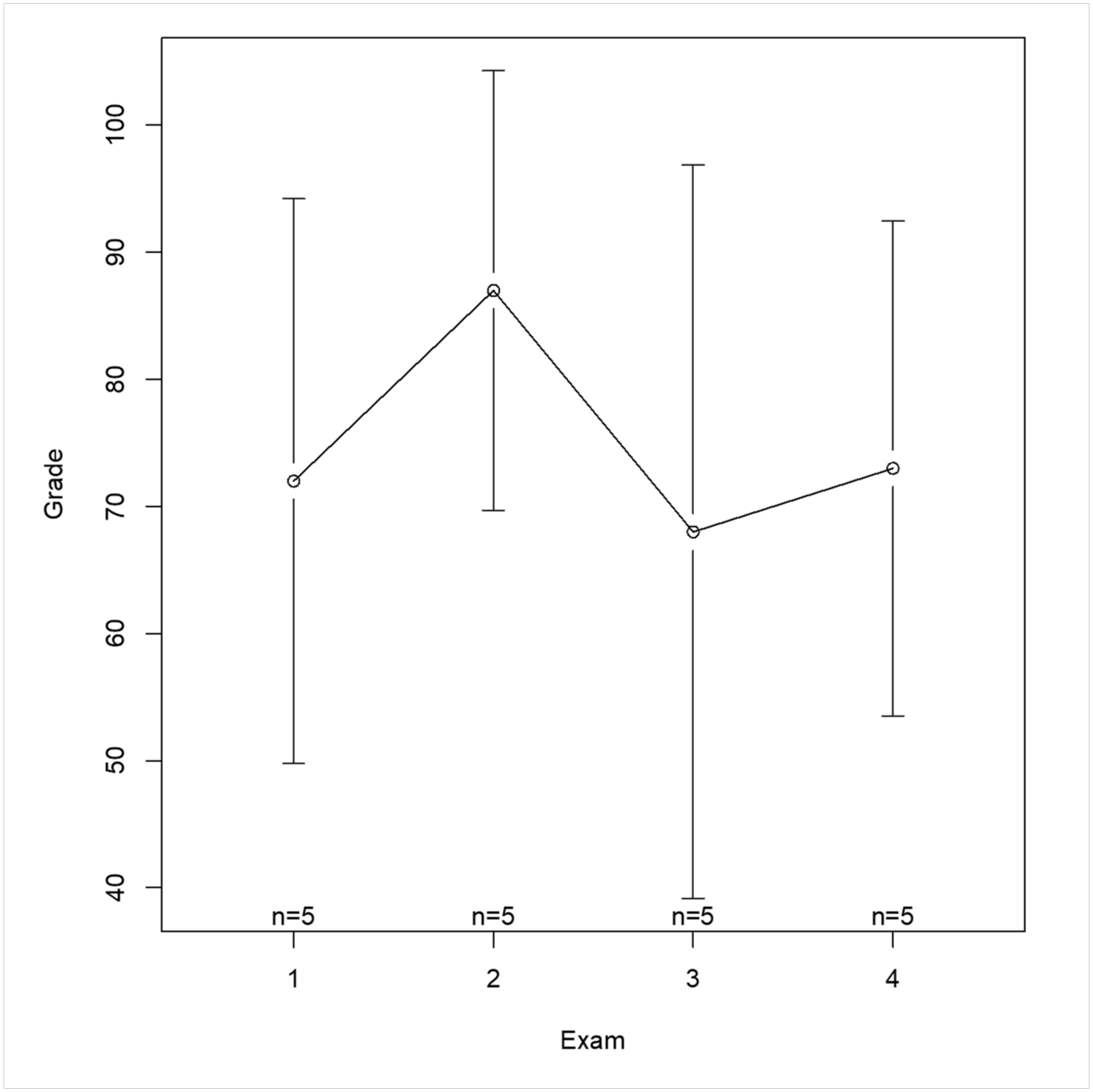
Small p-value ➔ Reject H_0 ➔ There is a evidence of a difference among the <u>population</u> means of the K groups.

ANOVA Output in *R*

```
> model=aov(Grade~as.factor(Exam))
> model
Terms:
                as.factor(Exam) Residuals
Sum of Squares             1030      5200
Deg. of Freedom               3        16

Residual standard error: 18.02776
Estimated effects may be unbalanced

> summary(model)
                Df Sum Sq Mean Sq F value Pr(>F)
as.factor(Exam)  3 1030.0   343.3  1.0564  0.395
Residuals       16 5200.0   325.0

> 1-pf(1.0564,3,16) #if the P-value hadn't been given
 [1] 0.3950020
```

After Installing Three Packages in *R*: gplots, gdata, gtools

```
> plotmeans(Grade~Exam)
```



95% CI's for each group mean shown in blue. Notice the substantial overlap.

Partition Variability (different formulas) + df

Between groups: (d.f. = K − 1)

$$SSGroups = n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + \cdots + n_K(\bar{y}_K - \bar{y})^2$$

Within groups: (d.f. = n − K)

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_K - 1)s_K^2$$

Total: (d.f. = n − 1)

$$SSTotal = \sum (y - \bar{y})^2 = (n-1)s_Y^2$$

$$SSTotal = SSGroups + SSE$$

Example: Four Exams

						n_k	Mean	S_i
Exam #1:	**62,**	**94,**	**68,**	**86,**	**50**	5	72.0	17.89
Exam #2:	**87,**	**95,**	**93,**	**97,**	**63**	5	87.0	13.93
Exam #3:	**74,**	**86,**	**82,**	**70,**	**28**	5	68.0	23.24
Exam #4:	**77,**	**89,**	**73,**	**79,**	**47**	5	73.0	15.68
					Overall	20	75.0	18.11

$$SSGroups = 5(72-75)^2 + 5(87-75)^2 + 5(68-75)^2 + 5(73-75)^2 = 1030$$

$$SSE = 4(17.89)^2 + 4(13.93)^2 + 4(23.24)^2 + 4(15.68)^2 = 5200$$

$$SSTotal = 19(18.11)^2 = 6230$$ (up to roundoff)

Alternate Form: ANOVA Model for Means

$$Y = \mu + \alpha_k + \varepsilon$$

Grand mean

Effect for k^{th} group

Random error

$$\mu_k = \mu + \alpha_k$$

$$\hat{\alpha}_k = \overline{Y}_k - \overline{Y}$$

Note: α_k sum to 0

H_0: $\mu_1 = \mu_2 = \ldots = \mu_K$
H_1: Some $\mu_k \neq \mu_j$

$\longleftrightarrow$

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_K = 0$
H_1: Some $\alpha_k \neq 0$

Estimating the common variance

$$\varepsilon \sim N(0, \sigma_{\varepsilon})$$

$$Y_{ik} \sim N(\mu_k, \sigma)$$

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_K - 1)s_K^2$$

$$\mathbf{MSE} = \frac{SSE}{n - k}$$

a weighted average of sample variances

MSE is an estimate of the (common) population variance

$$\mathbf{MSE} = \hat{\boldsymbol{\sigma}}^2$$

Example: Four Exams

						n_k	Mean	S_i
Exam #1:	**62,**	**94,**	**68,**	**86,**	**50**	5	72.0	17.89
Exam #2:	**87,**	**95,**	**93,**	**97,**	**63**	5	87.0	13.93
Exam #3:	**74,**	**86,**	**82,**	**70,**	**28**	5	68.0	23.24
Exam #4:	**77,**	**89,**	**73,**	**79,**	**47**	5	73.0	15.68
Overall						20	75.0	18.11

Four estimates of the population sd

MSE = 5200/16 = 325 = estimate of popn variance

$$\sqrt{MSE} = \sqrt{325} = 18.03$$

= estimate of population standard deviation

Section 5.2: Checking Conditions for ANOVA

$$\varepsilon \sim N(0, \sigma_{\varepsilon})$$ Check with residuals.

Zero mean: Always holds for sample residuals.

Constant variance:

Plots and numerical checks:

- Plot residuals vs. fits
- Plot Y versus group, or boxplot for each group
- Compare standard deviations of groups; check if largest is more than twice value of smallest.

Note: This is less crucial if the sample sizes are equal.

Checking Conditions, continued

Normality:

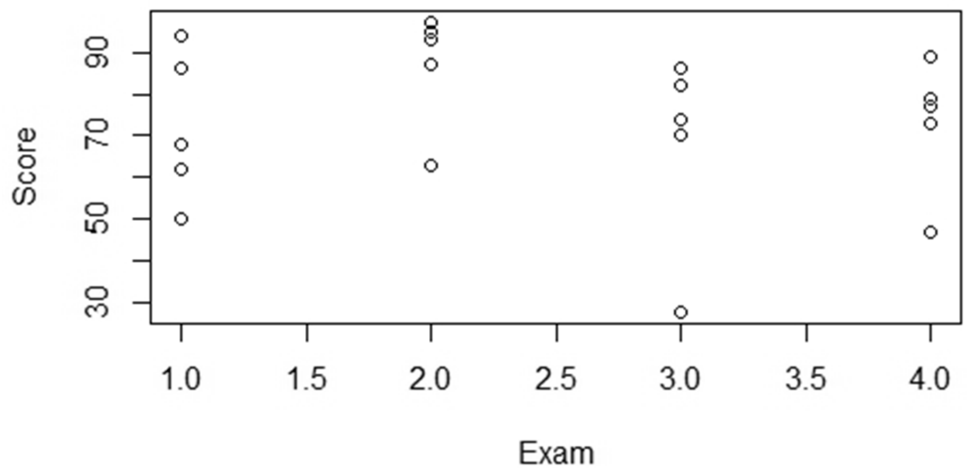
Histogram of residuals

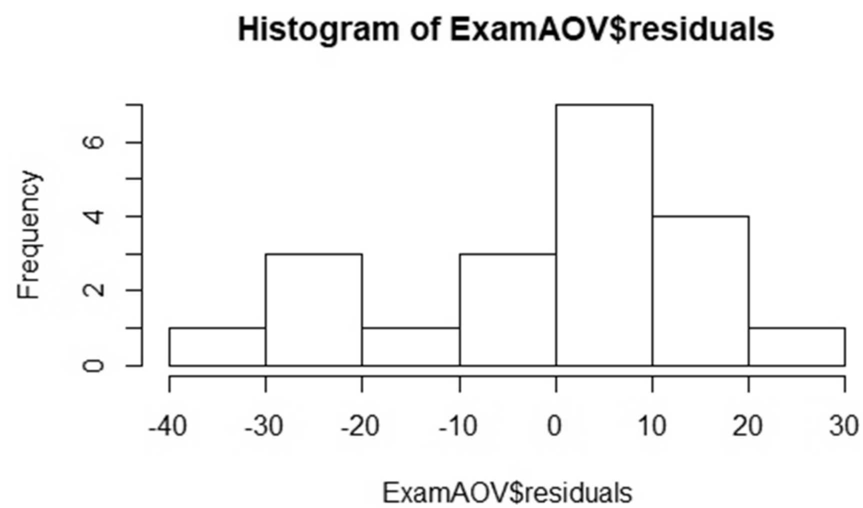
Normal probability plot of residuals

Independence:

Pay attention to data collection method. (See earlier slide.)

Plot of data and histogram of residuals





Section 5.3: Scope of Inference

Allocation of Units to Groups

Selection of Units	**Using Randomization**	Not using Randomization
At Random	***Random sample selected; units assigned randomly to treatment groups***	*Random samples selected from separate populations*
Not at Random	**Study units are found, then randomly assigned to treatment groups**	Available units from separate populations are studied

Inferences can be drawn to populations (← At Random row)

Causal inferences can be drawn (↑ Using Randomization column)

Some Examples

Exercise 5.19 – Life spans	Not random
Exercise 5.28 – Fenthion	Random samples
Exercise 5.30 – Blood pressure	Random samples, cause/effect?
Example 5.1 – Fruit flies	Random allocation

Now do example of seat location.