Chapter 15: Comparing means of more than two groups

1

More than two groups

- Thus far, the examples that we've looked at have had two groups (e.g., mutant vs wild type, drug vs placebo)
- How do you compare means of more than two groups?
  - Placebo, drug A, drug B, etc.
  - Species 1, species 2, species 3, etc.

2

Problem of multiple tests

- Tempting to do all possible pairwise comparisons
- Ex: species 1 vs species 2 vs species 3 vs species 4
  - 1 vs 2
  - 1 vs 3
  - 1 vs 4
  - 2 vs 3
  - 2 vs 4
  - 3 vs 4
- But the problem is that running multiple tests inflates the probability of getting at least one Type I error

3

ANOVA

- **Analysis of variance (ANOVA)** compares the means of multiple groups simultaneously in a single analysis
  - Tests for variation of means among groups
- H_0: $\mu_1 = \mu_2 = \mu_3 \ldots \mu_n$
- H_A: mean of at least one group is different from at least one other group

4

ANOVA in a nutshell

- Null assumption that all groups have the same true mean is equivalent to saying that each group sample is drawn from the same population
- But each group sample is bound to have a different mean due to sampling error
- ANOVA determines if there is more variance among sample means than we would expect by sampling error alone

5

ANOVA in a nutshell

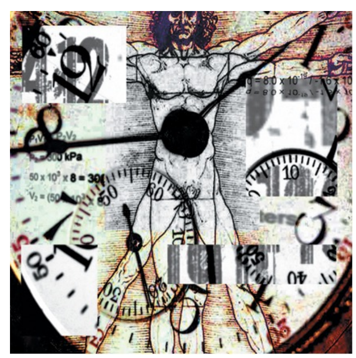
- Two measures of variation
- **Group mean square** (MS_{groups}) is proportional to the observed amount of variance among group sample means
  - Variation among groups
- **Error mean square** (MS_{error}) estimates the variance among subjects that belong to each group
  - Variation within groups
- Test statistic is a ratio:
  - True null: $MS_{groups} / MS_{error} = 1$
  - False null: $MS_{groups} / MS_{error} > 1$

6

7

Ex 15.1: The knees who say night

- Circadian clock
- Recovering from jet lag involves re-calibration of clock through light detection
- Study found that exposing back of knee to light can reset clock?
- Follow-up: participants awakened from sleep and exposed to three hour treatment of lights to eyes, lights to knees, or no light



8

Ex 15.1: The knees who say night

- H_0: $\mu_1 = \mu_2 = \mu_3$
- H_A: at least one μ_i is different from at least one other

TABLE 15.1-1 Raw data and descriptive statistics of phase shift, in hours, for the circadian rhythm experiment.

Treatment	Data (h)	$\bar{Y}$	*s*	*n*
Control	0.53, 0.36, 0.20, –0.37, –0.60, –0.64, –0.68, –1.27	–0.3088	0.6176	8
Knees	0.73, 0.31, 0.03, –0.29, –0.56, –0.96, –1.61	–0.3357	0.7908	7
Eyes	–0.78, –0.86, –1.35 –1.48 –1.52, –2.04, –2.83	–1.5514	0.7063	7

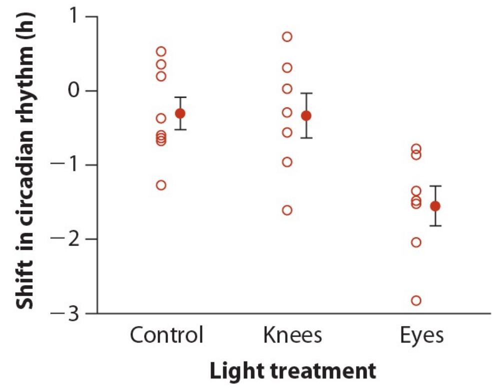


Fig 15.1-1

9

ANOVA calculations

Treatment	$\bar{Y}$	*s*	*n*
Control	–0.3088	0.6176	8
Knees	–0.3357	0.7908	7
Eyes	–1.5514	0.7063	7

- **Sums of squares (SS)** calculates two sources of variation (among and within groups)

$$SS_{groups} = \sum_i n_i \left(\bar{Y}_i - \bar{Y}\right)^2$$

$i = group$

$\bar{Y}_i = mean\ group\ i$

$\bar{Y} = mean\ all\ obs\ (grand\ mean)$

Grand mean

$$\bar{Y} = \frac{\sum_i n_i \bar{Y}_i}{N} = \frac{8(-0.3087) + 7(-0.3357) + 7(-1.5514)}{22} = -0.7127$$

10

ANOVA calculations

Treatment	$\bar{Y}$	*s*	*n*
Control	−0.3088	0.6176	8
Knees	−0.3357	0.7908	7
Eyes	−1.5514	0.7063	7

Grand mean = -0.7127

- **Sums of squares (SS)** calculates two sources of variation (among and within groups)

$$SS_{groups} = \sum_i n_i \left(\bar{Y}_i - \bar{Y}\right)^2$$

$i = group$

$\bar{Y}_i = mean\ group\ i$

$\bar{Y} = mean\ all\ obs\ (grand\ mean)$

$$SS_{groups} = 8[-0.3087 - (-0.7127)]^2 + 7[-0.3357 - (-0.7127)]^2 + 7[-1.5514 - (-0.7127)]^2$$

$$= 7.224$$

11

ANOVA calculations

Treatment	$\bar{Y}$	*s*	*n*
Control	−0.3088	0.6176	8
Knees	−0.3357	0.7908	7
Eyes	−1.5514	0.7063	7

Grand mean = -0.7127
SS_{groups} = 7.224

- **Sums of squares (SS)** calculates two sources of variation (among and within groups)

$$SS_{error} = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_i\right)^2 = \sum_i s_i^2 (n_i - 1)$$

$j = observation$

$$SS_{error} = (0.6176)^2(8-1) + (0.7908)^2(7-1) + (0.7063)^2(7-1)$$

$$= 9.415$$

12

ANOVA calculations

Treatment	$\bar{Y}$	s	n
Control	−0.3088	0.6176	8
Knees	−0.3357	0.7908	7
Eyes	−1.5514	0.7063	7

Grand mean = -0.7127

SS_{groups} = 7.224

SS_{error} = 9.415

SS_{total} = 16.639

13

Partitioning sum of squares

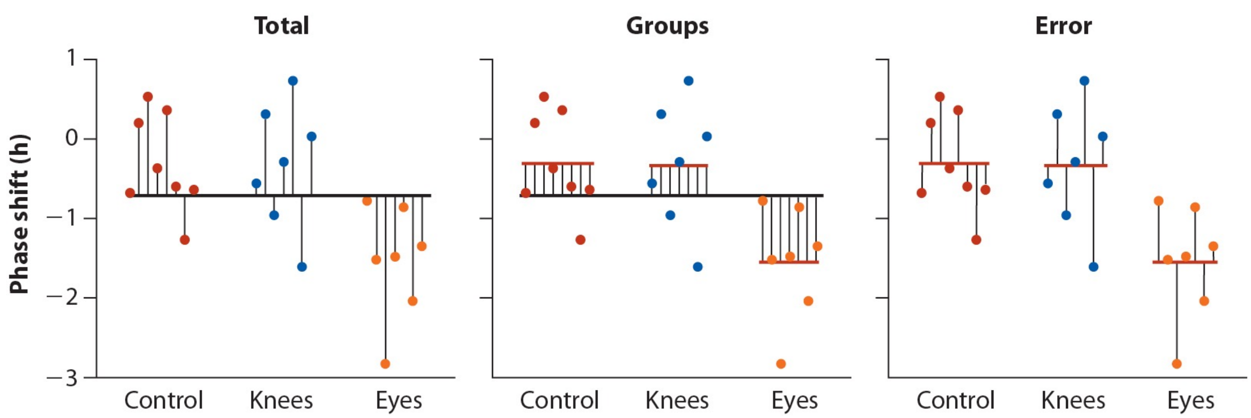


Fig 15.1-2

14

ANOVA calculations

Treatment	$\bar{Y}$	s	n
Control	−0.3088	0.6176	8
Knees	−0.3357	0.7908	7
Eyes	−1.5514	0.7063	7

Grand mean = -0.7127
SS_{groups} = 7.224
SS_{error} = 9.415
SS_{total} = 16.639

- **Group mean square (MS_{groups})** is variation among groups

$$MS_{groups} = \frac{SS_{groups}}{df_{groups}} \qquad df_{groups} = k - 1 \qquad k = number\ of\ groups$$

$$MS_{groups} = \frac{7.224}{3-1} = 3.6122$$

15

ANOVA calculations

Treatment	$\bar{Y}$	s	n
Control	−0.3088	0.6176	8
Knees	−0.3357	0.7908	7
Eyes	−1.5514	0.7063	7

Grand mean = -0.7127
SS_{groups} = 7.224
SS_{error} = 9.415
SS_{total} = 16.639
MS_{groups} = 3.6122

- **Group error square (MS_{error})** is variation among individuals in same group

$$MS_{error} = \frac{SS_{error}}{df_{error}} \qquad df_{error} = N - k \qquad k = number\ of\ groups \quad N = total\ number\ obs$$

$$MS_{error} = \frac{9.415}{22-3} = 0.4955$$

16

ANOVA calculations

Treatment	$\bar{Y}$	s	n
Control	−0.3088	0.6176	8
Knees	−0.3357	0.7908	7
Eyes	−1.5514	0.7063	7

Grand mean = -0.7127
SS_{groups} = 7.224
SS_{error} = 9.415
SS_{total} = 16.639
MS_{groups} = 3.6122
MS_{error} = 0.4955

- ***F*-ratio** test statistic

$$F = \frac{MS_{groups}}{MS_{error}}$$

$$F = \frac{3.6122}{0.4955} = 7.29$$

17

F test statistic

Treatment	$\bar{Y}$	s	n
Control	−0.3088	0.6176	8
Knees	−0.3357	0.7908	7
Eyes	−1.5514	0.7063	7

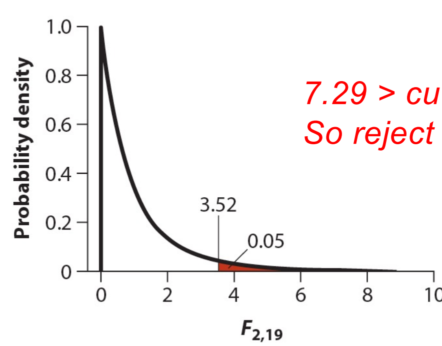
Grand mean = -0.7127
SS_{groups} = 7.224
SS_{error} = 9.415
SS_{total} = 16.639
MS_{groups} = 3.6122
MS_{error} = 0.4955

F = 7.29

- F statistic has pair of degrees of freedom
  - Numerator and denominator
  - $F_{2,19} = 7.29$
- Use F-distribution to calculate P-value
  - Stats table or computer

TABLE 15.1-3 An excerpt from Statistical Table D, with critical values of the *F*-distribution corresponding to the significance level $\alpha(1) = 0.05$.

	Numerator *df*									
Denominator *df*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35



7.29 > cutoff
So reject the null

Fig 15.1-1

18

ANOVA calculations

Treatment	$\bar{Y}$	s	n
Control	–0.3088	0.6176	8
Knees	–0.3357	0.7908	7
Eyes	–1.5514	0.7063	7

Grand mean = -0.7127
SS_{groups} = 7.224
SS_{error} = 9.415
SS_{total} = 16.639
MS_{groups} = 3.6122
MS_{error} = 0.4955
F = 7.29

- ANOVA table

TABLE 15.1-2 ANOVA table for the results of the circadian rhythm experiment (Example 15.1).

Source of variation	Sum of squares	*df*	Mean squares	*F*-ratio	*P*
Groups (treatment)	7.224	2	3.6122	7.29	0.004
Error	9.415	19	0.4955		
Total	16.639	21			

19

Variation explained

Treatment	$\bar{Y}$	s	n
Control	–0.3088	0.6176	8
Knees	–0.3357	0.7908	7
Eyes	–1.5514	0.7063	7

Grand mean = -0.7127
SS_{groups} = 7.224
SS_{error} = 9.415
SS_{total} = 16.639
MS_{groups} = 3.6122
MS_{error} = 0.4955
F = 7.29

- R^2 measures the fraction of variation in Y that is explained by group differences

$$R^2 = \frac{SS_{groups}}{SS_{total}}$$

$$R^2 = \frac{7.224}{16.639} = 0.43$$

20

Assumptions

- Measurements in every group represent a random sample from the corresponding population
- Variable is normally distributed in each of the k populations
  - Robust to deviations, particularly when sample size is large
- Variance is the same in all k populations
  - Robust to departures if sample sizes are large and balanced, and no more than 10x differences among groups

Alternatives

- Test normality with Shapiro-Wilk and test equal variances with Levene's test
- Data transformations can make data more normal and variances more equal
- Nonparametric alternative: Kruskal-Wallis test
  - Similar principle as Mann-Whitney U-test

Practice: indigobird repertoire size


Species	Mean	StDev	N
cha	15.7778	1.8329	18
cod	15.5238	2.2275	21
fun	13.9394	1.3906	33
pur	12.4231	1.2385	26


23

24

Which means are different?

- H_0: $\mu_1 = \mu_2 = \mu_3 \ldots \mu_n$
- H_A: mean of at least one group is different from at least one other group
- Reject the null, now what?

25

Planned comparisons

- A **planned comparison** is a comparison between means planned during the design of the study, identified before the data are examined
- In circadian clock follow-up study, the planned (*a priori*) comparison was difference in means between knee and control group

26

Knee vs control

Treatment	$\bar{Y}$	s	n
Control	–0.3088	0.6176	8
Knees	–0.3357	0.7908	7
Eyes	–1.5514	0.7063	7

Grand mean = -0.7127
SS_{groups} = 7.224
SS_{error} = 9.415
SS_{total} = 16.639
MS_{groups} = 3.6122
MS_{error} = 0.4955
F = 7.29

Difference in means

$$\bar{Y}_2 - \bar{Y}_1 = (-0.336) - (-0.309) = -0.027$$

Standard error of difference

$$SE = \sqrt{MS_{error}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad = 0.364$$

95% confidence interval of difference

$$\bar{Y}_2 - \bar{Y}_1 - SE \times t_{0.05(2),df} < \mu_2 - \mu_1 < \bar{Y}_2 - \bar{Y}_1 + SE \times t_{0.05(2),df}$$

$$-0.788 < \mu_2 - \mu_1 < 0.734$$

Range includes zero, so consistent with no diff in means!

27

Unplanned comparisons

- Comparisons are unplanned if you test for differences among all means
- Problem of multiple tests (increasing probability of Type I error) should be accounted for
- With the **Tukey-Kramer method** the probability of making at least one Type I error throughout the course of testing all pairs of means is no greater than the significance level α

28

Ex 15.4: Wood wide web

- Most plants have underground mutualistic interactions with mycorrhizae fungi
- Shaded trees might draw more carbon from mycorrhizae than non-shaded trees
- Experiment with three treatments:
  - Deep shade
  - Partial shade
  - No shade
- Transfer quantified using different carbon isotopes

Republished with permission of Nature Publishing Group, from Mycorrhizal fungi, David Read, Nature (vol. 388), issue 6642, Figure 1; 1997; permission conveyed through Copyright Clearance Center, Inc.

29

Ex 15.4: Wood wide web

TABLE 15.4-1 Summary of the net amount of carbon transferred from birch to Douglas-fir (Example 15.4).

Shade treatment	Sample mean $\overline{Y}_i$ (mg)	Sample standard deviation, s_i	n_i
Deep shade	18.33	6.98	5
Partial shade	8.29	4.76	5
No shade	5.21	3.00	5

TABLE 15.4-2 ANOVA table summarizing results of the Douglas-fir carbon-transfer data (Example 15.4).

Source of variation	Sum of squares	df	Mean squares	F-ratio	P
Groups (treatments)	470.704	2	235.352	8.784	0.004
Error	321.512	12	26.793		
Total	792.216	14			

Republished with permission of Nature Publishing Group, from Mycorrhizal fungi, David Read, Nature (vol. 388), issue 6642, Figure 1; 1997; permission conveyed through Copyright Clearance Center, Inc.

30

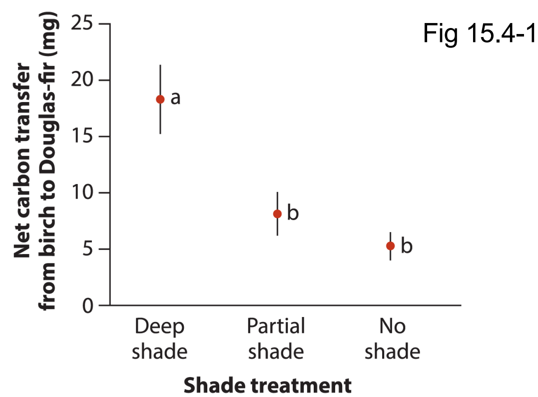
Tukey-Kramer method

- Works like a series of two-sample t-tests, but with a higher critical value to limit the Type I error rate
  - Because multiple tests are done, the adjustment makes it harder to reject the null

No shade	Partial shade	Deep shade
$\bar{Y}_3$	$\bar{Y}_2$	$\bar{Y}_1$
5.21	8.29	18.33

TABLE 15.4-3 Summary of Tukey–Kramer tests carried out on the results of Example 15.4.

Group i	Group j	$\bar{Y}_i - \bar{Y}_j$	SE	Test statistic q	Critical value $q_{0.05,3,12}$	Conclusion
Deep	No	13.12	3.2737	4.008	2.67	Reject H_0
Deep	Partial	10.04	3.2737	3.067	2.67	Reject H_0
Partial	No	3.08	3.2737	0.941	2.67	Do not reject H_0



31

Kruskal-Wallis post-hoc test?

- Suppose that your data…
  - Fail normality even after transformation
  - Generate a significant Kruskal-Wallis result
- So the interpretation is that the distribution of ranks differs for at least one group. But which one?
- Should not use Tukey-Kramer, which is a parametric test
- **Dunn's test** is the appropriate analysis for a post-hoc analysis of groups following a significant Kruskal-Wallis result
  - Will compare all possible pairs of groups while controlling for multiple tests

32

Notes

- Skipping sections 15.5 and 15.6