Chapter 12: Comparing two means

1

Lots of questions in biology compare two groups

- Wild type versus mutant
- Drug versus control
- Before versus after (some event)
- Male vs female
- Etc…

2

Different study designs

Two-sample design

- Two groups
- Each group composed of independent sample of units

Paired design

- Two groups
- Each sampled unit receives both treatments

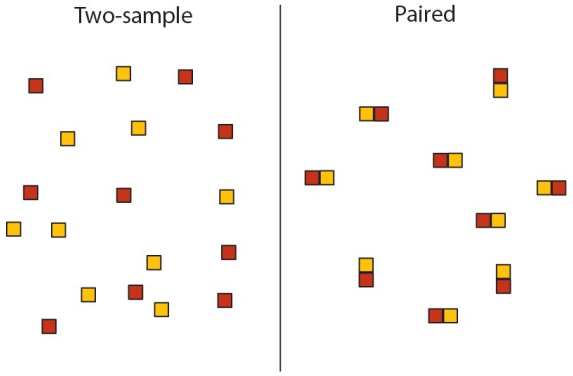


Fig 12.1-1

3

Different study designs

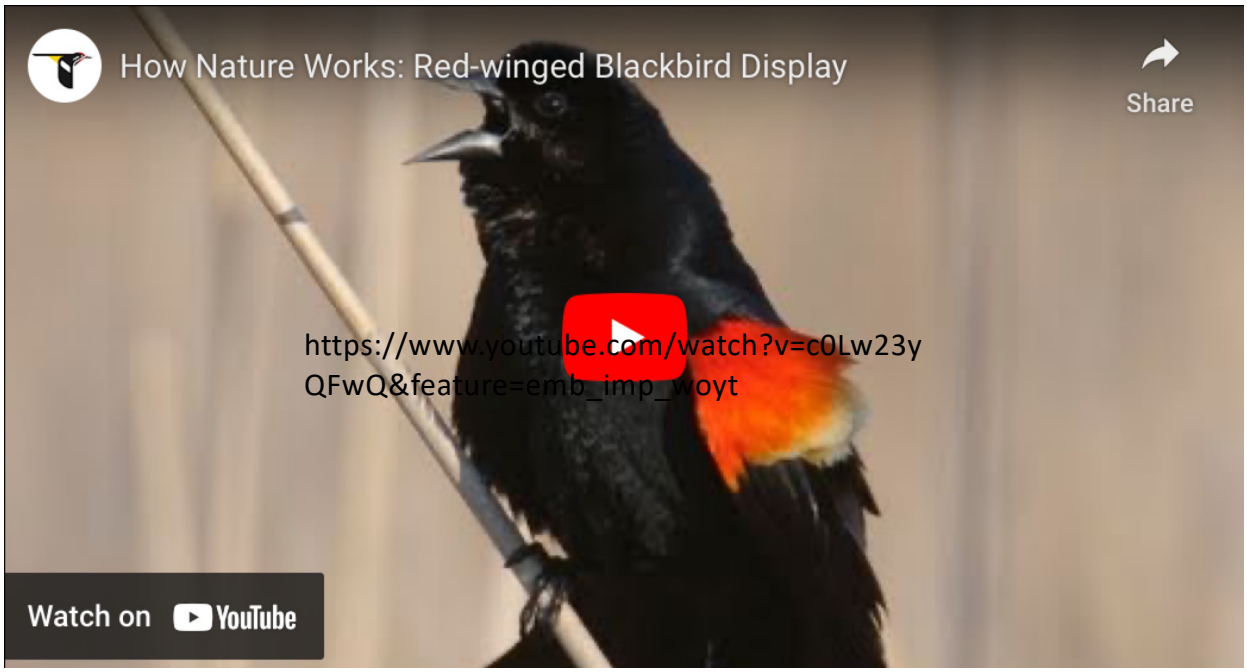
- In the **paired** design, both treatments are applied to every sampled unit. In the **two-sample** design, each treatment group is composed of an independent, random sample of units
- Paired designs are usually more powerful
  - Control for variation among sampling units
- But paired designs are often not possible

4

Paired design

- Examples…
  - Patient weight before and after hospitalization
  - Effects of sunscreen on one arm versus placebo on other arm
  - Effects of environment on identical twins raised under different socioeconomic conditions
- Two measurements from same sampling unit
- **Paired measurements are converted to a single measurement by taking the difference between them**.

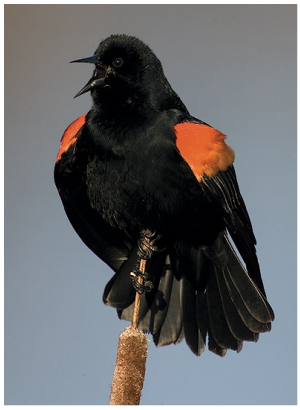
5



https://www.youtube.com/watch_popup?v=c0Lw23yQFwQ

6

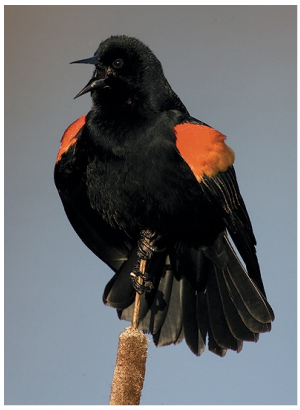
Ex 12.2: So macho it makes you sick?



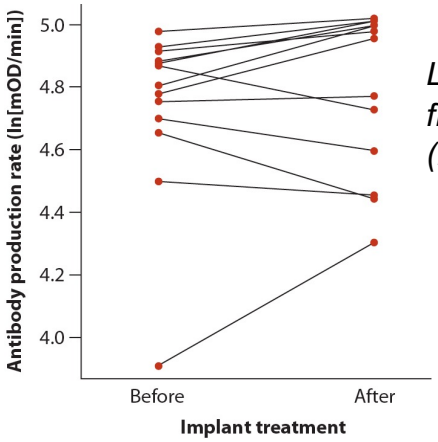
- In most species males compete with each other for access to females
- In many species, males with higher testosterone levels are more successful
- Does high testosterone have a trade-off of reduced immunocompetence?

7

Ex 12.2: So macho it makes you sick?



- Measured antibody production rate in 13 male red-winged blackbirds before and after a testosterone implant



Lines link two measurements from same unit (i.e., individual bird)

Fig 12.2-1

8

Difference between paired measurements

TABLE 12.2-1 Antibody production rate in blackbirds before and after testosterone implants. Each bird is represented by a single row and has a pair of antibody measurements; *d* is the difference ("after" minus "before") between the pair of measurements.

Male identification number	Before implant: Antibody production (ln[mOD/min])	After implant: Antibody production (ln[mOD/min])	*d*
1	4.65	4.44	–0.21
4	3.91	4.30	0.39
5	4.91	4.98	0.07
6	4.50	4.45	–0.05
9	4.80	5.00	0.20
10	4.88	5.00	0.12
15	4.88	5.01	0.13
16	4.78	4.96	0.18
17	4.98	5.02	0.04
19	4.87	4.73	–0.14
20	4.75	4.77	0.02
23	4.70	4.60	–0.10
24	4.93	5.01	0.08

d_i = (antibody production of male i after) − (antibody production of male i before).

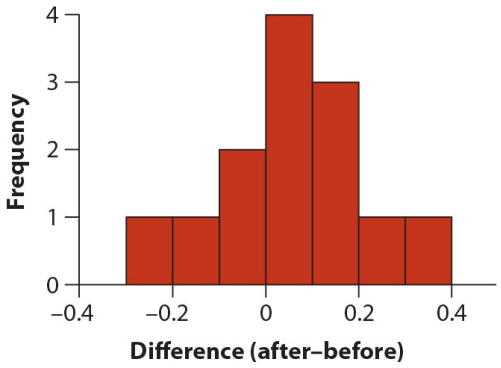


Fig 12.2-2

9

Estimating mean difference

- From sample of d_i
  - $\overline{d} = 0.056$ (after-before)
  - $s_d = 0.159$
  - $n = 13$
- With these estimates you can calculate the 95% CI for true mean difference (μ_d)

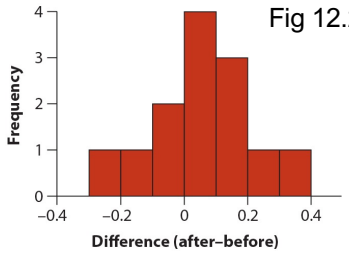


Fig 12.2-2

$$SE_{\overline{d}} = \frac{s_d}{\sqrt{n}}$$

$$SE_{\overline{d}} = \frac{s}{\sqrt{n}} = \frac{0.159}{\sqrt{13}} = 0.044$$

$$t_{0.05(2),12} = 2.18$$

$$\overline{d} - t_{0.05(2),df} SE_{\overline{d}} < \mu_d < \overline{d} + t_{0.05(2),df} SE_{\overline{d}}$$

$$0.056 - (2.18 \times 0.044) < \mu_d < 0.056 + (2.18 \times 0.044)$$

$$\mathbf{-0.040 < \mu_d < 0.152}$$

Range includes 0 (no difference), but also consistent with a slight drop or a modest increase in immunocompetence following testosterone implant

10

Paired t-test

- Used to test the null hypothesis that the mean difference of paired measurements equals a specific value
- H_0: mean change in antibody production after testosterone implants was zero ($\mu_d = 0$)
- H_A: mean change in antibody production after testosterone implants was not zero ($\mu_d \neq 0$)

11

Paired t-test statistic

- One paired samples are reduced to single measurement (d), the calculation of test statistic same as one-sample t-test

$$SE_{\overline{d}} = \frac{s_d}{\sqrt{n}} \qquad SE_{\overline{d}} = \frac{s_d}{\sqrt{n}} = \frac{0.159}{\sqrt{13}} = 0.044$$

$$t = \frac{\overline{d} - \mu_{d0}}{SE_{\overline{d}}} \qquad t = \frac{\overline{d} - \mu_{d0}}{SE_{\overline{d}}} = \frac{0.056 - 0}{0.044} = 1.27$$

$$df = n - 1 \qquad df = n - 1 = 13 - 1 = 12$$

12

Paired t-test statistic

- $t_{12} = 1.27$
- Can determine *P* with computer or statistical table
- *P*-value = 0.23
- Fail to reject the null hypothesis that the mean change in antibody production after the testosterone implant is zero

13

Assumptions of paired t-test

- Sampling units are randomly sampled from the population
- Paired differences have a normal distribution in the population

14

Formal tests of normality (Ch13.1)

- H_0: sample has normal distribution
- H_A: sample does not have normal distribution
- Should be used with caution
  - Small sample sizes lack power to reject a false null (Type II error)
  - Large sample sizes can reject null when the departure from normality is minimal and would not affect methods that assume normality

15

Shapiro-Wilk test (Ch13.1)

- A **Shapiro-Wilk test** evaluates the goodness of fit of a normal distribution to a set of data randomly sampled from a population
- Most commonly used formal test of normality
- Estimates mean and standard deviation using sample data
- Tests goodness-of-fit between sample data and normal distribution (with mean, sd of the sample)

16

17

Two-sample design

- Each treatment group is composed of an independent, random sample of units
- Wild type versus control, drug versus placebo, etc., where treatments are applied to separate and independent samples

18

https://www.youtube.com/watch_popup?v=okQYO10MT3c

19

Ex 12.3: Spike or be spiked

- Horned lizards and loggerhead shrikes
- Do horns protect lizards from being eaten?
- Measure horn lengths for 30 impaled lizards, and 154 live lizards
- Sampling units for each population are NOT "paired"

20

Ex 12.3: Spike or be spiked

TABLE 12.3-1 Summary statistics for lizard horn lengths.

Lizard group	Sample mean $\overline{Y}$ (mm)	Sample standard deviation s (mm)	Sample size n
Living	24.28	2.63	154
Killed	21.99	2.71	30

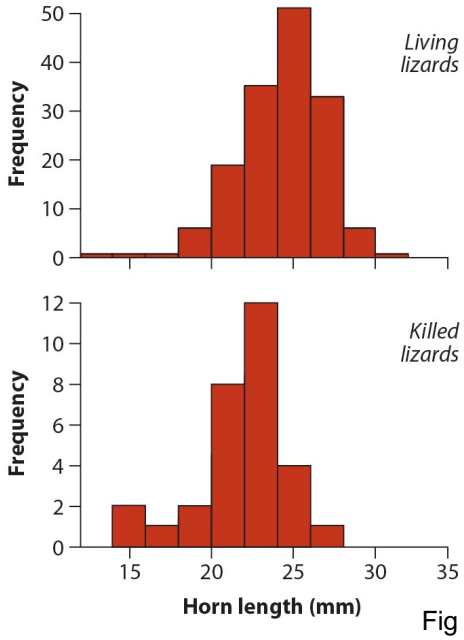


Fig 12.3-1

21

CI for difference between two means

- Statistic of interest: $\overline{Y}_1 - \overline{Y}_2$
- The **pooled sample variance s_p^2** is the average of the variances of the samples weighted by their degrees of freedom

$$s_p^2 = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} \qquad s_p^2 = \frac{153(2.63^2) + 29(2.71^2)}{153 + 29} = 6.98$$

$$df_1 = n_1 - 1 \qquad df_1 = 154 - 1 = 153$$

$$df_2 = n_2 - 1 \qquad df_2 = 30 - 1 = 29$$

22

CI for difference between two means

- Statistic of interest: $\overline{Y}_1 - \overline{Y}_2$
- Standard error of $\overline{Y}_1 - \overline{Y}_2$

$$SE_{\overline{Y}_1-\overline{Y}_2} = \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \qquad SE_{\overline{Y}_1-\overline{Y}_2} = \sqrt{6.98\left(\frac{1}{154}+\frac{1}{30}\right)} = 0.527$$

23

CI for difference between two means

- Statistic of interest: $\overline{Y}_1 - \overline{Y}_2$
- Confidence interval for $\overline{Y}_1 - \overline{Y}_2$

$$\left(\overline{Y}_1 - \overline{Y}_2\right) - t_{\alpha(2),df} SE_{\overline{Y}_1-\overline{Y}_2} < \mu_1 - \mu_2 < \left(\overline{Y}_1 - \overline{Y}_2\right) + t_{\alpha(2),df} SE_{\overline{Y}_1-\overline{Y}_2}$$

$$2.29 - 1.97(0.527) < \mu_1 - \mu_2 < 2.29 + 1.97(0.527)$$

$$\mathbf{1.25 < \mu_1 - \mu_2 < 3.33}$$

$df = df_1 + df_2 = 153 + 29 = 182$

$t_{0.05(2),182} = 1.97$

Range does not include 0 (no difference); consistent with horns on surviving lizards being 1.25-3.33 mm longer

24

Two-sample t-test

- Simplest test to compare the means of a numerical variable two independent groups
- Most commonly…
- H_0: $\mu_1 = \mu_2$
- H_A: $\mu_1 \neq \mu_2$

25

Two-sample t-test

- Test statistic

$$t = \frac{(\overline{Y}_1 - \overline{Y}_2)}{SE_{\overline{Y}_1 - \overline{Y}_2}}$$

$$t = \frac{2.29}{0.527} = 4.35$$

$$df = 153 + 29 = 182$$

P-value for $t_{182} = 4.35$?

P-value = 0.000023

Reject the null hypothesis of equal means

26

Assumptions of two-sample t-test

- Each of two samples is a random sample from its population
- Numerical variable is normally distributed in each population
  - Robust to minor deviations from normality
  - Need to run Shapiro-Wilk test on **both** samples
- Standard deviation (and variance) of the numerical variable is the same in both populations
  - Robust to some deviation from this if sample sizes of two groups are approximately equal

27

Formal tests of equal variance

- An *F*-test is sometimes used, but it is highly sensitive to departures from the assumption that the measurements are normally distributed in the population
- **Levene's test** performs better and is recommended
  - H_0: variances of the two groups are equal
  - H_A: variances of the two groups are not equal
  - Can be extended to more than two groups

28

What if variances in two groups are not equal?

- Standard *t*-test works well if both sample sizes are greater than 30 and there is less than 3-fold difference in standard deviations
- **Welch's *t*-test** compares the means of two groups and can be used even when the variances of the two groups are not equal
- Slightly less power compared to standard *t*-test

29

30

Correct sampling units

- When comparing the means of two groups, an assumption is that the samples being analyzed are random samples
- Often, repeated measurements are taken on each sampling unit
- Makes the identification of independent units more challenging
- An example…

31

Ex 12.4: So long; thanks for all the fish

- Brook trout is native to eastern North America

- Was introduced to west, competed with native chinook salmon
- Does presence of brook trout affect survivorship of chinook salmon?
- Marked salmon in streams with or without trout, estimated survivorship

32

Ex 12.4: So long; thanks for all the fish

TABLE 12.4-1 The numbers and proportion of chinook released and surviving in streams with and without brook trout. The study included 12 streams in total.

Brook trout	Number of salmon released	Number of salmon surviving	Proportion surviving
Present	820	166	0.202
Present	960	136	0.142
Present	700	153	0.219
Present	545	103	0.189
Present	769	173	0.225
Present	1001	188	0.188
Absent	467	180	0.385
Absent	959	178	0.186
Absent	1029	326	0.317
Absent	27	7	0.259
Absent	998	120	0.120
Absent	936	135	0.144
Total	9211	1865	

TABLE 12.4-2 Number of salmon surviving and not surviving in each trout treatment.

	Trout absent	Trout present
Survived	946	919
Did not survive	3470	3876

χ^2 analysis?

$P = 0.0071$

But this would be a flawed analysis!!

33

Ex 12.4: So long; thanks for all the fish

TABLE 12.4-1 The numbers and proportion of chinook released and surviving in streams with and without brook trout. The study included 12 streams in total.

Brook trout	Number of salmon released	Number of salmon surviving	Proportion surviving
Present	820	166	0.202
Present	960	136	0.142
Present	700	153	0.219
Present	545	103	0.189
Present	769	173	0.225
Present	1001	188	0.188
Absent	467	180	0.385
Absent	959	178	0.186
Absent	1029	326	0.317
Absent	27	7	0.259
Absent	998	120	0.120
Absent	936	135	0.144
Total	9211	1865	

- Individual salmon are NOT a random sample
- Grouped by stream
- *Stream* is the independently sampled unit
- Salmon within streams should be summarized by single measurement
  - Proportion surviving

34

Testing difference between means

- H_0: mean proportion of salmon surviving is the same in streams with and without trout
- H_A: mean proportion of salmon surviving is different in streams with and without trout

TABLE 12.4-3 Sample statistics for the proportion of salmon surviving in streams (Example 12.4), using the stream as the sampling unit.

Group	Sample mean	Sample standard deviation, s_i	Sample size, n_i
Brook trout present	0.194	0.0297	6
Brook trout absent	0.235	0.1036	6

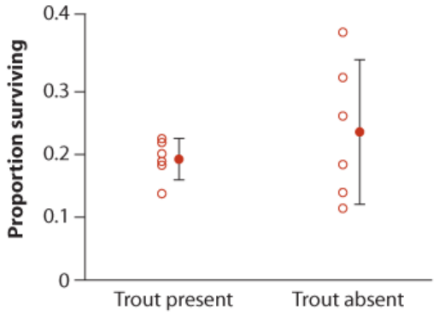


Fig 12.4-1

35

When sample standard deviations differ

- Since sample standard deviations are substantially different (> 3x difference) the Welch's t-test should be used
  - Levene's test *P*-value = 0.009
- Formulae are different than standard two-sample t-test

$$t_5 = 0.93$$

$$P = 0.39$$

Fail to reject the null, no difference in salmon survivorship with and without trout

36

37

Fallacy of indirect comparison

- Common error: compare each group mean to hypothesized value rather than comparing group means to each other
- Ex: since group 1 is significantly difference than zero, but group 2 is not, then groups 1 and 2 are significantly different from each other
- This error is called the **fallacy of indirect comparison**
- **Comparisons between two groups should always be made directly, not indirectly by comparing both to the same hypothesized value**

38

Ex 12.5: Mommy's baby, Daddy's maybe

- Mothers are assured maternity, but fathers are not assured paternity
- Do babies resemble their fathers, thus providing paternal assurance and increasing parental care by father?
- Asked volunteers to match picture of baby to one of three pictures of potential mothers/fathers

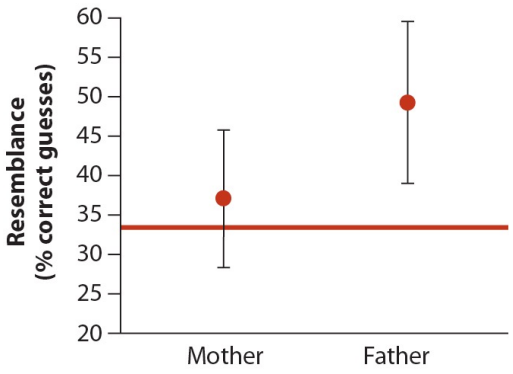


Fig 12.5-1

39

Ex 12.5: Mommy's baby, Daddy's maybe

- Study concluded that volunteers were not able to pick out mother (not different than null 33%)
- And that volunteers were able to pick out fathers (diff than 33%)
- But never directly compared mothers and fathers!
  - Test would have found no sig diff

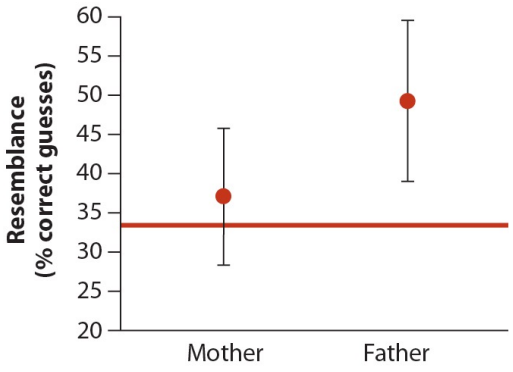


Fig 12.5-1

40

Interpreting overlap in confidence intervals

- Papers often report means and confidence intervals for two or more groups without running a two-sample t-test
- What can you infer?

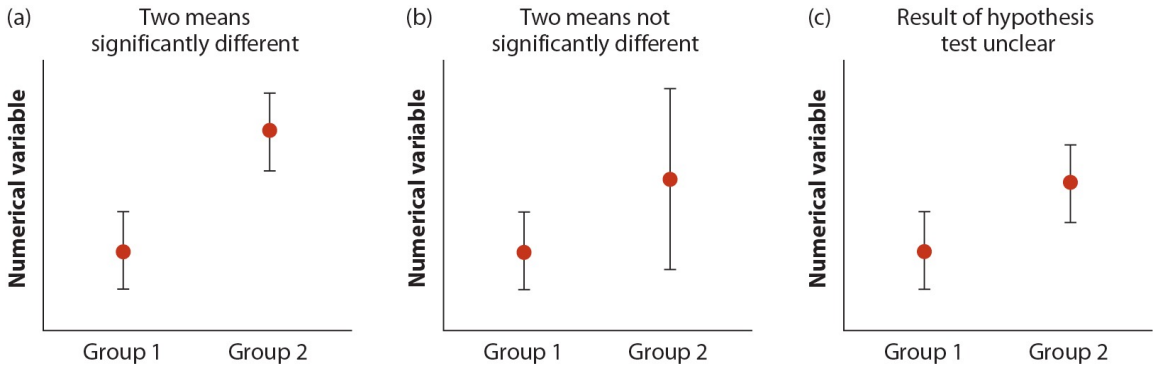


Fig 12.6-1

41