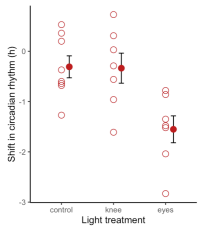
THE ANALYSIS OF VARIANCE (ANOVA)
for comparing multiple sample means (groups or treatments)

H_0: The samples come from statistical populations with the same mean, i.e., $\mu_{control} = \mu_{knee} = \mu_{eyes}$.

H_A: At least two samples come from different statistical populations with different means.



P-value (ANOVA) = 0.00447

Research conclusion: Light treatment influences shifts in circadian rhythm.

1

ANOVA

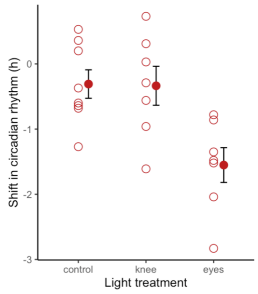
Research conclusion: Light treatment influences shifts in circadian rhythm.

How does light treatment influence shifts in circadian rhythm?

How do we know which group means differ from one another?

Why not simply not contrast all pairs of means using a two-sample mean t-test?

"The knees who say night"
Control vs. knee; control vs. eyes; knee vs. eyes?



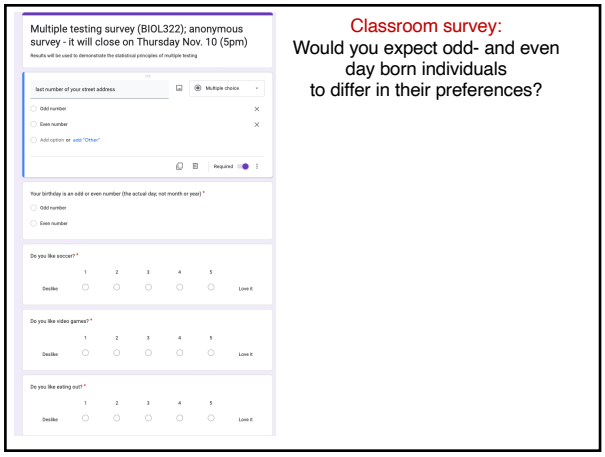
2

After ANOVA:

- Multiple testing and post hoc ("occurring or done after the event"; hoc = "not planned before it happens") tests.

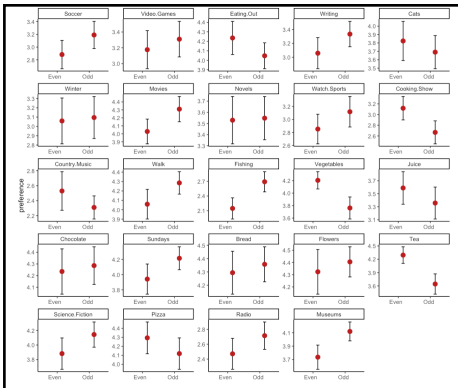
- The concept of family wise type I error and why we conduct ANOVAs first instead of two-sample t-tests!

3



Classroom survey:
Would you expect odd- and even day born individuals to differ in their preferences?

4

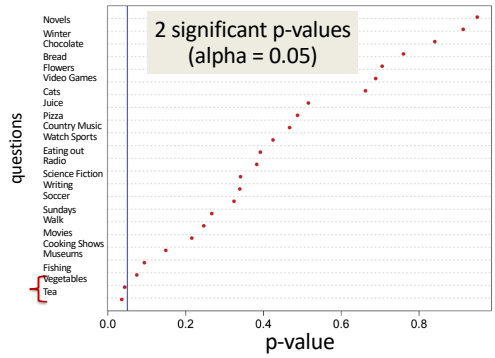


Birthday and preferences

One should not have any theoretical basis for preferences to vary among groups other than by chance alone!

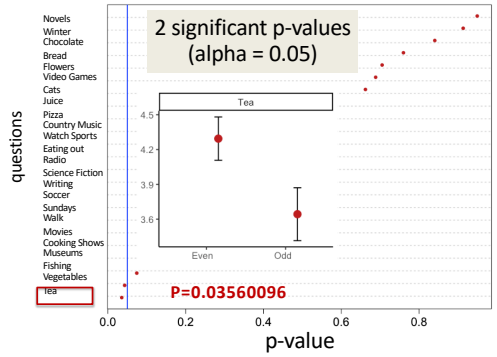
5

Contrast between odd and even-day born individuals - probability of rejection based on a two-sample t test (odd *versus* even)



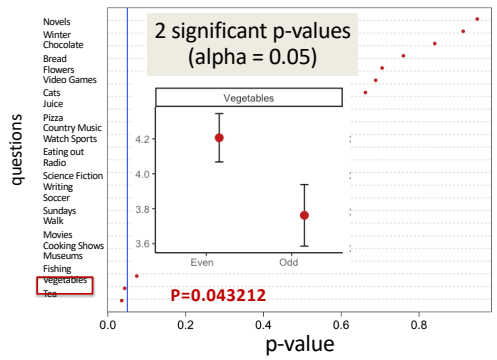
6

Contrast between odd and even-day born individuals - probability of rejection based on a two-sample t test (odd *versus* even)



7

Contrast between odd and even-day born individuals - probability of rejection based on a two-sample t test (odd *versus* even)

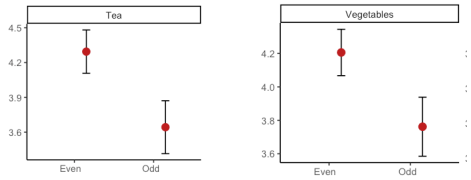


8

Birthday and Preferences:

We were even able to observe an association between liking tea and liking eating vegetables (in a plausible direction) simply by separating individuals according to their birthdays.

How can that be?



9

Another example of significance when there should be none

Lee, K.L. et al. (1980) Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. Circulation, 61: 508-515. DOI: 10.1161/01.cir.61.3.508

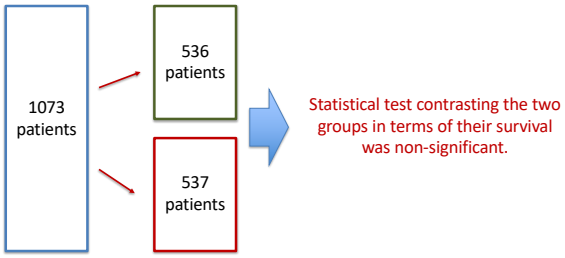
A simulated randomized clinical trial in coronary artery disease was conducted to illustrate the need for clinical judgment and modern statistical methods in assessing therapeutic claims in studies of complex diseases.

In this example, **1073 consecutive**, medically treated coronary artery disease patients from the Duke University data bank were randomized into **two groups**. The groups were reasonably comparable and, as expected, **there was no overall difference in survival between the two groups**.

10

Another example of significance when there should be none

1073 heart disease patients were **RANDOMLY** placed into two groups; no statistical difference was found in survival (not surprising given that they were randomly placed into groups as an exercise to demonstrate the issues with multiple testing) between the two groups.



11

Another example of significance when there should be none

Lee, K.L. et al. (1980) Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. Circulation, 61: 508-515. DOI: 10.1161/01.cir.61.3.508

A simulated randomized clinical trial in coronary artery disease was conducted to illustrate the need for clinical judgment and modern statistical methods in assessing therapeutic claims in studies of complex diseases.

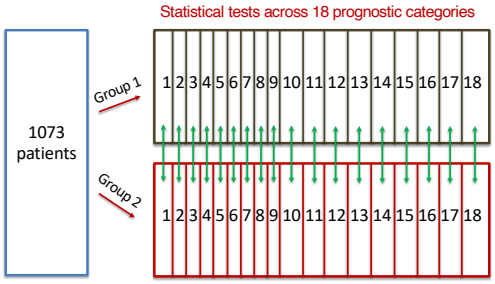
In this example, **1073 consecutive**, medically treated coronary artery disease patients from the Duke University data bank were randomized into **two groups**. The groups were reasonably comparable and, as expected, **there was no overall difference in survival between the two groups**.

But when patients were further subdivided into 18 prognostic categories, in a subgroup of 397 patients characterized by three-vessel disease and an abnormal left ventricular contraction, however, survival of group 1 patients was significantly different from that of group 2 patients.

12

Another example of significance when there should be none

The analysis of individuals divided into 18 prognostic categories based on heart morphology revealed a difference in survival between two groups in one of the categories. However, because the division of individuals into these categories was random, any observed difference in survival should be attributed to chance alone rather than an underlying causal factor.



13

Another example of significance when there should be none

Lee, K.L. et al. (1980) Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. Circulation, 61: 508-515. DOI: 10.1161/01.cir.61.3.508

A simulated randomized clinical trial in coronary artery disease was conducted to illustrate the need for clinical judgment and modern statistical methods in assessing therapeutic claims in studies of complex diseases.

In this example, **1073 consecutive**, medically treated coronary artery disease patients from the Duke University data bank were randomized into **two groups**. The groups were reasonably comparable and, as expected, **there was no overall difference in survival between the two groups**.

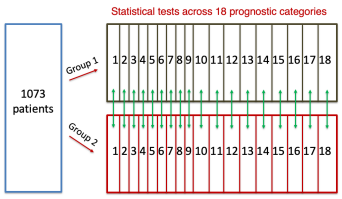
But when patients were further subdivided into 18 prognostic categories, in a subgroup of 397 patients characterized by three-vessel disease and an abnormal left ventricular contraction, however, survival of group 1 patients was significantly different from that of group 2 patients.

Multitest adjustment procedures indicated that the observed difference was likely the result of small imbalances in the distribution of several prognostic factors combined. This highlights the importance of clinicians exercising caution when interpreting such results. The differences could be attributable to chance or insufficient baseline comparability between groups, rather than a true effect of the therapy being evaluated.

14

Another example of significance when there should be none

- Patients grouped according as "three-vessel disease and an abnormal left ventricular contraction" were found to have differences between in survival between the two groups.
- However, patients were randomly assigned to each of the two groups in the beginning (i.e., survival *versus* non-survival).
- ***How did that happen?***



15

Another example of significance when there should be none

Lee, K.L. et al. (1980) Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. Circulation, 61: 508-515. DOI: 10.1161/01.cir.61.3.508

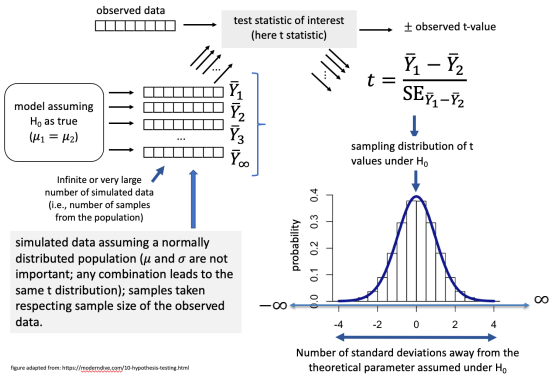
- 1073 heart disease patients were randomly placed into two groups; no difference was found in survival (not surprising) between the two groups.
  [akin to BIOL322 students divided according to their "birthdays"]
- Individuals within each group were then contrasted according to 18 prognostic categories (heart morphology used to predict the likely outcome of a heart condition). **[prognostics are akin to our 24 questions]**
- Individuals between the two groups were then contrasted for their differences in survival (any difference in survival should be due to chance alone as individuals were randomly divided into these categories). **[p-values for a test comparing the two groups]**
- Patients grouped according as "three-vessel disease and an abnormal left ventricular contraction" were found to have differences between in survival between the two groups. **[students differ in their preferences for drinking tea and eating vegetables]**
- However, patients were randomly assigned to each of the two groups in the beginning (i.e., survival versus non-survival). **[one should not expect differences related to odd/even birthdays]**
- **How did that happen?**

16

Let's take a break – 1 minute

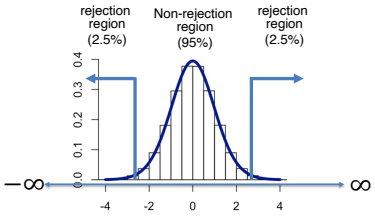
17

Remembering how the sampling distribution under the null hypothesis is built (conceptually)



18

Why when comparing multiple means, one should start with an ANOVA and not by two-sample t-tests?

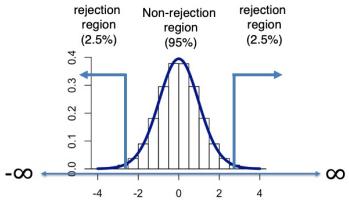


Under the null hypothesis (H_0), all possible t-values, including those in the rejection region, can occur. However, the probability of sampling a t-value that falls within the rejection region is equal to the chosen alpha level (e.g., 0.05). This reflects the likelihood of committing a Type I error when the null hypothesis is true.

If you conduct one test, the probability of committing a Type I error is equal to the chosen alpha level, such as 0.05. However, if you conduct multiple tests, the probability of committing at least one Type I error increases

19

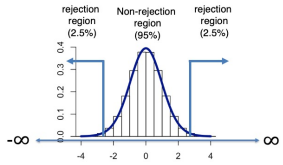
From tutorial 9: When making inferences from samples, we face a trade-off: to control the risk of making one type of error (Type I or false positives), we must accept a manageable risk of making another (Type II or false negatives).



All the infinite t values are possible under H_0, even the ones in the rejection region (they have a probability of alpha=0.05 to be sampled).

20

For each test conducted, the probability of committing a Type I error, which is rejecting the null hypothesis when it is actually true, is equal to the chosen alpha level (e.g., 0.05 or 5%). This remains true for individual tests, but when multiple tests are conducted, the cumulative probability of committing at least one Type I error increases unless appropriate adjustments are made.



All the infinite t values are possible under H_0, even the ones in the rejection region (they have a probability of alpha=0.05 to be sampled).

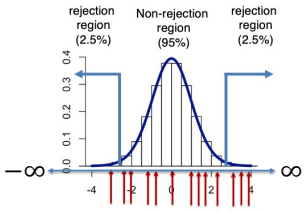
There is a high likelihood (95% chance) that the test will not yield a significant result (i.e., p-value ≥ 0.05) purely by chance. This scenario can be compared to throwing a dart without aiming at the target distribution, where most outcomes do not land in the rejection region.

Let's assume that the null hypothesis is indeed true (like in our student survey and the heart study).

21

For each test conducted, the probability of committing a Type I error, which is rejecting the null hypothesis when it is actually true, is equal to the chosen alpha level (e.g., 0.05 or 5%). This remains true for individual tests, but when multiple tests are conducted, the cumulative probability of committing at least one Type I error increases unless appropriate adjustments are made.

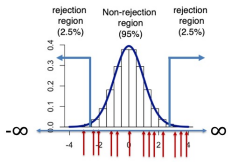
If multiple darts are thrown without aiming at the distribution, there is still a 5% chance of landing in the rejection region (i.e., obtaining a p-value < alpha) by pure chance. This reflects the probability of a Type I error for each individual test.



All the infinite t values are possible under H_0, even the ones in the rejection region (they have a probability of alpha=0.05 to be sampled)

22

Why when comparing multiple means one should start with an ANOVA and not multiple t-test – because they inflate the number of false positive tests

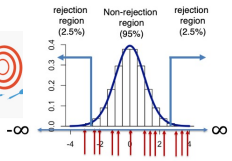


If you conduct too many tests, you will eventually, by chance alone, obtain a t-value that falls in the rejection region. Remember that the sampling distribution includes all possible values for the t-statistic measuring the difference between two sample means, assuming that the null hypothesis (H_0) is true.

Because we eliminate implausible (low-probability) values in the sampling distribution under the assumption that the null hypothesis is true, and use an alpha value to define the rejection area, it is clear that conducting too many tests will eventually lead to Type I errors for a given alpha. In other words, the more tests you conduct, the higher the likelihood of false positives (rejecting the null hypothesis when it should not be rejected).

23

Would you expect odd- and even day born individuals to differ in their preferences?



If we set an alpha of 0.05, i.e., acceptance area of 95% (0.95), then the chance of finding at least one significant test when you should not (i.e., false positive) out of 24 tests (groups) is:

$1\text{-}0.95^{24} = 0.708$

71% chance of finding at least 1 significant difference between odd and even born individuals in their preferences when H_0 is true!

24

Would you expect odd- and even day born individuals to differ in their preferences?

odd-day / even-day born	dislike 1	2	3	4	Love it 5
			X		
1) Do you like soccer?	X				
2) Do you like playing video games?			X		
3) Do you like eating out?					
4) Do you enjoy writting?					
5) Do you like cats?			X		
6) Do you like to watch movies?					X
7) Do you like to read novels?					

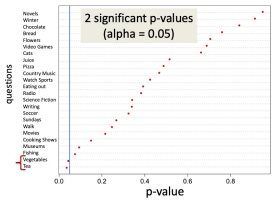
.....

21) Do you like science fiction?	X				
22) Do you like pizza?		X			
23) Do you like to listen to the radio?				X	
24) Do you like museums?			X		

$1\text{-}0.95^{24}=0.708$

70.1% chance of finding at least 1 significant test when all H_0 are true!

2 tests were in fact significant.



25

Let's assume that 100 tests were conducted:

If we set an alpha of 0.05, i.e., acceptance area of 95% (0.95), then the chance of finding at least one significant test when you should not (i.e., false positive) out of 100 tests (groups) is:

$$1\text{-}0.95^{100} = 0.994$$

99.4% chance of finding at least 1 significant difference between group 1 and group 2 when H_0 is true!

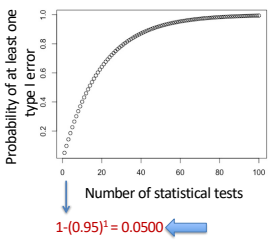
SO, 100% chance if you conduct 100 tests on samples that are expected to vary just due to chance alone (i.e., for which the null hypothesis H_0 is true).

26

If we set an alpha of 0.05, i.e., acceptance area of 95% (0.95), then the chance of finding at least one significant test when you should not (i.e., false positive) out of 1 test is obviously the original alpha:

$1\text{-}0.95^{1} = 0.05$

5% chance of finding at least 1 significant test when H_0 is true!



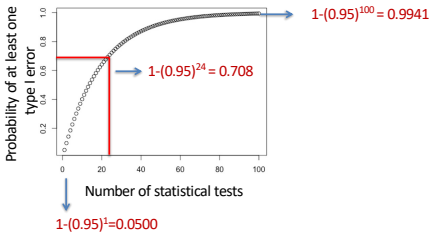
$1\text{-}(0.95)^{1} = 0.0500$

27

If we set an alpha of 0.05, i.e., acceptance area of 95% (0.95), then the chance of finding at least one significant test when you should not (i.e., false positive) out of 24 tests is:

$1\text{-}0.95^{24}=0.708$

70.1% chance of finding at least 1 significant test when H_0 is true!



28

Let's take a break – 1 minute

29

The purpose of performing an ANOVA beforehand is to protect against inflated Type I errors that can arise from conducting multiple pairwise comparisons.

When ANOVA yields a significant result, the next step is to determine which pairs of means can be considered genuinely significant.

To address this, we need a method to control for the increased likelihood of Type I errors due to multiple testing.

The Tukey's honest test.

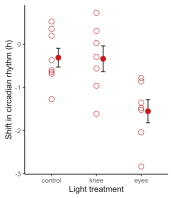
30

THE ANALYSIS OF VARIANCE (ANOVA) for comparing multiple sample means (groups)

H_0: The samples come from statistical populations with the same mean, i.e., $\mu_{control} = \mu_{knee} = \mu_{eyes}$.

H_A: At least two samples come from different statistical populations with different means.

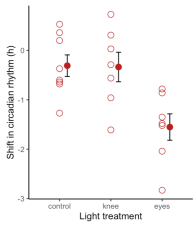
When ANOVA is significant, which pairs of means can be "honestly" considered significant?



31

How many pairs of means are possible to be contrasted (i.e., differences between means)?

$$\binom{r}{2} = \frac{r!}{2!(r-2)!} = \frac{r(r-1)}{2}$$



$$\frac{3(3-1)}{2} = 3$$

Control – Knee
Control – Eyes
Knee - Eyes
} 3 mean pairs (contrasts)

32

The post-hoc (after ANOVA) - Tukey's honest test

There is a pair of hypotheses for each pair of means as follows:

H_0: $\mu_i = \mu_j$ for each pair $i \neq j$

H_A: $\mu_i \neq \mu_j$ for each pair

i and *j* stand for the subscripts of the groups (treatments) being compared.

Control – Knee
Control – Eyes
Knee - Eyes
} 3 mean pairs (contrasts)

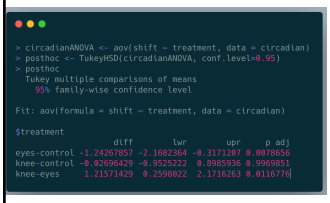
33

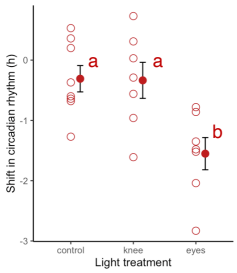
Tukey's honest test in R

```
> circadianANOVA <- aov(shift ~ treatment, data = circadian)
> posthoc <- TukeyHSD(circadianANOVA, conf.level=0.95)
> posthoc
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = shift ~ treatment, data = circadian)

$treatment
                   diff        lwr        upr     p adj
eyes-control -1.24267857 -2.1682364 -0.3171207 0.0078656
knee-control -0.02696429 -0.9525222  0.8985936 0.9969851
knee-eyes     1.21571429  0.2598022  2.1716263 0.0116776
```

34

Tukey's honest test in R: we often use letters (a, b, c., etc) to show on graphs the means that are different and similar.





35

The test statistic for the Tukey Test is calculated as:

$$Q = \frac{|X_i - X_j|}{SE}$$

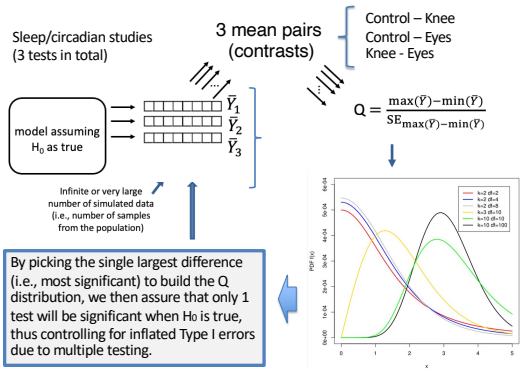
$$SE_{i-j} = \sqrt{\frac{s^2_{p(i,j)}}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$$s^2_{p(i,j)} = \frac{df_i s_i^2 + df_j s_j^2}{df_i + df_j}$$

The quantity s_p^2 is called the pooled sample variance and is the average of the sample variances weighted by their degrees of freedom.

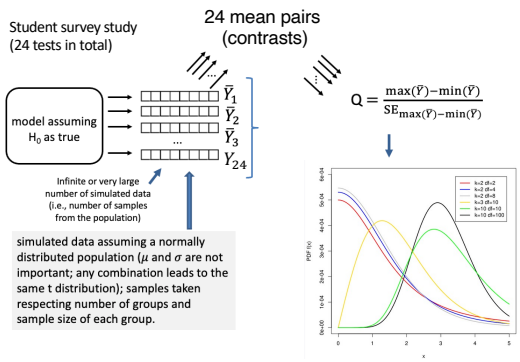
36

Q is then contrasted against a distribution built from the largest possible difference between the two-sample means given the number of two-sample tests and assuming H_0 as true.



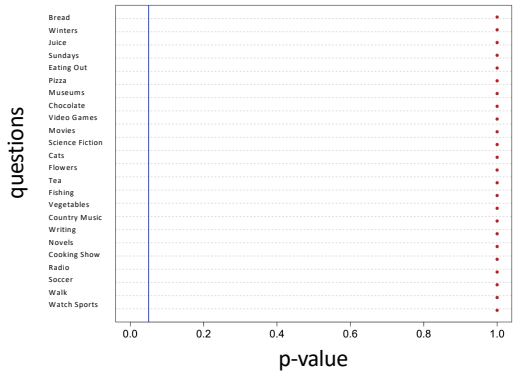
37

Q is then contrasted against a distribution built from the largest possible difference between the two-sample means given the number of two-sample tests and assuming H_0 as true.



38

No difference from the survey detected as significant after the Tukey test



39

ANOVA & the Tukey-test:

Assumptions:

- Each of the samples (observations within groups) is a random sample from its population.

- The variable (shift in circadian rhythm) is normally distributed in each (treatment) population.

- The variances are equal among all statistical populations from which the treatments were sampled.

40

Testing for differences in variances among populations can be done using Levene's test. While its calculation may be too complex for the BIOL322 level, it is important to understand its existence, its utility, and how to apply it in R.

H_0: $\sigma^2_{control} = \sigma^2_{knee} = \sigma^2_{eye}$

H_A: At least one population variance (σ^2) is different from another population variance or other population variances.

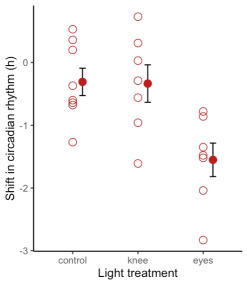
We need to generate evidence towards H_0 to apply an ANOVA to the data at hands.

41

Testing differences in variances among populations - The Levene's test

H_0: $\sigma^2_{control} = \sigma^2_{knee} = \sigma^2_{eye}$

H_A: At least one population variance (σ^2) is different from another population variance or other population variances.



42

Levene's test:

Assumptions:

- Each of the samples (observations within groups) is a random sample from its population.

- The variable (shift in circadian rhythm) is normally distributed in each (treatment) population.

43

Testing for differences in variances among populations can be done using Levene's test. While its calculation may be too complex for the BIOL322 level, it is important to understand its existence, its utility, and how to apply it in R.

```
leveneTest(shift ~ factor(treatment), data=circadian)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.1586 0.8545
      19
```

P = 0.8545. Based on an alpha = 0.05, we should not reject the null hypothesis that: $\sigma^2_{control} = \sigma^2_{knee} = \sigma^2_{eye}$

Therefore, we should feel confident to conduct a standard ANOVA to the data (there is a Welch-like ANOVA).

44