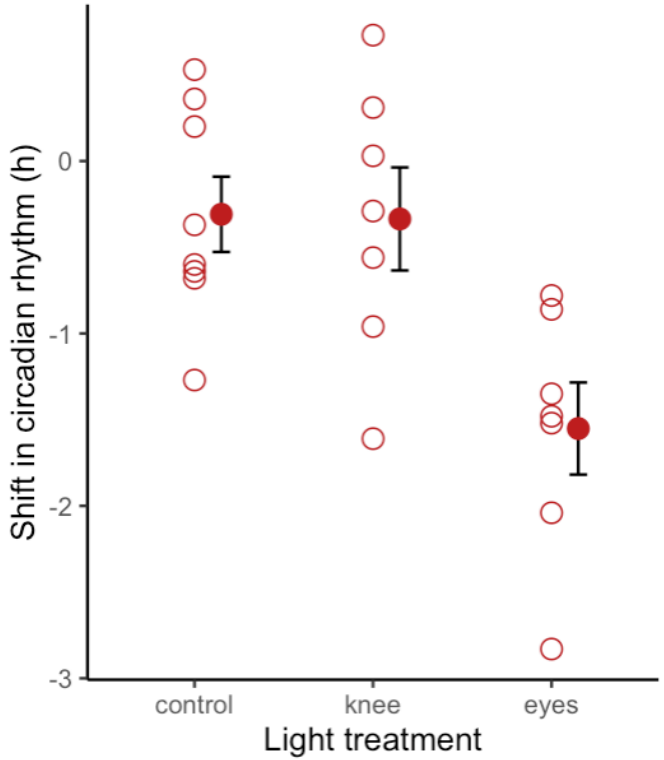
THE ANALYSIS OF VARIANCE (ANOVA)
for comparing multiple sample means (groups or treatments)

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.

ANOVA

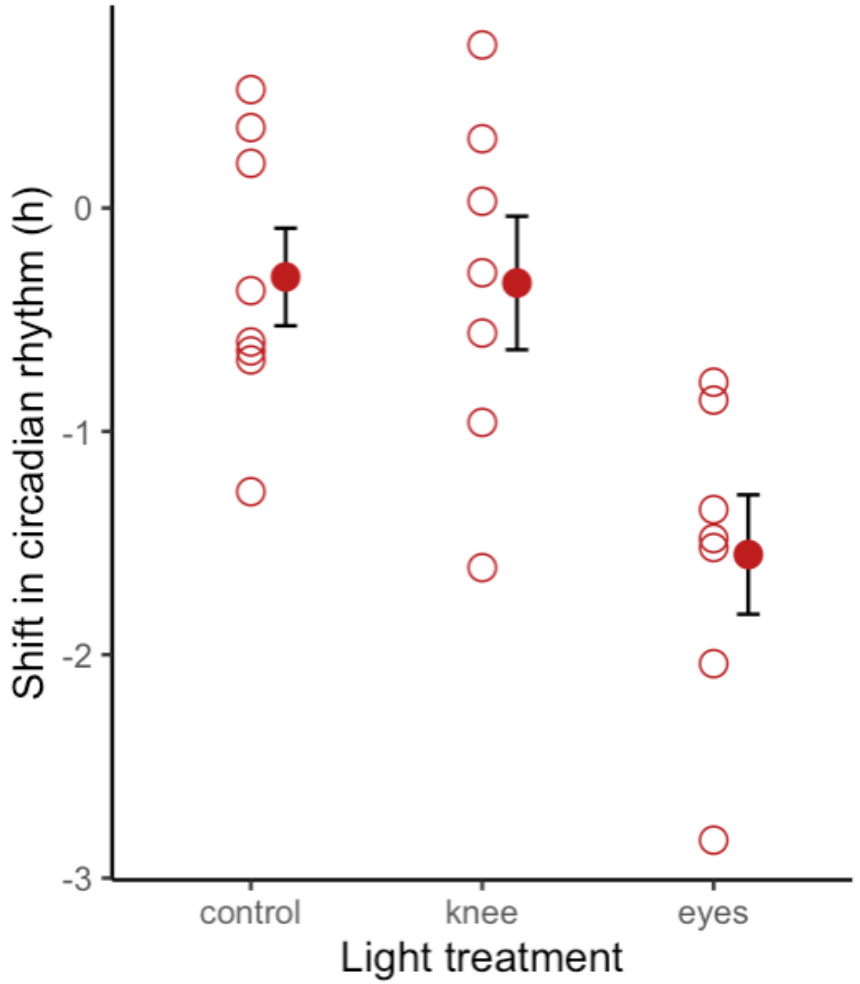
Research conclusion: Light treatment influences shifts in circadian rhythm.

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?



Why ANOVA before pairwise two sample mean t-tests

Pairwise t-tests ask **many small questions**:

- Is shift in circadian rhythm in the control group different from the eye group?
- Is shift in circadian rhythm in the control group different from the knee group?
- Is shift in circadian rhythm in the eye group different from the eye group?

ANOVA first asks the **global question**:

Is there any evidence that the group means differ at all?

This is often the scientifically meaningful starting point. If the global test shows **no overall difference**, then doing multiple pairwise comparisons becomes less justified.

Why ANOVA before pairwise two sample mean t-tests

ANOVA protects against "data dredging" (today's lecture)

If one runs pairwise tests without a prior ANOVA, you may end up:

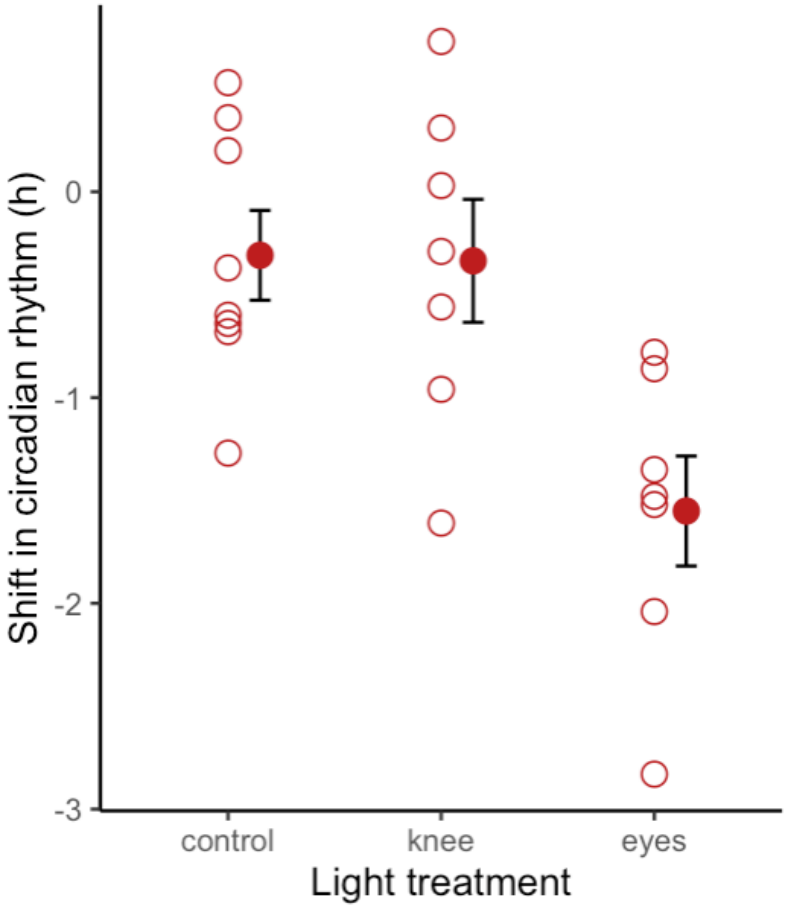
Running many comparisons
Interpreting a few significant ones
Ignoring the overall pattern

ANOVA: Is there any difference at all?
Post-hoc (pairwise) tests: Where are the differences?

Data dredging is the practice of testing many hypotheses or searching through data without a prior plan until statistically significant results are found, increasing the likelihood of false positives.

After ANOVA:

Multiple testing and post hoc (“after the event”; Latin: post hoc) tests, which are conducted after an overall test (e.g., ANOVA) to determine where differences occur.



Classroom survey:

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

Multiple testing survey (BIOL322); anonymous survey - it will close on Tuesday March 31 (5pm)

Results will be used to demonstrate the statistical principles of multiple testing

last number of your street address *

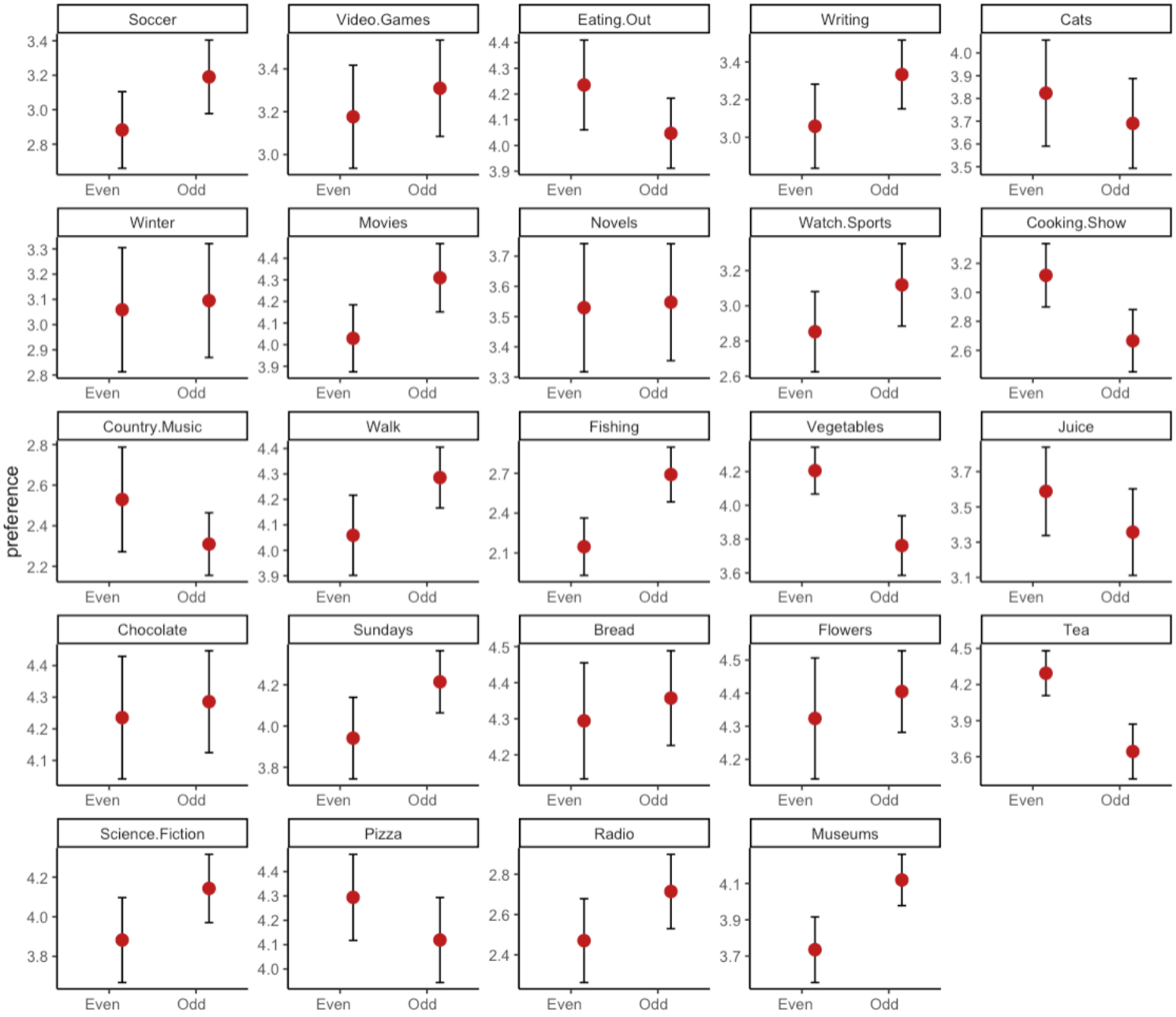
- Odd number
- Even number

Your birthday is an odd or even number (the actual day; not month or year) *

- Odd number
- Even number

Do you like soccer? *

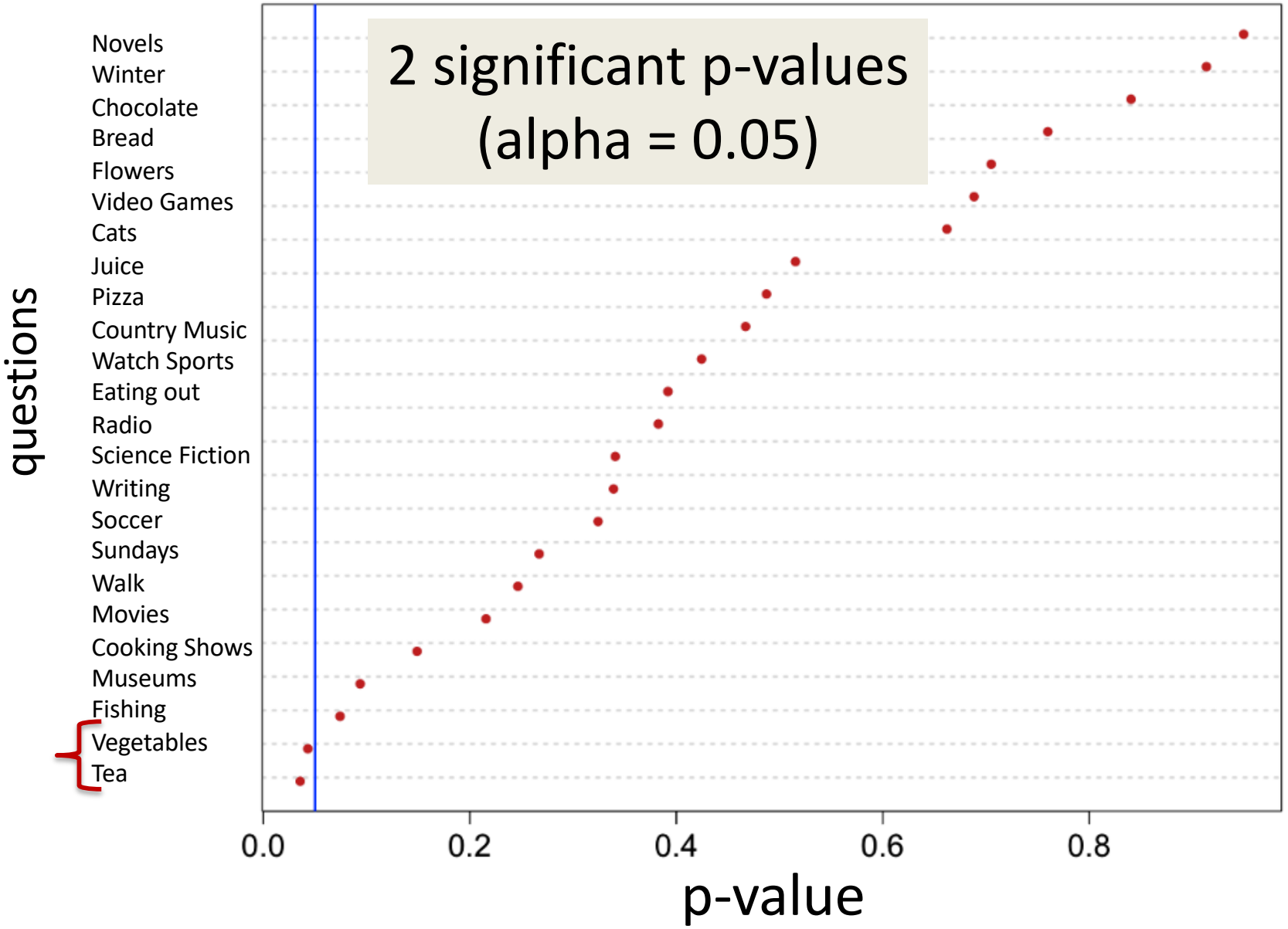
	1	2	3	4	5	
Deslike	○	○	○	○	○	Love it



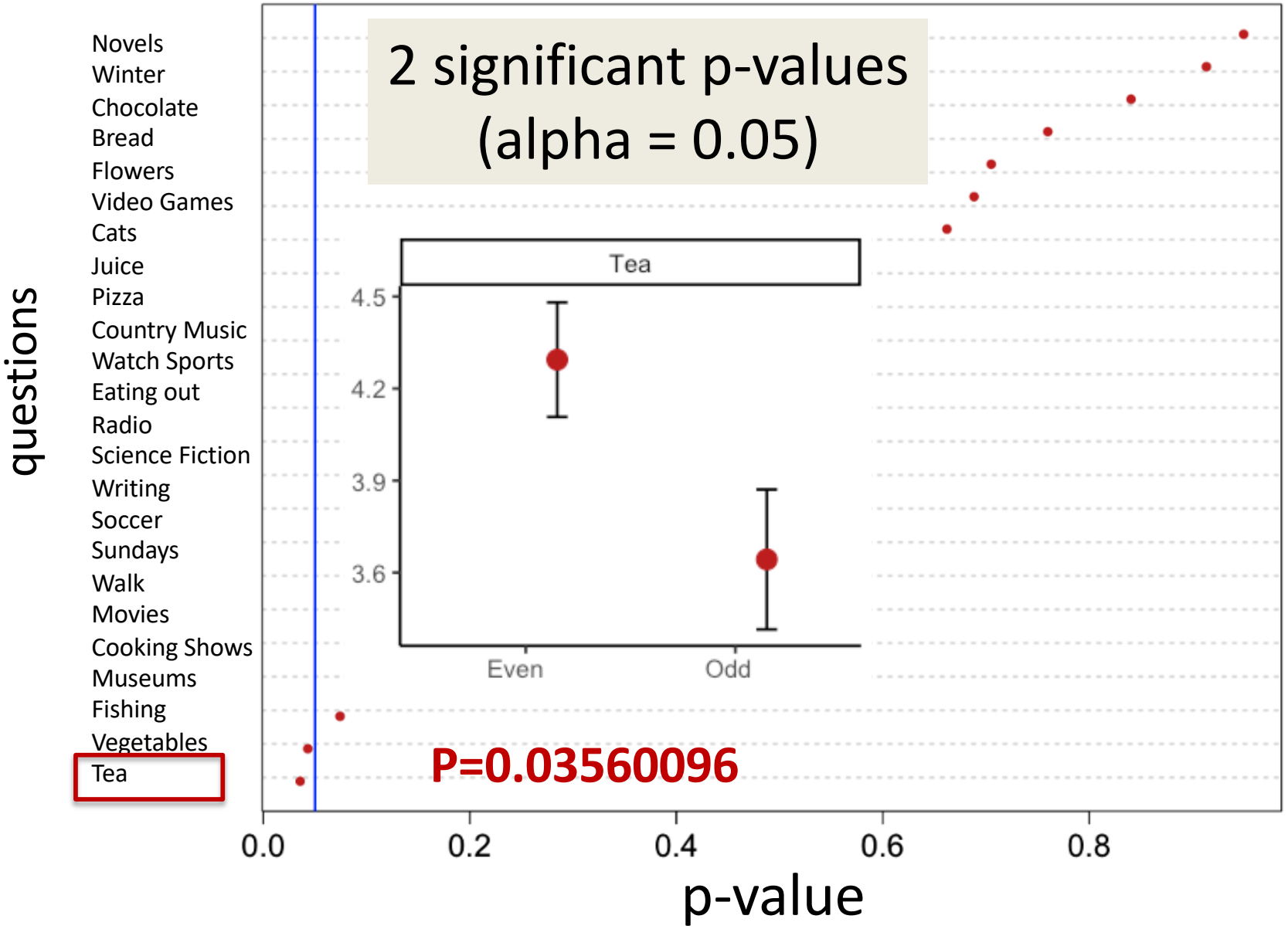
Birthday and preferences

One should have no theoretical reason to expect preferences to differ among groups beyond what would occur by chance alone.

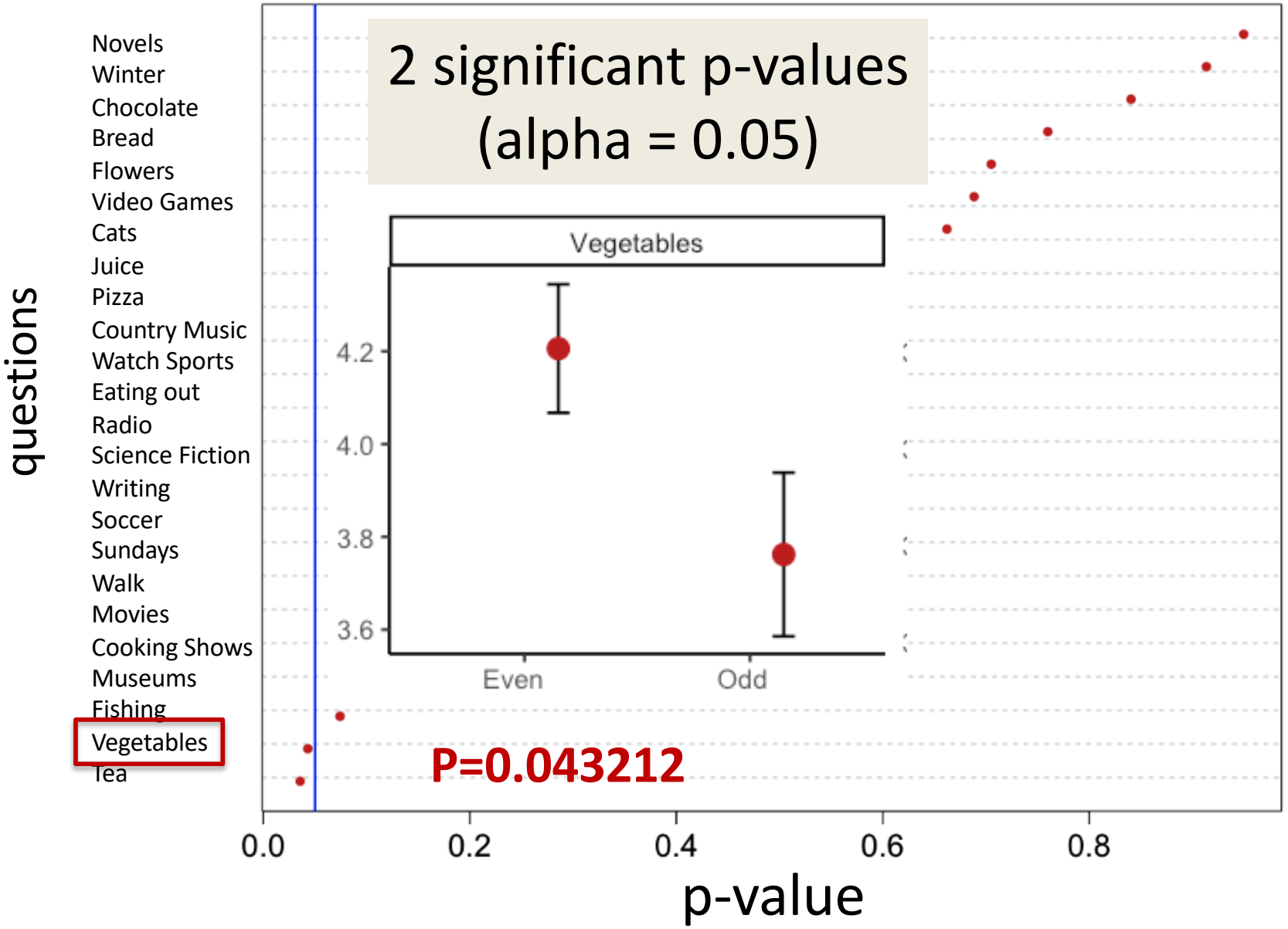
Contrast between individuals born on odd versus even days - probability of rejecting the null hypothesis based on a two-sample t-test.



Contrast between individuals born on odd versus even days - probability of rejecting the null hypothesis based on a two-sample t-test.



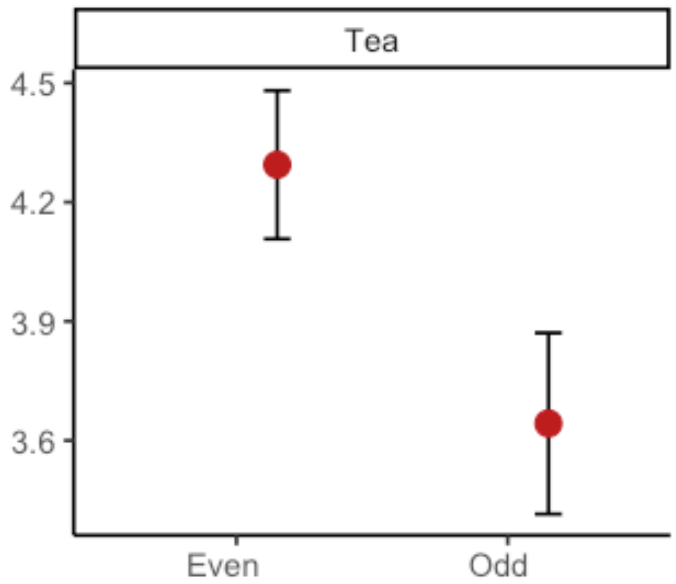
Contrast between individuals born on odd versus even days - probability of rejecting the null hypothesis based on a two-sample t-test.

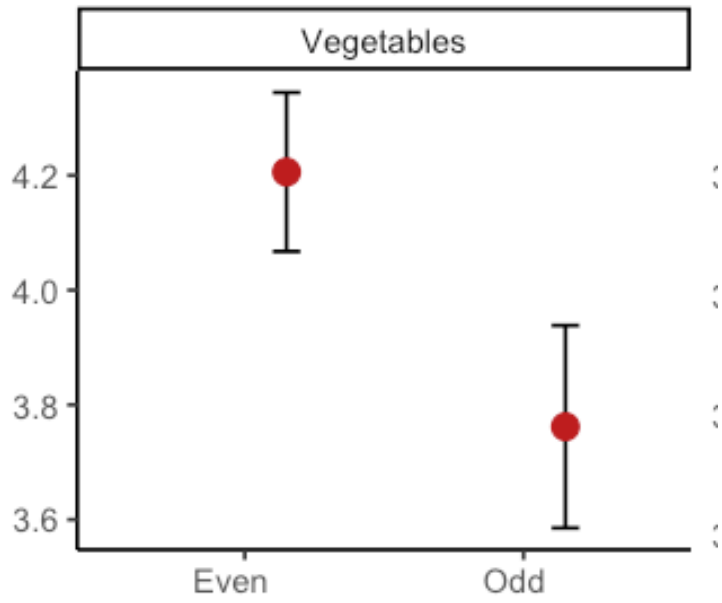


Birthday and Preferences:

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

How can that be?



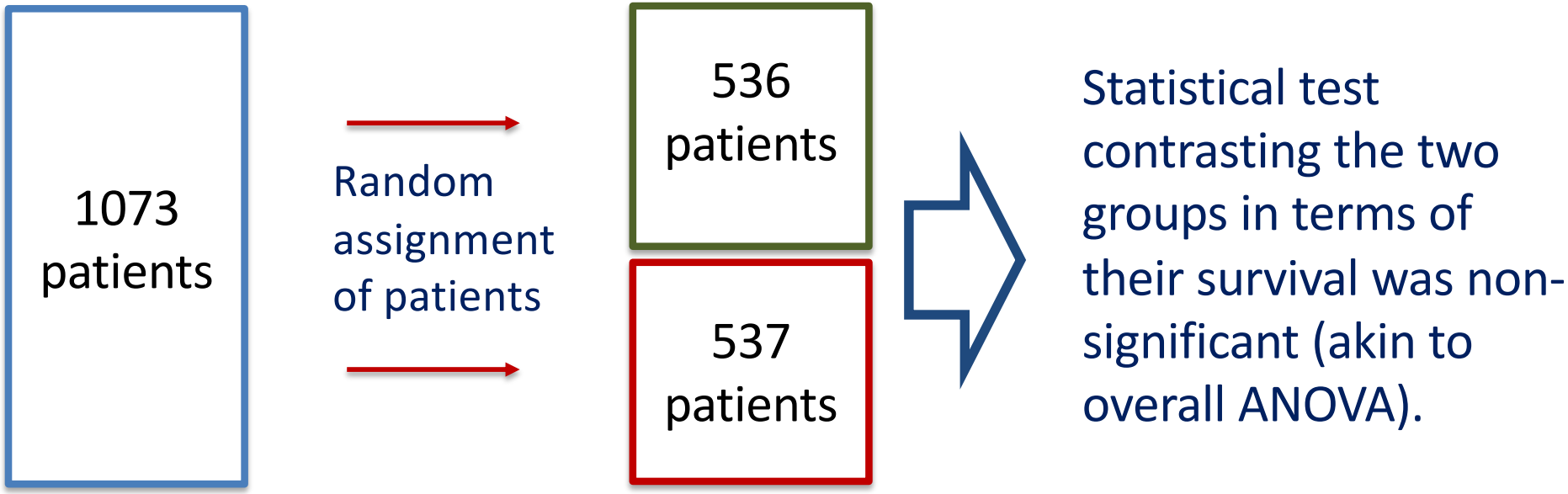


Another example of finding statistical significance when none should exist

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 used to illustrate the need for clinical judgment and appropriate statistical methods when assessing therapeutic claims in complex diseases.

In this example, 1,073 medically treated patients from the Duke University databank were randomly assigned to two comparable groups. As expected, there was no overall difference in survival between the groups.



Another example of finding statistical significance when none should exist

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

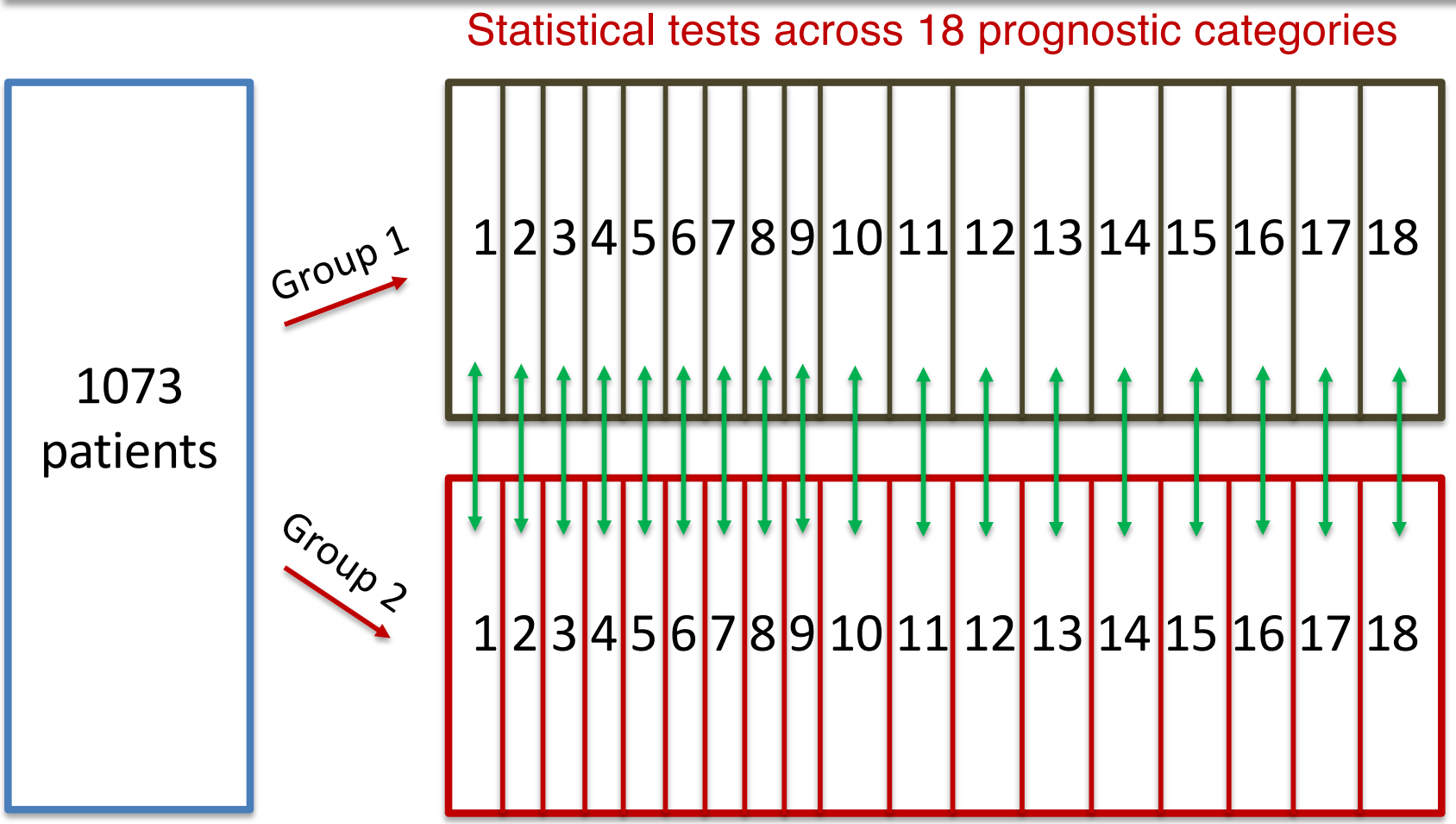
A simulated randomized clinical trial in coronary artery disease was used to illustrate the need for clinical judgment and appropriate statistical methods when assessing therapeutic claims in complex diseases.

In this example, 1,073 medically treated patients from the Duke University databank were randomly assigned to two comparable groups. As expected, there was no overall difference in survival between the groups.

However, when patients were further subdivided into **18 prognostic categories**, a one subgroup of (randomly chosen) patients with three-vessel disease and abnormal left ventricular contraction showed a statistically significant difference in survival between the two groups.

Another example of finding statistical significance when none should exist

However, when patients were further subdivided into **18 prognostic categories**, a one subgroup of (randomly chosen) patients with three-vessel disease and abnormal left ventricular contraction showed a statistically significant difference in survival between the two groups.



Another example of finding statistical significance when none should exist

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

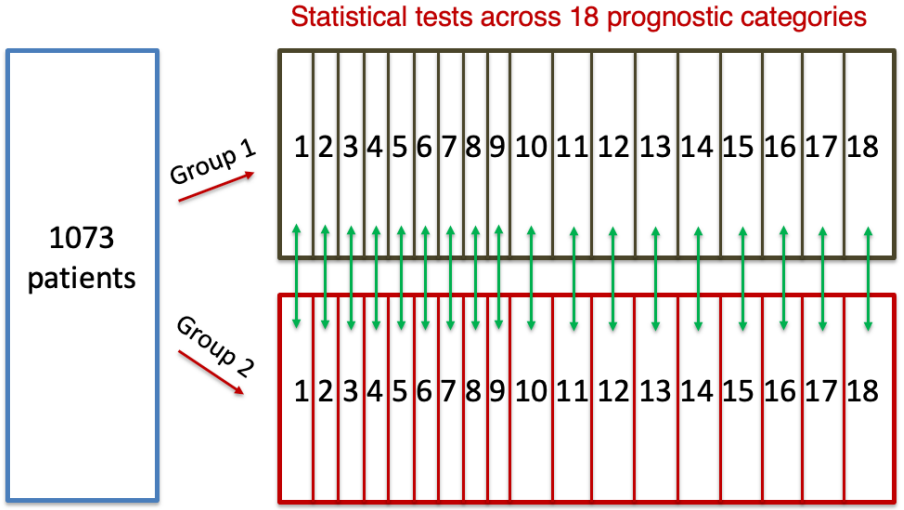
However, when patients were further subdivided into **18 prognostic categories**, a one subgroup of (randomly chosen) patients with three-vessel disease and abnormal left ventricular contraction showed a statistically significant difference in survival between the two groups.

Multiple testing adjustment procedures indicated no statistical differences in prognostic factors.

This study highlights the importance of clinical judgment when interpreting subgroup results. The apparent difference may therefore reflect chance, rather than a true treatment effect.

Another example of finding statistical significance when none should exist

- Patients classified as having three-vessel disease and abnormal left ventricular contraction showed differences 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?***



Another example of finding statistical significance when none should exist

- 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?**

Recall what the sampling Distribution Under H_o represents:

If we flip a fair coin **20 times**, the number of heads follows a **Binomial distribution**:

Most likely outcomes: **8–12 heads**
Less likely outcomes: **0–3 or 17–20 heads** (rare events)

This distribution represents the **sampling distribution under the null hypothesis**.

Type I error: occurs when we **reject a true null hypothesis**.

In this example: The coin **is fair,** but if we observe **an extreme result** (e.g., 18 heads out of 20), we would conclude wrongly that the coin is biased → **Type I error.**

However, if the coin is truly biased, then the sampling distribution changes, making extreme outcomes more likely. In that case, observing a large number of heads would no longer be a rare event under the true distribution, and rejecting the null hypothesis would be the correct decision rather than a Type I error.

In this example: The coin **is fair,** but if we observe **an extreme result** (e.g., 18 heads out of 20), we would conclude wrongly that the coin is biased → **Type I error.**

However, if the coin is truly biased, then the sampling distribution changes, making extreme outcomes more likely. In that case, observing a large number of heads would no longer be a rare event under the true distribution, and rejecting the null hypothesis would be the correct decision rather than a Type I error.

But since we do not know the truth, we rely on the sampling distribution under the null hypothesis (H_0).

This means that, on rare occasions, we may observe an extreme outcome purely by chance and incorrectly reject the null hypothesis, resulting in a Type I error.

And the probability of observing such a rare event purely by chance, when H_0 is true, is defined by alpha (α), the Type I error rate.

Recall how the sampling distribution under the null hypothesis is constructed (conceptually) for a continuous variable

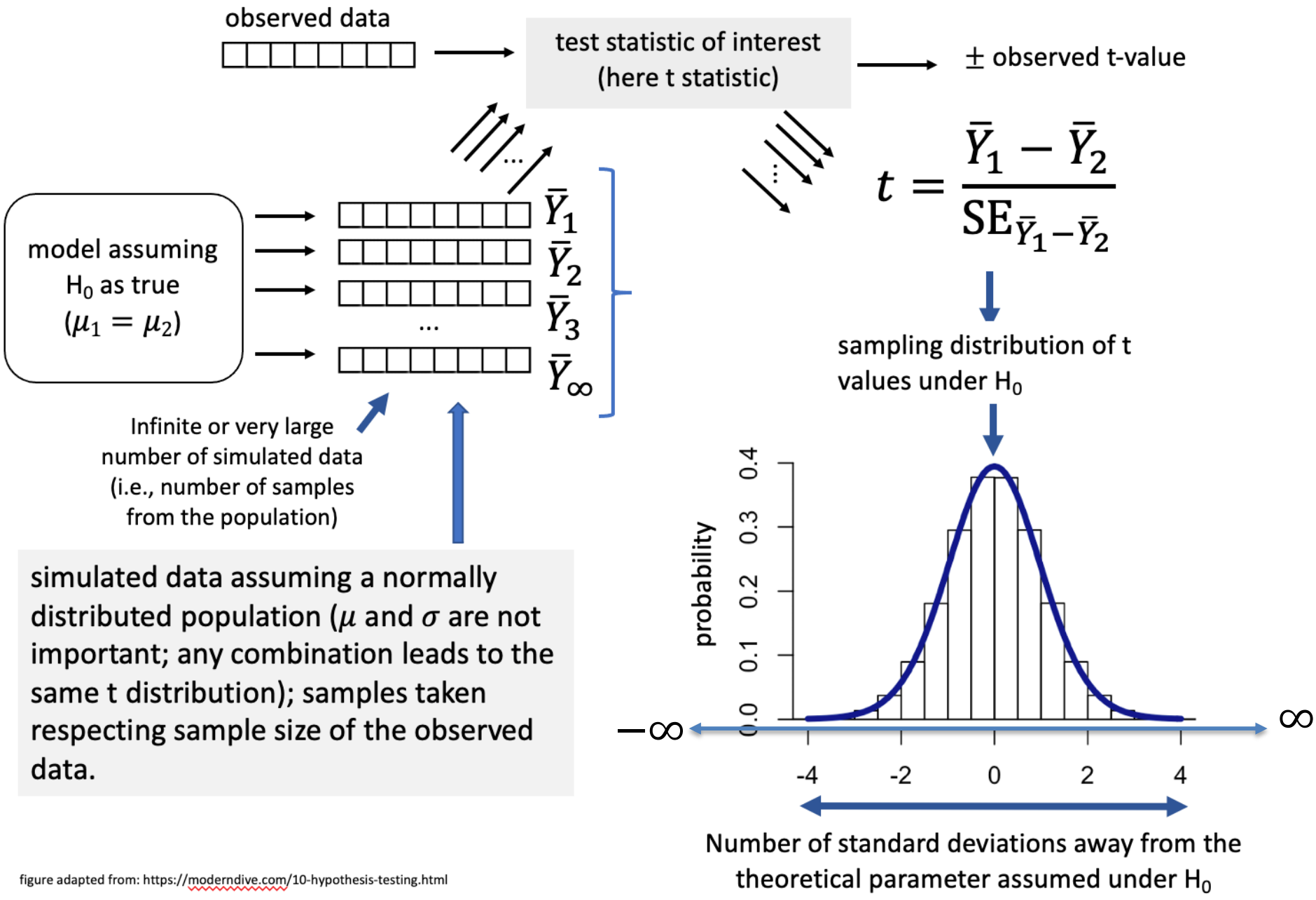
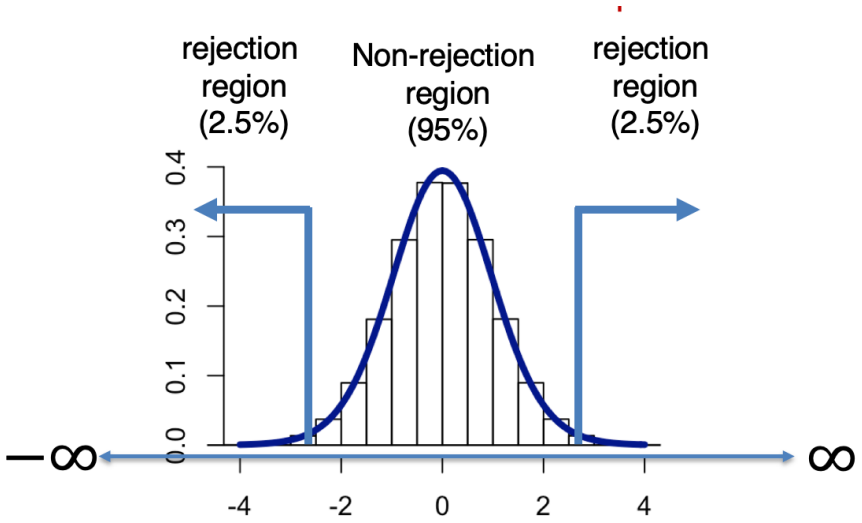


figure adapted from: https://moderndive.com/10-hypothesis-testing.html

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

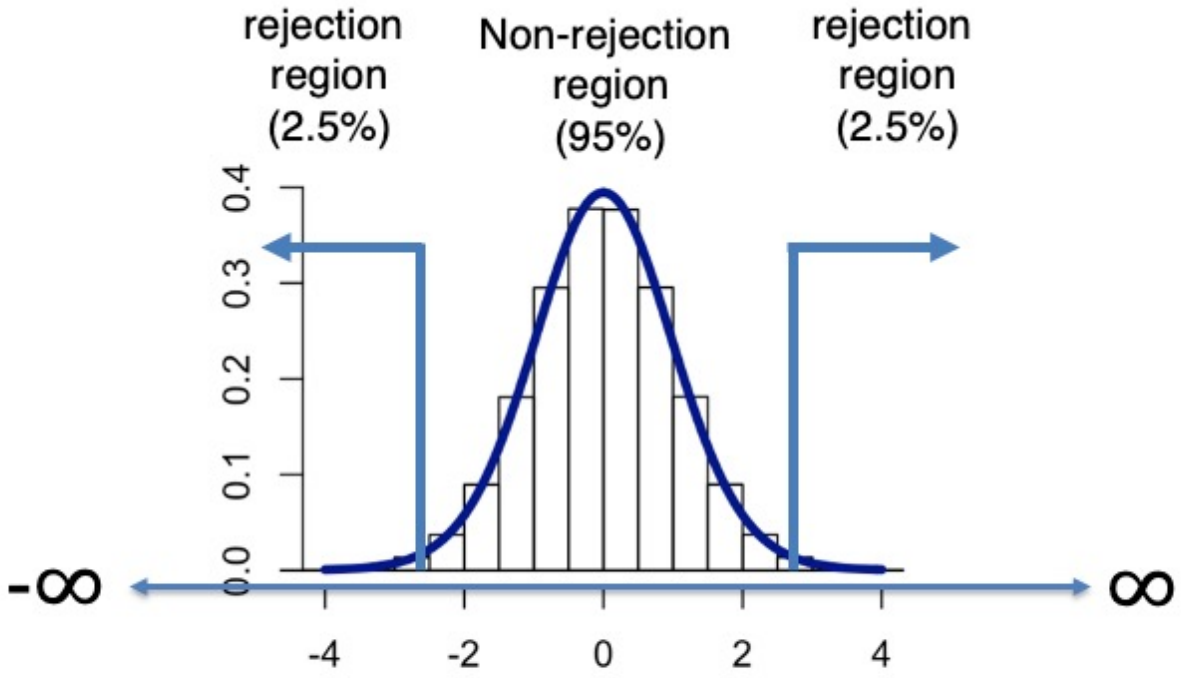


p-values are calculated under the assumption that H_0 is true, a small p-value provides evidence against H_0

Recall: under the null hypothesis (H_0), all t-values are possible, although some are more likely than others. The rejection region is defined so that the probability of observing a t-value within this region equals the chosen significance level (e.g., $\alpha = 0.05$). This corresponds to the probability of committing a Type I error when the null hypothesis is true.

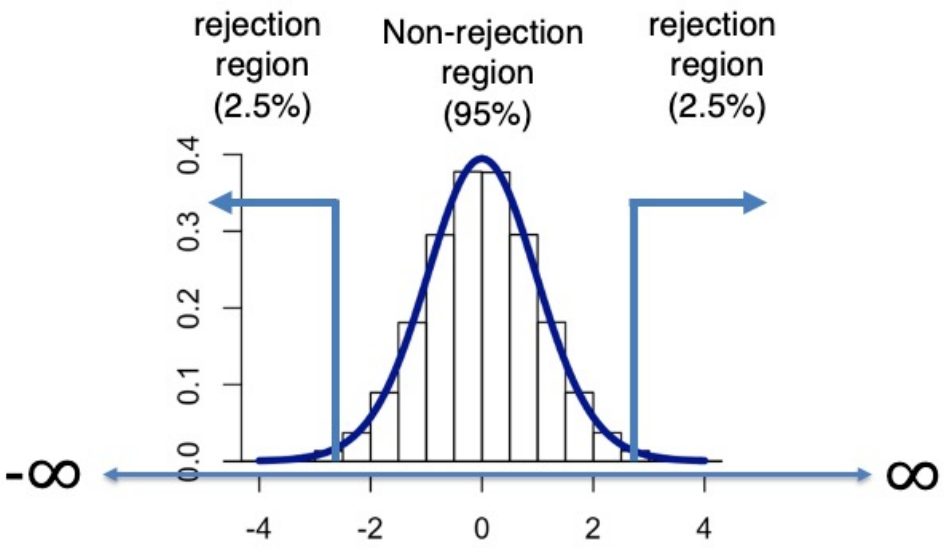
If we conduct a single test, the probability of committing a Type I error is equal to the chosen significance level (e.g., $\alpha = 0.05$). However, when multiple tests are conducted, the probability of committing at least one Type I error increases, often substantially, depending on the number of tests performed.

From a tutorial: When making inferences from samples, we face an inherent trade-off: reducing the risk of one type of error (Type I, or false positives) generally increases the risk of another (Type II, or false negatives).



All the infinite t-values are possible under H_0, including those in the rejection region; however, the probability of sampling a value in the rejection region is equal to the chosen significance level (e.g., $\alpha = 0.05$).

For each test conducted, the probability of committing a Type I error (i.e., rejecting the null hypothesis when it is actually true) is equal to the chosen significance level (e.g., α = 0.05, or 5%). This applies to each individual test. However, when multiple tests are performed, the probability of committing at least one Type I error increases unless appropriate adjustments are made.

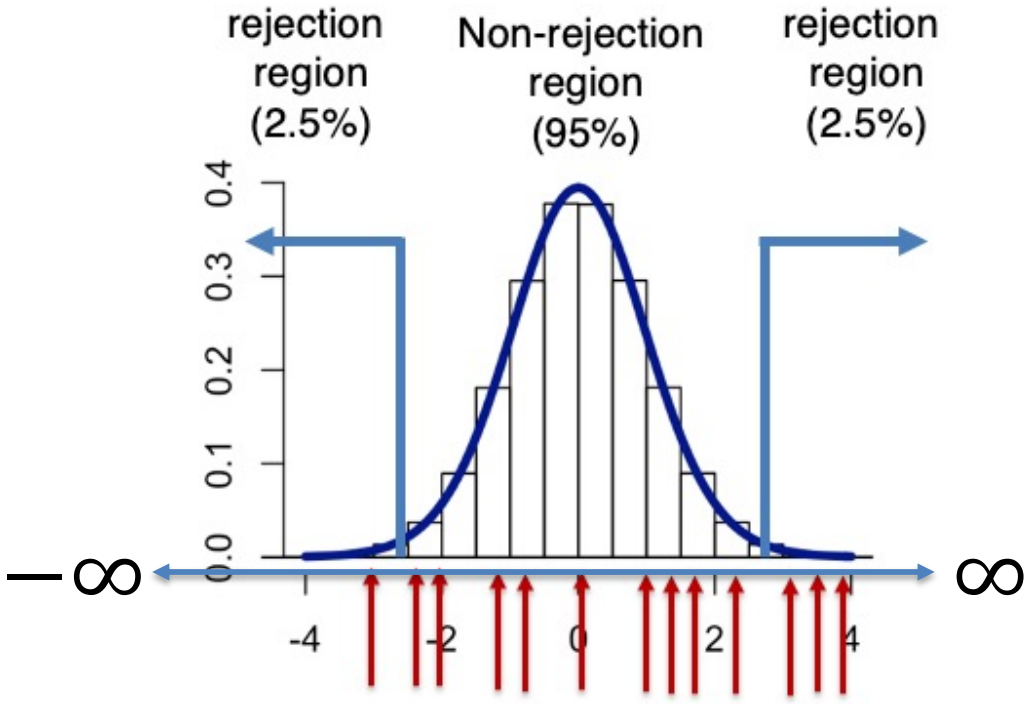


All the infinite t values are possible under H_0, even the ones in the rejection region (they have a probability of alpha=0.05 to be sampled).

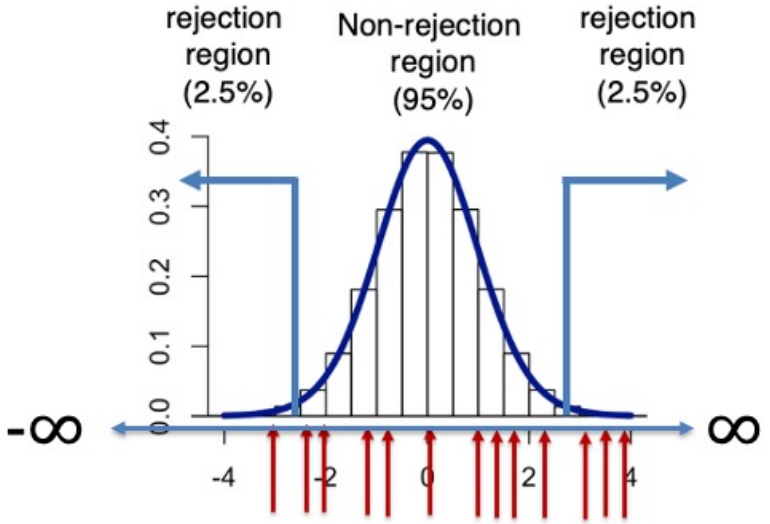
Let's assume that the null hypothesis is indeed true, as in our student survey and the heart study.

There is a high likelihood (95%) that the test will not yield a significant result (i.e., $p \geq 0.05$) when the null hypothesis is true. This can be likened to throwing a dart at a target representing the sampling distribution, where most outcomes fall inside the non-rejection region, i.e., by chance along, the probability of committing a type I error is 5%

If all null hypotheses are true when comparing multiple means (pairwise), throwing multiple darts at the sampling distribution means that each dart still has a 5% chance of landing in the rejection region (i.e., obtaining a p-value < α) purely by chance. This represents the probability of committing a Type I error for each individual test.



All the infinite t-values are possible under H_0, including those in the rejection region; however, the probability of sampling a value in the rejection region is equal to the chosen significance level (e.g., $\alpha = 0.05$).



If you conduct many tests, you will eventually (by chance alone) obtain a t-value that falls in the rejection region. Recall that the sampling distribution under the null hypothesis includes all possible values of the t-statistic when there is no true difference between means.

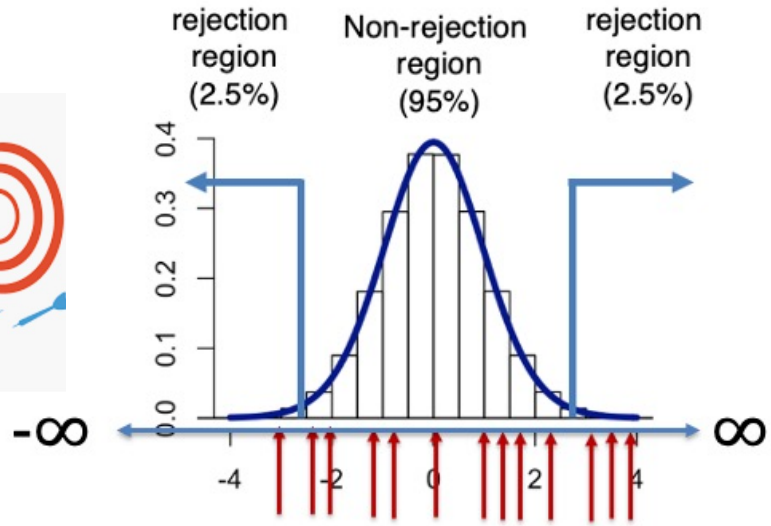
Because the rejection region is defined by low-probability outcomes (based on the chosen α level), each test carries a fixed probability of producing a Type I error when the null hypothesis is true. As the number of tests increases, the probability of observing at least one such low-probability outcome also increases. In other words, the more tests you conduct, the greater the likelihood of false positives (rejecting the null hypothesis when it should not be rejected).

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

	dislike				Love it
odd-day even-day born	1	2	3	4	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		



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:

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

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

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

	dislike				Love it
odd-day even-day born	1	2	3	4	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?					

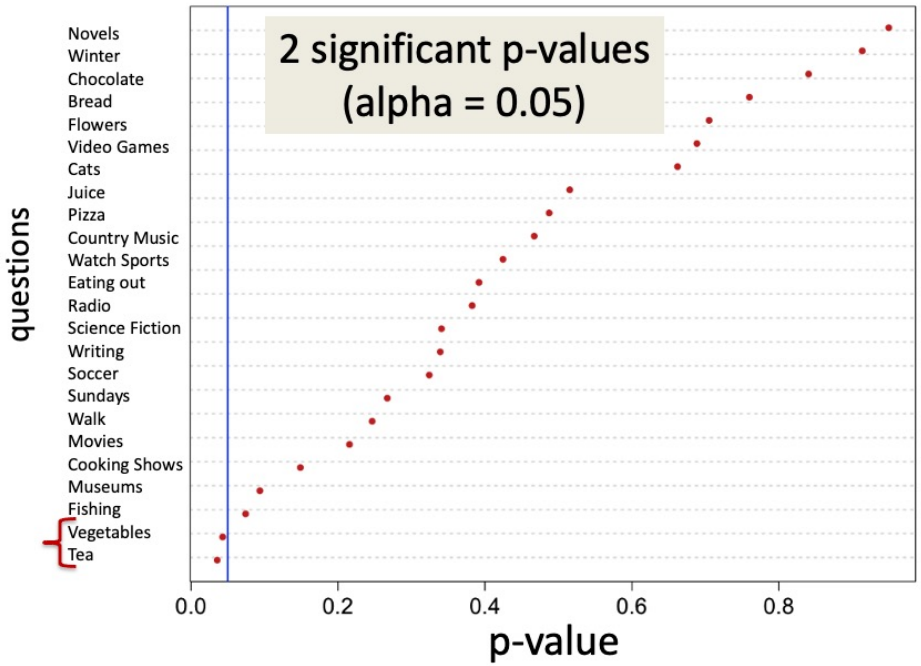
……

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		

$1-0.95^{24}=0.708$

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

2 tests were in fact significant.



Let's assume that 100 statistical 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:

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

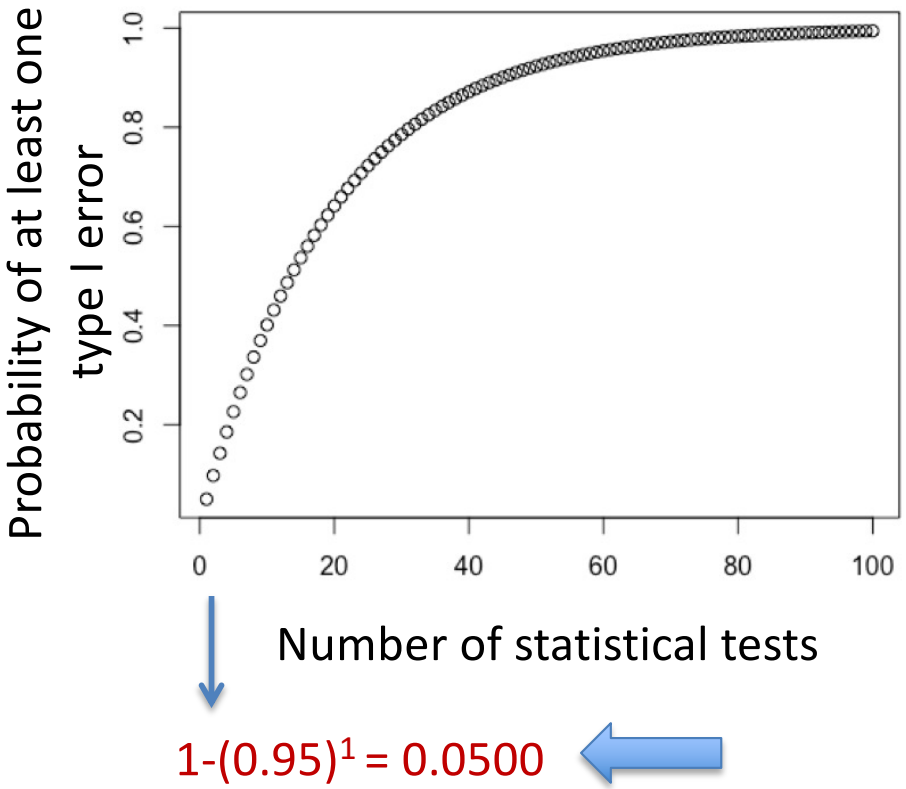
99.4% chance of finding at least 1 significant difference between group 1 and group 2 when H_0 is true!

If you conduct many tests on samples that vary only by chance (i.e., when the null hypothesis is true), the probability of obtaining at least one significant result becomes very high. For example, if you conduct 100 independent tests with $\alpha = 0.05$, the probability of obtaining at least one false positive is approximately 99.4%.

If we set α = 0.05 (i.e., a 95% acceptance region), then the probability of obtaining at least one significant result when the null hypothesis is true—based on a single test—is simply α, or 0.05. This corresponds to the probability of committing a Type I error for one test.

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

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

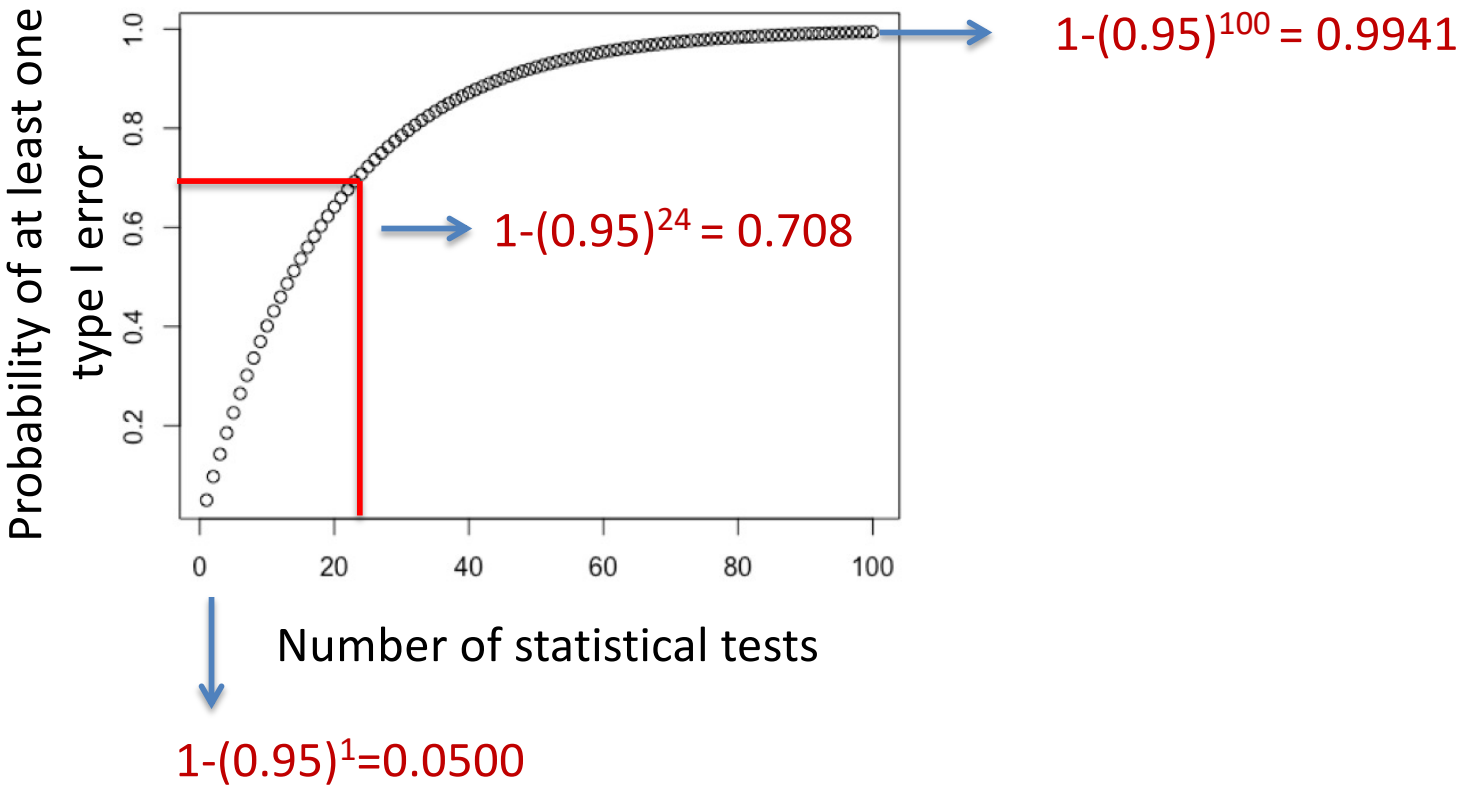


$1\text{-}(0.95)^1 = 0.0500$

If we set α = 0.05 (i.e., a 95% acceptance region), then the probability of obtaining at least one significant result by chance when conducting 24 independent tests is:

$1-0.95^{24} = 0.708$

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



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.

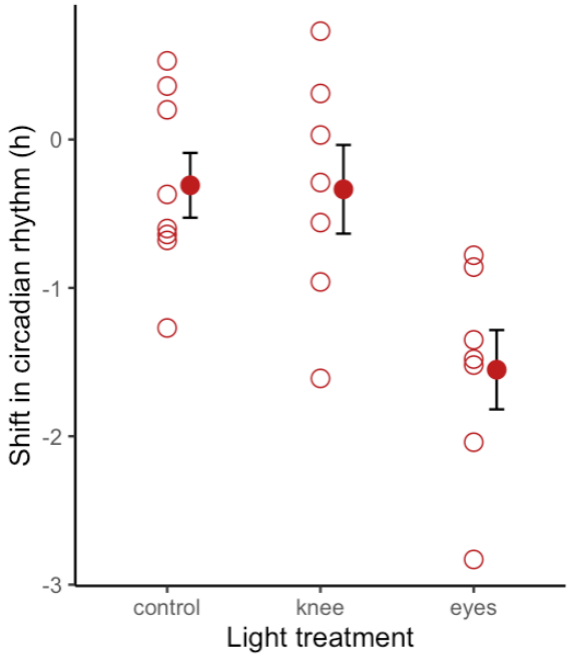
The Tukey's honest test.

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., $\mu_{control} = \mu_{knee} = \mu_{eyes}$.

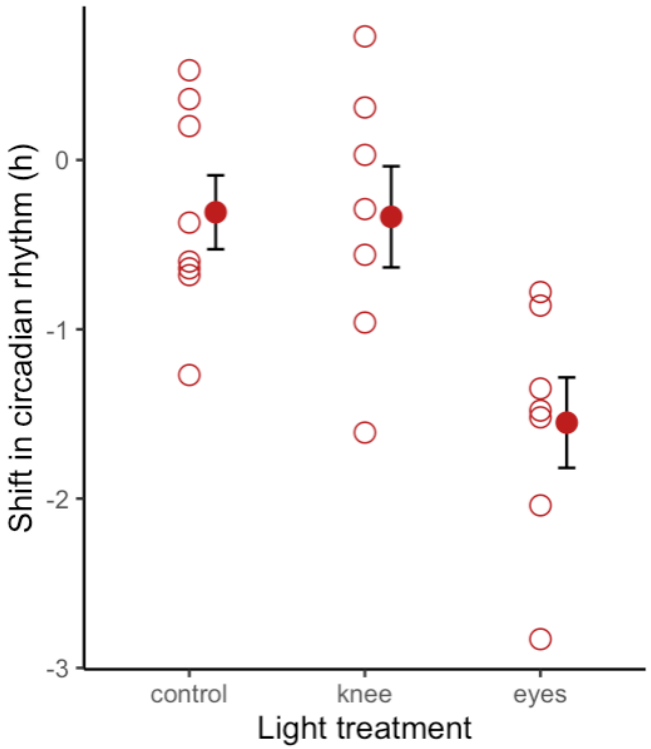
H_A: At least two samples come from different statistical populations with different means.

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



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)

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)

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
```

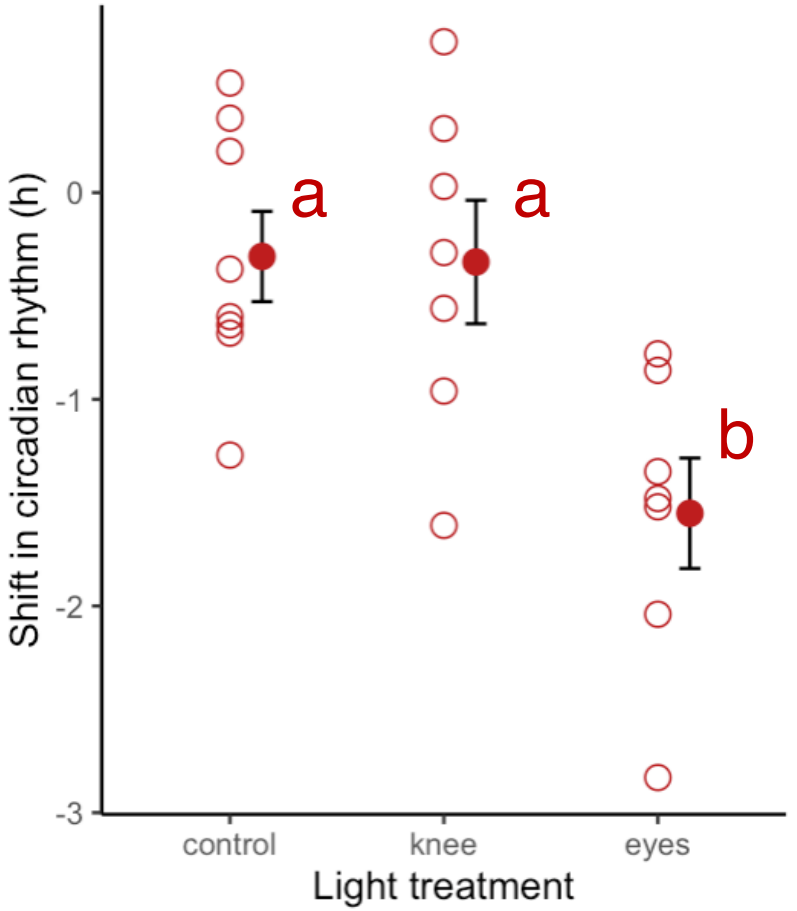
In Tukey's Honest Significant Difference (HSD) test in R, letters (a, b, c, etc.) are often used on graphs to indicate which group means are significantly different or not.

Groups that share the same letter are not significantly different, whereas groups with different letters are significantly different.

```
> 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
```



The test statistic for the Tukey Test is calculated as:

$$Q = \frac{\left|\bar{X}_i - \bar{X}_j\right|}{SE}$$

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

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

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.

The statistic Q is then compared with a distribution based on the largest expected difference between sample means, given the number of pairwise comparisons being made and assuming that H_0 is true.

Sleep/circadian studies
(3 tests in total)

3 mean pairs
(contrasts)

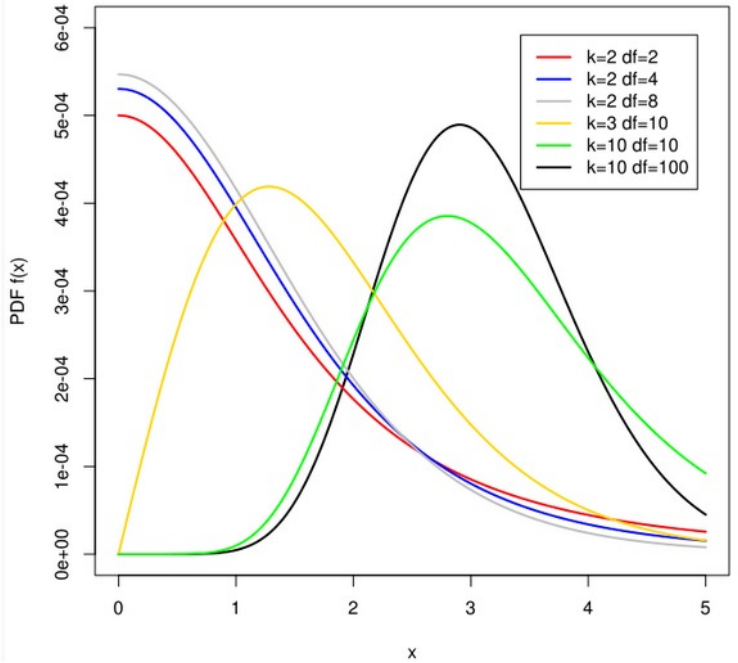
Control – Knee
Control – Eyes
Knee - Eyes

model assuming
H_0 as true

$\bar{Y}_1$
$\bar{Y}_2$
$\bar{Y}_3$

$$Q = \frac{\max(\bar{Y}) - \min(\bar{Y})}{SE_{\max(\bar{Y}) - \min(\bar{Y})}}$$

Infinite or very large number of simulated data (i.e., number of samples from the population)

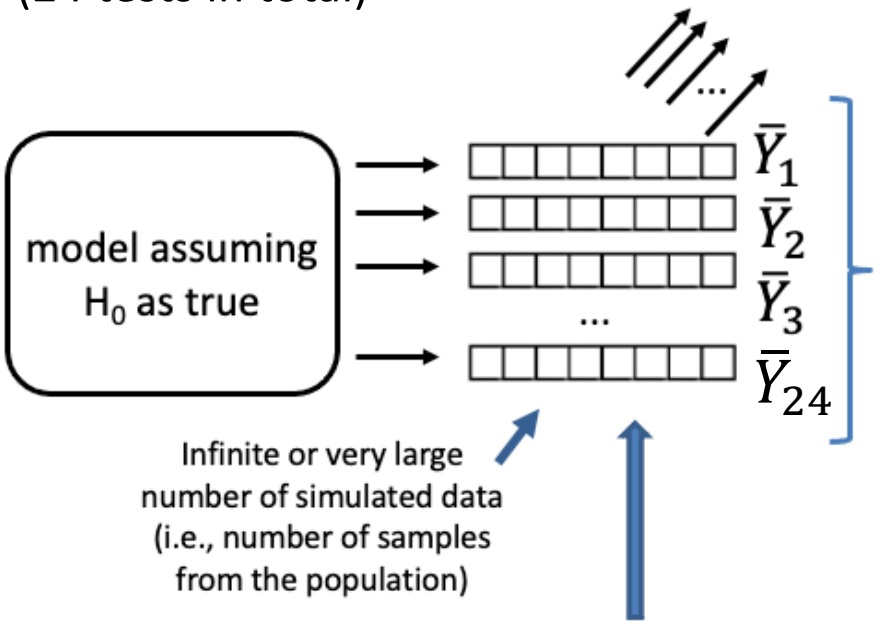


By basing the Q distribution on the largest expected difference among sample means, Tukey's test ensures that the probability of declaring any comparison significant when H_0 is true is controlled. This approach limits inflated Type I errors that arise from multiple testing by controlling the family-wise error rate.

The statistic Q is then compared with a distribution based on the largest expected difference between sample means, given the number of pairwise comparisons being made and assuming that H_0 is true.

Student survey study
(24 tests in total)

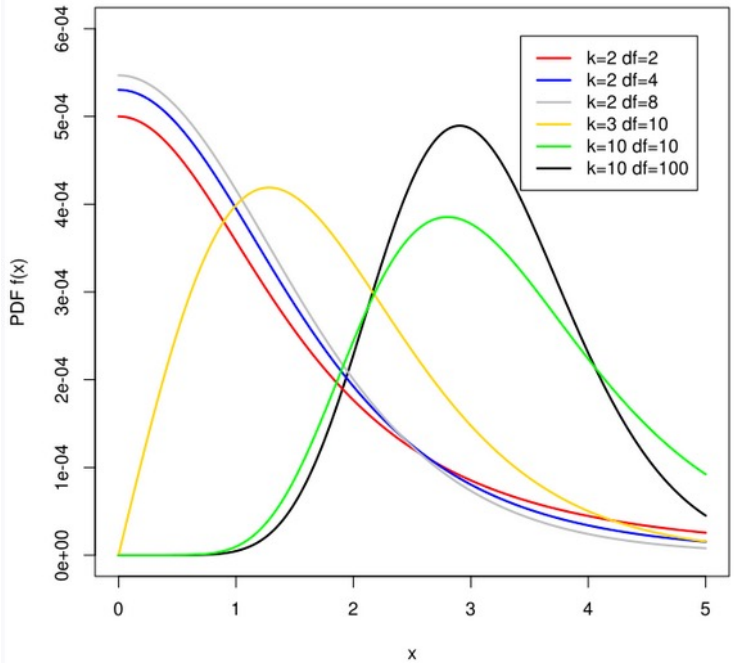
24 mean pairs
(contrasts)



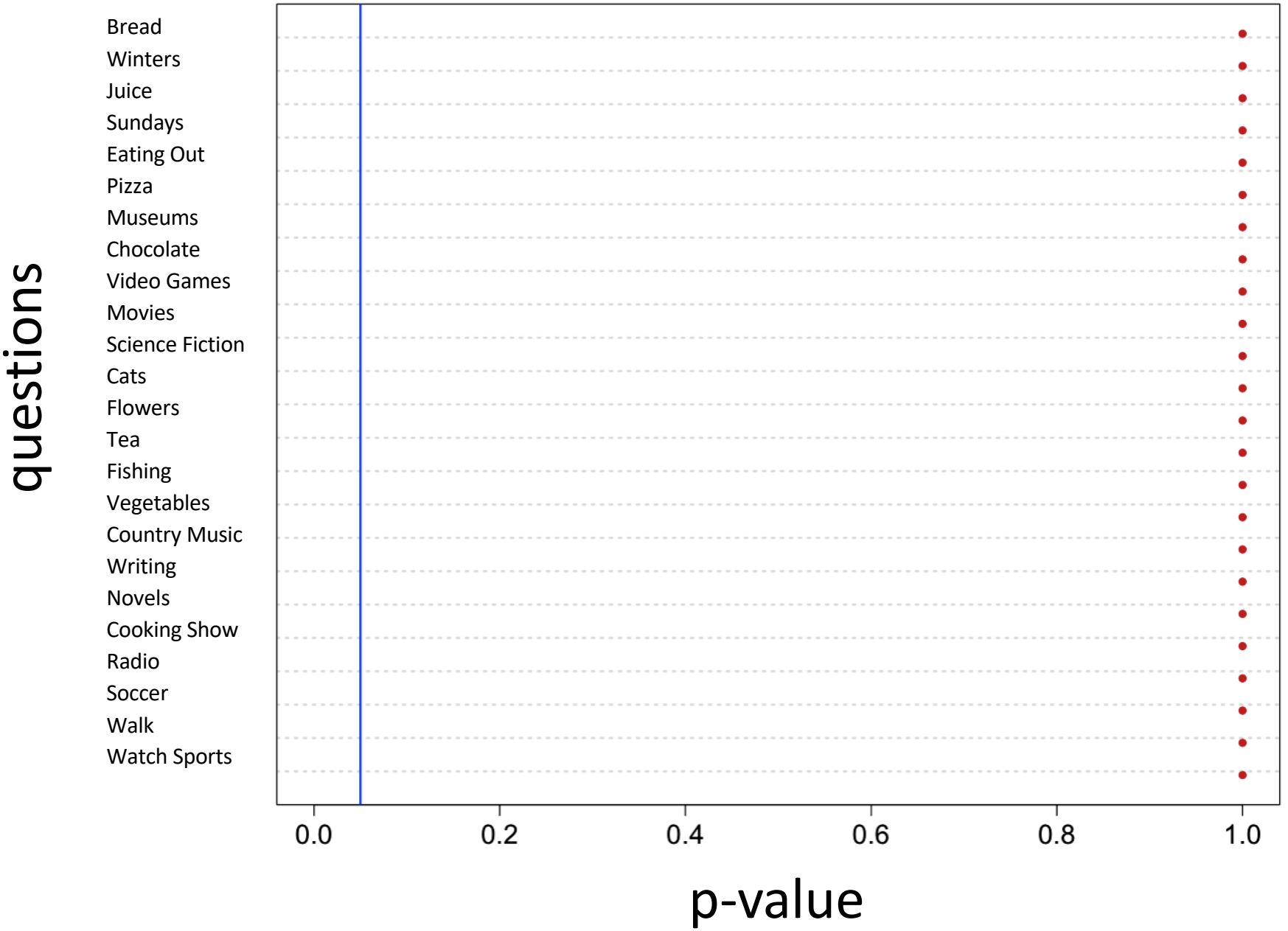
$$Q = \frac{\max(\bar{Y}) - \min(\bar{Y})}{SE_{\max(\bar{Y}) - \min(\bar{Y})}}$$

Infinite or very large number of simulated data (i.e., number of samples from the population)

simulated data assuming a normally distributed population (μ and σ are not important; any combination leads to the same t distribution); samples taken respecting number of groups and sample size of each group.



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



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.

Differences in variances among populations can be assessed using Levene's test. Although its calculation is beyond the scope of BIOL-322, it is important to understand its purpose, when to use it, and how to implement it in R.

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

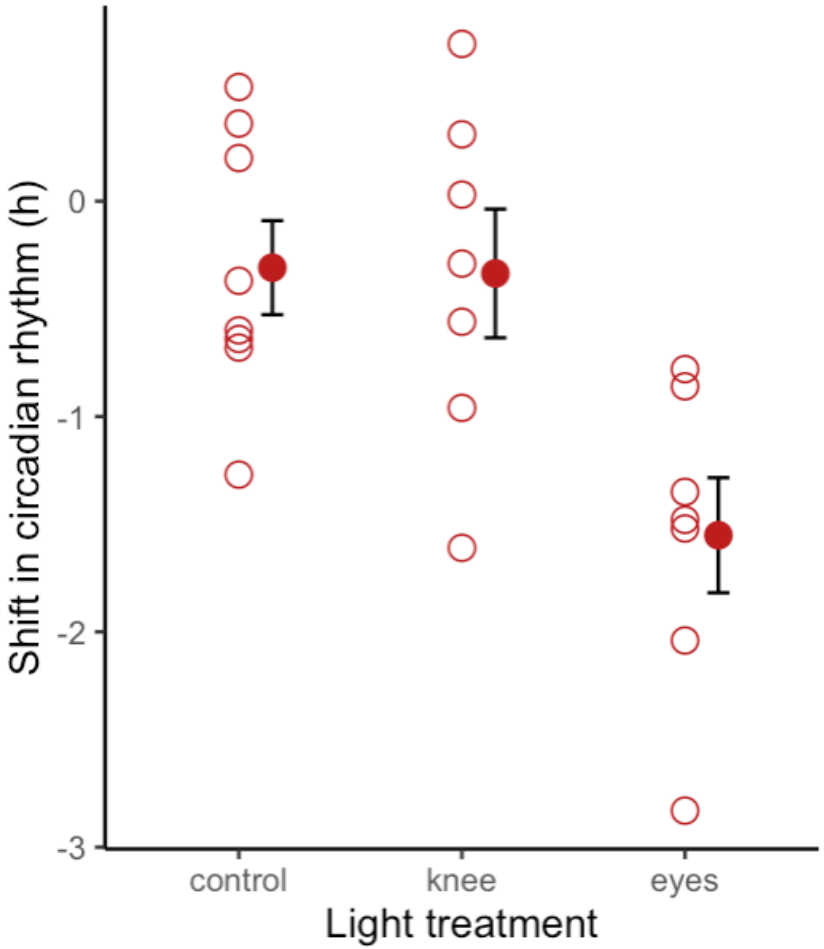
H_A: At least one population variance (σ^2) is different from another population variance or other population variances.

We need to verify that the assumptions of ANOVA are reasonably met before applying it to the data at hand.

Testing for differences in variances among populations: 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.



Levene’s test - assumptions:

Observations within each group are independent and represent random samples from their respective populations.

The test is relatively robust to departures from normality, although extreme deviations may still affect results.

Differences in variances among populations can be assessed using Levene's test. Although its calculation is beyond the scope of BIOL-322, it is important to understand its purpose, when to use it, and how to implement 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 can feel confident applying a standard ANOVA to the data (although a Welch-type ANOVA is available when variances differ).