Statistics Review

2020-07-16

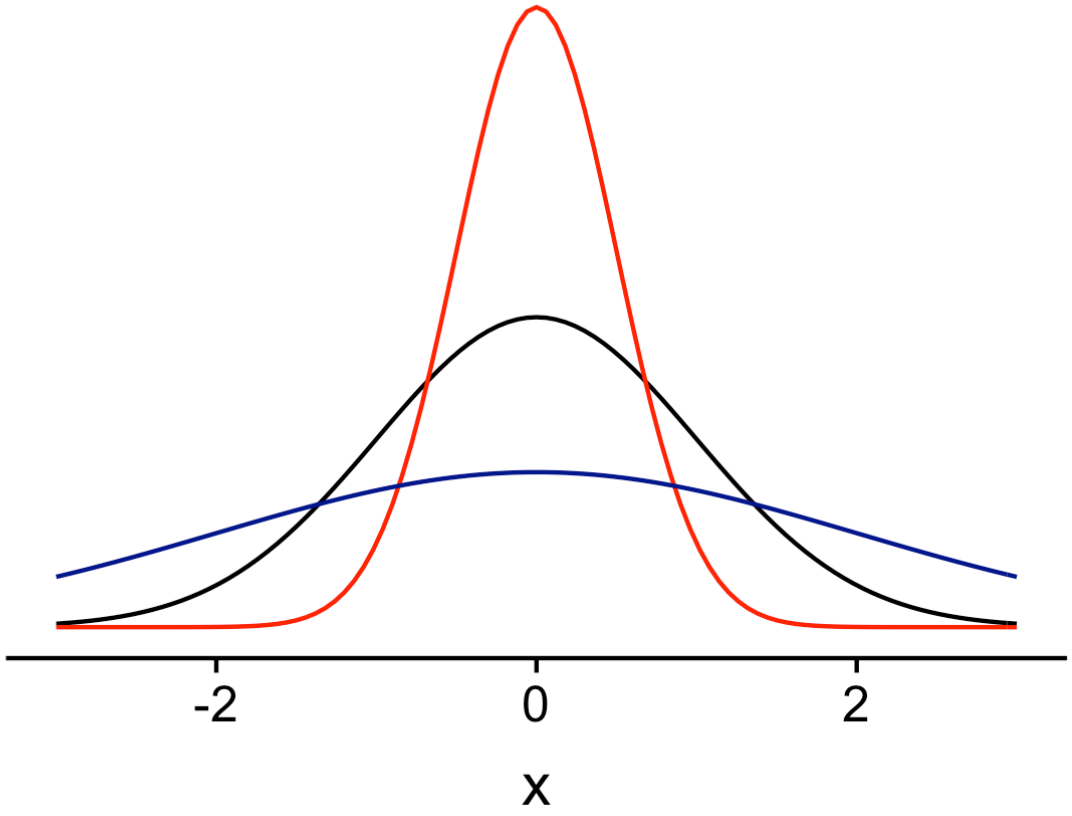
Basic Summary Values

Measures of Central Tendency

- Mean - average
- Median – central value
- Mode – most repeated value

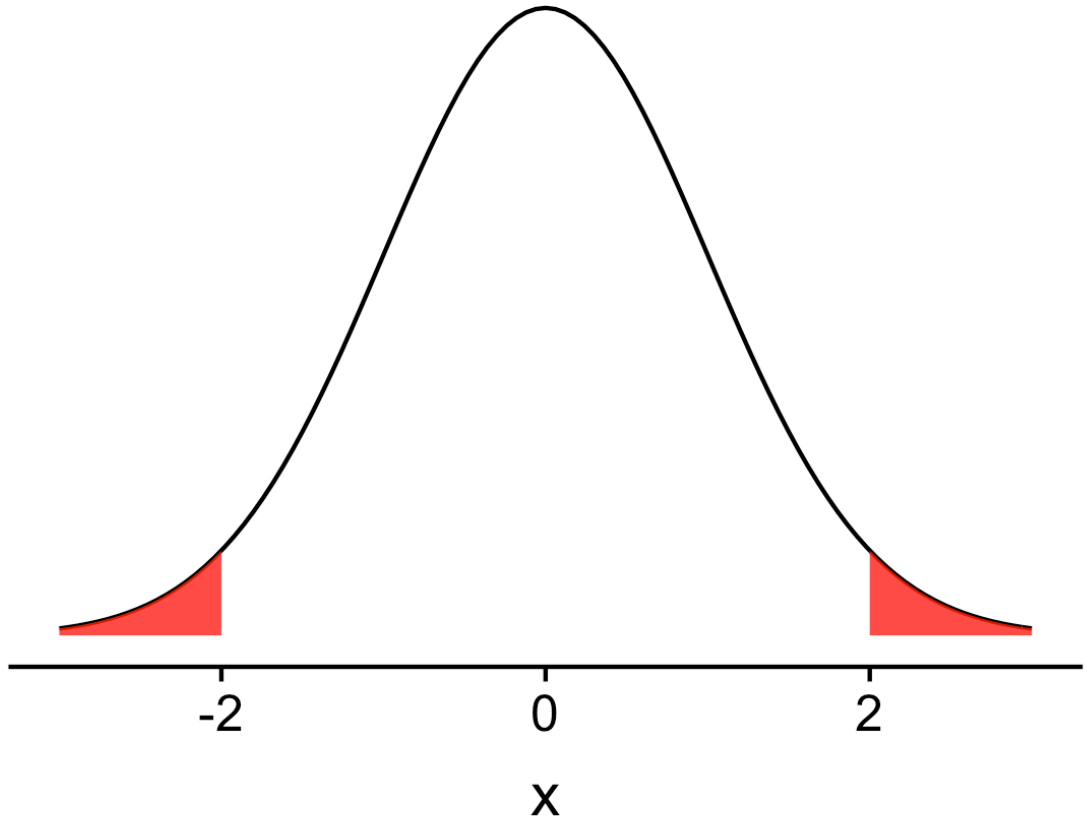
Measures of Spread

- Range – difference between the highest and lowest value
- Standard deviation – measures the dispersion of the data



Hypothesis Testing

- Hypothesis testing compares your data to a pre-determined null distribution (usually the normal distribution). You state a null and alternative hypothesis and calculate the probability your observations happened ***under the null hypothesis.***
- Null hypothesis, **H0**: Everything happened by random chance.
- Alternative hypothesis, **H1**: My observations happened because of my idea.
- Saying p-value = 0.05 means that there's a 5% chance the observation happened randomly under the null distribution.



Test for Continuous Data: one sample t-test

- For testing continuous values against some known mean
- I have an iris with a sepal length of 7 inches and I think that it's because of my new iris fertilizer. Is that iris' sepal length abnormally large?
  - **H0:** There's nothing different about the fertilizer.
  - **H1:** The fertilizer does increase iris sepal length.

```
> t.test(iris$Sepal.Length, mu = 5.8)

One Sample t-test

data:  iris$Sepal.Length
t = 0.64092, df = 149, p-value = 0.5226
alternative hypothesis: true mean is not
equal to 5.8
95 percent confidence interval:
 5.709732 5.976934
sample estimates:
mean of x
 5.843333
```

Test for Continuous Data: two sample t-test

- For testing 2 continuous values against each other
- Is there a difference between the sepal lengths of versicolor and virginica irises?
  - **H0:** There's no difference in the mean sepal lengths.
  - **H1:** There is a difference in the mean sepal lengths.

```
> t.test(iris[iris$Species == 'versicolor',1],
iris[iris$Species == 'virginica', 1])

Welch Two Sample t-test

data:  iris[iris$Species == "versicolor", 1]
and iris[iris$Species == "virginica", 1]

t = -5.6292, df = 94.025, p-value = 1.866e-07

alternative hypothesis: true difference in
means is not equal to 0

95 percent confidence interval:

 -0.8819731 -0.4220269

sample estimates:

mean of x mean of y

    5.936     6.588
```

Test for Continuous Data: paired two sample t-test

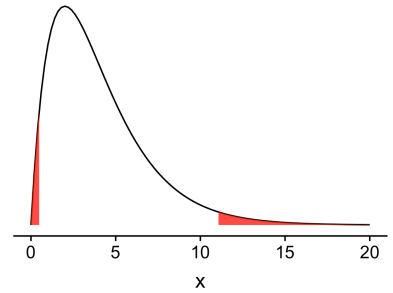
- For testing 2 continuous values against each other *when there is some natural pairing between the samples*
- The sleep dataset in R has data on the amount of time patients sleep on two different sleep medications compared to control. Is there a difference between the two medications?
  - **H0:** There is no difference in the amount of time patients sleep.
  - **H1:** There is a difference in the amount of time patients sleep.

```
> t.test(extra ~ group, data = sleep,
                paired = TRUE)

Paired t-test

data:  extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference
in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

Test for Discrete Data: chi-square



- Test for when you have counts of discrete data; test expected counts against observed counts
- Are babies more likely to be born on one day of the week over other days of the week?
  - **H0:** There is an equal chance of babies being born every day
  - **H1:** There isn't an equal chance of babies being born every day

```
> chisq.test(birth_days$num_births,
p = birth_days$exp_prob_birth)

Chi-squared test for given
probabilities

data:  birth_days$num_births
X-squared = 15.057, df = 6, p-value
= 0.01982
```

The Multiple Testing Problem

- If you do enough tests, you expect to see significant results, just *by random chance*
- Say you flip a coin 10 times and record the number of heads you get. Then you repeat the "experiment" 10 times. You expect to get about heads about 5 times

5 5 6 4 2 4 5 5 4 4

- Now let's do it 100 times

5 6 5 4 5 7 6 5 5 5 5 5 7 3 5 6 6 4 5 6

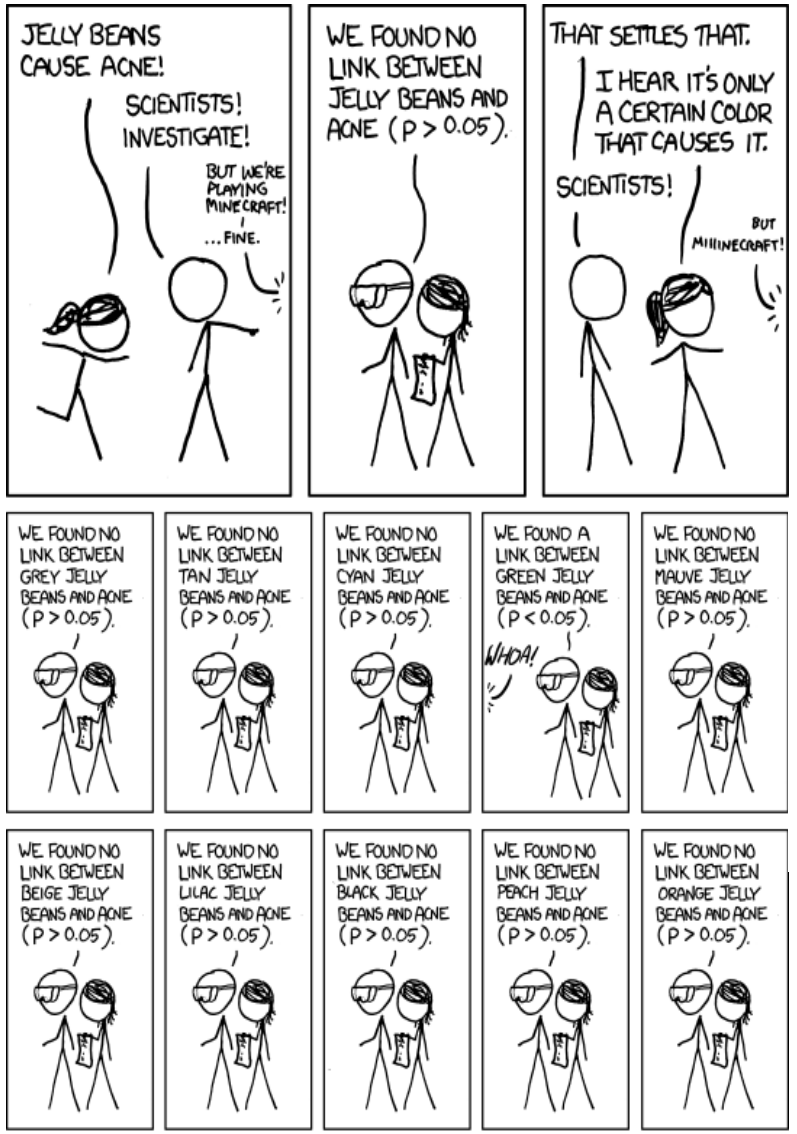
4 3 6 5 6 5 5 6 6 2 5 5 3 6 **9** 6 6 3 6 4

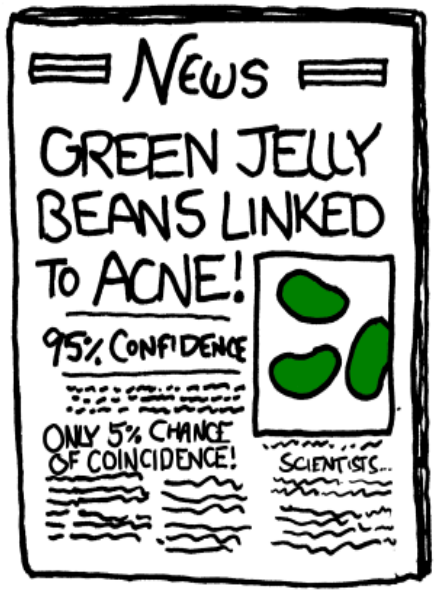
6 5 3 3 4 2 4 4 4 4 7 7 4 3 7 3 3 1 6 4

5 6 3 4 5 6 4 **8** 5 5 7 2 4 4 7 6 4 3 5 5

4 4 7 4 5 4 3 4 5 4 **8** 5 6 2 6 6 4 5 3 7

- Have to correct for multiple testing when you test, for example, all 20,000 genes in the human genome for differences





Pairwise Test for Multiple Conditions: ANOVA

- For testing more than continuous values against all combinations of each other
- Is there a difference in sepal length between the three species of iris in the iris dataset?
  - **H0:** There is no difference
  - **H1:** There is a difference between at least one group

```
> aov(Sepal.Length ~ Species, data = iris) %>%
TukeyHSD()

  Tukey multiple comparisons of means

    95% family-wise confidence level


Fit: aov(formula = Sepal.Length ~ Species, data =
iris)


$Species

                          diff       lwr       upr p adj
versicolor-setosa    0.930 0.6862273 1.1737727     0
virginica-setosa     1.582 1.3382273 1.8257727     0
virginica-versicolor 0.652 0.4082273 0.8957727     0
```

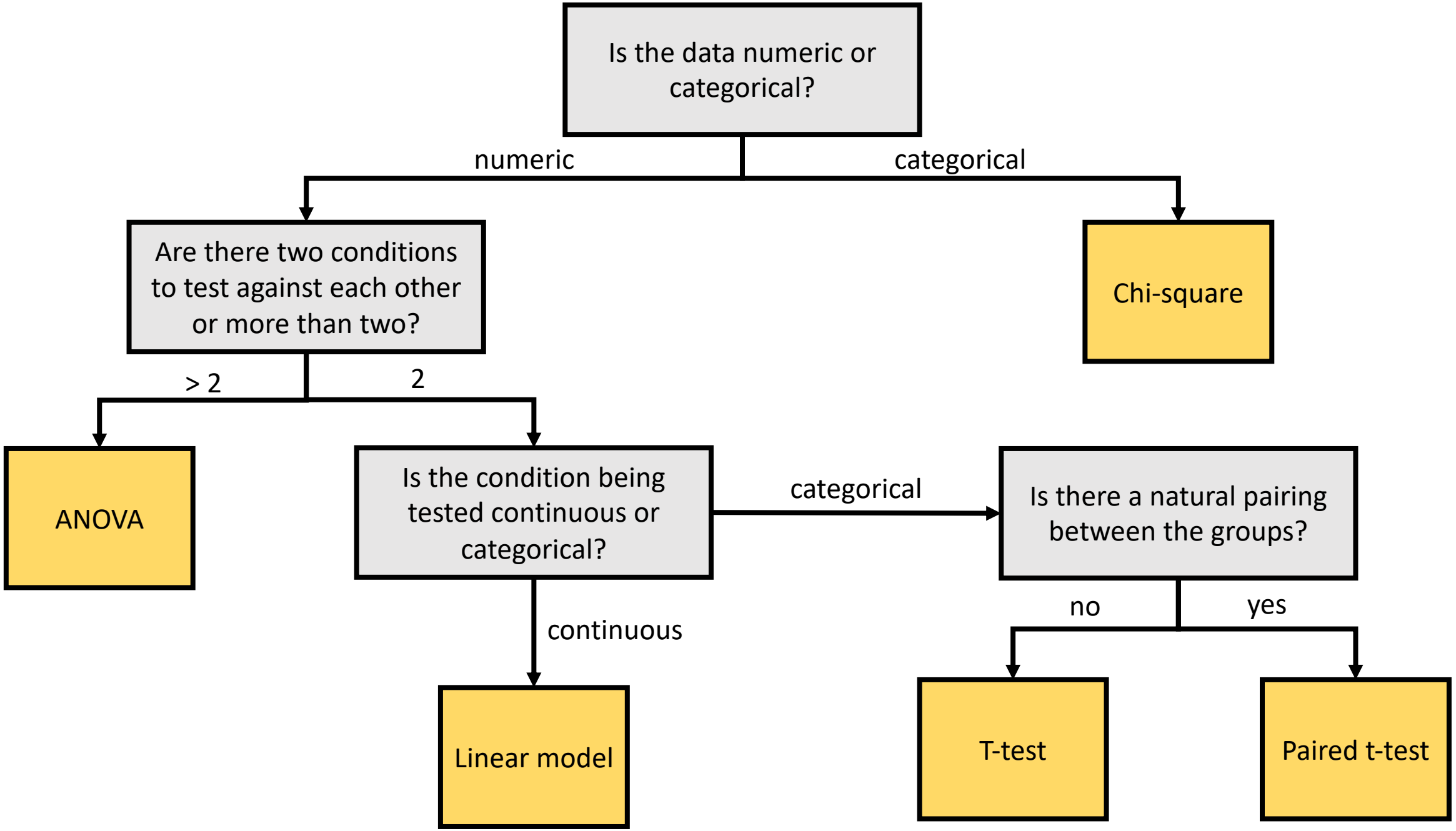
Test for Continuous Conditions: Linear Model

- For testing continuous variables over a continuous condition (like DNA methylation over time)
- AKA finding a line of best fit
- Is there an association between sepal width and sepal length in the iris dataset?
  - **H0:** There is no relationship
  - **H1:** There is a relationship

```
> lm(Sepal.Length ~ Sepal.Width, data = iris)

Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = iris)


Coefficients:
(Intercept)  Sepal.Width
     6.5262      -0.2234
```


Is the data numeric or categorical?
numeric
categorical
Are there two conditions to test against each other or more than two?
Chi-square
> 2
2
ANOVA
Is the condition being tested continuous or categorical?
categorical
Is there a natural pairing between the groups?
continuous
no
yes
Linear model
T-test
Paired t-test

DEMO