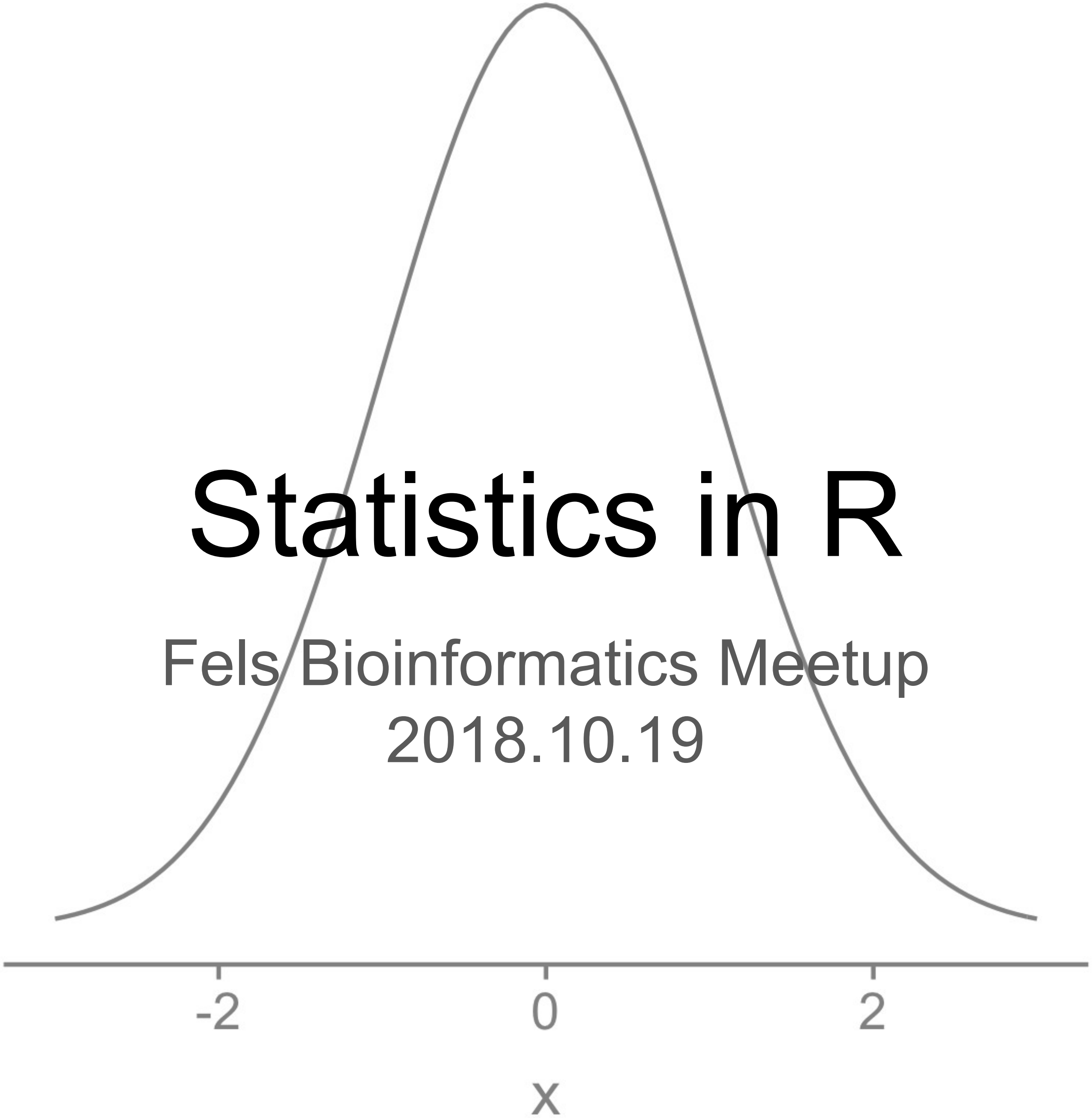
# Statistics in R

Fels Bioinformatics Meetup
2018.10.19



# Descriptive and Summary Statistics

# Measures of Central Tendency

mean

```
mean(iris$Sepal.Length)

[1] 5.843333
```

# Measures of Central Tendency

mean

```
mean(iris$Sepal.Length)

[1] 5.843333
```

median

```
median(iris$Sepal.Length)

[1] 5.8
```

# Measures of Central Tendency

mean

```
mean(iris$Sepal.Length)

[1] 5.843333
```

median

```
median(iris$Sepal.Length)

[1] 5.8
```

mode

No native mode function! There is a function `mode()`, but it doesn't find the statistical mode

# Measures of Central Tendency

mean

```
mean(iris$Sepal.Length)

[1] 5.843333
```

median

```
median(iris$Sepal.Length)

[1] 5.8
```

mode

No native mode function! Can find anyway, for example:

```
iris %>% group_by(Sepal.Length) %>% count()
%>% arrange(desc(n))
   Sepal.Length     n
          <dbl> <int>
 1            5    10
 2          5.1     9
```

# Measures of Central Tendency

mean

median

mode

Remember that you can apply these tests using `summarize()`:

```
> iris %>% group_by(Species) %>%
summarize(avg_sepal_len = mean(Sepal.Length),
median_sepal_len = median(Sepal.Length))

# A tibble: 3 x 3
  Species    avg_sepal_len median_sepal_len
  <fct>              <dbl>            <dbl>
1 setosa              5.01              5
2 versicolor          5.94              5.9
3 virginica           6.59              6.5
>
```

# Measures of Spread

range

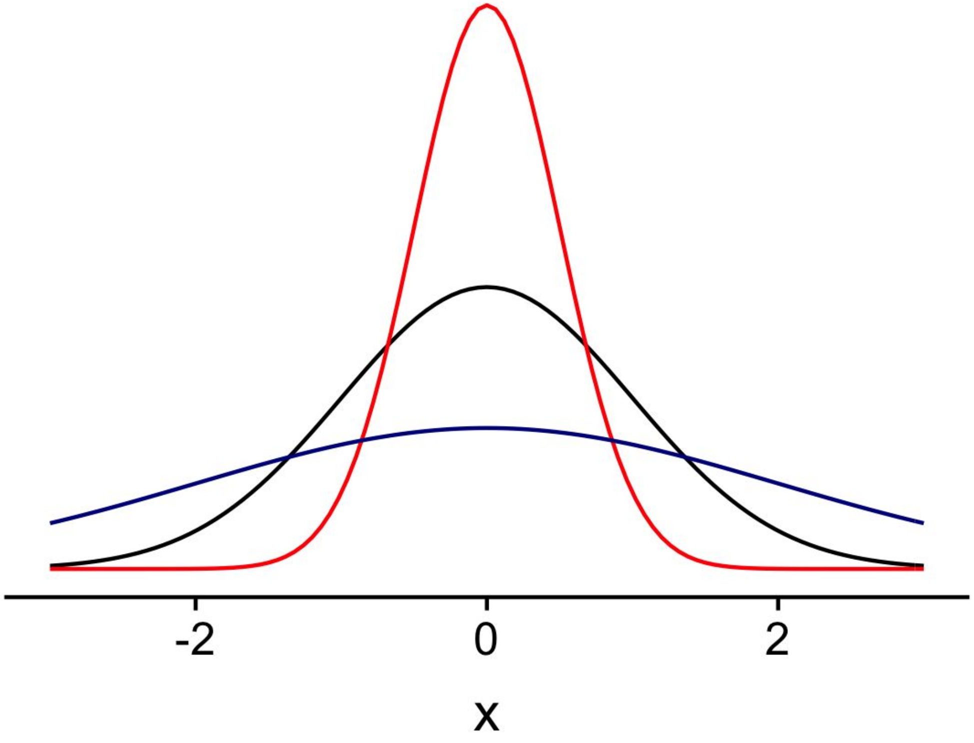
```
range(iris$Sepal.Length)

[1] 4.3 7.9
```

# Measures of Spread

range

```
range(iris$Sepal.Length)

[1] 4.3 7.9
```

standard deviation



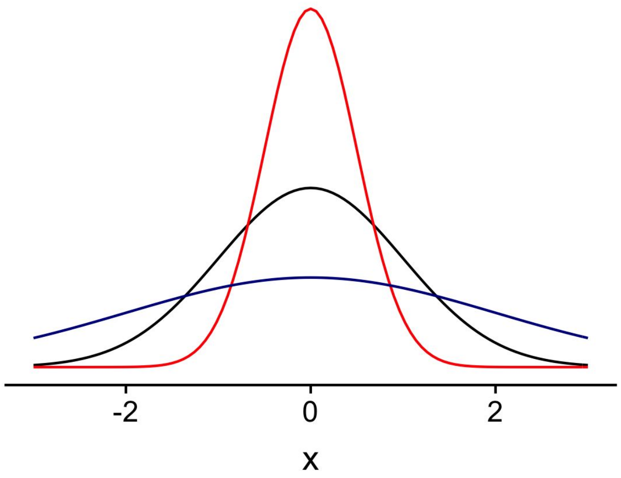
# Measures of Spread

range

```
range(iris$Sepal.Length)

[1] 4.3 7.9
```

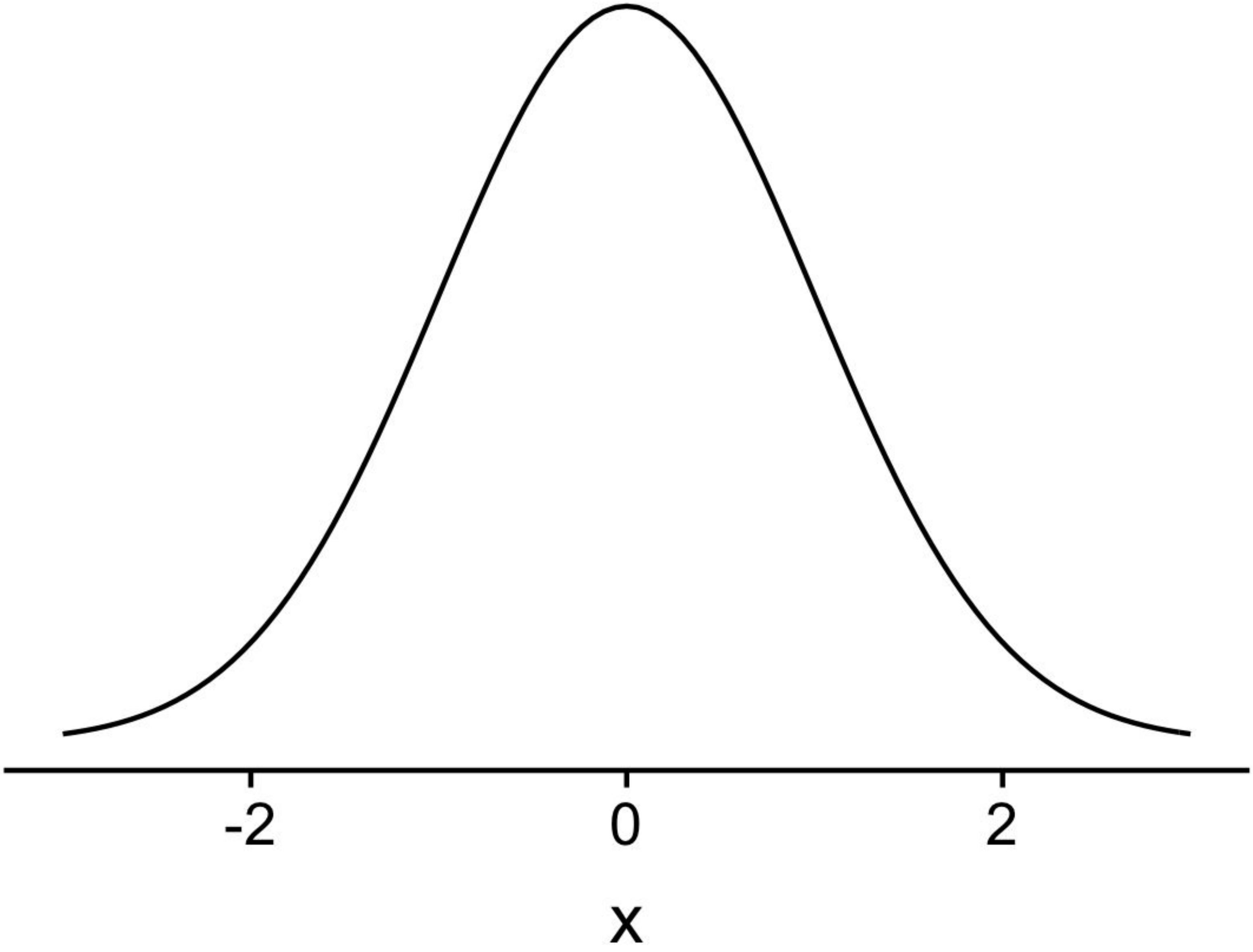
standard deviation

```
sd(iris$Sepal.Length)

[1] 0.8280661
```



# Hypothesis Testing for Continuous Data

# Normal Curve



# 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.***

# 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.

# 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.

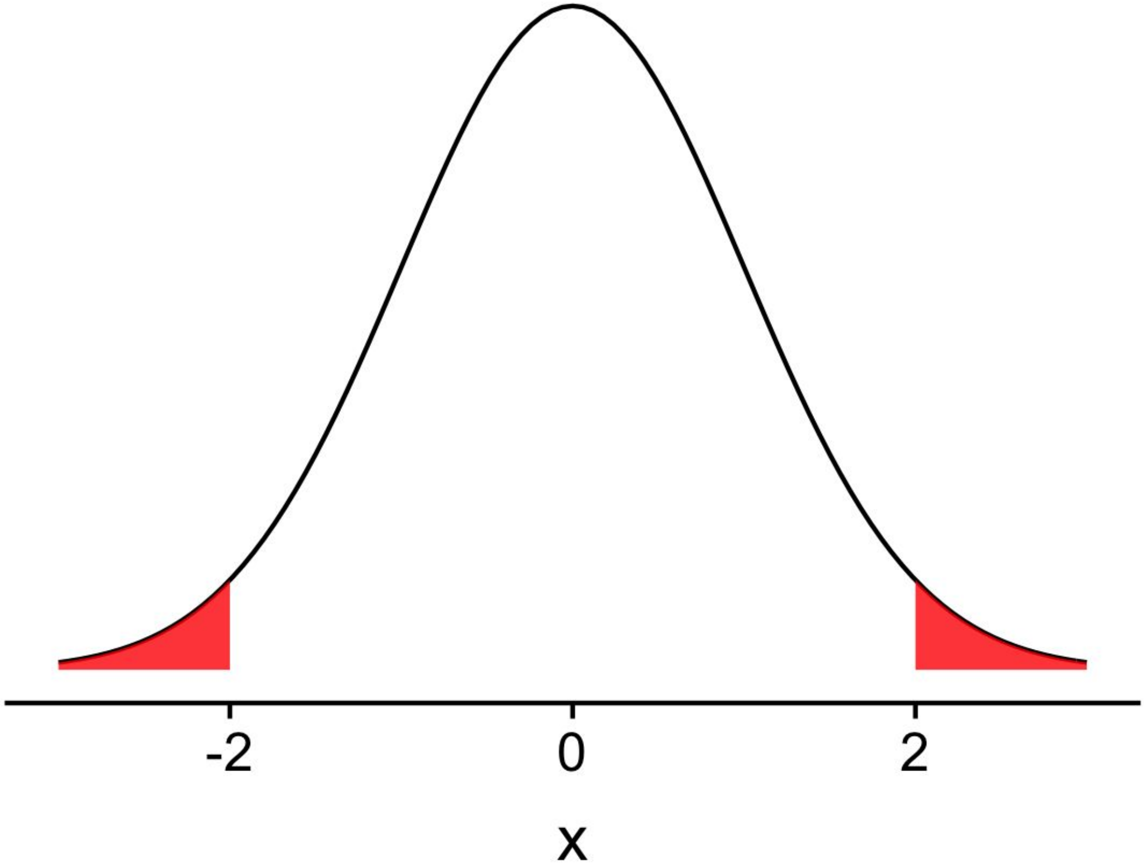
Alternative hypothesis, **H1**: My observations happened because of my idea.

# What does a hypothesis test tell you?

Null hypothesis, **H0**:
Everything happened by random chance.

Alternative hypothesis, **H1**:
My observations happened because of my idea.

Pvalue = 0.05 means that there's a 5% chance the observation happened randomly under the null distribution.



# One Sample t test

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
```

# Two Sample t test

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
```

# Two Sample t test (alternative tidier syntax)

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.

```
> iris %>% filter(Species != 'setosa') %>%
    t.test(Sepal.Length ~ Species, data = .)

    Welch Two Sample t-test

data:  Sepal.Length by Species
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 in group versicolor  mean in group virginica
                   5.936                    6.588
```

# Paired Two Sample t test

The sleep dataset in R has data on the amount of time patients sleep on two different sleep medications compared to control.

H0: There is no difference in the amount of time patients sleep.
H1: There is a difference in the amount of time patients sleep.

```
> sleep
   extra group ID
1    0.7     1  1
2   -1.6     1  2
3   -0.2     1  3
4   -1.2     1  4
5   -0.1     1  5
6    3.4     1  6
7    3.7     1  7
8    0.8     1  8
9    0.0     1  9
10   2.0     1 10
```

# Paired Two Sample t test

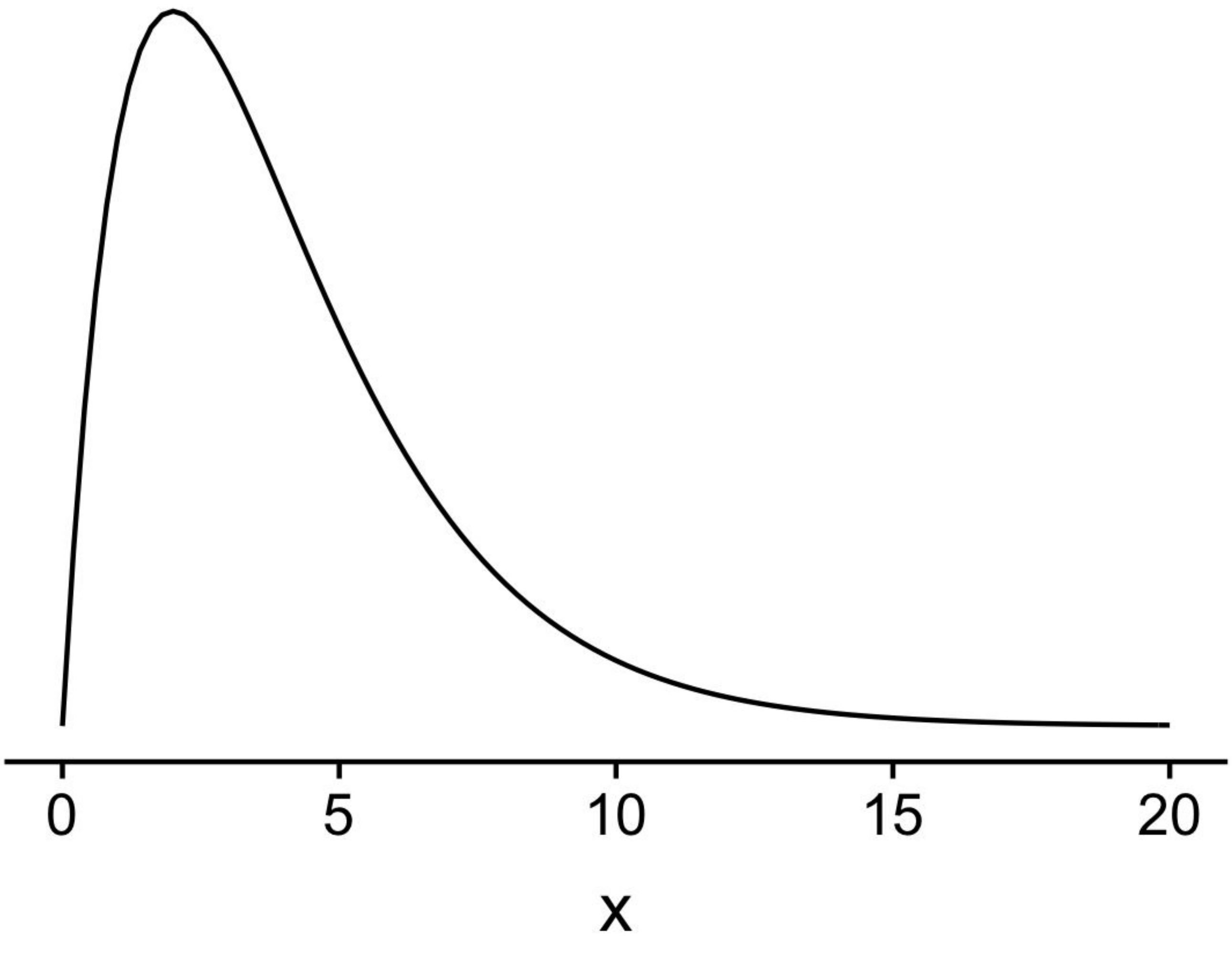
The sleep dataset in R has data on the amount of time patients sleep on two different sleep medications compared to control.

H0: There is no difference in the amount of time patients sleep.
H1: There is a difference in the amount of time patients sleep.

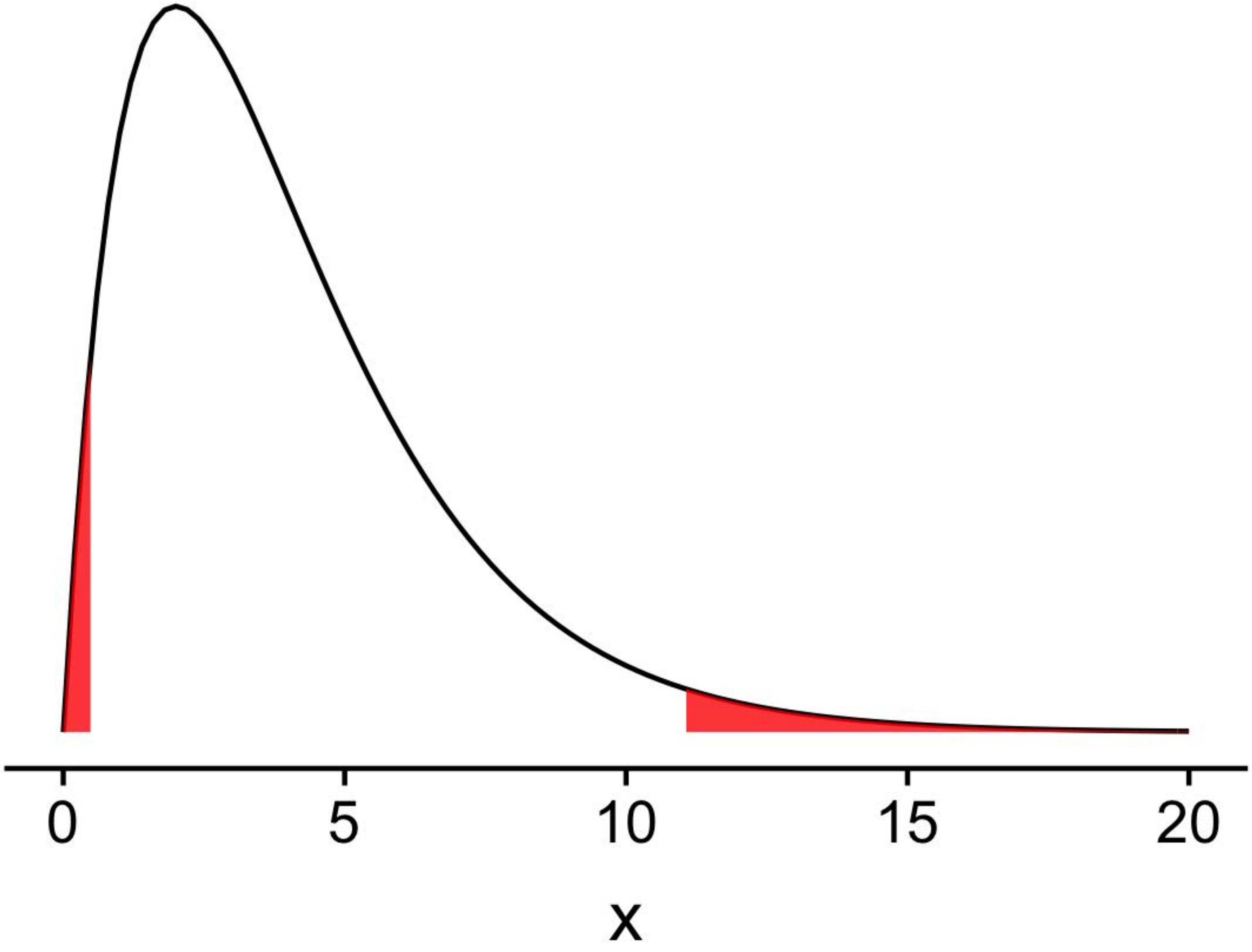
```
> t.test(extra ~ group, data = sleep, paired = TRUE)

    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
```

# Hypothesis Testing for Discrete Data

# Chi square distribution



# Chi square distribution



# Chi-square goodness-of-fit test

You flip a coin 10 times and it comes up heads 8 times. You repeat the experiment 10 times and come up with high numbers all ten time. Is the coin weighted?

H0: The coin isn’t weighted
H1: The coin is weighted

```
> coin_tosses
# A tibble: 10 x 2
   head_count prob_head
        <dbl>     <dbl>
 1          8       0.5
 2          7       0.5
 3          9       0.5
 4          7       0.5
 5          6       0.5
 6          6       0.5
```

# Chi-square goodness-of-fit test

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

```
> birth_days
# A tibble: 7 x 4
  day       num_births num_days exp_prob_birth
  <chr>          <dbl>    <dbl>          <dbl>
1 Sunday            33       52          0.142
2 Monday            41       52          0.142
3 Tuesday           63       52          0.142
4 Wednesday         63       52          0.142
5 Thursday          47       52          0.142
6 Friday            56       53          0.145
7 Saturday          47       52          0.142
```

# Chi-square goodness-of-fit test

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
```

# Chi-square contingency table test

Is there a difference in the number of tasks Wives vs Husbands complete?

H0: There’s no difference in the number of tasks completed
H1: There is a difference in the number of tasks completed

```
> housetasks
# A tibble: 13 x 5
   task        Wife Alternating Husband Jointly
   <chr>      <int>       <int>   <int>   <int>
 1 Laundry      156          14       2       4
 2 Main_meal    124          20       5       4
 3 Dinner        77          11       7      13
 4 Breakfeast    82          36      15       7
 5 Tidying       53          11       1      57
 6 Dishes        32          24       4      53
 7 Shopping      33          23       9      55
 8 Official      12          46      23      15
 9 Driving       10          51      75       3
10 Finances      13          13      21      66
```

# Chi-square contingency table test

Is there a difference in the number of tasks completed when you consider alternating and joint tasks as well?

H0: There’s no difference in the number of tasks completed
H1: There is a difference in the number of tasks completed

```
> housetasks %>% select(-task) %>% chisq.test(.)

	Pearson's Chi-squared test

data:  .
X-squared = 1944.5, df = 36, p-value < 2.2e-16
```

# Tidying the Test

# broom package

The broom package tidies common statistical tests and models for you.

```
> 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
```

# broom package

The broom package tidies common statistical tests and models for you.

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

# A tibble: 1 x 8
  estimate statistic p.value parameter conf.low conf.high
     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl>
1    -1.58     -4.06 0.00283         9    -2.46    -0.700

  conf.high method   alternative
  <chr>              <chr>
1 Paired t-test      two.sided
```

# Correcting for Multiple Testing

# The Problem with Multiple Testing

If you do enough tests, you expect to see significant results, just *by random chance* **.**

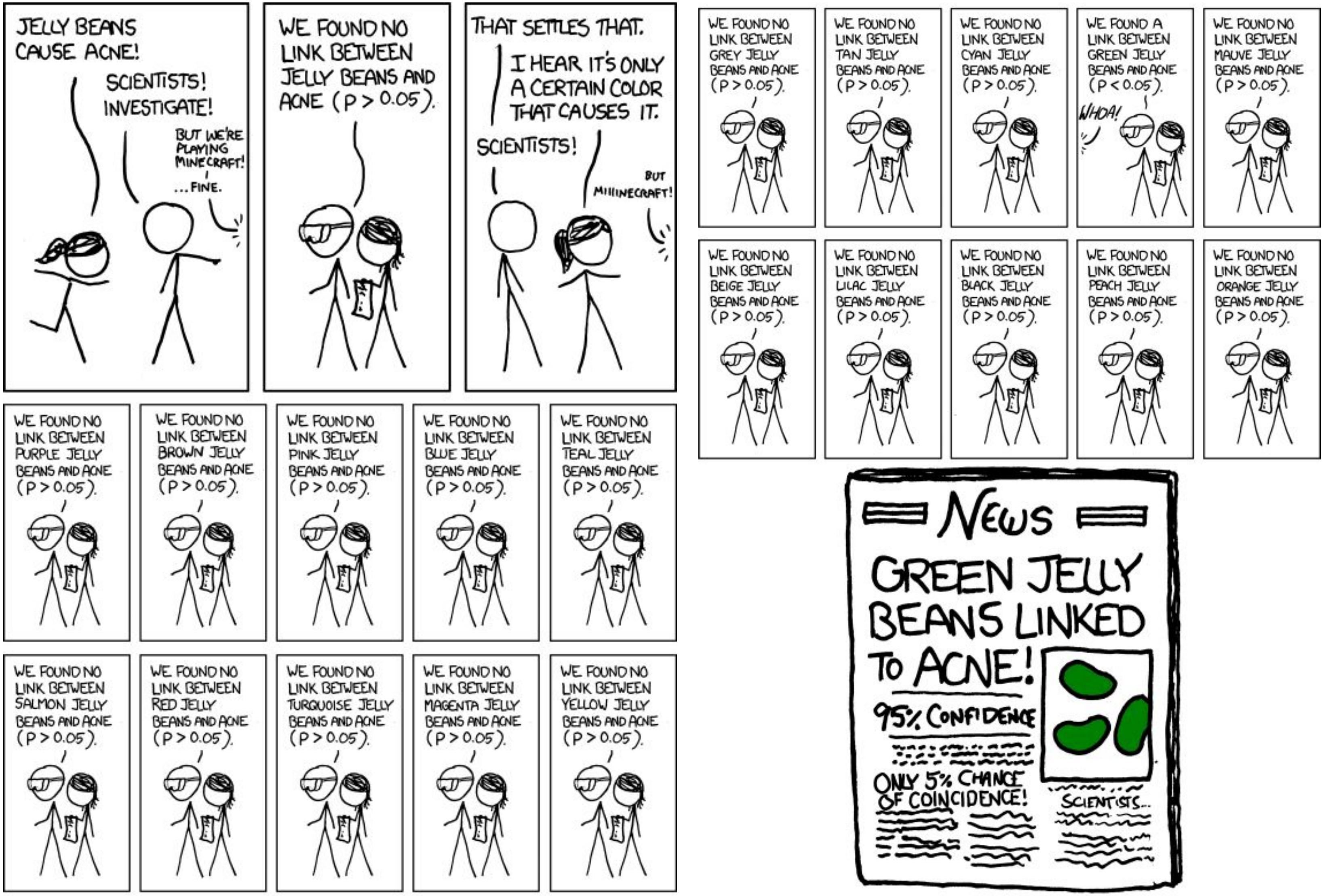
Say you flip a coin ten times. Then you repeat the experiment ten times.

```
> num_heads_10
 [1] 5 5 6 4 2 4 5 5 4 4
```

Now flip a coin ten times and repeat the experiment a hundred times:

```
> num_heads_100
  [1] 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
 [46] 2 4 4 4 4 7 7 4 3 7 3 3 1 6 4 5 6 3 4 5 6 48 5 5 7 2 4 4 7 6
      4 3 5 5 4 4 7 4 5 4 3 4 5 4
 [91] 8 5 6 2 6 6 4 5 3 7
```

# The Problem with Multiple Testing


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

# Correcting for Multiple Testing

Can correct anything with `p.adjust()`.

```
> p_values
 [1] 0.050 0.100 0.008 0.060 0.150 0.030 0.090 0.001 0.010 0.020
```

# Correcting for Multiple Testing

Can correct anything with `p.adjust()`.

```
> p_values
 [1] 0.050 0.100 0.008 0.060 0.150 0.030 0.090 0.001 0.010 0.020
```

# Correcting for Multiple Testing

Can correct anything with `p.adjust()`.

```
> p_values
 [1] 0.050 0.100 0.008 0.060 0.150 0.030 0.090 0.001 0.010 0.020
```

With correction for multiple testing:

```
> p.adjust(p_values)
 [1] 0.250 0.270 0.072 0.250 0.270 0.180 0.270 0.010 0.080 0.140
```

# Correcting for Multiple Testing

Can correct anything with `p.adjust()`.

```
> p_values
 [1] 0.050 0.100 0.008 0.060 0.150 0.030 0.090 0.001 0.010 0.020
```

With correction for multiple testing:

```
> p.adjust(p_values)
 [1] 0.250 0.270 0.072 0.250 0.270 0.180 0.270 0.010 0.080 0.140
```

Can pick what correction you’d like, some are harsher than others:

```
> p.adjust(p_values, method = 'fdr')
 [1] 0.083 0.111 0.033 0.085 0.150 0.060 0.111 0.010
 [9] 0.033 0.050

> p.adjust(p_values, method = 'bonferroni')
 [1] 0.50 1.00 0.08 0.60 1.00 0.30 0.90 0.01 0.10 0.20
```

# Correcting for Multiple Testing

Can correct anything with `p.adjust()`.

```
> p_values
 [1] 0.050 0.100 0.008 0.060 0.150 0.030 0.090 0.001 0.010 0.020
```

With correction for multiple testing:

```
> p.adjust(p_values)
 [1] 0.250 0.270 0.072 0.250 0.270 0.180 0.270 0.010 0.080 0.140
```

Can pick what correction you'd like, some are harsher than others:

```
> p.adjust(p_values, method = 'fdr')
 [1] 0.083 0.111 0.033 0.085 0.150 0.060 0.111 0.010
 [9] 0.033 0.050

> p.adjust(p_values, method = 'bonferroni')
 [1] 0.50 1.00 0.08 0.60 1.00 0.30 0.90 0.01 0.10 0.20
```

**FDR is a good default choice**

# Built-in Multiple Testing Correction

Without correction for multiple testing:

```
> pairwise.t.test(iris$Sepal.Length, iris$Species) %>% tidy()
# A tibble: 3 x 3
  group1     group2      p.value
* <chr>      <chr>         <dbl>
1 versicolor setosa     1.75e-15
2 virginica  setosa     6.64e-32
3 virginica  versicolor 2.77e- 9
```

With correction for multiple testing:

```
> pairwise.t.test(iris$Sepal.Length, iris$Species, p.adj = 'fdr') %>%
tidy()
# A tibble: 3 x 3
  group1     group2      p.value
* <chr>      <chr>         <dbl>
1 versicolor setosa     1.32e-15
2 virginica  setosa     6.64e-32
3 virginica  versicolor 2.77e- 9
```