module_4_2

Keaton Wilson

4/16/2020

Site-specific data, permutation tests

We're going to use another data set on penguins for the rest of this module. Remember that there are a few goals here:
1. To explore the penguin nesting data set and generate some conclusions about what factors are related to nesting success
2. Take these conclusions to inform our decisions about where the newly developed road should go. We're trying to mitigate environmental impact as much as possible by understanding what factors drive penguin nesting success.

Exploring the new data

Let's take a look at the new data set.

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------- tidyverse
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts -------------------------------------------------------------------------- tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
penguins_2 = read_csv("https://tinyurl.com/yardhofj")
```

```
## Parsed with column specification:
## cols(
##   site_id = col_double(),
##   year = col_double(),
##   tussocks = col_double(),
##   dist_to_water = col_double(),
##   stone_size = col_double(),
##   num_nests = col_double()
## )
```

```
glimpse(penguins_2)
```

```
## Rows: 200
## Columns: 6
## $ site_id       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ year          <dbl> 1971, 1971, 1971, 1971, 1971, 1971, 1971, 1971, 1971,...
## $ tussocks      <dbl> 4.219239, 3.932796, 6.263532, 4.116507, 4.118727, 7.1...
## $ dist_to_water <dbl> 27.86114, 25.33264, 28.93208, 28.39797, 28.75344, 28....
## $ stone_size    <dbl> 45.04550, 39.73190, 37.47425, 45.87984, 45.69094, 38....
## $ num_nests     <dbl> 18, 8, 3, 25, 15, 7, 15, 19, 13, 15, 25, 11, 23, 22, ...
```

Independent Exploration of Data

Spend 3-5 minutes on your own exploring the data set. Focus on the following questions:
1. How many individual sites are represented here?
2. What is similar from the first data set we examined?
3. What is different from the first data set we examined?
4. Are there any columns/variables that you have questions about?

Comparisons of means - another approach...

One of the main questions we're interested in with these data is whether or not there is a difference in our variables across all sites but between time intervals (e.g. have things changed between 1971 and 2011). Today, we're going to specifically focus on number of tussocks. What tools have we talked about so far that would address this question?

We're going to explore another option today that is a little more flexible - there are a bunch of assumptions associated with t-tests that can really mess things up - and the type of test we're working on today gives us a way to still make comparisons, even when some of those assumptions aren't met. This method also allows us to go over hypothesis testing, p-values and some other statistical fundamentals that are part of the data science toolkit.

The permutation test - but with alpacas

Check out this link: https://www.jwilber.me/permutationtest/

Spend about 5 minutes on your own walking through the logic and example and then work in your groups to answer the following discussion questions (remember to designate someone as the reporter for reporting out!):
1. What is the conclusion in the example scenario(from a reality-perspective, not a statistical one) - what actionable insights were gained from this experiment?

2. Can you explain what the big difference is between this method and a t-test?

3. Can you give a brief outline of each step for the permutation test?

4. Can you think of any problems with running through all the permutations possible, as the author talks about?

5. What is the hypothesis we want to test with our data? What is the null hypothesis associated with this question?

Coding permutation tests in R

```
library(perm)
glimpse(penguins_2)
```

```
## Rows: 200
## Columns: 6
## $ site_id       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ year          <dbl> 1971, 1971, 1971, 1971, 1971, 1971, 1971, 1971, 1971,...
## $ tussocks      <dbl> 4.219239, 3.932796, 6.263532, 4.116507, 4.118727, 7.1...
## $ dist_to_water <dbl> 27.86114, 25.33264, 28.93208, 28.39797, 28.75344, 28....
## $ stone_size    <dbl> 45.04550, 39.73190, 37.47425, 45.87984, 45.69094, 38....
## $ num_nests     <dbl> 18, 8, 3, 25, 15, 7, 15, 19, 13, 15, 25, 11, 23, 22, ...
```

```
test = perm::permTS(x = penguins_2 %>%
                    filter(year == 1971) %>%
                    pull(tussocks),
             y = penguins_2 %>%
                    filter(year == 2011) %>%
                    pull(tussocks))
```

Practice - does stone size change with year?

By yourselves, generate code that tests the following hypothesis: there is a difference in the mean stone size between 1971 and 2011.

What is the null hypothesis? What does the output of the permutation test tell us?