Homework 7: New York City Trees
Comparing a Numerical Variable Across Two Groups

New York City Trees

New York City's trees shade us in the summer, beautify our neighborhoods, help reduce noise, and support urban wildlife. Beyond these priceless benefits, our urban forest provides us a concrete return on the financial investment we put into it. This return includes stormwater interception, energy conservation, air pollutant removal, and carbon dioxide storage. Our publicly owned trees are as much of an asset to us as our streets, sewers, bridges, and public buildings.

The dataset includes a random sample of 672 Norway Maple and Honeylocust trees from every tree in New York City and includes the tree's location by borough, species by common name `spc_common`, size in diameter inches `tree_dbh`, health, and issues with the tree's roots, trunk, and branches.

```
head(nyc_trees)
```

```
# A tibble: 6 x 6
  tree_id tree_dbh status spc_common    sidewalk problems
    <dbl>    <dbl> <chr>  <chr>         <chr>    <chr>
1  515625       14 Alive  honeylocust   NoDamage None
2   49463        7 Alive  honeylocust   NoDamage None
3   35410       10 Alive  honeylocust   NoDamage TrunkOtherBranchOther
4  639748       17 Alive  Norway maple  NoDamage None
5  133934       15 Alive  honeylocust   NoDamage None
6   38232       18 Alive  honeylocust   NoDamage None
```

Research question Is the average tree diameter of a Norway Maple different than the tree diameter of a Honey Locust?
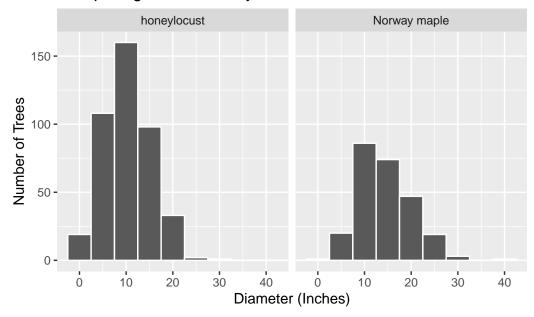
```
favstats(tree_dbh ~ spc_common, data = nyc_trees)
```

```
   spc_common min Q1 median Q3 max      mean      sd   n missing
1  honeylocust   1  6     10 14  30 10.31116 4.968529 421       0
2 Norway maple   1 11     14 18  41 14.22709 5.725052 251       0
```
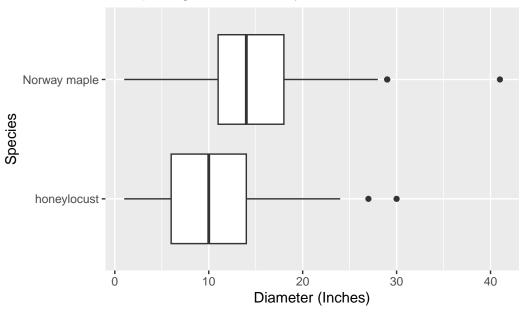
```
ggplot(data = nyc_trees,
       mapping = aes(x = tree_dbh)) +
  geom_histogram(binwidth = 5, color = "white") +
  facet_wrap(~spc_common) +
  labs(title = "Comparing New York City Trees",
       x = "Diameter (Inches)",
       y = "Number of Trees")
```



```
ggplot(data = nyc_trees,
       mapping = aes(x = tree_dbh,
                     y = spc_common)) +
  geom_boxplot() +
  labs(title = "Comparing New York City Trees",
       x = "Diameter (Inches)",
       y = "Species")
```
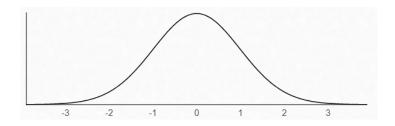
Comparing New York City Trees

1. Identify the variables and data type.

- Explanatory:

- Response:

2. Is this an experimental study or observational study? Explain.

3. Discuss the visual comparison of the the two tree species. Describe their distributions (recall: 4 characteristics). Which tree species appears to have a higher average diameter?

4. Identify the observed mean, standard deviation, and sample size of each species in the study.

3

5. Calculate the observed difference in average diameter (Norway Maple - Honeylocust). Assign an appropriate symbol.

6. State the parameter in words $(\mu_{\text{Norway Maple}} - \mu_{\text{Honeylocust}})$.

7. Write the null and alternative hypotheses using the appropriate mathematical symbols.

8. Check the assumptions necessary to conduct a two-sample independent t-test.

9. Calculate the t-statistic.

10. Using the t-distribution below, show how you would calculate/estimate the p-value. *Yes, this is going to be a little funky given the magnitude of the t-statistic.*

11. The following R code conducts a two-sample t-test for the data. What do you think the `order =` argument does? What do you hypothesize would happen to the results if we used `order = c("honeylocust", "Norway maple")` instead?

```
t_test(x = nyc_trees,
       response = tree_dbh,
       explanatory = spc_common,
       order = c("Norway maple", "honeylocust"),
       alternative = "two-sided",
       conf_int = FALSE)
```

```
# A tibble: 1 x 5
  statistic  t_df  p_value alternative estimate
      <dbl> <dbl>    <dbl> <chr>          <dbl>
1      9.00  469. 5.56e-18 two.sided       3.92
```

12. Using the output above, write a conclusion in the context of the problem.

13. Where can we generalize our results (i.e. what is our population in which we obtained the observed data)?

14. Construct a 95% confidence interval for the true difference in mean diameter between Honeylocust and Norway Maple trees.

Paired t-test vs Two-Sample Independent t-test

For each of the following scenarios, indicate whether a paired t-test analysis (dependent groups) or a two-sample independent t-test analysis (independent groups) would be appropriate based on the details given.

a. You wonder whether undergraduate or graduate students tend to go to more football games. You take a random sample of 20 undergraduate students and then another random sample of 20 graduate student, and ask how many games they attended last fall.

    PAIRED   TWO-SAMPLE

b. On average, is there a significant difference in one's weight before and after the holidays?

    PAIRED   TWO-SAMPLE

c. Is autism marked by different brain growth patterns in early life, even before an autistic diagnosis is made? Studies have linked brain size in infants and toddlers to a number of future ailments, including autism. One study looked at the brain sizes of 30 autistic boys and 13 non-autistic boys who all had received an MRI scan as toddlers. The whole-brain volume (in ml) was recorded for each boy.

    PAIRED   TWO-SAMPLE

d. Woo and McKenna (British Journal of Dermatology, 2003) investigated the effect of broadband ultraviolet B (UVB) therapy and topical calcipotriol cream used together on area of psoriasis. One of the outcome variables from the study was the Psoriasis Area and Severity Index (PASI). The PASI score of twenty subjects were measured at baseline and after eight treatments.

PAIRED   TWO-SAMPLE

e. A researcher interviews 64 students in their freshman year and then again in their senior year to determine if their opinion has changed regarding their ideal number of children.

PAIRED   TWO-SAMPLE

f. Some archaeologists theorize that ancient Egyptians interbred with several different immigrant populations over thousands of years. To see if there is any indication of changes in body structure that might have resulted, researchers measured the breadth (measured in mm) of a random sample of male Egyptian skulls dated from 4000 B.C. and from 200 B.C. (Ancient Races of the Thebaid, 1905).

PAIRED   TWO-SAMPLE

> 🔥 Canvas Quiz
>
> Make sure to complete the Homework Quiz on Canvas.