Homework_0

Oviya Mohan & Yue Zhang

7/19/2022

Complete and submit this Rmd by emailing it to either yzh131@u.rochester.edu or omohan@ur.rochester.edu by 12 am on Aug 22, 2022.

R Markdown

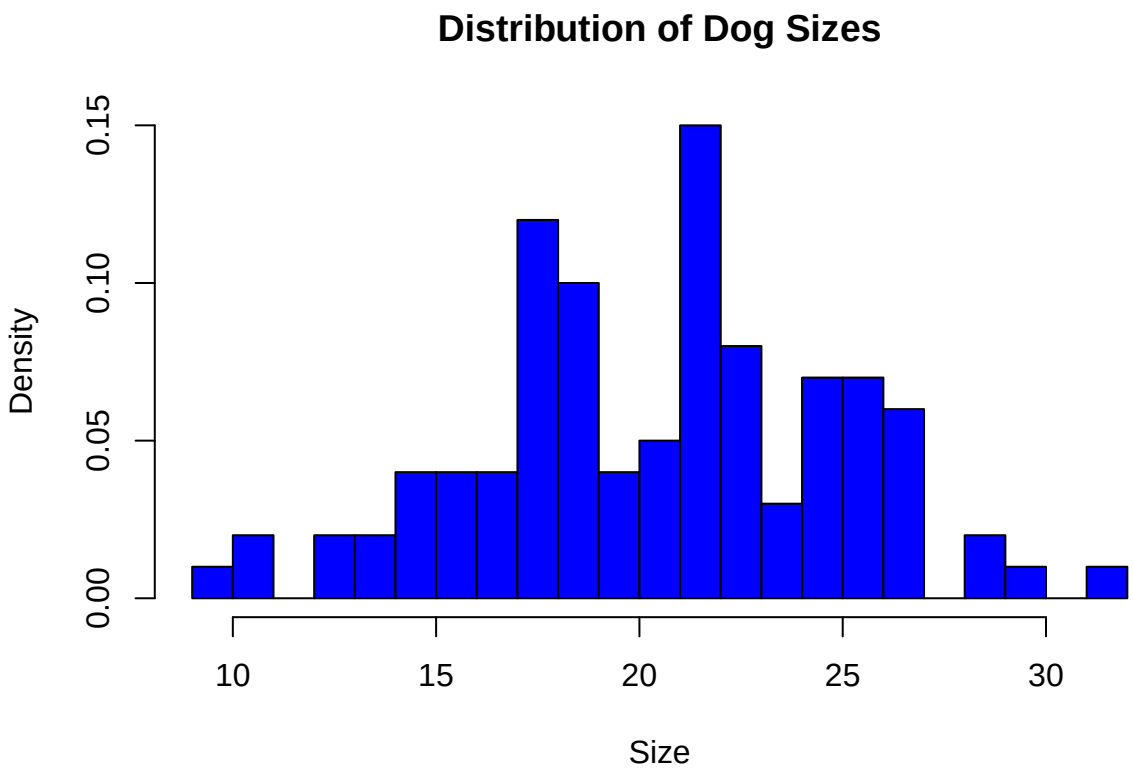
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

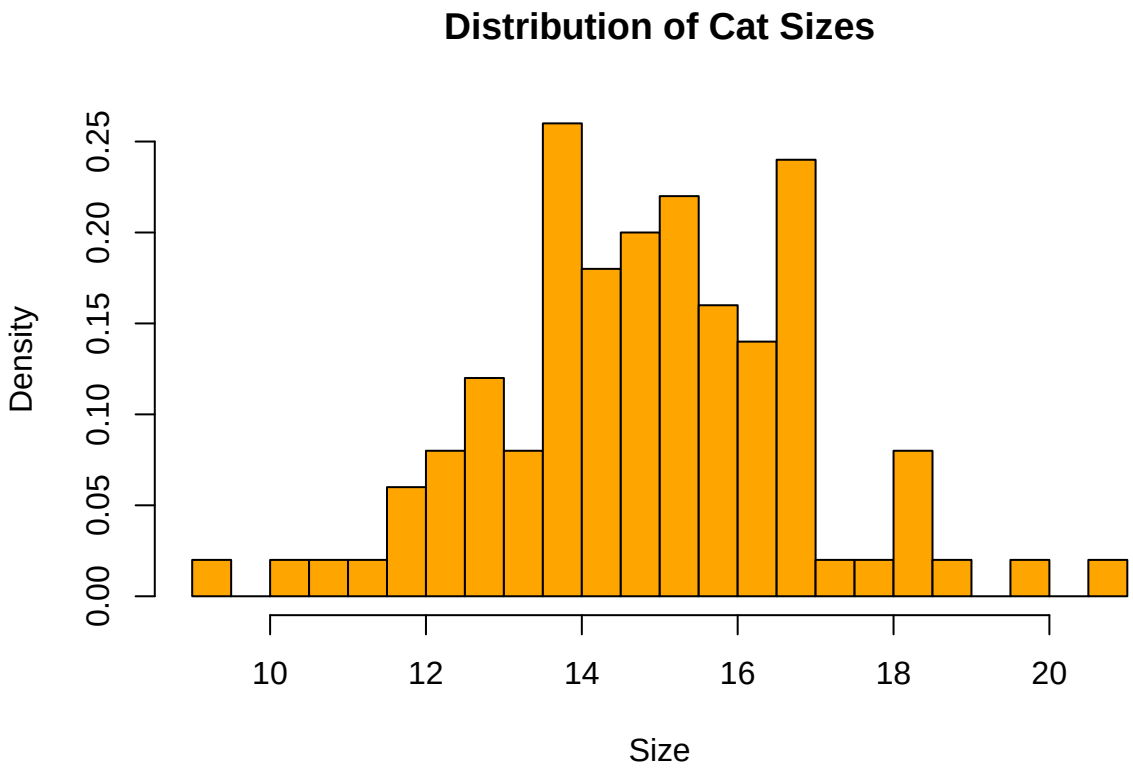
Intro to R - Resources

Simulating data to work with

```
#nothing is ever truly random when "randomly" generated on a computer - always psuedo random
#set.seed(124) #can be used to get the same "random" sample consistently
dog_sample = rnorm(100,20,5)
cat_sample = rnorm(100,15,2)

#random R tip #1: to view any variable, highligh it and Cmd+enter/Ctrl+enter

#summarizing our data
dog_mean <- mean(dog_sample)
dog_sd <- sd(dog_sample)

cat_mean <- mean(cat_sample)
cat_sd <- sd(cat_sample)

#visulizing data
hist_dog <- hist(dog_sample,
main="Distribution of Dog Sizes",
xlab="Size",
col="blue",
alpha = 0.6,
freq=FALSE,
breaks = 20
)
```



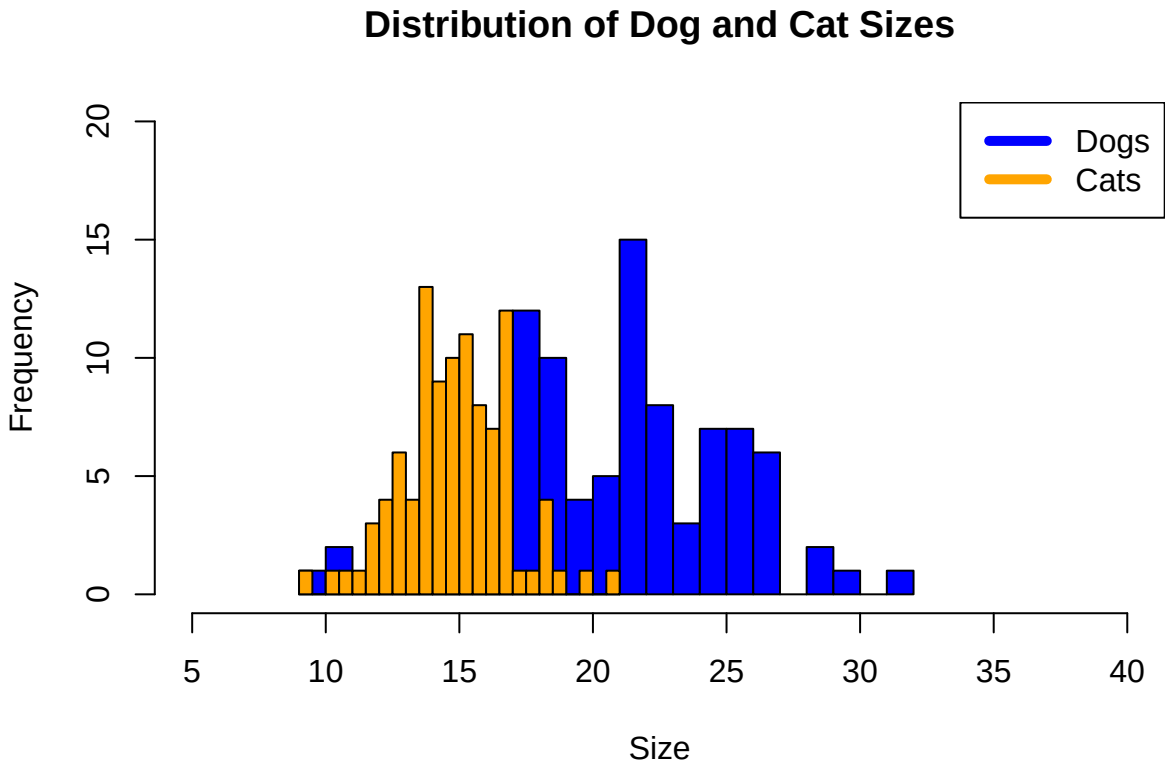

```
hist_cat <- hist(cat_sample,
main="Distribution of Cat Sizes",
xlab="Size",
col="orange",
alpha = 0.4,
freq=FALSE,
breaks = 20
)
```




```
#in the same plot
plot(hist_dog, xlim = c(5, 40), ylim = c(0,20), col = "blue", main = "Distribution of Dog and Cat Sizes")
```

```
## integer(0)
```

```
legend("topright", c("Dogs", "Cats"), col=c("blue", "Orange"), lwd=5)
```



############### Q1. Report the mean and sd of the two sample:

Null Hypothesis Signifiance Testing (NHST)

Readings:

Null Hypothesis Significance Testing: a short tutorial

Storyline for data simulated above: What do we want to test? Null Hypothesis: Alternate Hypothesis:

Q2. How do you think changing the Null Hypothesis affects the tests conducted, reported results and inferences?

Two-Sample t-test

The two-sample t-test is a parametric test that compares the location parameter of two independent data samples.

The test statistic is

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}},$$

where $\overline{x}$ and $\overline{y}$ are the sample means, s_x and s_y are the sample standard deviations, and n and m are the sample sizes.

In the case where it is assumed that the two data samples are from populations with equal variances, the test statistic under the null hypothesis has Student's t distribution with $n + m -$ freedom, and the sample standard deviations are replaced by the pooled standard deviation

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}.$$

In the case where it is not assumed that the two data samples are from populations with equal variances, the test statistic under the null hypothesis has an approximate Student's t dist number of degrees of freedom given by Satterthwaite's approximation. This test is sometimes called Welch's t-test.

T-test Quick Summary:
Types:

```
#one line code for each type of test so that after assumptions, they can choose which one to use and us
```

We are using: INDEPENDENT TWO SAMPLE T-TEST - make this a Q?

Assumptions of a t-test

1. Independence Well, you can either be a cat or a dog... Unless you're Muffin the dog-cat or Pico the cat-dog but let's assume that all our cats are only cats and all our dogs are only dogs.
2. Normality For N >= 100, Shapiro-Wilk test for normality Insert resource for other kinds of tests?

Report Normality

3. Homogeneity - eqaulity of variances Levene's Test

Report Homogeniety

Q3. Based on the results of testing the assumptions, what t-test should we do? Do it here and report the results - APA style! - Should we given an example of how to report?

```
#t-test
```

The p-value monster Q4. What is p-hacking?

Confidence Intervals

Some food for thought List some extra interesting resources

Below list three questiosn/comments you have for discussion in the first class based on what you have done above and the readings. Atleast two of the questions should be conceptual. If you have any syntax/code

based doubts, do NOT include them here - ask on Slack. The questions you list here will be voted on along with questions from others and we will be disucssing the most popular questions.