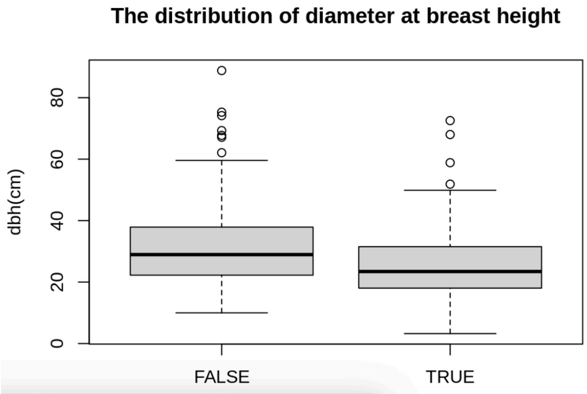
(1) Explain in 1-2 sentences what it means that the researchers “stratified by disease status within each species.” (Hint: It may help to make a contingency table for the two variables; 5 points)

“Stratifies by disease status within each species” means that the researchers firstly divided the species into two groups: species with disease and species without disease. Then, they sample from each group. The researchers stratified each species by disease status since there is a possibility that species that have disease would share some similarities that the species without disease don’t share with. In this way, balance and independence of the results can be guaranteed if the researchers stratified the samples.

(2) `View` the tree disease dataset. Choose a continuous response variable and make a chart showing its distribution, separated by disease status (5 points).

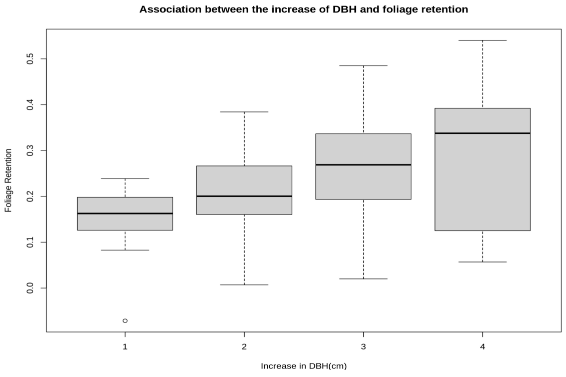
I choose to use box plot to illustrate the distribution separated by disease status. Box plot can both present a numerical response variable, which is DBH(continuous response variable), and a categorical predictor variable, which is disease status in this case.



The distribution of diameter at breast height

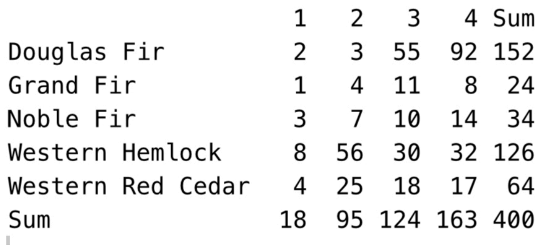
(3) The researchers believe that DBH growth is primarily modulated by foliage retention. Make a plot showing the association between the two. Explain in 1-2 sentences whether the plot is consistent with the researcher’s hypothesis (5 points). Choose 1 additional striking characteristic of the plot and hypothesize in 1-2 sentences what might cause this effect. (5 points)

I choose to use box plot to show the association between the DBH growth and foliage retention. I think the researchers’ hypothesis is correct, since as the foliage retention rises, the mean and the median of the increase in DBH also heightened. One striking characteristic of this plot is that the first plot that has the shortest whiskers, which means that the max and the min of this plot are limited in a smallest range compared with other three box-and-whisker plots, also



Association between the increase of DBH and foliage retention has an outlier. However, the outlier would not affect the observation of distribution since the whiskers can exclude the outliers.

(4) Make a contingency table, with margins, showing the relationship between species and foliage retention. Calculate the following probabilities (2 points each). Remember that answers must be in full, professionally formatted sentences to receive credit; just numbers is insufficient. You must also explain which rules and formulas you have used and why.



```
                     1   2   3   4 Sum
Douglas Fir          2   3  55  92 152
Grand Fir            1   4  11   8  24
Noble Fir            3   7  10  14  34
Western Hemlock      8  56  30  32 126
Western Red Cedar    4  25  18  17  64
Sum                 18  95 124 163 400
```

a. The probability that a randomly chosen tree is a Douglas Fir.
The probability that a randomly chosen tree is a Douglas Fir is 38%.
152/400=0.38
Formula: Number of Douglas Fir/all trees
Rule: equiprobability rule

b. The probability that a randomly chosen tree is some variety of Fir.
The probability that a randomly chosen tree is some variety of Fir is 52.5%.
210/400=0.525
Formula: All of the Firs(Douglas Firs+Grand Firs+Noble Firs)/All trees
Rule: basic additive rule

c. The probability that a randomly chosen tree has 4 years of retained foliage.
The probability that a randomly chosen tree has 4 years of retained foliage is 40.75%.
163/400=0.4075
Formula: Trees that has 4 years of retained foliage/All trees
Rule: equiprobability rule

d. The conditional probability that a Douglas Fir tree will have 4 years of retained foliage.
The conditional probability that a Douglas Fir tree will have 4 years of retained foliage is 60.5%.
(92/400)/(152/400)=0.6052632

Formula: Douglas Fir trees that have 4 years of retained foliage/all Douglas Fir trees
Rule: Definition of conditional probability

e. The conditional probability that a tree with less than 3 years of retained foliage is a Douglas Fir.
The conditional probability that a tree with less than 3 years of retained foliage is a Douglas Fir is 4.42%.
[(2+3)/400]/[(18+95)/400]=0.04424779
Formula: Douglas Firs that with less than 3 years of retained foliage(first year + second year)/all trees that with less than 3 years of retained foliage
Rule: Definition of conditional probability

(5) Make a contingency table, with margins showing the relationship between foliage.retention and disease status. Assuming the two variables are independent, use the **multiplicative rule** to calculate the probability that a diseased tree will have 4 years of retained foliage. Compare this to the **empirical probability**. Explain in 1-2 sentences whether you think disease status and foliage retention are independent and why (5 points).

P(4 year retained foliage) = 0.4075
P(Diseased) = 0.5
According to multiplicative rule{P(A) * P(B) = P(AB)}, the probability that a diseased tree will have 4 years of retained foliage is 0.4075 * 0.5 = 0.2035. From the table, we can get the empirical probability is 47/400 = 0.1175 which doesn't equal to 0.2035. Thus, the disease status and foliage retention are not independent.

(6) Calculate the empirical probability that a tree experiences less than 0.2 cm of DBH growth, conditional on the fact that that tree is diseased. Compare this to the empirical probability conditioned on the fact that the tree is not diseased (Hint: use booleans; 5 points).
The empirical probability that a tree experiences less than 0.2 cm of DBH growth, conditional on the fact that that tree is diseased is 68%(136/200). The empirical probability that a tree experiences less than 0.2 cm of DBH growth, conditional on the fact that that tree is not diseased is 2%(4/200). Comparing those two, the tree without disease but increase less than 0.2cm in DBH is 66% less than the trees with disease and increase less than 0.2cm in DBH.

Independent Learning Problem:

(1) Give an example of a leading question and explain in 1-3 sentences which answer it may be leading towards and why (5 points).

"More than 50% people said that their favorite language is English, is your favorite language English?" can be understood as a leading question since it doesn't use the neutral wording. People have a great possibility to answer "yes" rather than "no" even though their favorite language may not be English but some other subjects, because the question implies that the majority of people said "yes" for this question. So this may lead to a biased result. A question that is not leading question can be, "what is your favorite language?"

(2) Give an example of a question that might be subject to response bias and explain in 1-3 sentences which answers might be favored and why (5 points).
An example of response bias is that:
Do you prefer warm water? Agree or disagree
Do you prefer cold water? Agree or disagree
In this situation, question one and question two contradicts one another, but people are likely to answer "agree"(or "disagree") for both questions. Those can be considered as a subtle leading questions. It can be better phrasing like:
Do you prefer drink:
A: warm water
B: cold water
In this way, contradictory answers are not allowed. Also, in order to not to have order bias, the order of the answers should be randomized.

(3) Refer to the SAT dataset from the Lab 2 practice problems. The UC system defines URM inconsistently. In other circumstances, they define URM to include Filipino, Hmong, Vietnamese, and Native Hawaiian students. However, in this dataset, they group these students under "Asian/Pacific Islander." Explain in 2-4 sentences at least one way in which this choice of variables might bias the results (5 points).

According to the dataset, the retention probability of non-URM students was 0.98. However, students from URM had a retention probability of 0.93 primarily due to financial difficulties, Professor micro aggressions and imposter syndrome. Those are the characteristics that non-URM don't share with. If the UC system include URM into a bigger group, " Asian/Pacific Islander", the retention probabilities for URM and Asian/ Pacific Islander"(non-URM) are different, and this would cause social bias. Social bias happened when a situation involves prejudice towards particular groups of people, and in this case, URM is being prejudiced.