Summarizing and visualizing data SOLUTIONS

BIOL01104 Spring 2020, Dr. Spielman

This worksheet uses a dataset of information about olives collected from Italy. The data contains information about the fatty acid content for 572 olives from Northern Italy, Southern Italy, and Sardenia (a Mediterranean island off the coast of Italy). Answer questions about each plot with your table. **For all plots, identify the type of data (quantitative/numeric or categorical) on each axis. If numeric, identify if it is discrete or continuous**

Histograms

The histograms below show the **distribution of oleic acid percentages** for olives collected from the three regions.

1. Describe the shape each distribution:
   - Are they unimodal or bimodal?
     - Northern Italy is unimodal, Sardinia and Southern Italy are bimodal.
   - Roughly symmetric or asymmetric?
     - All are asymmetric
   - Are there any outliers?
     - Northern Italy has an outlier around 84, and it looks like Southern Italy *might* have some outliers at the right tail around 81.
2. What is (roughly) the minimum, maximum, and mode for each distribution?

Region	Minimum	Maximum	Mode
Northern Italy	~73	~84	~79.5
Sardenia	~68	~75	~74.5
Southern Italy	~63.5	~81	~70

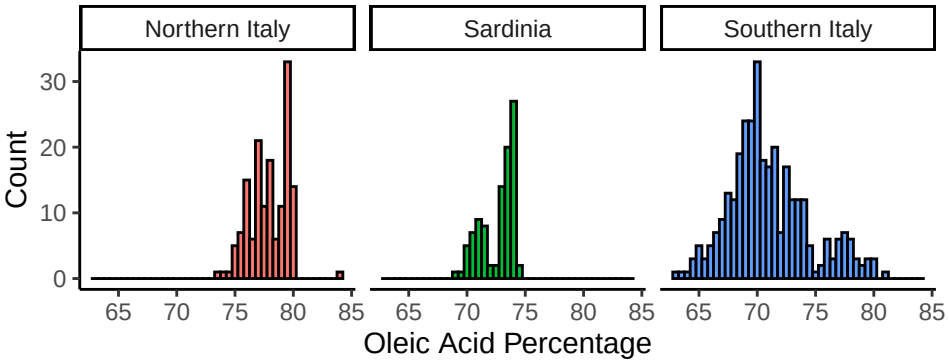
3. Rank the regions in terms of their *means*: Which region has the highest mean oleic acid percentage? The second highest? The lowest? (See next answer)
4. Rank the regions in terms of their *variation*: Which region has the most variation for oleic acid percentage? The second highest? The lowest? *Calculate each distribution's coefficient of variation (COV) to answer this question!* Note: the *actual means* and standard deviations are...

Region	Mean	Standard Deviation	COV
Northern Italy	77.9	1.6	0.021
Sardenia	72.7	1.4	0.019
Southern Italy	71.0	3.5	0.05

5. Determine which distribution *most accurately reflects the TRUE POPULATION distribution of its regional olives.* To answer this question, you will need to calculate the *standard error*, which tells you how close the sample mean is to the population mean. Note: there are 151 olives from Northern Italy, 98 olives from Sardinia, and 323 olives from Southern Italy.

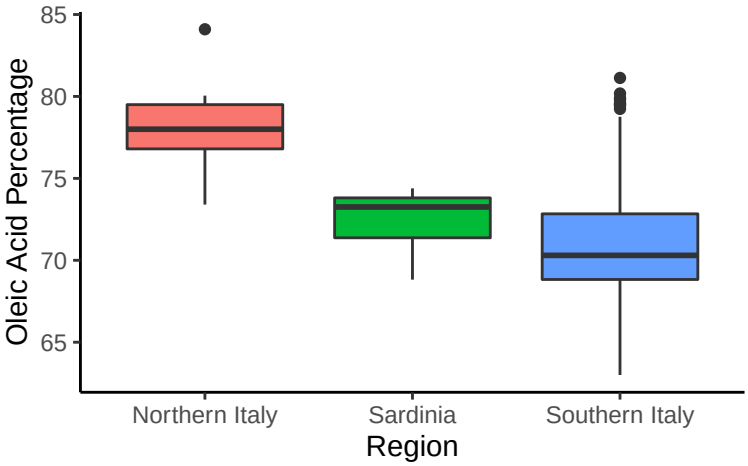
Region	SE
Northern Italy	0.13
Sardenia	72.7
Southern Italy	0.19

0.13 is the lowest, so the Northern Italy sample likely best approximates its true population mean.



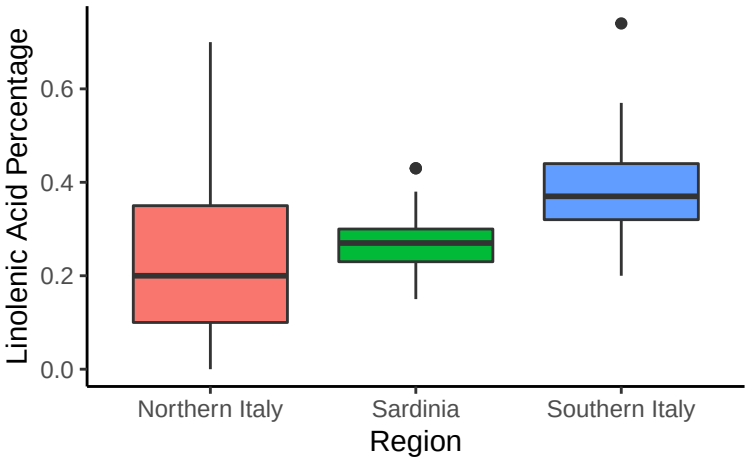
Boxplots

The boxplots below again shown the **distribution of oleic acid percentages** for olives collected from the three regions. Answer questions 1, 2, and 3 (although consider the *median* rather than mean!) from the Histogram section again, but using the boxplots are your guide. Were boxplots or histograms more informative for answering these questions?



Median relationships the same as the mean. Much easier to see outliers (indeed Southern Italy DOES have a couple outliers above 80!), but impossible to tell if unimodal/bimodal, and very difficult to assess symmetry compared to histograms.

The boxplots below shown the **distribution of linolenic acid percentages** for olives collected from the three regions. Answer the same questions (mean, standard deviation). Then, compare the trends you observed for the previous boxplots of oleic acid to these linolenic acid boxplots. What are the similarities and differences between these two types of acids in olives?

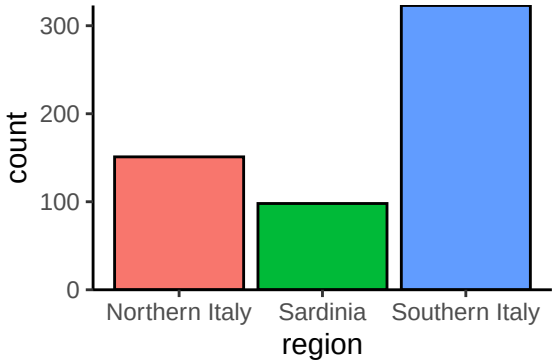


Relationship is opposite to oleic acid - Southern has highest median, Northeran has lowest median. It is probably fair to say Sardinia has least amount of variation, Nothern Italy the largest (although we'd have to confirm this with COV).

Finally, draw by hand (with your table!) a *bar plot* showing the mean oleic acid percentage for each region (refer to Histogram question #4 for the actual values). Include error bars for *standard deviation* (draw roughly the entire error bar length to equal the standard deviation).

Barplots

The barplot below shows the *number of olives* collected from each region. Roughly determine how many olives were collected from each region. What is the *mode* of the distribution of olive counts?

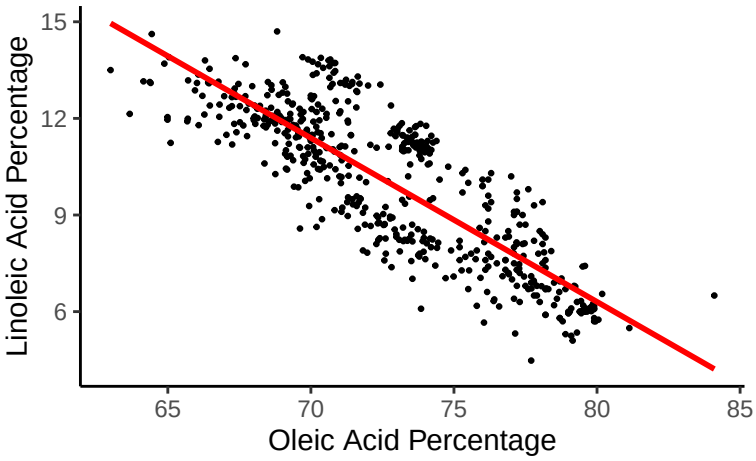


Mode is Southern Italy. Roughly, 150 from Northern Italy, 100 from Sardinia, 310 from Southern Italy.

Scatterplots, Part 1

The scatterplot below shows the *relationship* between percentages of oleic aand linoleic acid in each olive. Thus, there are 572 points in this figure (1 point per olive).

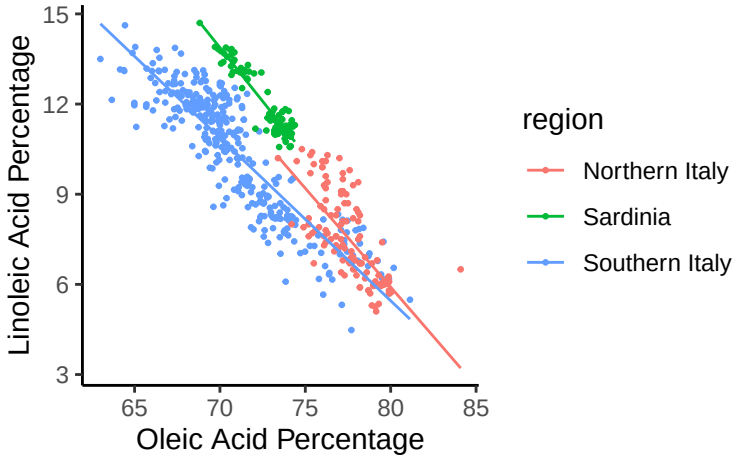
Is there a strong, moderate or weak correlation? Is the correlation positive or negative? Draw a rough estimate of the line-of-best-fit for these points.



Moderate-to-strong negative relationship

The scatterplot below is a modified version of the above scatterplot, where this time points are colored according to which region the olive comes from.

What differences and similarities do you observe among regions? Draw a rough estimate of the line-of-best-fit for each region SEPARATELY and compare their trends: Which region appears to have the highest correlation? The lowest? Are all correlations the same direction (positive or negative)?



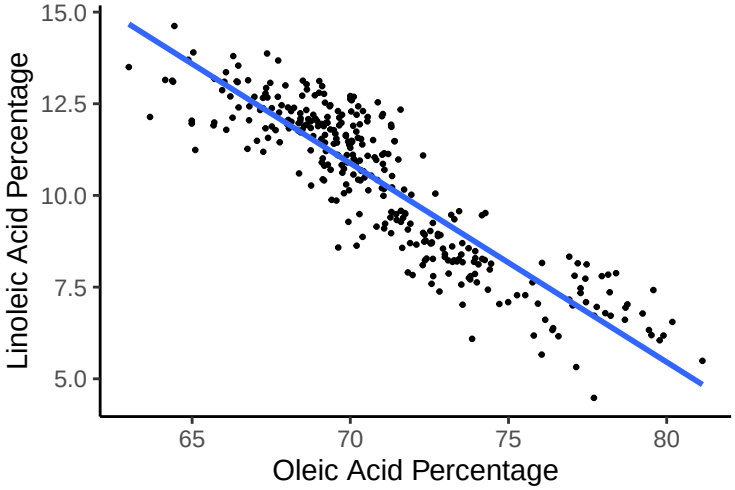
All are negative. Sardinia has the strongest relationship, and Northern Italy likely has the weakest.

Scatterplots, part 2

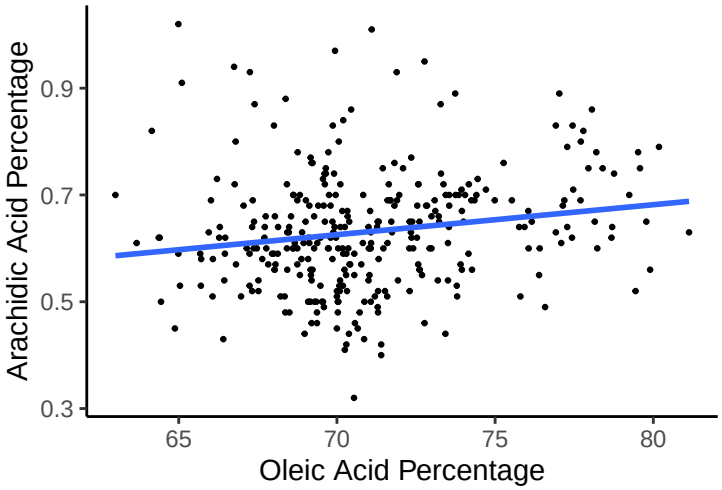
Below are four different scatterplots each with a *line of best fit* showing the overall relationship between X and Y variables. Each plot shows the relationship between two different types of fatty acids, for Southern Italy olives only.

Characterize each plot as having a strong, moderate or weak correlation, as well as whether the correlation is positive or negative.

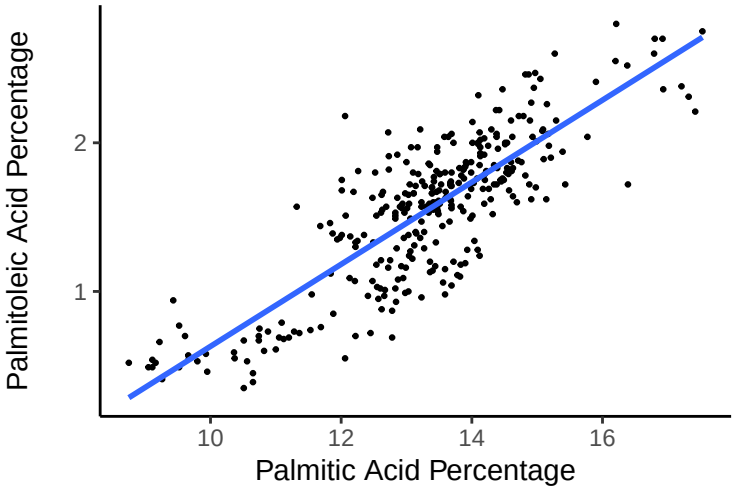
STRONG NEGATIVE



WEAK POSITIVE



MODERATE/STRONG POSITIVE



WEAK/MODERATE POSITIVE