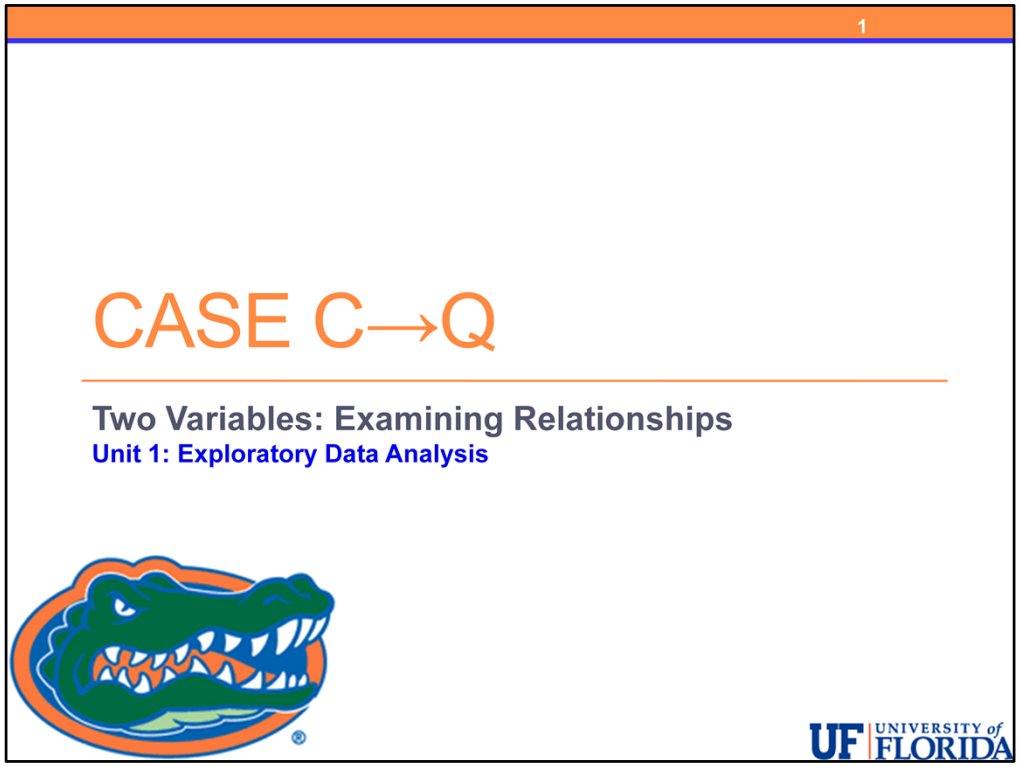


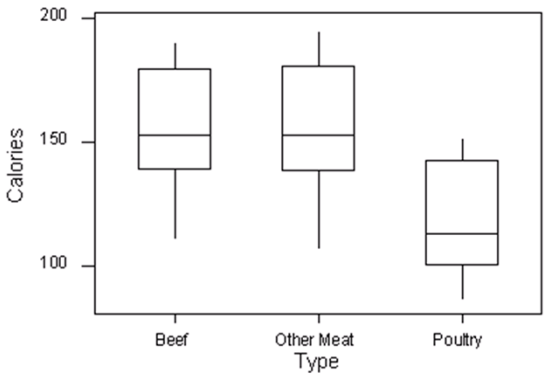
Case C-Q. Again, this works for Case Q-C also - we can conduct the same analyses for Case Q-C, but we will have difficulty addressing our actual research question.

In either case, or in a generic case with one quantitative and one categorical variable where the role of the variable as explanatory or response is not clear, the methods we present here allow us to investigate the relationship or association between these two variables.

Soon we will see that statistical analysis alone can never determine the causal nature of the relationships we pose in our research questions.

Case C-Q

Compare distributions of quantitative response for each category of explanatory variable using side-by-side boxplots supplemented by descriptive statistics



Statistic	Beef	Meat	Poultry
min	111	107	86
Q1	139.5	138.5	100.5
Median	152.5	153	113
Q3	179.75	180.5	142.5
Max	190	195	152

UF UNIVERSITY of FLORIDA

Here, in Case C-Q, we want to compare the distributions of a quantitative response for each category of the explanatory variable.

Therefore we are doing the same thing that we learned for one quantitative variable (visual displays and numerical measures) but we're doing this one time for each group.

We're going to get a boxplot for each group, displayed side-by-side, and numerical measures for each group.

Then we're going to talk about what we see.

And what do we see here?

There don’t seem to be any outliers. The variation is similar between the three distributions.

The pattern or shape of the three distributions, as displayed in the boxplots, is also very similar. The median splits each box on the low end so that the middle 50% shows evidence of being skewed right, however, the lower tails tend to be longer than the upper tails which would indicate the opposite – skewed left. Therefore, these distributions don’t have a easily defined shape based upon the boxplots alone.

There's not much difference between beef and other but poultry tends to have a lower

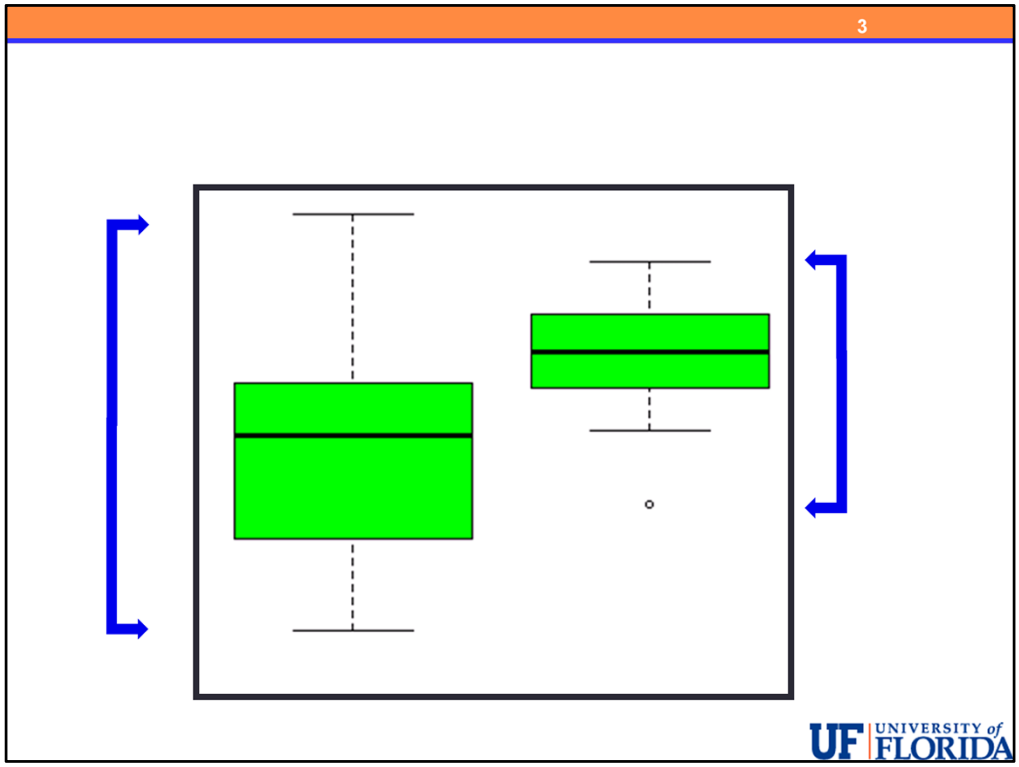
calorie count than the other two. We might want to estimate the difference. Here, based upon what we have in this table, we might say: based upon the medians, in our sample, poultry hot dogs have approximately 40 less calories than beef or other.

However, observe that we could clearly find Beef or Other hot dogs which have a lower calorie count than certain poultry hot dogs. There is overlap between top half of the distribution for poultry hot dogs and the bottom halves of the distributions for Beef and Other.

In summary, in this case, Case C-Q (and possibly Q-C) anything I could ask you to do for one quantitative variable - just calories not split up - I can ask here but three times.

We talk about the components addressed for one quantitative variable - shape, center, spread, outliers and provide appropriate visual displays and numerical measures to go with that. Then compare and contrast the distributions for the groups under study.

Let's review a few other graphs that we have seen so far.

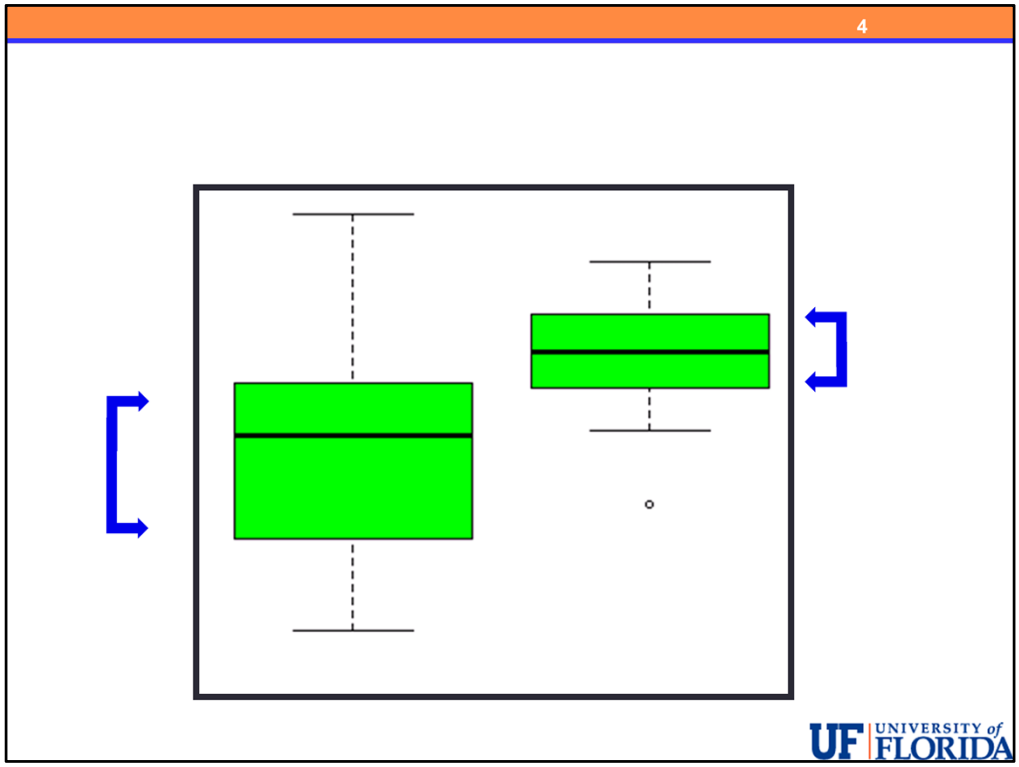


Here we have two unlabeled boxplots with an illustration of the range.

For these two distributions, both the center and spread are different.

The distribution on the right has a higher center (the center of the right distribution is above the third quartile of the left distribution). The entire distribution on the right, except for the single low outlier, lies in the upper 50% of the left distribution. So clearly, on average, values in the right distribution are larger.

The spread of the distribution on the right is much smaller than that of the distribution on the left. In this display we have the range illustrated.



And here we have the IQR.

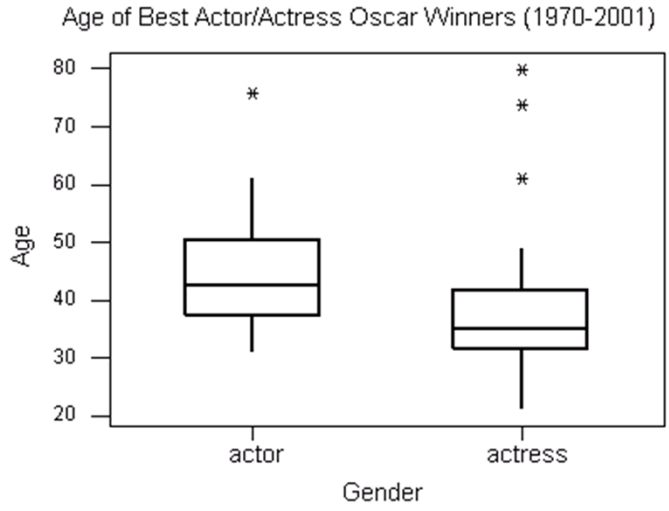
The standard deviations would also agree.

We mentioned the one outlier in the distribution on the right.

What about shape? The right distribution is very close to symmetric with an outlier.

The left distribution gives conflicting information in that the upper tail is longer than the lower tail but the upper half of the box is shorter than the lower half of the box. This distribution is not clearly skewed in either direction. Possibly if we looked at the histogram we would feel comfortable in calling it “reasonably symmetric” but based solely on this boxplot it is difficult to tell.

Side By Side Boxplots



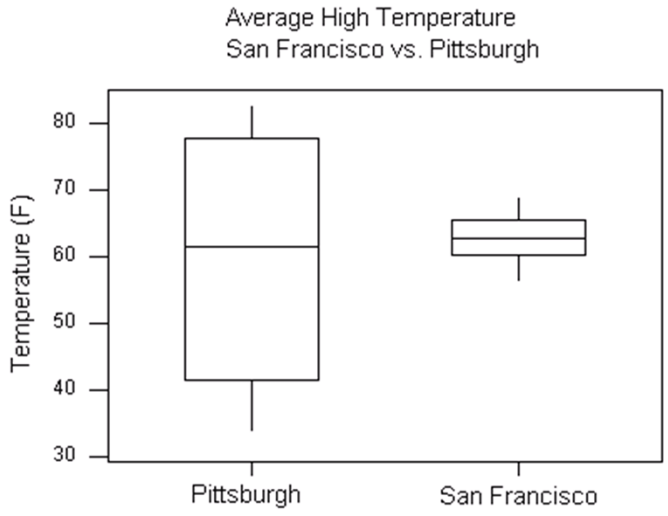
UF UNIVERSITY of FLORIDA

Here, again, is the comparison of the distributions of ages of best actor and actress Oscar winners.

The distribution of ages is higher for actors than actresses. Males had one high outlier and females three high outliers. The shapes might be slightly skewed right but this is difficult to verify from the boxplots alone.

The variation is somewhat similar. Females may have a slightly less variable distribution, ignoring outliers.

Side By Side Boxplots



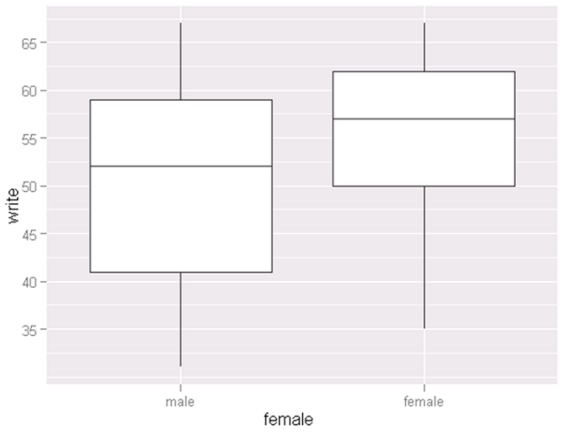
UF UNIVERSITY of FLORIDA

As we previously discussed, the primary difference here is in the variation. In both cities, the median is in the low 60's however, in Pittsburgh, the temperature varies from 35 to 80 versus in San Francisco, it only varies from the 50-70.

Both distributions are reasonably symmetric.

It isn't clear how these "average high temperatures" were calculated but … we won't delve any deeper into this scenario.

Side By Side Boxplots

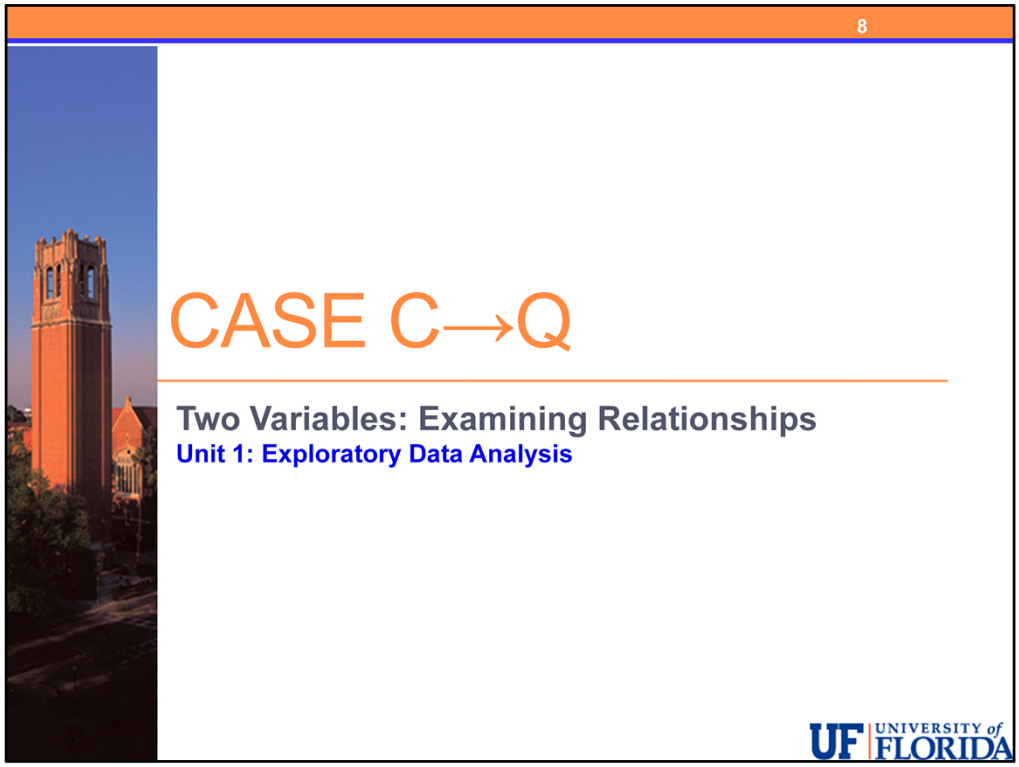


UF UNIVERSITY of FLORIDA

In this plot comparing the writing scores of males and females, Males seem slightly more variable than females. Both distributions are somewhat skewed left, the female distribution seems more clearly skewed left than that for males.

The median of the distribution for females is approximately 5 points higher than that for males. The min, Q1, and Q3 are also higher for females than males. The maximum is about the same for both distributions. Notice that 75% of females scored higher than 50 which is only a few points lower than the median for males.

Therefore, on average, females tended to have slightly higher writing scores than males.



In Case C-Q (and often Case Q-C), we visually display the data using side-by-side boxplots.

Although we might not always discuss the shape of our data in our final presentation of results, minimally it should be observed at the exploratory data analysis stage. If the boxplots are not enough, you can obtain histograms for each group to supplement the side-by-side boxplots.

For numerical measures we should provide appropriate measures of center and spread.

Although we might not always provide all values in our final presentation of results, commonly at this stage we would determine for each group:

- Sample size
- Sample mean
- Sample median
- Sample standard deviation.
- Five-number summary
- Find, determine, and/or note any outliers

Later we will add values such as the estimated standard error of the mean and/or the confidence interval for the mean and you may see that the software we are using already provides these values!

That concludes our conceptual discussion of exploratory data analysis when we have one

categorical variable and one quantitative variable and we wish to examine the relationship or association between them.