Summarizing Data: Relationships Between Variables

Relationships Between Quantitative and Categorical Variables

In previous lectures we have seen various kinds of data. The quantitative variables include the life expectancy of different countries and the age difference of skeletons. The categorical variables include regions of the world and sex of a skeleton.

Relationships in some sense are at the heart of statistics and we are going to study them a lot. In this section, we are going to get a first look at how to consider the relationships between these different variables.

Example 1
Consider the 197 different countries and territories around the world and their life expectancies. Each of these countries and territories fit into some region around the world, and we divided the world into 6 different regions. We are interested in the relationship between these countries' life expectancies and the region of the world that they are in.

We will begin by comparing East Asia & Pacific with Sub-Saharan Africa using the summary statistics below:

East Asia & Pacific	Sub-Saharan Africa
30 countries and territories	49 countries and territories
median = 72.95	median = 55.44
mean = 73.09	mean = 56.80
min = 62.48	max = 77.65
first quartile = 68.77	third quartile = 59.40

In the East Asia & Pacific region, there are 30 countries and their median and mean life expectancy are both approximately 73 years. In Sub-Saharan Africa there are 49 different countries and territories. Their median and mean are both pretty close to 56 years. We can immediately say that the center of the data is quite a bit higher in East Asia & Pacific as compared to Sub-Saharan Africa. But that's not necessarily the whole story.

If we look at the minimum life expectancy in East Asia & Pacific, it is only equal to 62.48 years. Whereas, if we look at the maximum life expectancy in Sub-Saharan Africa, it is equal to 77.65 years, which is a lot more. So it means that the highest life expectancies in Sub-Saharan Africa are certainly a lot higher than the lowest life expectancies in East Asia & Pacific.

Looking a little further, the first quartile of the life expectancies in East Asia & Pacific is just under 69 years. If we look at the third quartile of the life expectancies in Sub-Saharan Africa it is a little over 59 years. In other words, the first quartile in East Asia & Pacific is

quite a bit higher than the third quartile in Sub-Saharan Africa. This provides some pretty clear evidence that most of the life expectancies in East Asia & Pacific are indeed larger than most of the life expectancies in the Sub-Saharan Africa.

We can see this more clearly if we make boxplots for each of the two regions and put them side by side. In Figure 1 we see a box plot for the life expectancies of the countries and territories in East Asia & Pacific and right next to it a box plot for the life expectancies for the countries in Africa. Looking at these two box plots we can say that most of the box plot for East Asia & Pacific is quite a bit higher than most of the box plot for Sub-Saharan Africa. It seems for the most part life expectancies in East Asia & Pacific are indeed higher than life expectancies in Sub-Saharan Africa.

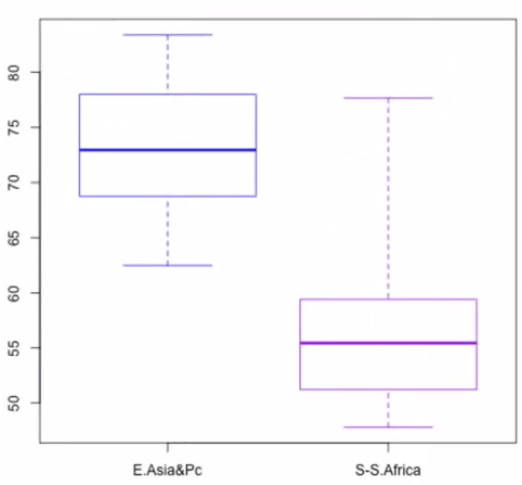

Figure 1: Side-by-side box plots of life expectancies for East Asia & Pacific and Sub-Saharan Africa

Example 2

We can consider the relationship between a quantitative variable and a categorical variable using the skeleton data. Remember that we have a data set of 400 skeletons and for each one we have the difference in the estimated age of death as compared to the actual age of death which is a quantitative variable. We also have the mass category, a categorical variable with the following possible values: underweight, normal, overweight, and obese. Now we can ask: "Is there a relationship between these two variables?" or "Does the mass category of the skeleton have an effect on this difference?"

Let's go straight to considering side-by-side boxplots of the difference in ages for all four classifications, shown below in Figure 2.

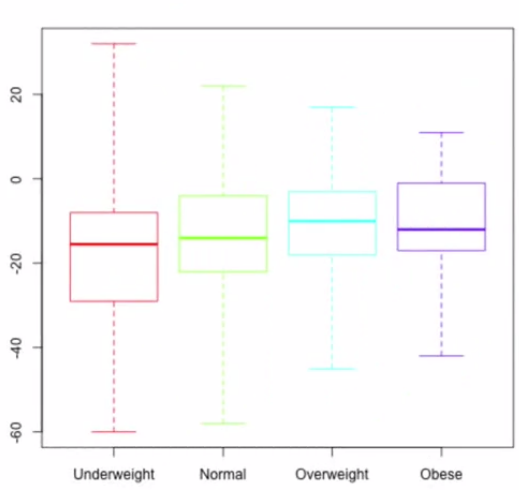

Figure 2: Side-by-side box plots of difference in estimated age at death for each mass category

Once again, we can see that they are fairly similar indicating that there is not a strong relationship between the mass category of the skeleton and the difference in the estimated versus actual ages. But there might be a small effect! For example, we can see that the underweight skeletons are slightly more negative. This suggests that maybe the differences are slightly more negative for the underweight skeletons as opposed to the other ones. Is this difference significant? We will have to learn about that in subsequent lectures.