Chapter 5

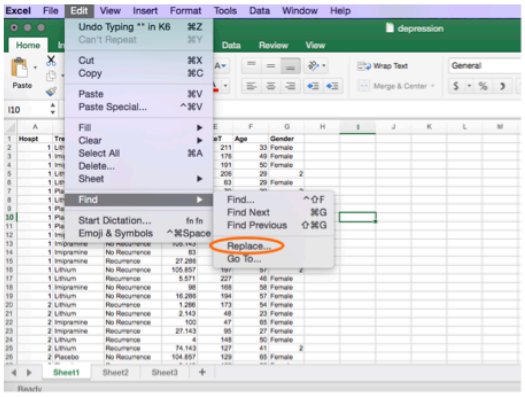
Deciphering between categorical and quantitative variables

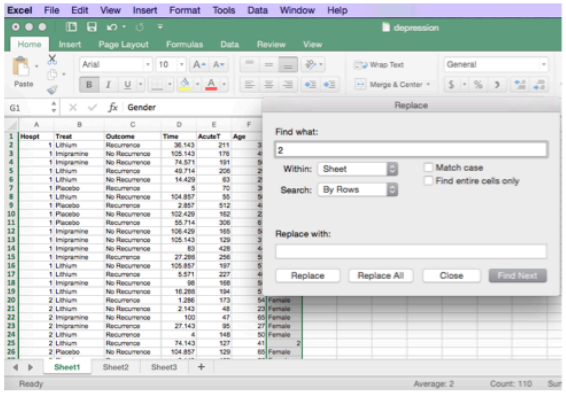
A quantitative variable may be made categorical by defining a range of values. For example, we could be comparing annual income.

- Annual income could be measured in dollars, making it a quantitative variable.
- Alternatively, we could categorize income into three categories: high, medium, and low. We would define the range of dollars earned for each category (bins). For instance, high annual income could be defined as any income over six digits. In this case, income is categorical.

Find and replace

In OLI they tell you to use the find and replace Excel function. They say to pull it up from the Editing group in Home tab. I was able to do this by going to the Edit drop-down and selecting Find and then over to replace. Also, control F pulled this menu up, while command F only searched through the document.



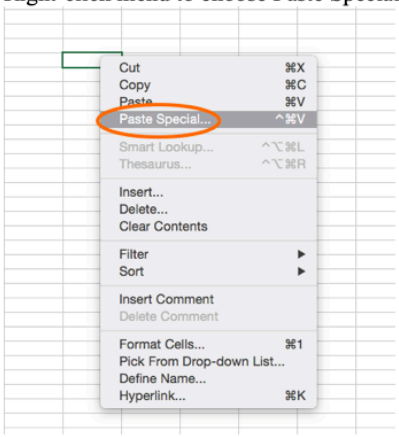


Sort

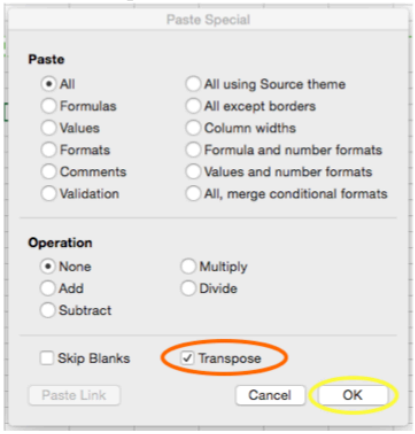
Excel is set to sort columns, not rows. So, if you try to sort a row it probably won't work. You could change the settings. Its faster to paste the data as a column and then sort.

Turning a row into a column is called transposing. Also, turning a column into a row is also called transposing.
When you copy a row of data from OLI use paste special to paste it as a column in Excel. After copying the data paste special is found by right clicking in Excel. Also, the Excel sort-cut "command+control+v" brings up paste special.

Right-click menu to choose Paste Special

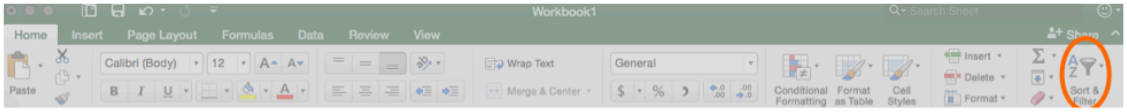


Check Transpose and then hit OK



For me, for whatever reason, copying the row in OLI and then trying to paste special into Excel did not work. I had to copy the row from OLI, regular paste the row into Excel as a row, and then copy that row in Excel and use paste special to transpose.

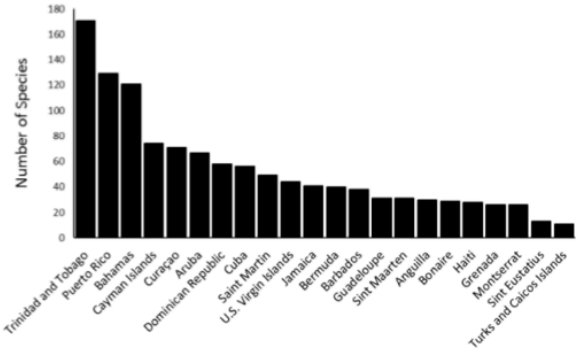
Now that we have a column of data we can sort the data. Select the column and choose the Select and Sort icon.



Histograms

To make a histogram we bin our quantitative variable by defining a range for each bin. This is what was described above when we made a quantitative variable into a categorical one. In this case, our variable is still a quantitative variable, but we analyze it by categorizing it into bins. The point here is that you can't make a histogram out of a quantitative variable directly. First, you have to separate a quatitative variable into bins and then you can make a histogram. With a categorical variable you don't have to separate the variable into bins to make a histogram.

The following histogram shows the number of bird species found in different countries in the Carribean in 2016. In this histogram the x-variable is country and this is a categorical variable. Our y-variable is the number of bird species.



Data from: Global Big Day. *The Cornell Lab of Ornithology*. 2016.
Image found at: https://www.birdscaribbean.org/category/news/caribbean-birdwatch/

Image credit to: www.linda-jackson.co.uk Trinidad & Tobago Tourism Development Company Ltd, Antigua & Barbuda Tourist Office, and featured books Publishers/Authors

We could then be interested in looking at the size of the habitat of some bird species. Although we do not have this data, lets pretend that we do. Lets say we are interested in the habitat size of the Blue Honeycreeper (seen above), the Golden Pheasant (below, left) and the Kakapo (below, right). FYI. The Blue Honeycreeper is both found in Trinidad and Tobago (in the Carribean), while the Golden Pheasant is found in china and the Kakapo is found in New Zeland. The range of habitat for the Golden Pheasant is shown.

Image credit to: Utah's Hogle Zoo. 2018.
https://www.hoglezoo.org/meet_our_animals/animal_finder/golden_pheasant/

Image credit to: Mike Bodie. 2009.
https://www.flickr.com/photos/docnz/4015891720/

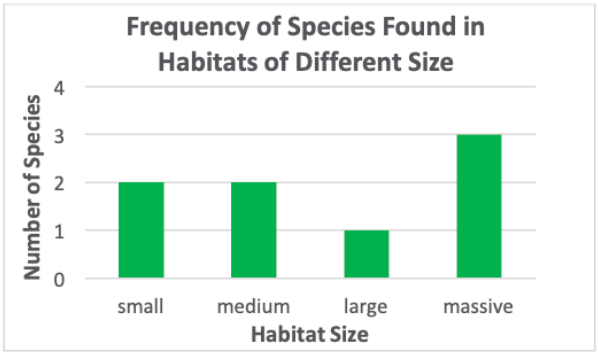
Lets say that we determine that the habitats of these and some other bird species.
The following data is confabulated! Although the habitats of some of these birds are quite small, they are not just a few acres.

Species	Habitat Size
Blue Honeycreeper	8 acres
Golden Pheasant	5 acres
Kakapo	2 acres
Nightingale	10 acres
Canary	20 acres
Seagull	25 acres
Sparrow	30 acres
Pigeon	40 acres

This is quantitative data. We could then separate it into bins. We could specify that a small habitat is between 0 and 6 acres, a medium habitat is between 6 and 10 acres, a large habitat is between 11 and 20 acres, and a massive habitat is anything over 20 acres. Then, the number of bird species that falls into these categories is as follows.

Habitat Size	Number of Bird Species
Small	2 bird species (Golden Pheasant and Kakapo)
Medium	2 bird species (Blue Honeycreeper and Nightingale)
Large	1 bird species (Canary)
Massive	3 bird species (Seagull, Sparrow, Pigeon)

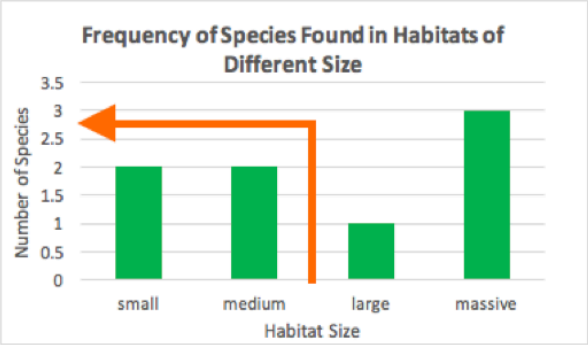
The histogram that corresponds to this data is as follows:



The total number of bird species is 8. In statistics, the total number is often denoted by n or N.

If you want to know the percent of bird species found in habitats that are medium or smaller, you would look at the small and medium bars in the histogram. You then determine the number of bird species that live in medium and in small habitats. In our case, the small habitat includes 2 birds (Golden Pheasant and Kakapo) and the medium habitat includes 2 birds (Blue Honeycreeper and Nightingale). That makes a total of 4 birds that live in small and medium habitats.

Looking at the histogram,



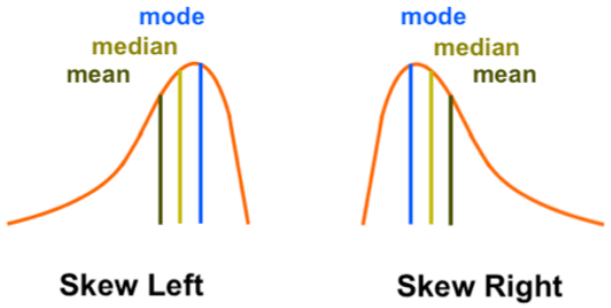
Birds found in medium or smaller habitats includes birds found in medium habitats + birds found in small habitats.

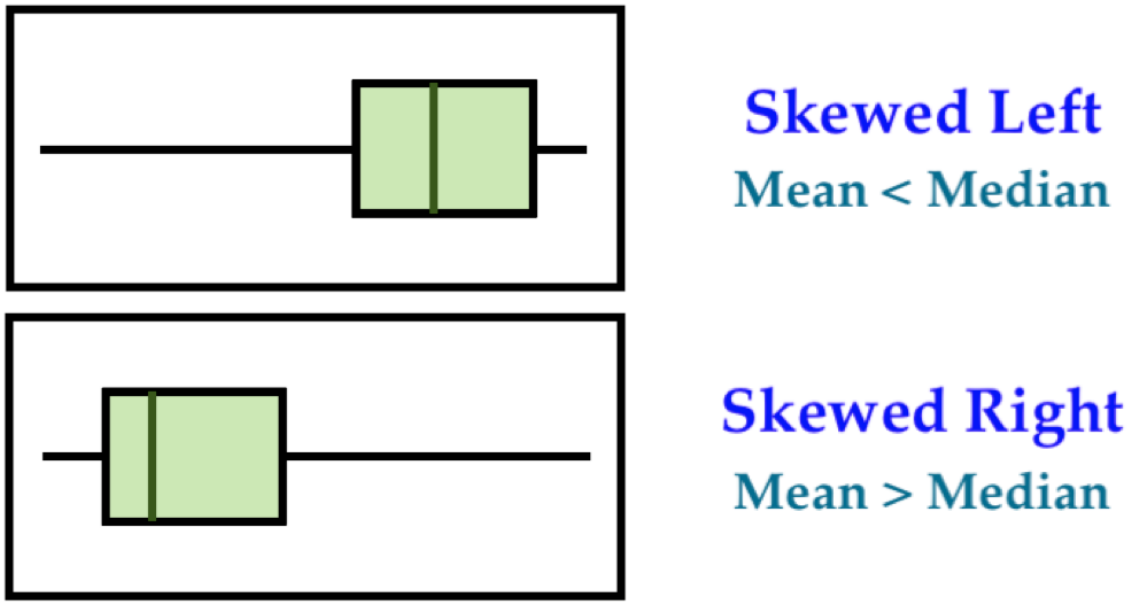
Then, the percent of birds found in habitats that are medium or smaller=

$$\frac{Birds\ in\ medium\ habitats + Birds\ in\ small\ habitats}{Total\ number\ of\ birds} * 100 = \frac{4}{8} * 100 = .5 * 100 = 50\%$$

Skewness

As a note about skewness: to me it seems counterintuitive that the main area of the distribution is on the right in left skewed distributions, while it is on the left in right skewed distributions. Right skewed distributions are called so because the mean of right skewed distributions is on the right of the median and mode. Also, the tail of right skewed distributions is on the right.





When a distribution is skewed left, the mean is less than the median, while when a distribution is skewed right, the mean is greater than the median.

This Lecture was prepared by Elizaveta Latash