1/22/2023

Summarizing data in tables
Statistical graphs

Summarizing data in table: we are typically talking about categorical variables.
Frequency tables: one column lists the values of the variables, and the second column lists the count (frequency)

In Excel, these are called pivot tables.

Instead of just frequency (count), we can also make the second column a relative frequency (percent of the total). This produces a column with proportions rather than counts. Can be expressed any way that proportions can be expressed: fractions, percentages, decimals.

Cumulative frequency table: each count in the table is the sum of the counts of the current category and any category that preceded it in the table. When you get to the bottom of the table, the total and the count for the last category are the same.

Relative Cumulative frequency table: take the counts in the cumulative frequency table and convert those to percentages (proportions). The last line of the table would be 100%.

If we have numerical data that we want to put in a table, we have to “bin” the data. We have to force it into categories. These are harder to do in Excel. We have to do the binning mostly manually in Excel.

Making Statistical Graphs
Graphs for categorical variables are made from the summary tables (frequency tables)

Pie chart – display the relative frequencies of the data
Bar chart – can display counts or relative frequencies

Pie charts require that the percentages add up to 100%. (Never include the total in the graph.) And you can only plot one variable.

Rule of thumb for pie charts: generally stay at 7 or fewer categories. Don’t make pie charts from variables with LOTS of values.

What makes for a good graph?

Graphs need to stand by themselves. Can’t depend on people being able to access the original data. Legend if you have lots of data. Axis labels/titles on the bar graph. Pie charts should have values on each slice. Descriptive title.

Bad graphs: are missing some of this data. Avoid 3D affects. 3D can make the graphs more difficult to read, and perspective can distort relative sizes. Don’t be too busy.

A note about Pareto charts:

The categories are ordered on the graph by the size of the bars (largest to smallest). (Other programs may order from smallest to largest.) In Excel, they also add a relative cumulative frequency line (the percentages are marked in the right side axis).

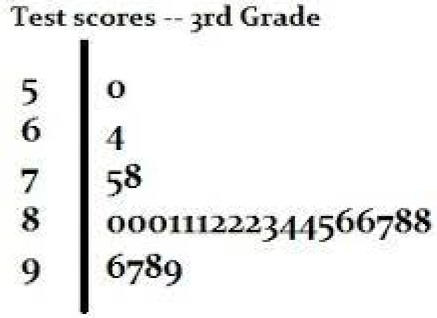
Don't use this for variables that have an inherent order.

If your data is ordered (like our income categories, or time), then you can make a line graph, with the ordered variable on the horizontal axis.

Line graphs are our transition to graphs for numerical data.

For other kinds of numerical data:

Stemplot –



Test scores -- 3rd Grade

5	0
6	4
7	58
8	000111222344566788
9	6789

Stems on the left side of the bar, and leaves (for individual observations) are on the right.

Histograms are essentially bar graphs for numerical data: the data is binned (put into categories) and then graphed using frequencies or relative frequencies.

(histograms don't have spaces between the bars).

The number of bins depends on how much data you have: 5 bins to 20 bins
Too few bins makes the graph look blocky and doesn't reveal any detail. Too many bins can pancake the data or leave gaps. All bins should have the same width (be the same size).

Boxplot:
Is a simple graph that uses 5 statistics from the data to plot: median, minimum and maximum, and the quartiles (1/4 and ¾ points of the data).

Data can be symmetric – roughly (think bell curves)
Can be skewed – long tail on one side. The tail is actually determines the name of the skewness. A long tail on the right: right-skewed. And a long tail on the left, is called left-skewed.

Comparing two variables:
2 categorical variables: make a two-way table to summarize the data, use a bar graph (clustered). Not a pie chart.

One categorical and one numerical: comparative box plot.

If we have two numerical variables:
If one is ordered: then we can use a line graph (one observation per value)
If not ordered in time, then we can use a scatterplot.

We will encounter these types in later chapters, so we will build them at that time.

Next week: descriptive statistics of numerical data.