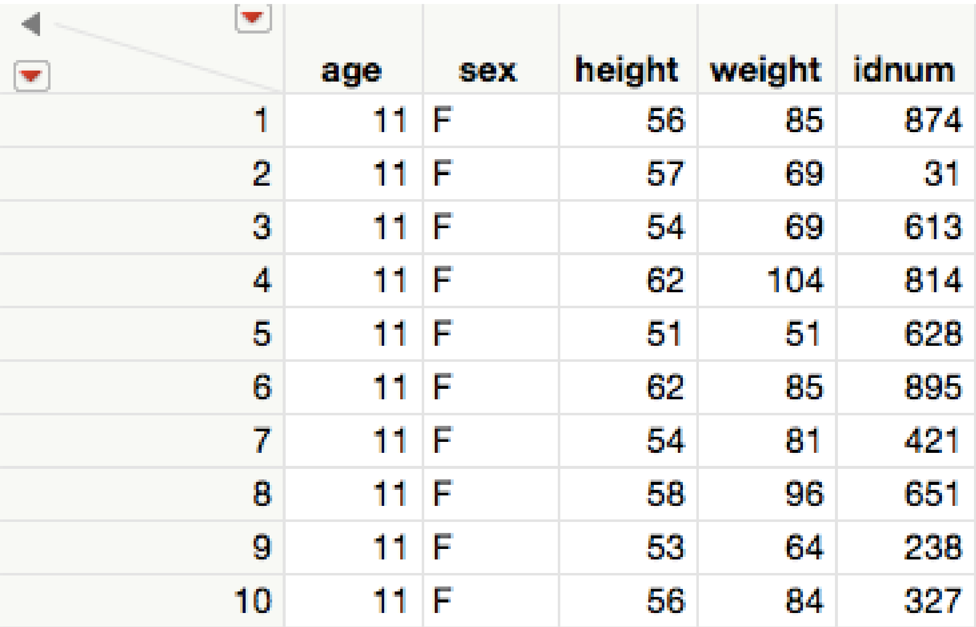
Basic Biostatistics: What Your Medical School SHOULD Have Taught You But (Probably) Didn’t

This is a brief overview of statistical topics that are important for physicians to understand. Every doctor (this includes **you**!) should have a general idea of these things, if only to arm yourself with the ability to read journal articles in your specialty. If you don’t keep up with journal articles, at least peripherally, then your practice will become hopelessly out-of-date in just a few years.

Basic Vocabulary

First, let’s establish a basic vocabulary. A lot of epidemiology and statistics seems harder than it really is because the terms are unfamiliar, so let’s settle this issue right away.

Data are ideally organized into **rows** and **columns**, like in a spreadsheet. The rows are normally called **observations** and the columns are **variables**. This gives you a structure like this:

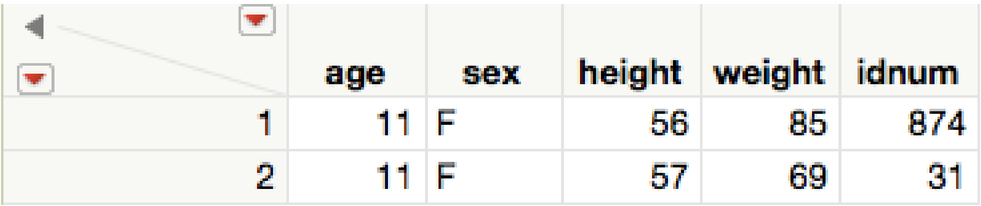


	age	sex	height	weight	idnum
1	11	F	56	85	874
2	11	F	57	69	31
3	11	F	54	69	613
4	11	F	62	104	814
5	11	F	51	51	628
6	11	F	62	85	895
7	11	F	54	81	421
8	11	F	58	96	651
9	11	F	53	64	238
10	11	F	56	84	327

This organization of data into a spreadsheet format makes doing any analysis easier. When data look like this we call them “tidy”.

Another important distinction is whether variables are **quantitative** or **qualitative.** Quantitative variables are either **continuous** (think of an analog radio tuner your grandparents adjusted by dial) or **discrete** (think of a digital radio in your car). A clinical example of a quantitative/continuous variable is *blood pressure*, while an example of a quantitative/discrete variable is *number of pillows used by a patient with orthopnea.* (A classic example is that of a piano playing discrete tones, while a violin plays continuous ones.)

The distinction between qualitative and quantitative variables is far more important than the distinction between continuous and discrete. Qualitative (or **categorical)** variables consist of names of categories. Whether a patient in your study has diabetes or not is **categorical**, it is a “yes” or “no” thing, right? This holds even if the variable is coded in the spreadsheet as a number (like 0 for “no diabetes” or 1 for “has diabetes”), as is often the case practically—just because it’s coded as a number doesn’t necessarily mean it’s quantitative! Categorical data can be subdivided into **binary** (like diabetes=0 for a patient without diabetes or diabetes=1 for a patient who has it) or **multicategorical** (systolic blood pressure <100 mmHg, SBP 100-120 mmHg, SBP >120 mmHg). In any case, these are **categories** being represented. While blood pressure COULD be expressed quantitatively, if you sort the patients into categories like this—which you could call low, average, and high blood pressure—you are expressing them QUALITATIVELY.



	age	sex	height	weight	idnum
1	11	F	56	85	874
2	11	F	57	69	31

In our example data set (above), **which variables are quantitative and which are qualitatitive?** (Feel free to ignore *idnum***,** which is simply a variable listing an id number for each patient.)

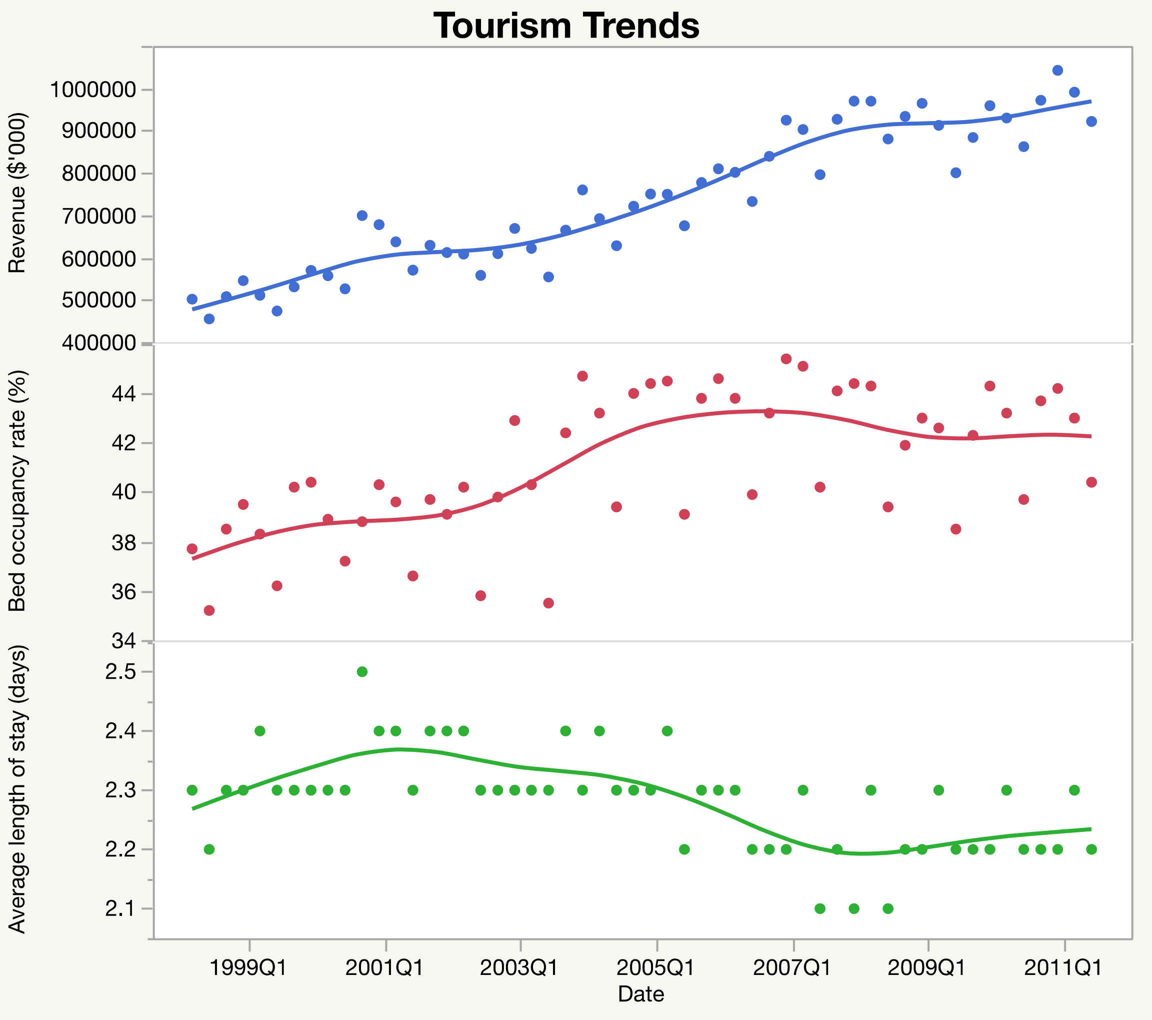
Visualizing Data

Having looked at some basic definitions, the next task should be to consider how to visualize our data. This is topic is HUGELY important to our discussion, since these kinds of visualizations 1) help you really understand what is really going on in an analysis and (as a result) 2) show up in papers all the time.

Scatterplot

Scatterplots are a simple way of comparing variables. These are composed of **quantitative** variables on both the x AND the y axes. In the example plots below, note the **smooth line** showing the relationship between the points on both axes.

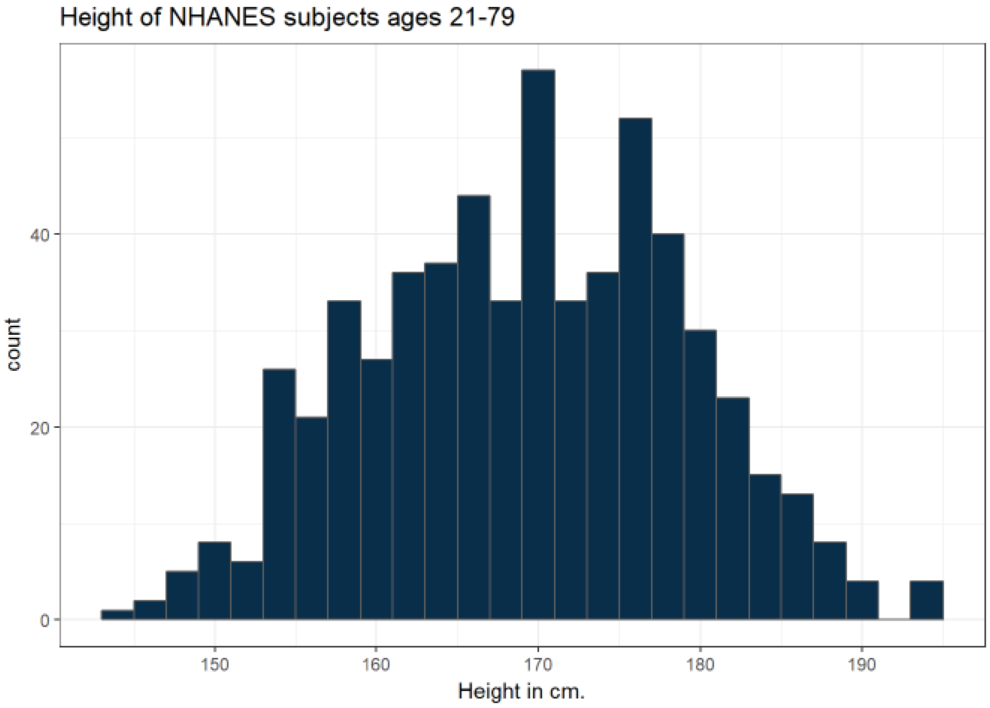
<<SPOILER ALERT: this is really just a simple **regression** line>>



The above graphic illustrates some nice properties that are invariably seen in **good visualizations**. Notice how you can look at this image and instantly understand what is going on . . . no need for a caption or someone to walk you through it out loud. With the **title** and the **clearly labelled** axes, the image is fairly self-explanatory. For more on this topic, check this book out.

Histogram

Another simple type of visualization is the histogram. At the most basic level, histograms describe the **distribution** (or “spread”) of single variables—in the below example, the single variable being addressed is height, and the distribution is showing how many people of each height there are. (Notice how you can figure out exactly what is going on here just by looking at the labelled axes.)

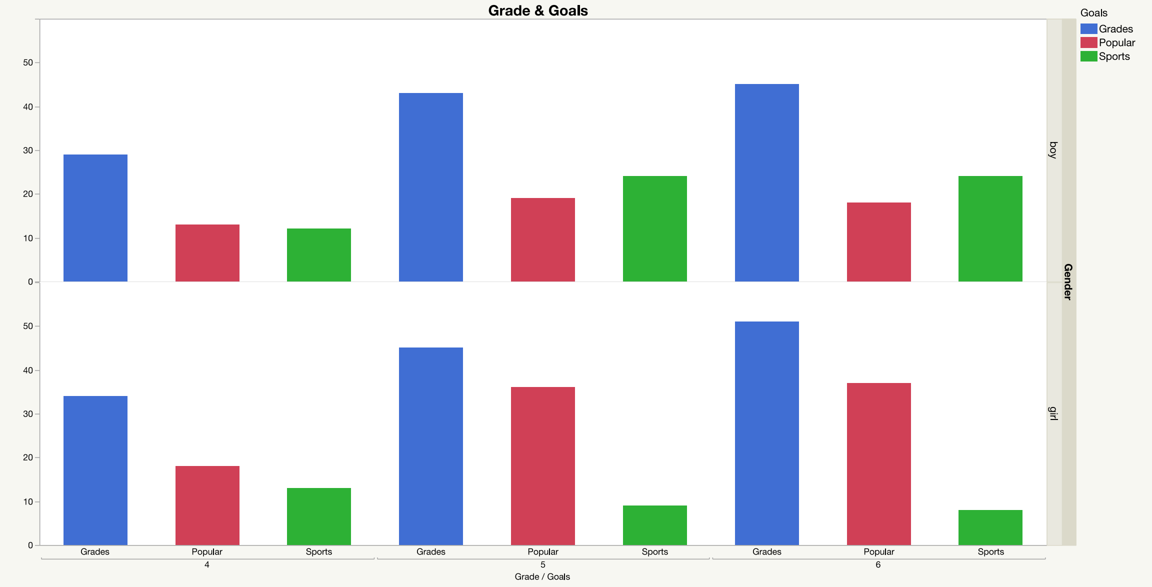


If you really wanted to, couldn’t you figure out how many total patients are represented in this graphic? Wouldn’t you just have to add up all the counts in each bar (each of which represents the number of people at each height)?

Note that this distribution has the unique property of being (relatively) **normal.** **“**Normal” distributions have a sort of **bell-shaped curve**, with most observations being in the middle, and either side appearing fairly symmetric. This concept is good to file away for the time being, since a number of statistical tests we will look at later make an assumption that the data are distributed this way.

Bar Chart

Bar charts are close relatives of the histogram, but are a bit more flexible. Where the histogram always describes a single variable and its distribution, bar charts can describe other things. Look at this example:

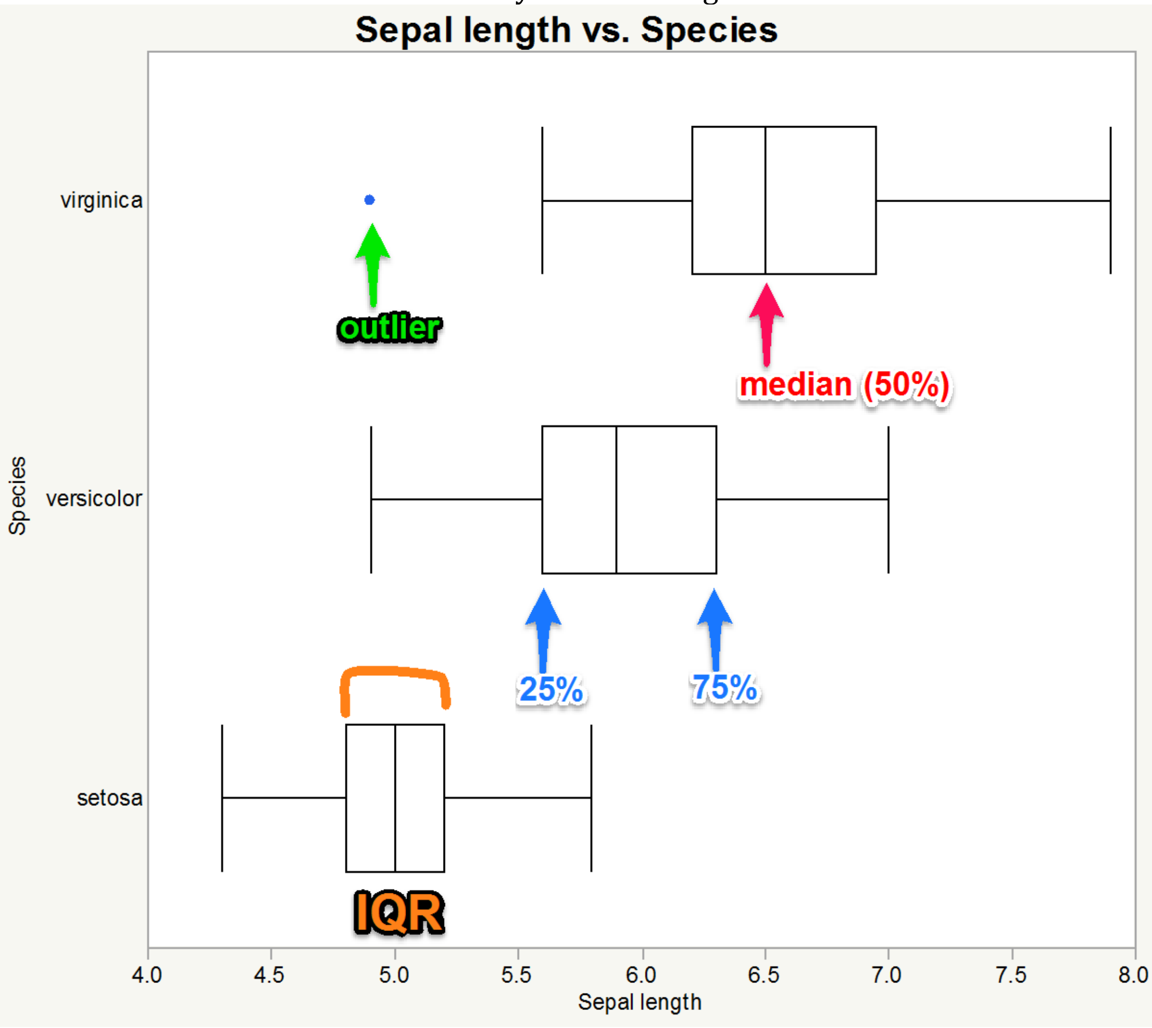


If this bar chart is hard to read on your computer you may have to zoom in. What's going on here? It looks like the bar chart is examining what is important to kids in grades 4-6. Both boys (on the top) and girls (on the bottom) care a lot about grades in each year, but it seems like the girls care more about popularity—especially in grades 5 and 6—and the boys care more about sports. Do you agree?

Note: if you are objecting that this chart doesn't exactly follow the criteria for good visualizations I mentioned above, due to the y axis not being labelled, you are absolutely right!

Box Plot

Finally, we come to the box plot. Box plots are simple graphics that contain a tremendous amount of information if you are willing to learn their conventions.



Note that "IQR" stands for interquartile range—the range of data from the 25th percentile to the 75th. The space between the "whiskers" (the vertical lines at either end of each box plot) contains the data that haven't been identified as "outliers" based on widely-recognized statistical convention.

Which of these iris species contains the tightest and most symmetric distribution of sepal length? Is there a term you can think of for data that looks symmetric like this? (Hint: we mentioned it in the histogram section.) Which species is the most skewed—meaning there are outliers in one direction but not the other? Which has the highest median? Can you see how data represented this way could be useful in summarizing information?

For more information on box plots, click on the picture of John Tukey. →