perceptual edge

Distribution Displays, Conventional and Potential

Stephen Few, Perceptual Edge

Visual Business Intelligence Newsletter
July/August/September 2014

We can graphically display how a set of quantitative values is distributed from lowest to highest in various ways for various purposes. In this article, we'll look at five conventional distribution graphs and examine the strengths, weaknesses, and uses of each. We'll also imagine ways in which these conventional graphs could be enhanced to improve their usefulness.

Conventional Distribution Displays

When we examine distributions, we typically focus on four characteristics:

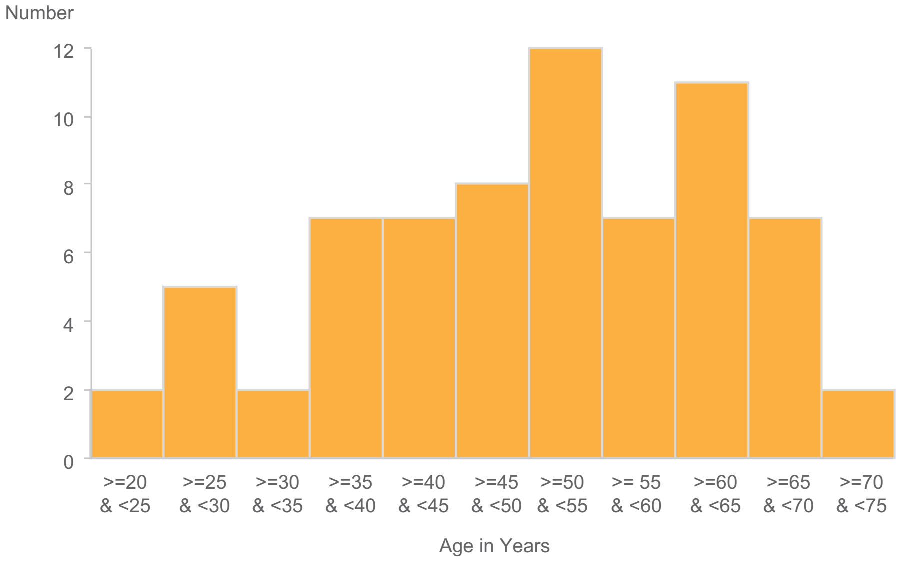
- Central Tendency (a single measure of the distribution's center, usually the median or mean)
- Spread (the quantitative range across which the values are distributed, from lowest to highest)
- Shape (the pattern in which the values are distributed across the quantitative range)
- Outliers (values that are significantly different from the norm)

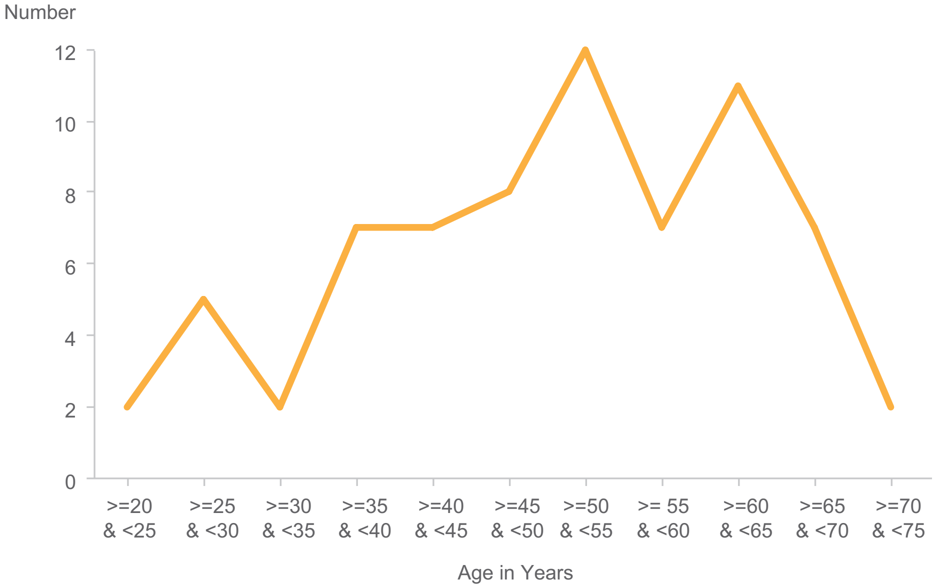
Graphs that we use for examining distributions ought to make most, if not all, of these characteristics visible. In addition to examining individual distributions, we often need to compare distributions to one another, so some of these graphs should support easy comparisons as well.

Different graphs have different strengths. No one graph does everything equally well.

Intervals Along the Quantitative Scale

Two of the most common graphs for displaying distributions are *histograms* and *frequency polygons*. They differ in that a histogram represents a distribution using bars and a frequency polygon does so using a line, but they both do this by displaying the number (or percentage) of values (a.k.a., frequency) that fall into each of a series of equally sized intervals into which the full quantitative range has been divided, from lowest to highest. These intervals are consistent in range while the frequency of values varies. In the following two examples we see the distribution of a group's ages, divided into intervals of five years each.



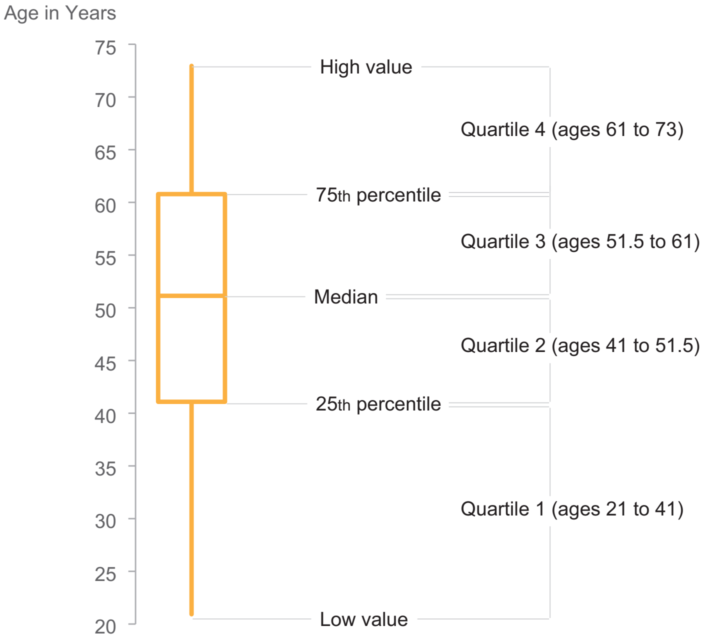


By grouping the values into intervals and displaying the number of values in each, both histograms and frequency polygons provide a simple overview of a distribution's shape. The line of a frequency polygon represents the shape a little more simply and clearly than the bars of a histogram, but the histogram provides a slightly easier way to see differences in frequencies among intervals by comparing the heights of the bars.

Despite their popularity, these two graphs fall short in a few ways:

- They don't show the distribution's central tendency
- They only provide an approximate measure of spread, for they don't give us the lowest and highest values precisely
- They don't identify outliers or provide their precise values

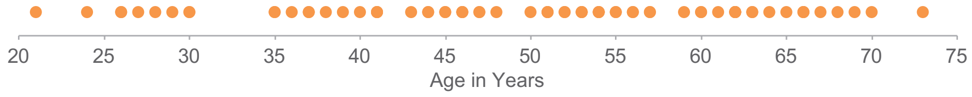
In contrast to histograms and frequency polygons, box plots represent distributions as the ranges across which equal proportions of values are located (e.g., four ranges that each contain 25% of the values, called *quartiles*). In this case the values that fall into each quartile are consistent in proportion and number (25% each) but vary in the quantitative ranges that they span.



When distributions are normal (i.e., bell-shaped), the mean may be used as the measure of central tendency and standard deviation may be used rather than percentiles to display dispersion around the mean. The median, however, can be used in all cases.

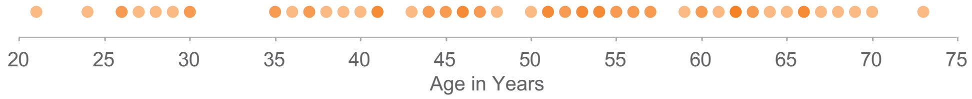
Summary vs. Individual Values

Another distribution graph, called a *strip plot*, does not conventionally group values into intervals at all, but instead displays each value positioned along a single quantitative scale.

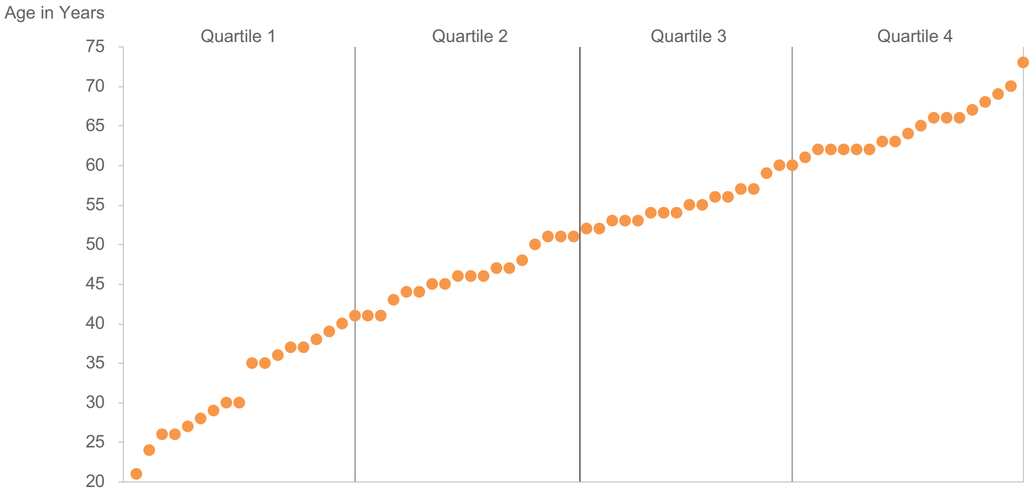


By displaying the individual values, a strip plot makes it possible in this case to see the exact ages of each individual in the data set. For example, we now know that no individuals in the 20-24 age group are 20, 22, or 23 years old. This isn't visible in graphs that group ranges of ages together into intervals, such as histograms and frequency polygons. A potential downside of the strip plot, however, is the fact that it doesn't show the shape of the distribution nearly as well as histograms and frequency polygons. Every form of display has its strengths and weaknesses.

Another potential problem with the strip plot in this example is over-plotting—dots that are positioned on top of one another. When multiple individuals are the same age, the dots that represent them appear in the same space, so we can't tell that there's more than one. This problem can be alleviated fairly well, however, by making the dots transparent. As you can see in the example below, with transparent dots, the more dots there are in one location, the darker their color.



Similar to the strip plot, a *quantile plot* also displays individual values but does so in a way that eliminates the over-plotting problem. In a quantile plot, the quantitative scale resides on the y-axis and the values are arranged horizontally from left to right, beginning with the lowest and continuing in order to the highest, and they're divided into quartiles.



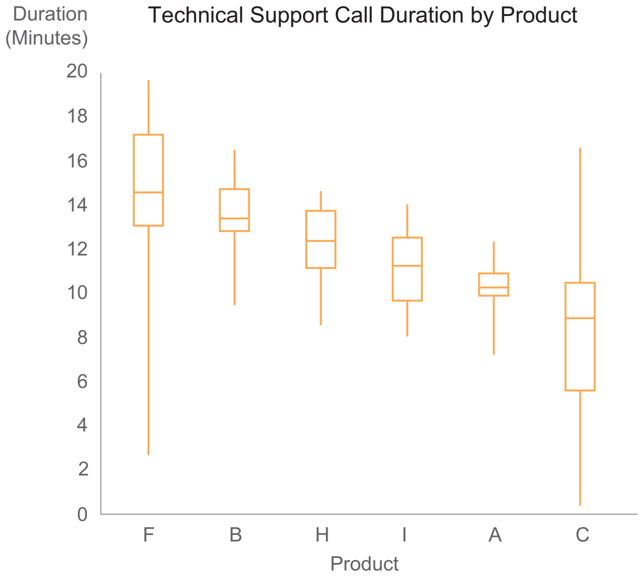
Because each value has its own position along the x-axis, values never overlap. By solving the over-plotting problem in this manner, however, a quantile plot is limited in the number of values that it can contain compared to a strip plot. You can only place so many dots next to one another in the limited space of a page or screen.

When a distribution is divided into intervals that contain equal proportions of values, the intervals are called *quantiles*. Consequently, a *quartile* is a quantile that contains a proportion of 25% (four in total), a *decile* is a quantile that contains a proportion of 10% (10 in total), and a *percentile* is a quantile that contains a proportion of 1% (100 in total).

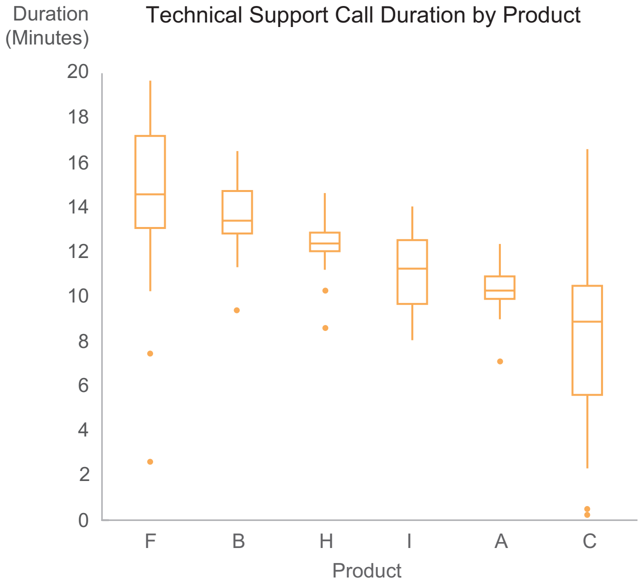
Quantile plots display the shape of the distribution, but not in the familiar and intuitive way that histograms and frequency polygons do. As a result, it takes time learn how to interpret a quantile plot's representation of shape. With practice, though, we can learn to wrap our heads around this less intuitive representation, as many statisticians have learned to do.

Single vs. Multiple Distributions

We use some distribution graphs primarily for examining the distribution of a single set of values (histograms, frequency polygons, strips plots, and quantile plots) and others primarily for comparing multiple distributions (box plots) but this distinction isn't rigid. When the Princeton statistician John Tukey originally created the box plot, he probably used the proportional approach because it provides a simple display that lends itself to easy comparisons among many distributions at once. As you can see in the following example, these simple objects consisting of boxes (i.e., rectangles) and lines provide a great deal of information that can be easily compared.

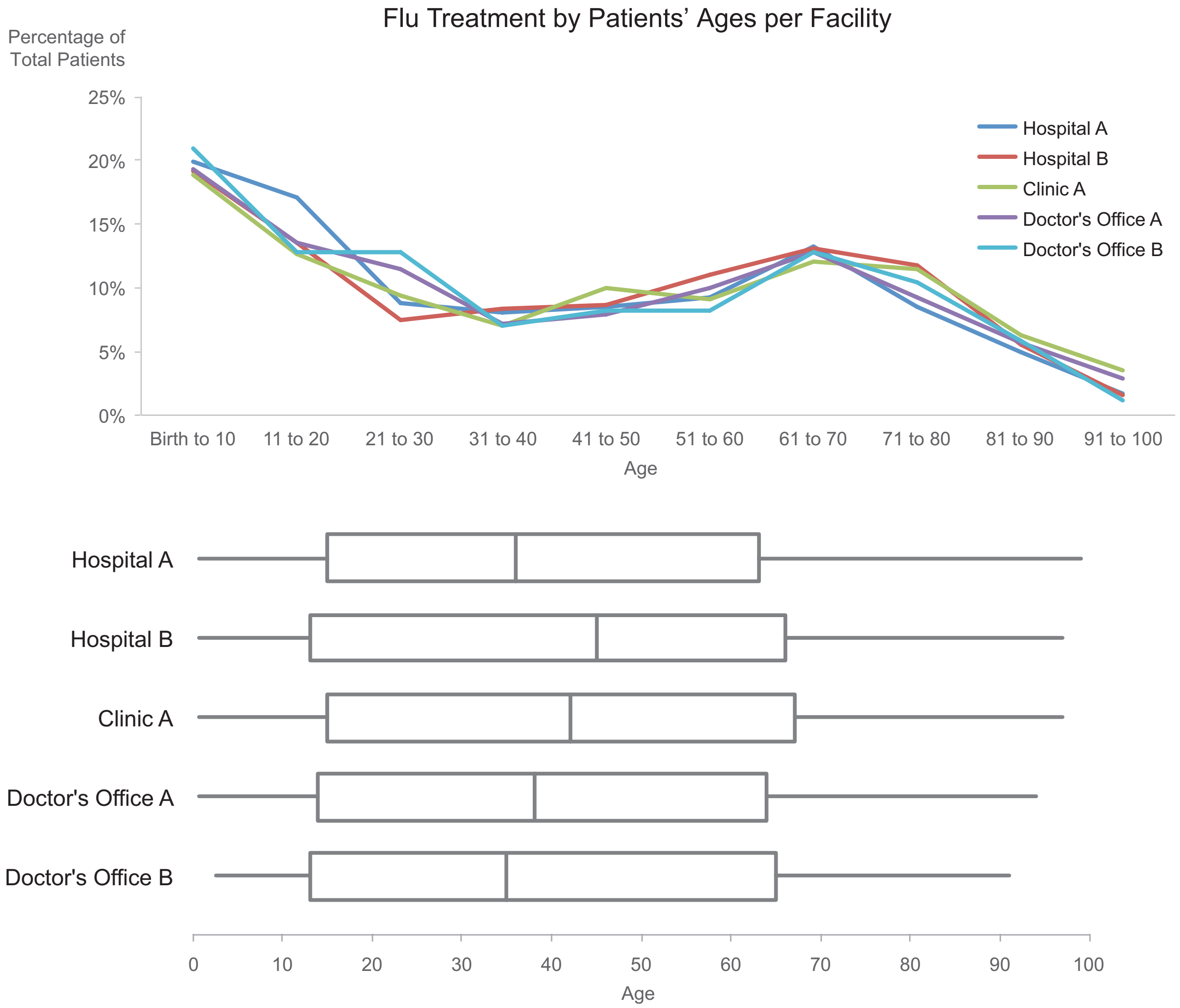


Box plots typically support another feature that I haven't illustrated yet: they can display outliers individually. This feature appears in the following graph with outliers displayed as individual dots:



Now, rather than only seeing a long line representing the first quartile at the bottom of product F's values, we can see that most of the shortest support calls lasted longer than 10 minutes, with only two that were shorter. Seeing the outliers displayed separately in this manner can be quite enlightening. When outliers appear separately in box plots, the line does not extend to the lowest or highest value in the set, but instead to the lowest or highest value that isn't an outlier. The example above contains six distributions, but a box plot could contain many more.

The simplicity of these boxes supports quick and easy comparisons, but it comes at a price: they provide less information about the shape of the distribution than we can see in histograms and frequency polygons. In part, this is because frequency displays divide the range of values into several intervals, not just four quartiles, and in part because the tops of the bars in a histogram and the path of the line in a frequency polygon display the shape of the distribution directly in a way that requires less cognitive effort. In the example below, a frequency polygon and a box plot display the same five distributions.



Even though the frequency polygon gives us a better sense of a distribution's shape, placing more lines in a frequency polygon would produce even more over-plotting and clutter than the one above already exhibits, but many distributions can appear in a box plot without clutter.

The features of distribution graphs that we've considered can be summarized as follows:

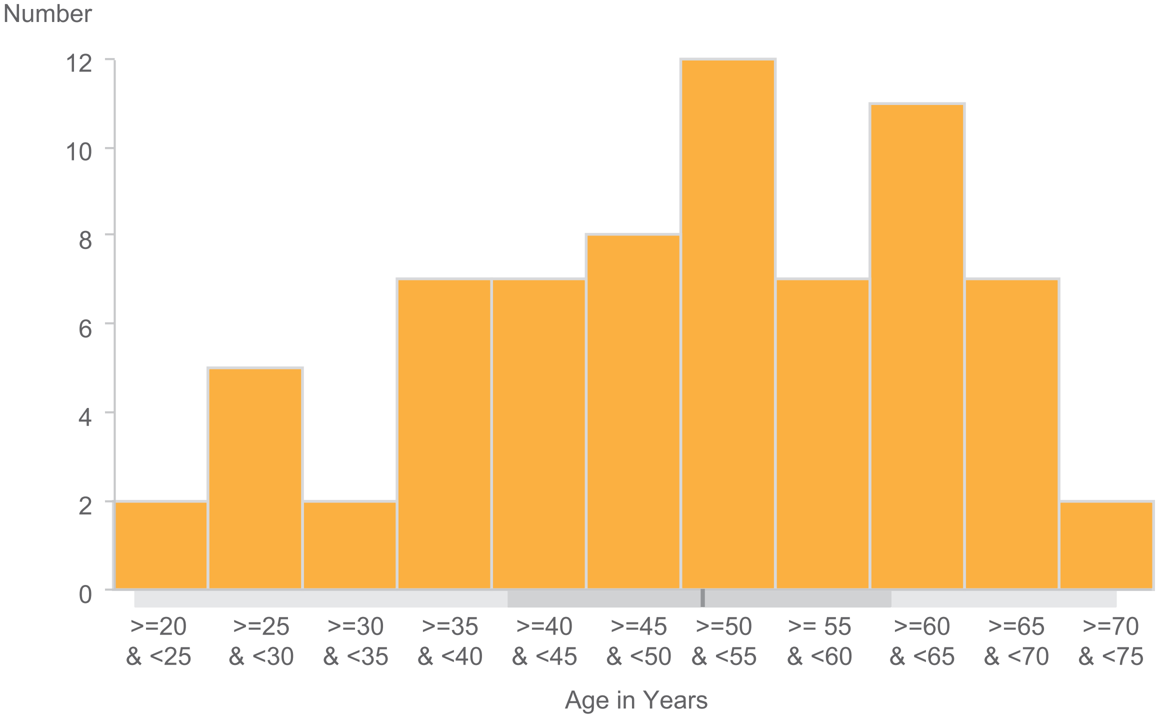
Feature	Histogram	Frequency Polygon	Box Plot	Strip Plot	Quantile Plot
Interval ranges	Consistent	Consistent	Variable	NA	Variable
Values per interval	Variable	Variable	Consistent	NA	Consistent
Displays the overall shape	Well	Very well	Satisfactorily	Not well	Not intuitively
Displays individual Values	No	No	Outliers only	Yes	Yes
Displays the central tendency	No	No	Yes	No	Yes
Displays the spread	Approximately	Approximately	Yes	Yes	Yes
Displays outliers	Approximately	Approximately	Yes, explicitly	Yes	Yes
Supports comparisons	Not well	Only a few	Well and many	Not well	Not well

Useful Enhancements to Conventional Distribution Displays

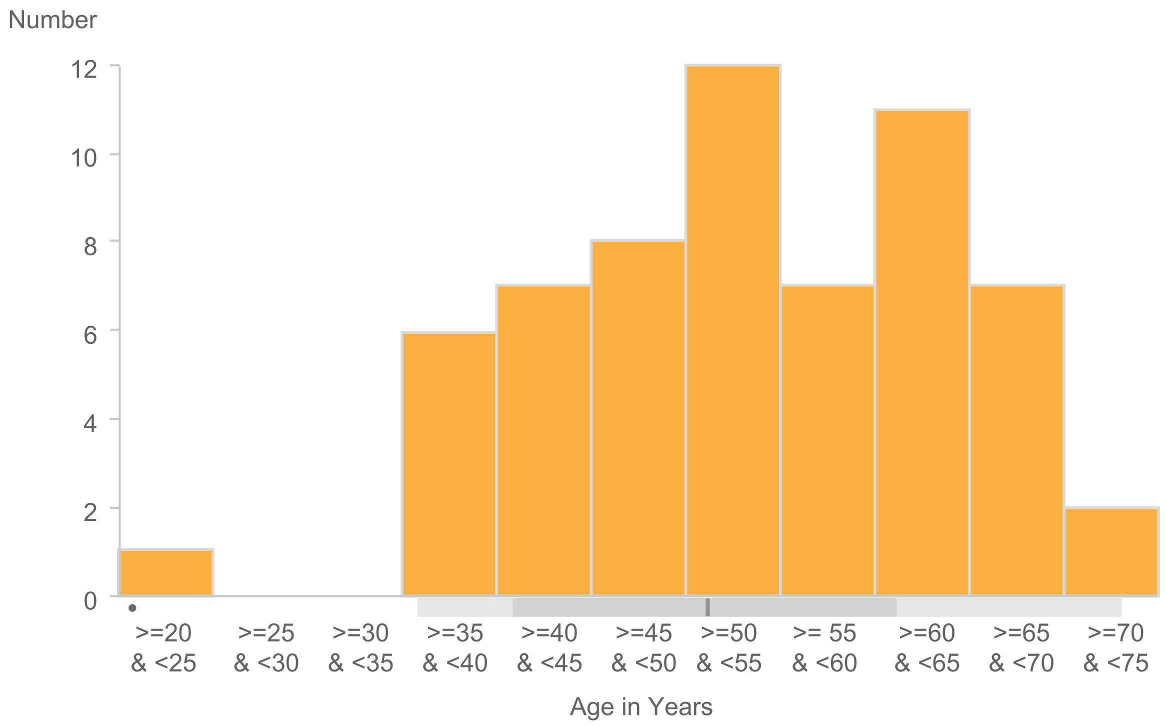
The varying strengths of these five distribution graphs lend them to different uses, leading to potentially different insights. For this reason, visual analysis tools should allow us to quickly shift between these graphs easily with relatively little loss of orientation. They should also incorporate as many useful features into each graph as possible without over-complicating them, even if doing so strays from convention.

Unconventional Additions of Information

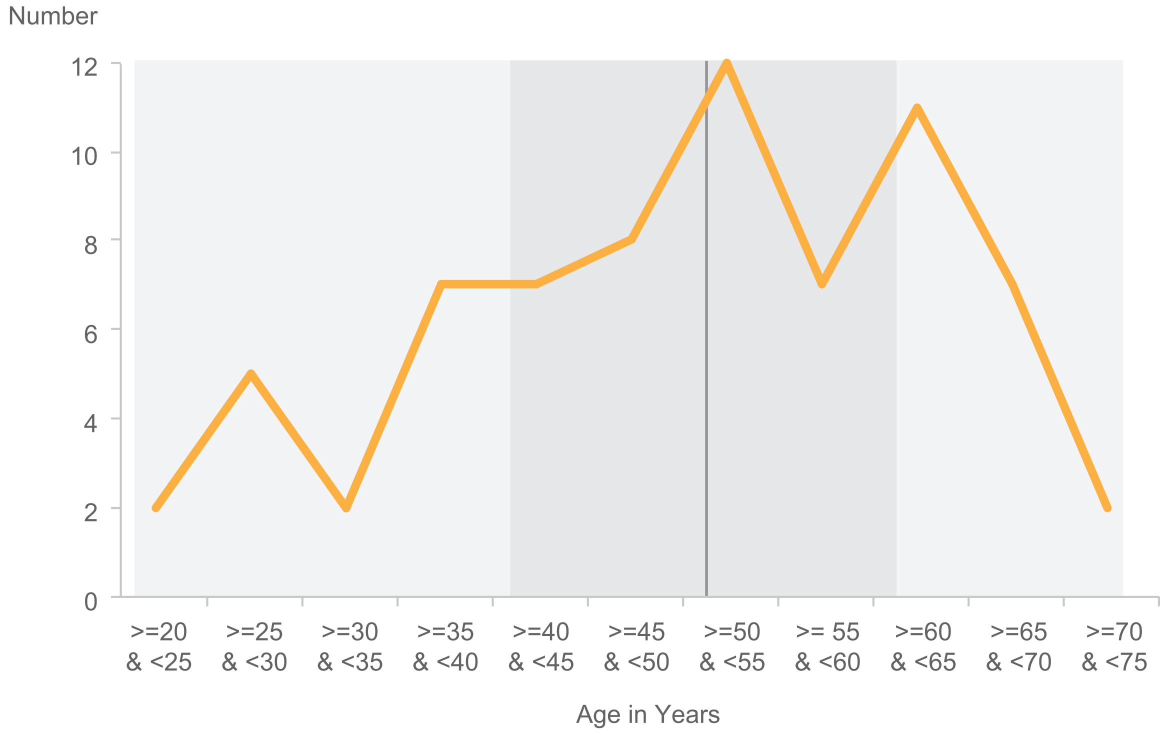
Remember that histograms and frequency polygons don't show the central tendency, show the spread only approximately, and don't identify or precisely show the values of outliers. Despite these conventional losses in information, this could be remedied easily. For instance, the median as a measure of central tendency, the precise lowest and highest values for an accurate view of the spread, and even quartiles could be displayed along the x-axis. Here's an example of how this could be done without overcomplicating or cluttering the graph:

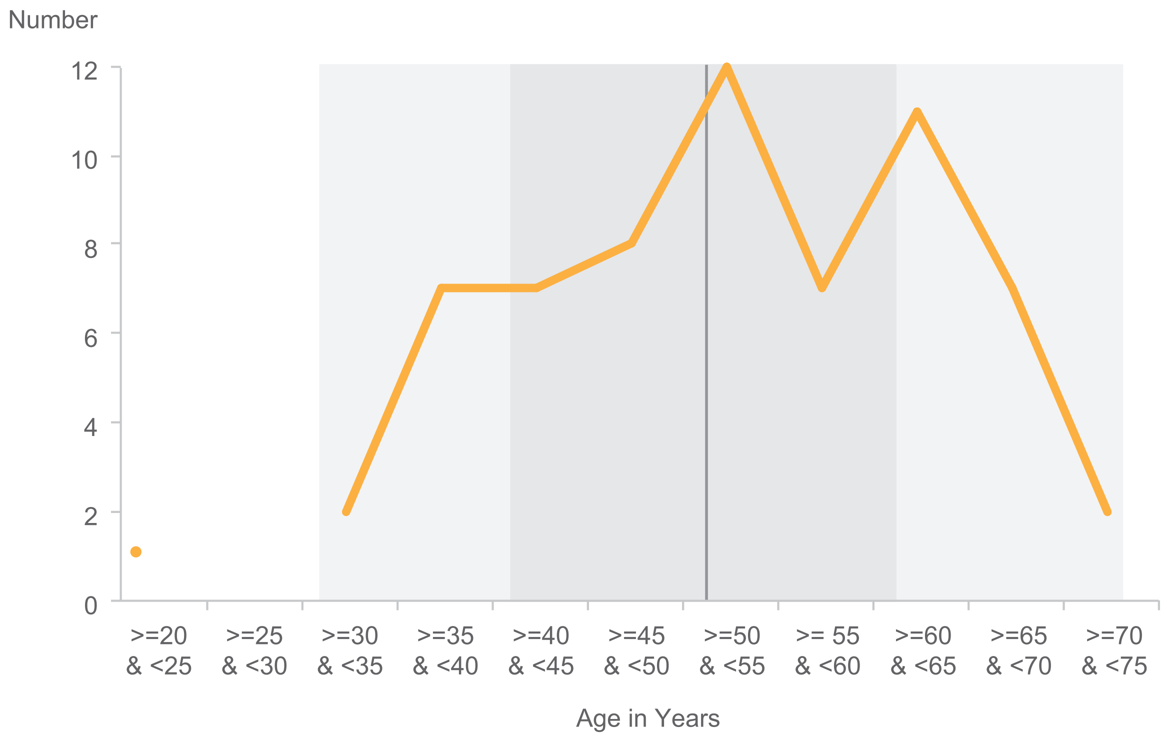


This particular data set doesn't include any outliers, but if it did, such as in the following example, outliers could be displayed separately as illustrated.

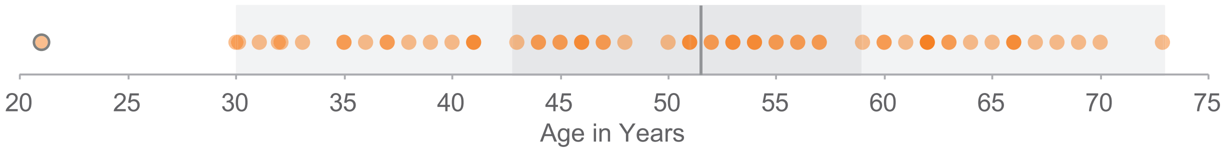


Frequency polygons could incorporate this information in a similar manner, or, because the background of the plot area is almost entirely visible, the information could be displayed as illustrated in the two next examples, one without and one with outliers.





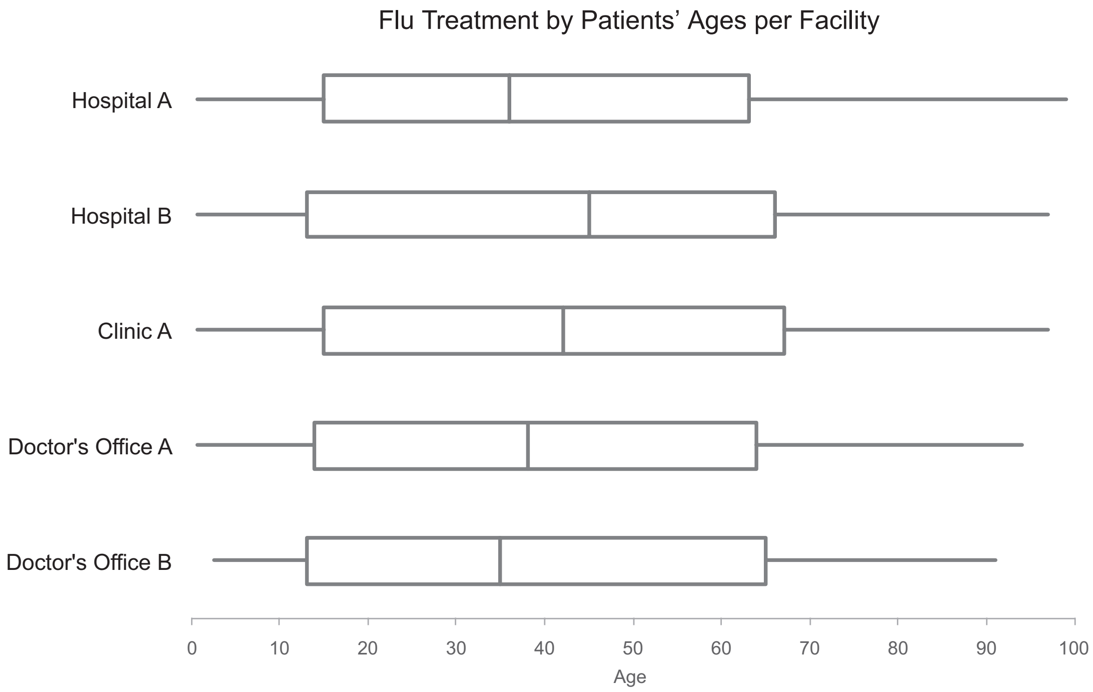
Strip plots don't typically display the central tendency or much information about a distribution's shape, such as quartiles, and even though they show outliers, they don't identify them as such. This can be easily remedied, however, such as in the manner illustrated below.



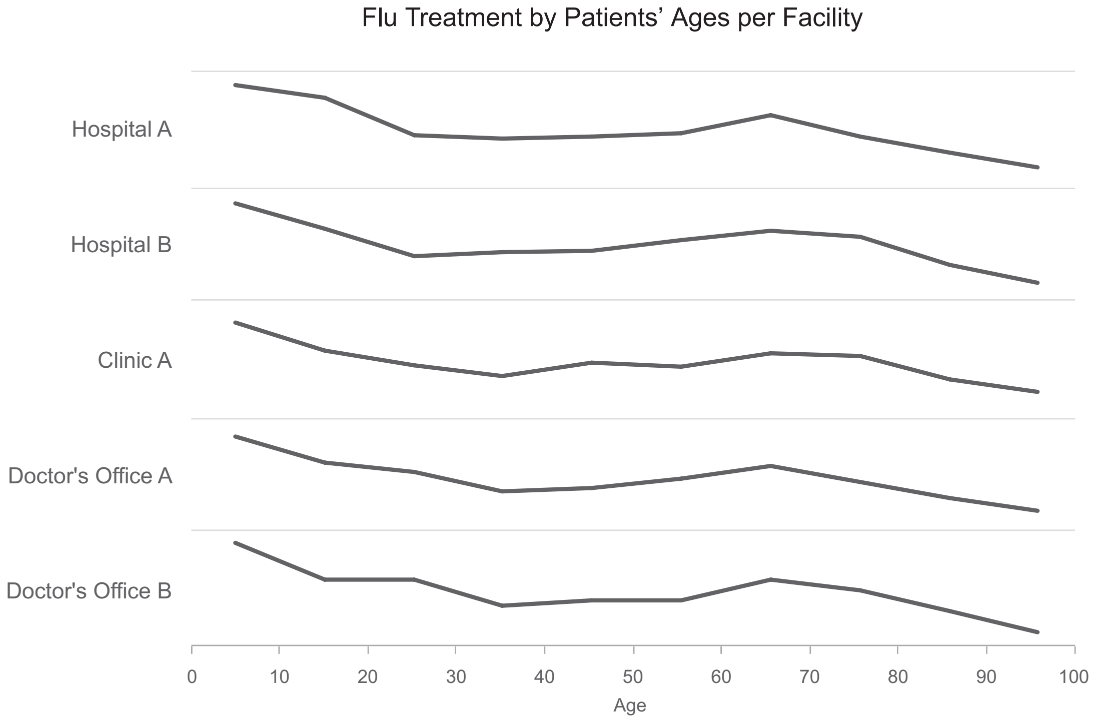
These examples don't illustrate all of the ways that these graphs could be enhanced to display additional information. With a little imagination and an understanding of visual perception and cognition, other designs could be created that work just as well or better.

Momentary Switching Between Proportional and Frequency Displays

Quite often, when we use box plots to compare distributions, there are moments when it would be useful to see the shape of the distributions in greater detail. Wouldn't it be nice if we could press a mouse button while hovering over a control and have that action causes the boxes to be temporarily replaced with histograms or frequency polygons, then switch immediately back to the boxes when the button is released? Imagine using the box plot on the following page:



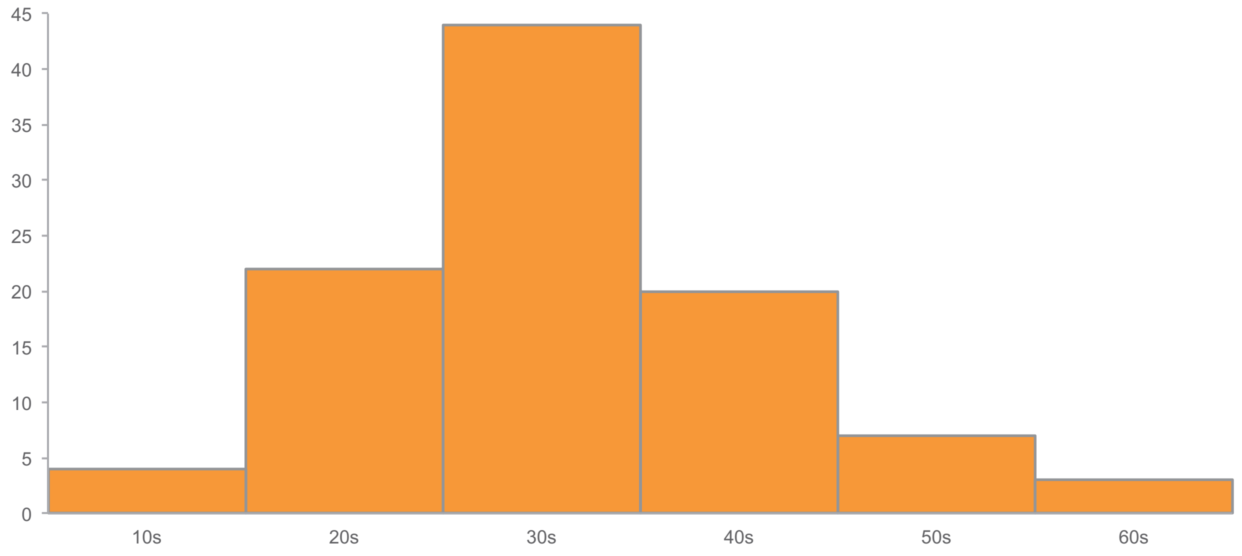
Now, to check for more subtle differences in the shape of Clinic A's and Doctor's Office A's distributions, imagine holding down a mouse button to have the following set of frequency polygons temporarily appear.



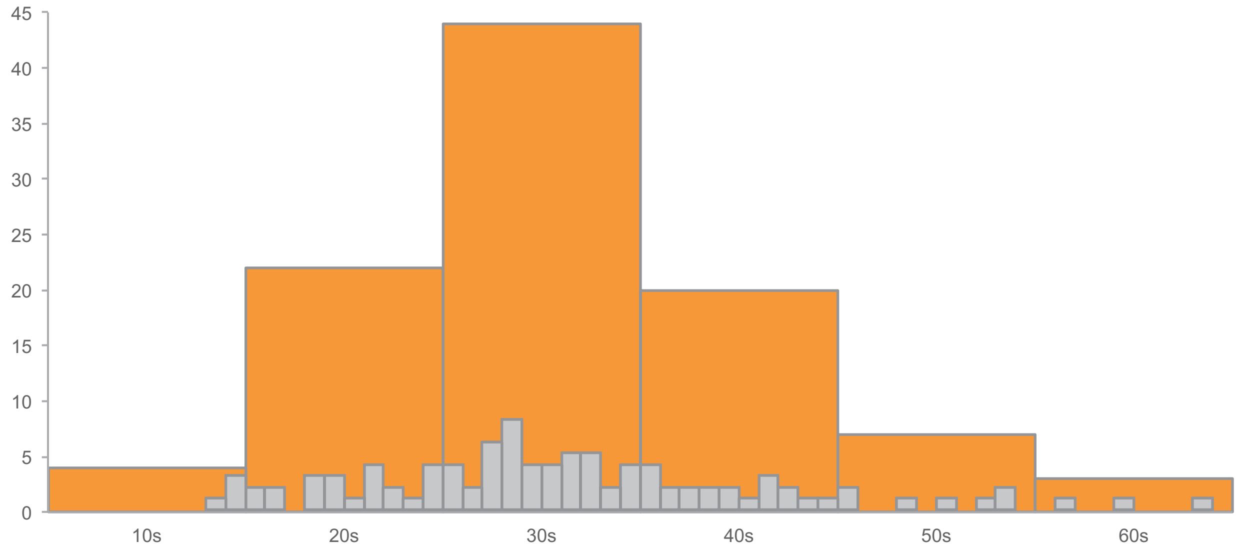
What couldn't be seen in the box plot can now be seen with ease. I'd love this feature, but I've not yet found it in a product.

Hierarchical Histograms

Histograms provide a nice overview of a distribution, but interesting details in the shape can be lost when viewing information that has been aggregated into large intervals. This could be partially remedied by equipping histograms to display multiple hierarchical levels of intervals simultaneously. For example, a histogram with age intervals of 10 years each could be subdivided to simultaneously show intervals of 1 year each within the 10-year intervals. Here's how it would look with 10-year intervals only:



And it might look like this with the addition of a lower level of detail of one-year intervals:



The ability to turn this on and off at will with the click of a button would make hierarchical binning in histograms like this quite useful.

Even though histograms, frequency polygons, box plots, strip plots, and quantile plots have been around for a long time, with a little effort we can make them more effective, and should.

Discuss this Article

Share your thoughts about this article by visiting the Distribution Displays, Conventional and Potential thread in our discussion forum.

About the Author

Stephen Few has worked for nearly 30 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the quarterly *Visual Business Intelligence Newsletter*, and speaks frequently at conferences. He is the author of three books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten*, Second Edition, *Information Dashboard Design: Displaying Data for at-a-Glance Monitoring*, Second Edition, and *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. You can learn more about Stephen's work and access an entire library of articles at www.perceptualedge.com. Between articles, you can read Stephen's thoughts on the industry in his blog.