MSKTC
SCI · TBI · BURN
Model Systems Knowledge Translation Center

Effective Use of Histograms

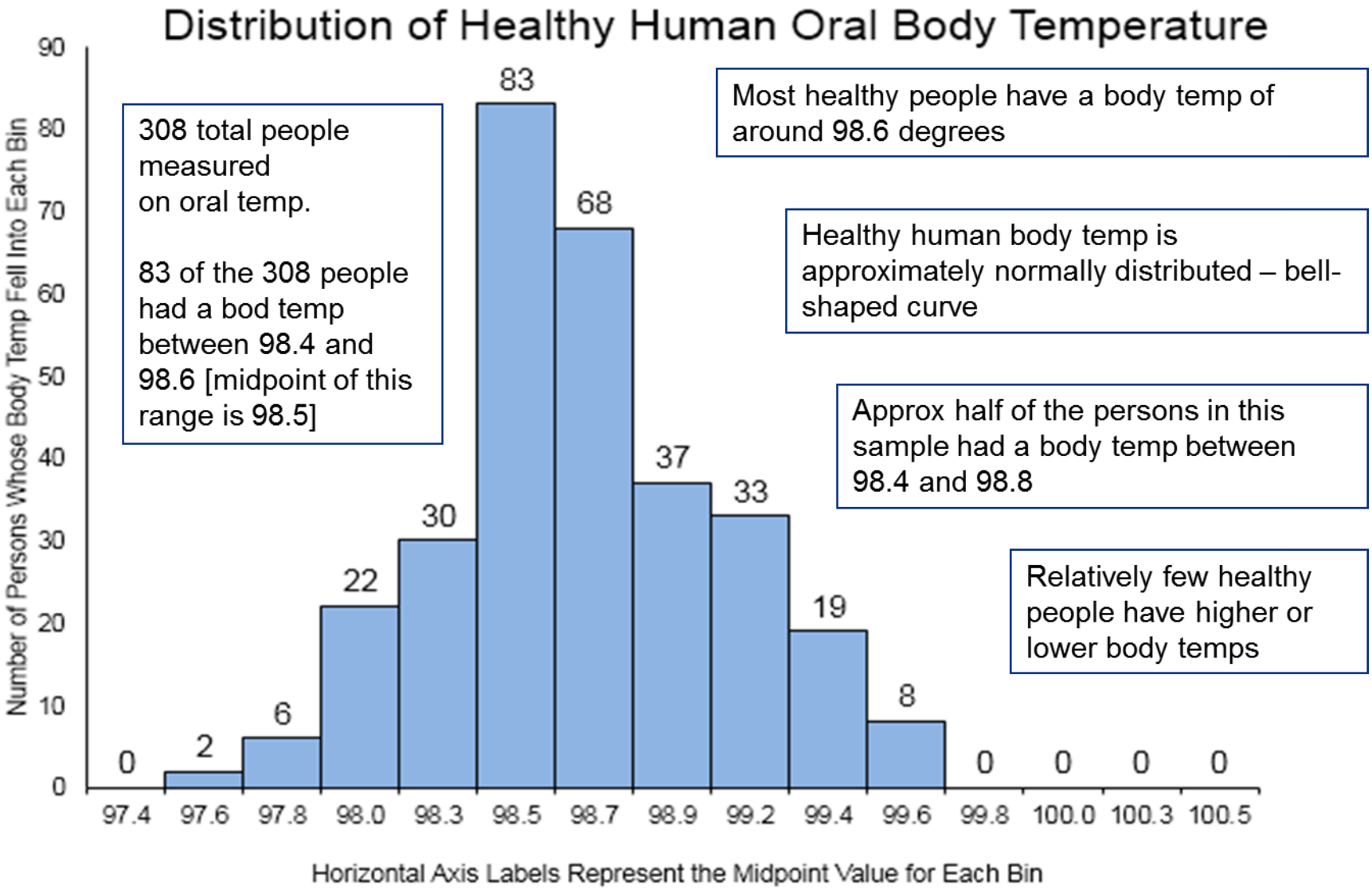
Purpose	This tool provides guidelines and tips on how to effectively use histograms to communicate research findings.
Format	This tool provides guidance on histograms and their purposes, and shows examples of preferred practices and practical tips for histograms.
Audience	This tool is designed primarily for researchers from the Model Systems that are funded by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR). The tool can be adapted by other NIDILRR-funded grantees and the general public.

The contents of this tool were developed under a grant from the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90DP0012-01-00). The contents of this fact sheet do not necessarily represent the policy of Department of Health and Human Services, and you should not assume endorsement by the Federal Government.

Histograms

- The primary use of a Histogram Chart is to display the distribution (or "shape") of the values in a data series.
- For example, we might know that normal human oral body temperature is approx 98.6 degrees Fahrenheit. And we might presume that the range of healthy body temperature is approximately normally distributed, with most people having body temps close to 98.6 and progressively fewer healthy people with body temps lower or higher than 98.6.
- To test this, we might sample 300 healthy persons and measure their oral temperature.
- For analysis we may decide to count the number of people whose body temp was between, say, 98.40 and 98.59, and count the number of people whose body temp was between 98.60 and 98.79, and the number of people whose body temp was between 98.80 and 98.99, and so on, using counting bins approximately 0.20 degrees wide, above and below the average body temp.
- When done, we will have a chart of normal body temp with a peak in the 98.6 degree range and progressively fewer people with lower body temps and progressively fewer people with higher body temps.

Histograms



308 total people measured on oral temp.

83 of the 308 people had a bod temp between 98.4 and 98.6 [midpoint of this range is 98.5]

Most healthy people have a body temp of around 98.6 degrees

Healthy human body temp is approximately normally distributed – bell-shaped curve

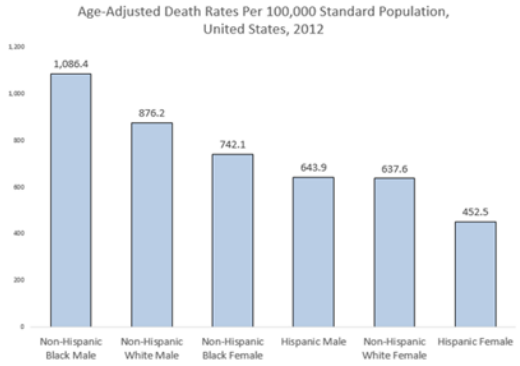
Approx half of the persons in this sample had a body temp between 98.4 and 98.8

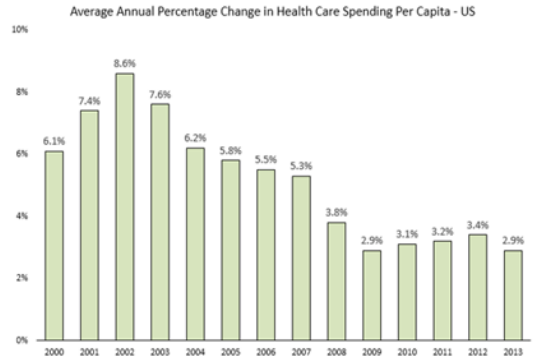
Relatively few healthy people have higher or lower body temps

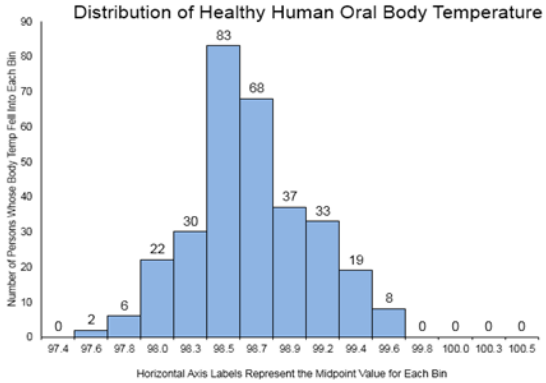
Each counting bin is approx 0.2 degrees wide

Histograms

- Histograms often look similar to column charts but there are some important differences.
- Column charts often display data with categorical groupings along the horizontal axis (such as diabetes rates for different race-ethnicity groups).





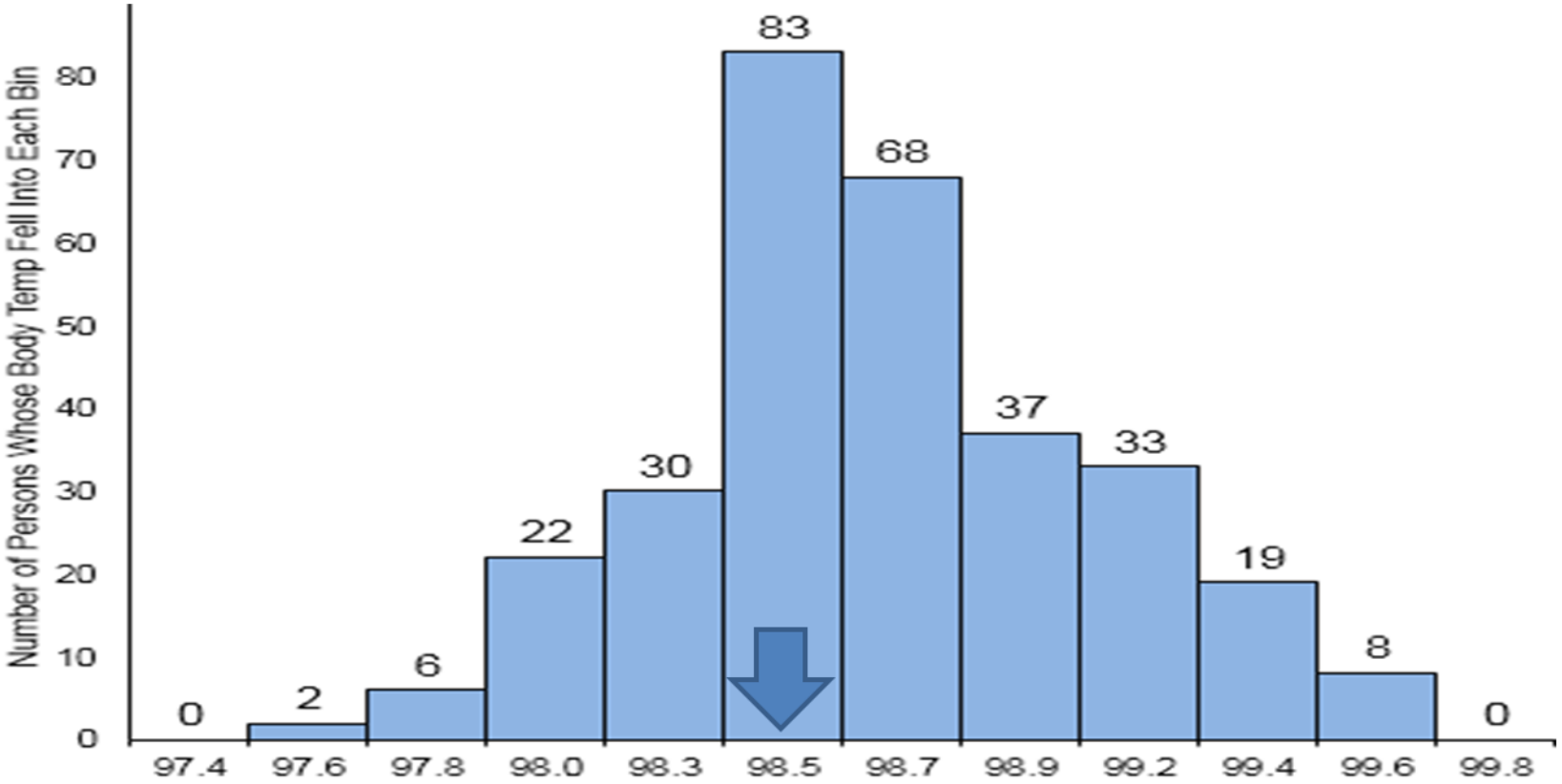


- Column charts are also used to display data over time – with time represented along the horizontal axis (such as the number of clients receiving services by fiscal year).
- Histograms, however, use quantitative, continuous data (such as the continuous range of human body temperature) which has been "binned" into equal-size segments (such as 0.20 degree range segments) along the horizontal axis.

Histograms

- The varying vertical height of each histogram rectangle represents the number of people, things, or events that fall into each of the histogram counting bins.
- The bins are positioned on the horizontal axis to be adjacent and the resultant vertical rectangles - representing the count of persons, things, events in each bin - will touch each other.
- The adjacent bins and rectangles are intended to convey the fact that the horizontal axis represents continuous data (not categorical data as in a column chart) .
- Column charts also have gaps between the vertical columns to convey that the horizontal axis for column charts is categorical, not continuous. Thus, it is also important that the columns in an actual categorical column chart not touch each other - to prevent users from mistakenly interpreting a categorical column chart as a histogram.

Histograms



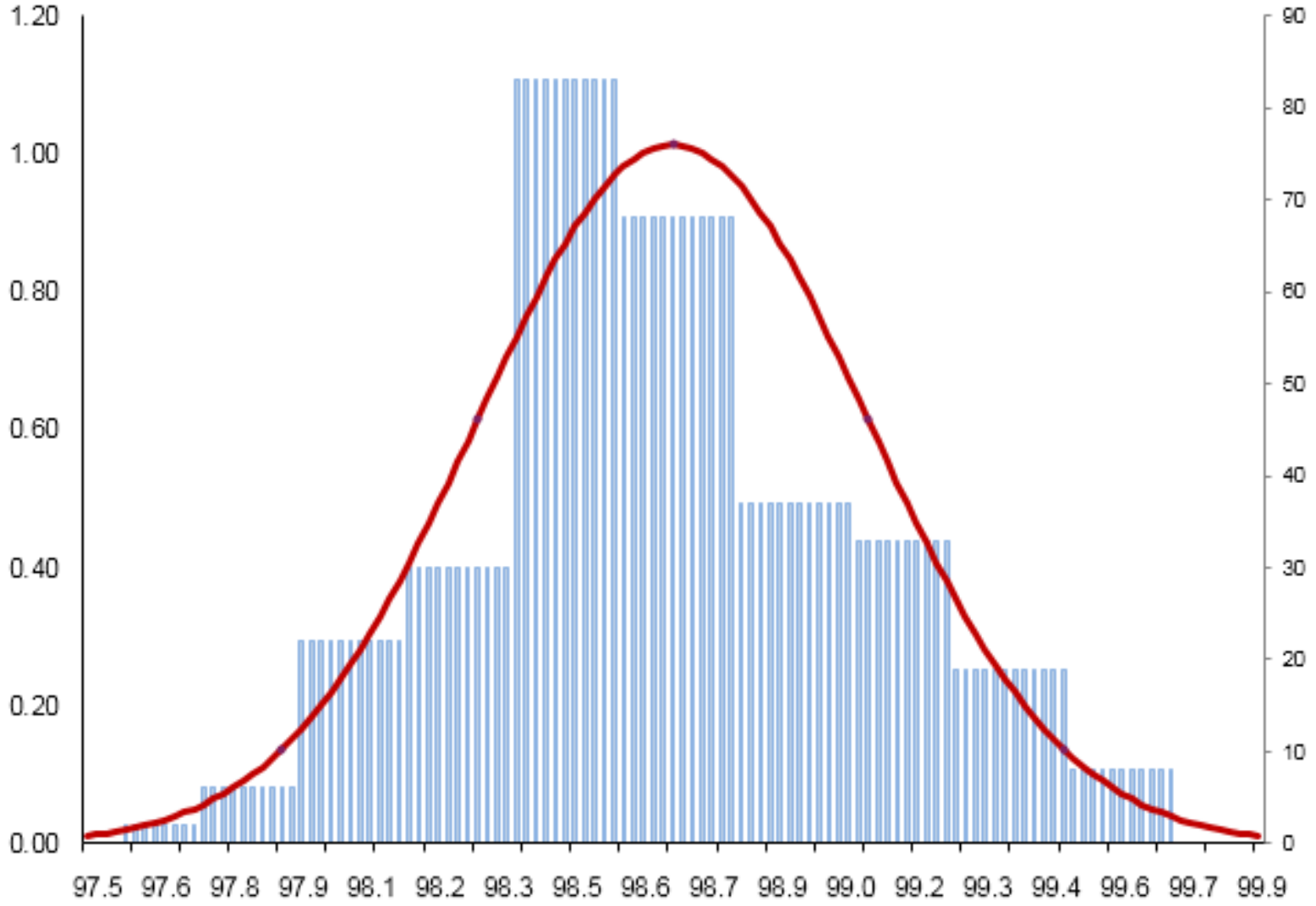
To prevent the horizontal axis from appearing too cluttered, the value 98.5, for example, represents the midpoint of a counting bin range from 98.39 to 98.60. The width of each bin in this example is approx 0.22 degrees wide. You could label the bins on the horizontal axis as 98.39-98.60 98.61-98.82 etc. But such would be more difficult to display and read.

You can also force your bin ranges to "cleaner" categories if you wish, such as bins precisely 0.20 degrees wide, though the number of displayed bins is generally more important than the width of each bin.

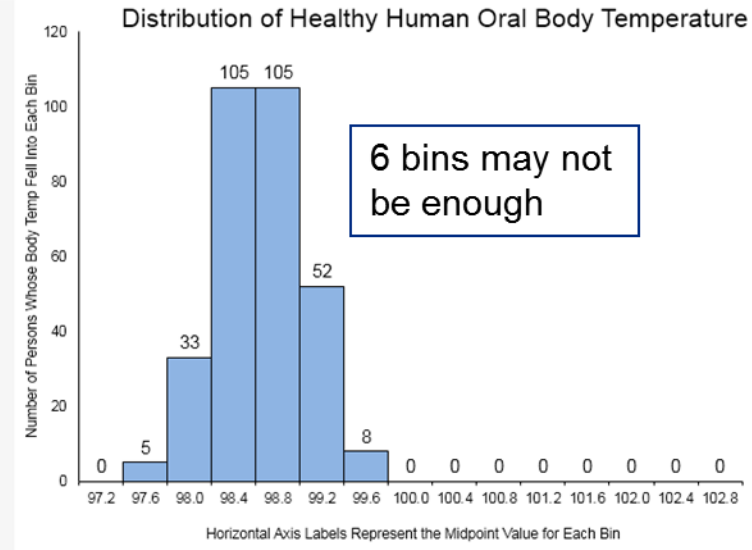
Histograms

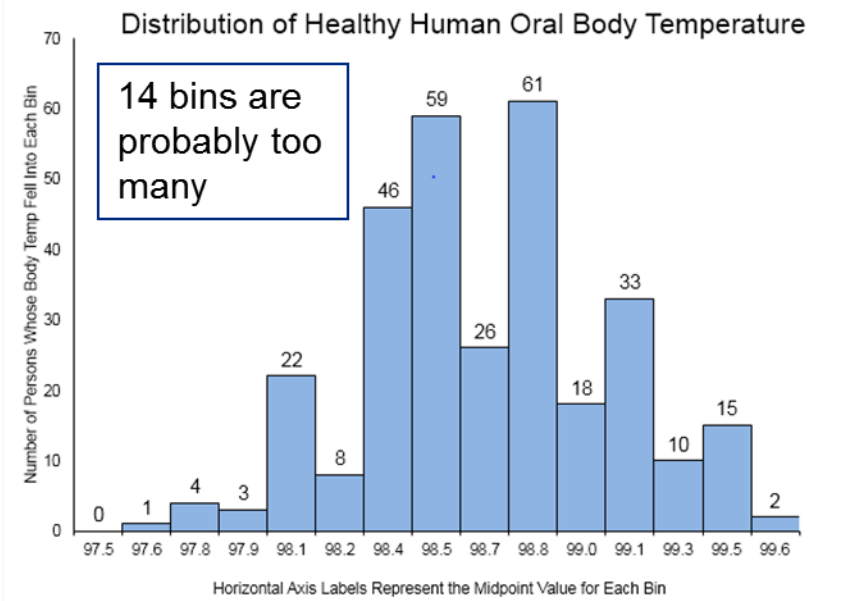
- Histograms are usually displayed as adjacent vertical rectangles of varying heights which convey the “shape” of the distribution.
- Optionally, a histogram may include a smooth “best fit” curve overlaying the vertical rectangle display to better summarize the general shape of the data distribution.
- In the case of the histogram of normal body temperature discussed above, the “shape” of any overlaid curve would resemble a “bell-shaped curve” (or normal distribution).
- In other cases, the distribution of the histogram (and any overlying smooth curve) may be “skewed” in one direction or the other, with more cases bunched up at the lower end of the distribution or at the higher end of the distribution (instead of the majority of cases bunched up in the middle as you would expect in a normal distribution).

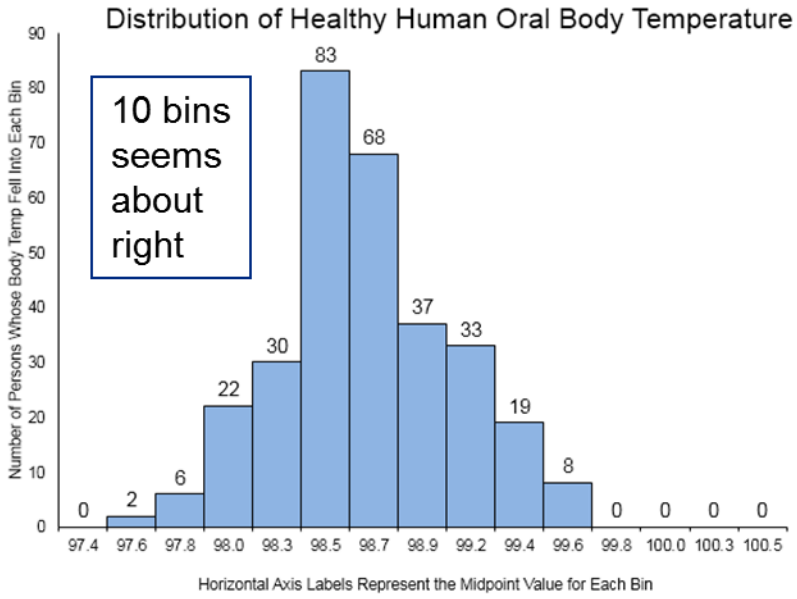
Histogram – With Smooth Curve Overlay



Histograms – How Many Bins?



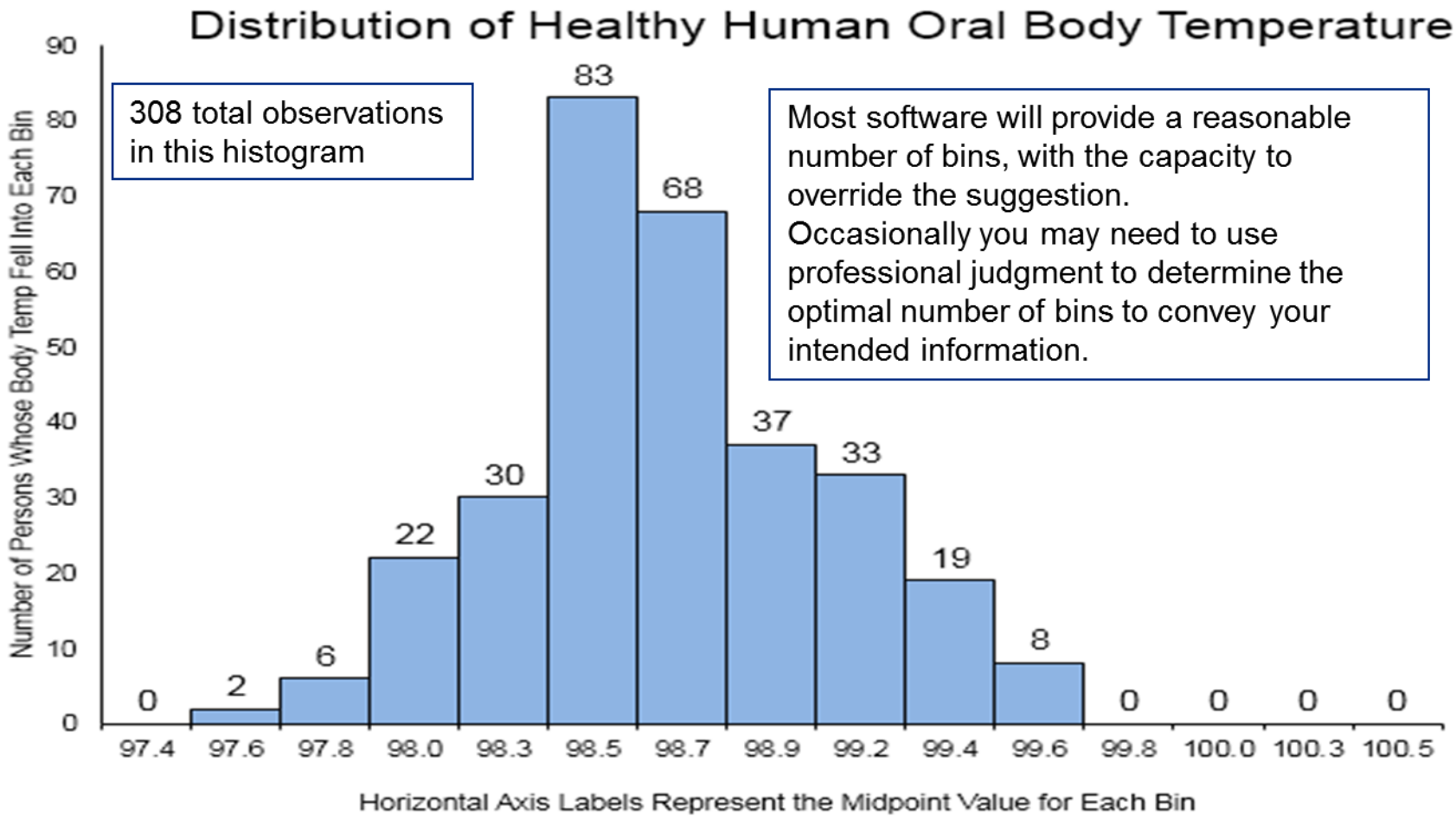




These three histograms all use the same 308 observations, just bucketed into a different number of bins of different width.

The number of bins you select for tabulating your observations can change the shape and perception of your data distribution (either unintentionally or purposefully).

Histograms – How Many Bins?

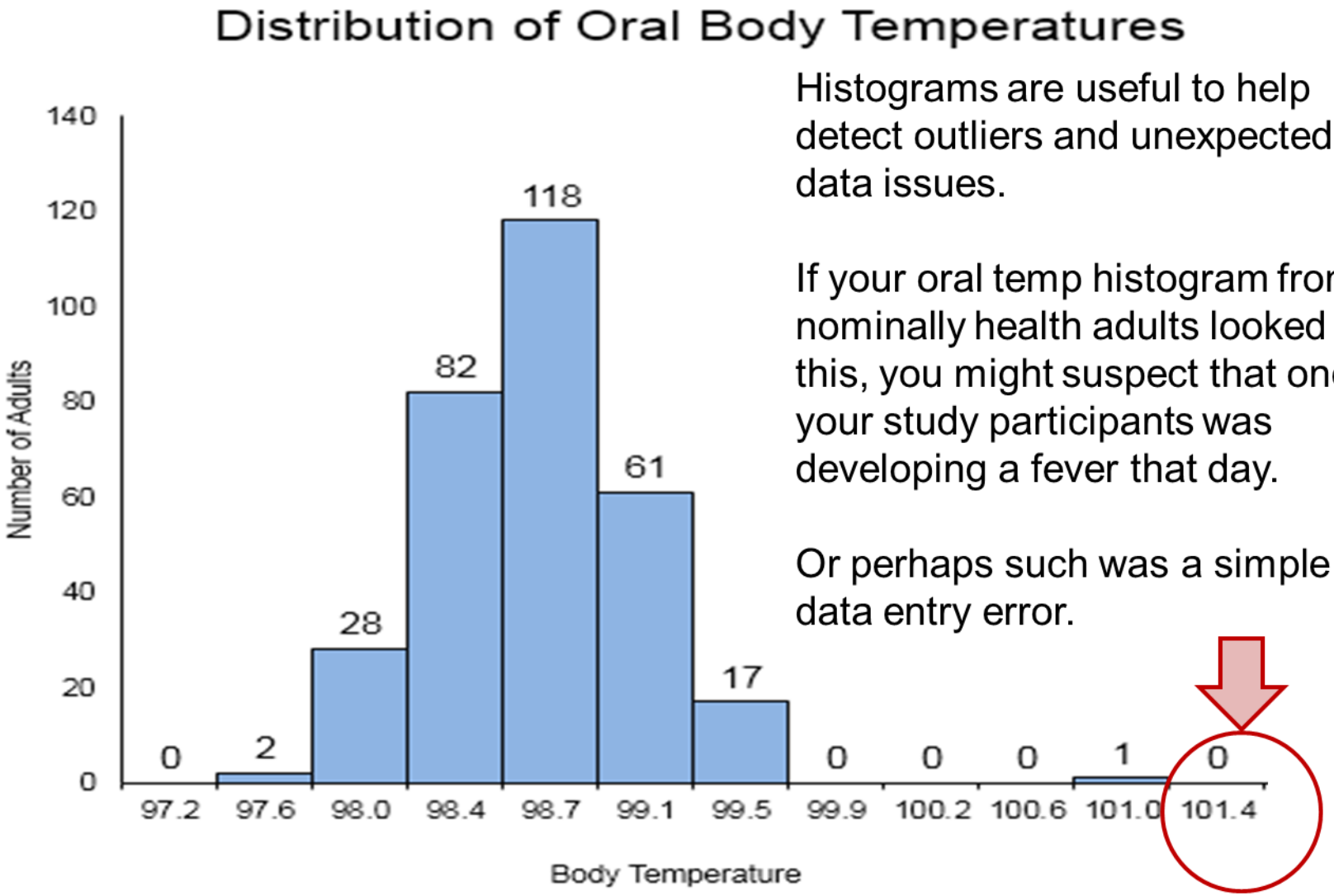


308 total observations in this histogram

Most software will provide a reasonable number of bins, with the capacity to override the suggestion.
Occasionally you may need to use professional judgment to determine the optimal number of bins to convey your intended information.

Various formulas exist to determining the "ideal" number of non-zero bins. One formula: Min(13,1+(3.322*LOG(Total Obs))+1)
In this example: Min(13, 1+(3.322*LOG(308))+1) = 10.27 bins or 10 bins, rounded. Generally recommended that never exceed 13 bins for a histogram.

Histograms – Outlier Detection

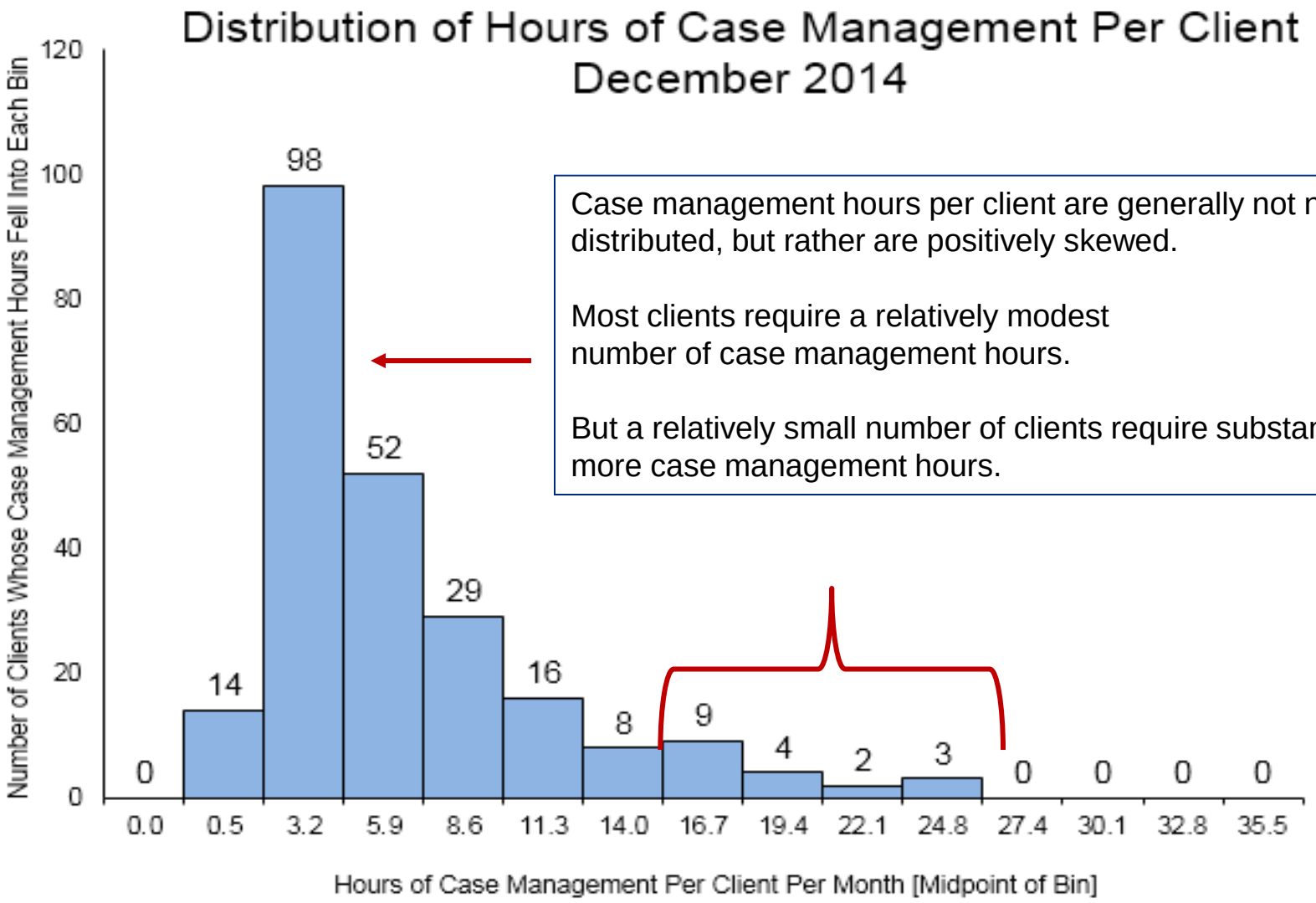


Histograms are useful to help detect outliers and unexpected data issues.

If your oral temp histogram from nominally health adults looked like this, you might suspect that one of your study participants was developing a fever that day.

Or perhaps such was a simple data entry error.

Histograms – Positively Skewed Distribution

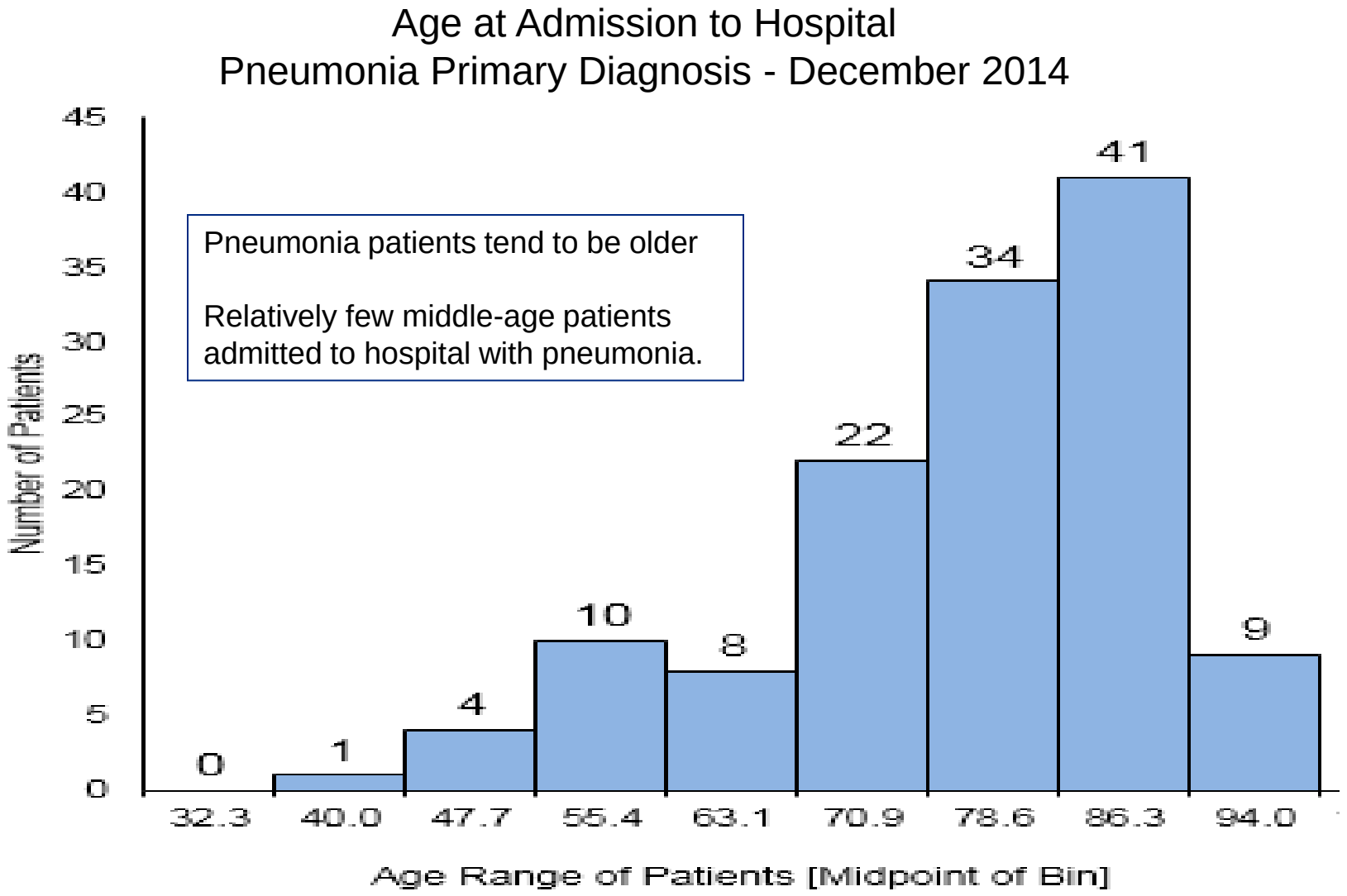


Case management hours per client are generally not normally distributed, but rather are positively skewed.

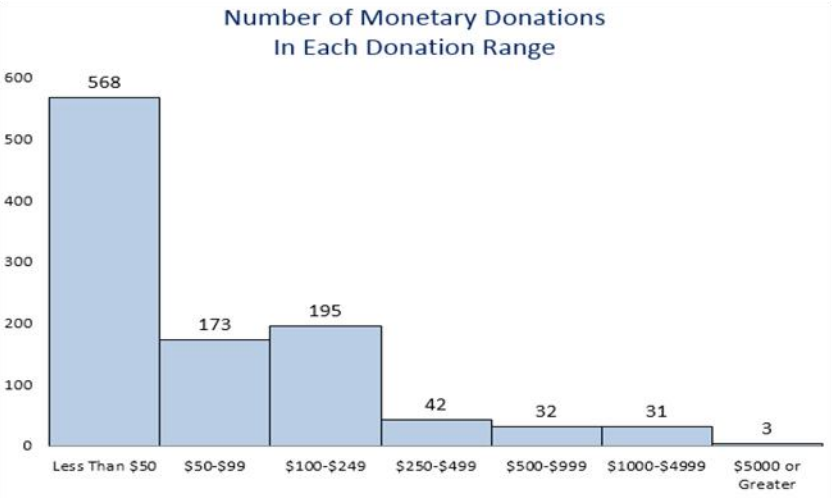
Most clients require a relatively modest number of case management hours.

But a relatively small number of clients require substantially more case management hours.

Histograms – Negatively Skewed Distribution



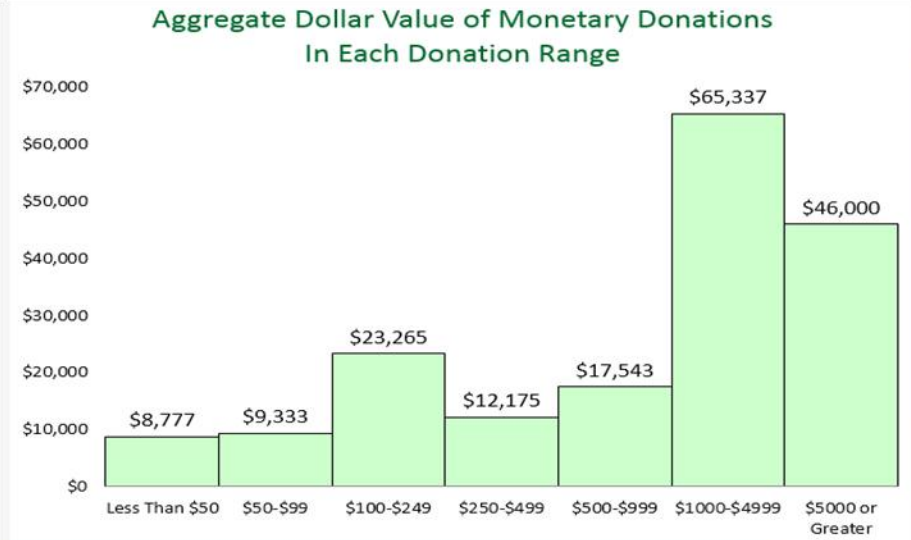
Histograms – Charitable Donations



Most donors give less than $50.

A very small number of donors give $5,000 or more.

Positive skew.



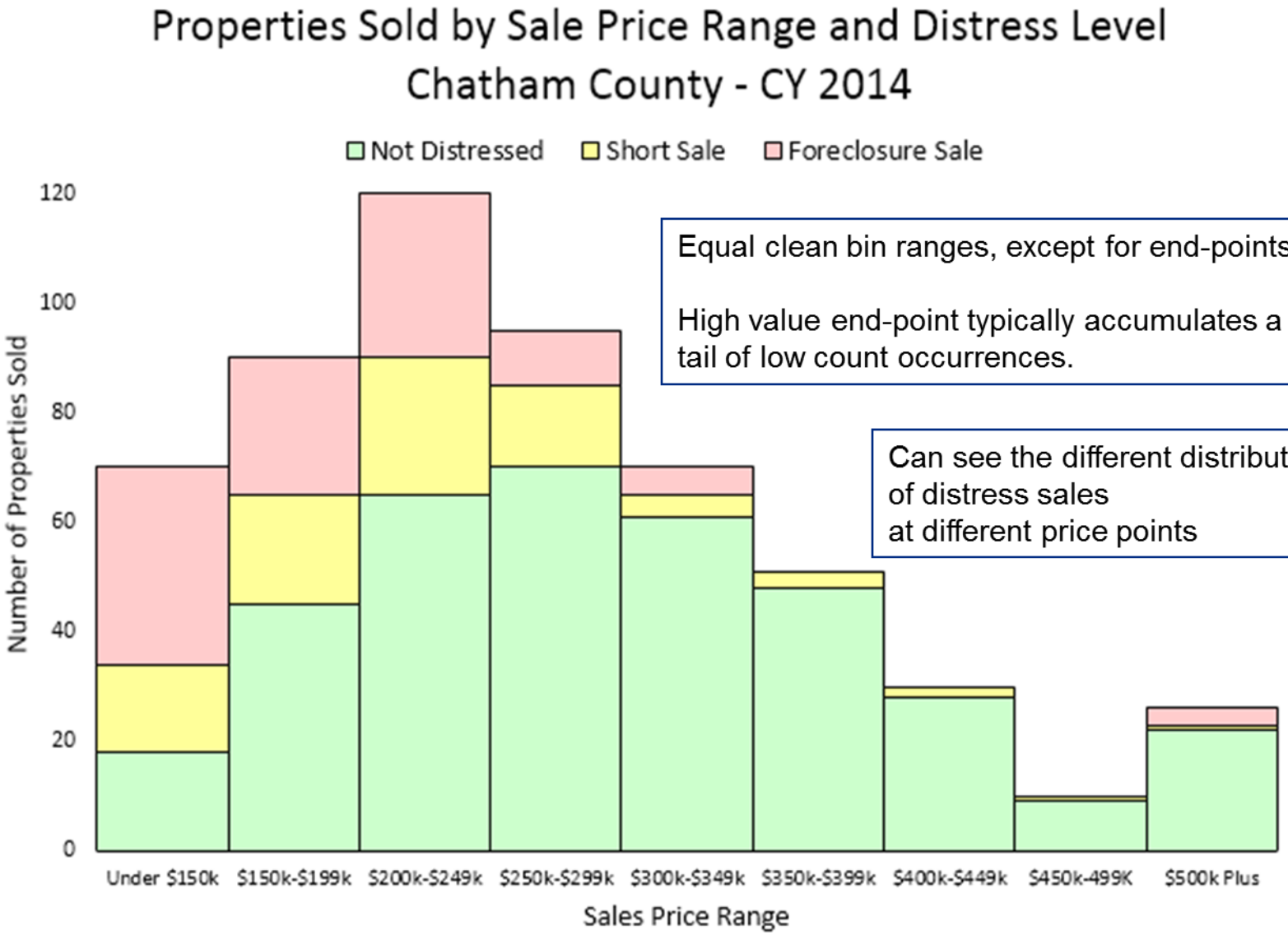
But the relatively few larger donors provide the majority of the dollar value of the donations received.

Make sure you send the warmest thank you notes to these donors.

Negative skew.

Note that used “cleanly” defined but unequal-sized bins in this example for purposes of clear communication with the target audiences.

Histograms – Stacked Histograms



Equal clean bin ranges, except for end-points.

High value end-point typically accumulates a long tail of low count occurrences.

Can see the different distributions of distress sales at different price points

Source: Mock Data