Chapter 2: Summarizing Data

“Numerical quantities focus on expected values, graphical summaries on unexpected values.”

–John Tukey

Preview:

- Displaying stationary data distributions
  - Bar charts, needle plots, frequency histograms
  - Analysis of same
  - Causes of common patterns
- Summary measures for stationary data distributions and when each is appropriate
- Boxplots and outliers
- Resistant summary measures

$\approx$

CAUTION

The graphs and measures presented in this chapter are meant to summarize the pattern of variation of data from stationary processes, and may not make sense in other settings. Therefore, data taken over time should always be checked for stationarity before using them.

$\approx$

Some Terminology

- **Variable:** Name of what is being counted, measured or observed.
- **Variable Types**
  - **Quantitative/Categorical**
  - **Discrete/Continuous**

Displaying Data Distributions

Example: A company manufactures knobs for appliances (washing machines, dishwashers, etc.). In one step of the manufacturing process for a certain knob, the spindle diameter, having a nominal value of 3mm, is measured. If it is too small (below 2.9mm), the part is rejected. If it is too large (above 3.1mm), the part is sent back for reworking. Otherwise, it is accepted. Here are data on a production lot of 12 parts:

$\approx$

Displaying Data Distributions

Diameter	Action[1]
2.8	X
2.9	A
3.2	R
3.0	A
3.0	A
3.0	A
2.9	A
2.7	X
2.9	A
3.2	R
3.3	R
3.1	A

[1] Accept (A), Rework (R), Reject (X).

Displaying Data Distributions

In these data, the variable **diameter** is quantitative and continuous, **action** is categorical.

Suppose we have data on 100 production lots and create a new variable, **count**, which for each lot counts how many of the knobs are accepted. Then **count** is quantitative and discrete.

$\approx$

Displaying Data Distributions

Use a **Bar chart** for categorical data. Figure 1 shows a bar chart of action (the data are found in sasdata.knobs and the SAS code for this figure is found here).

$\approx$

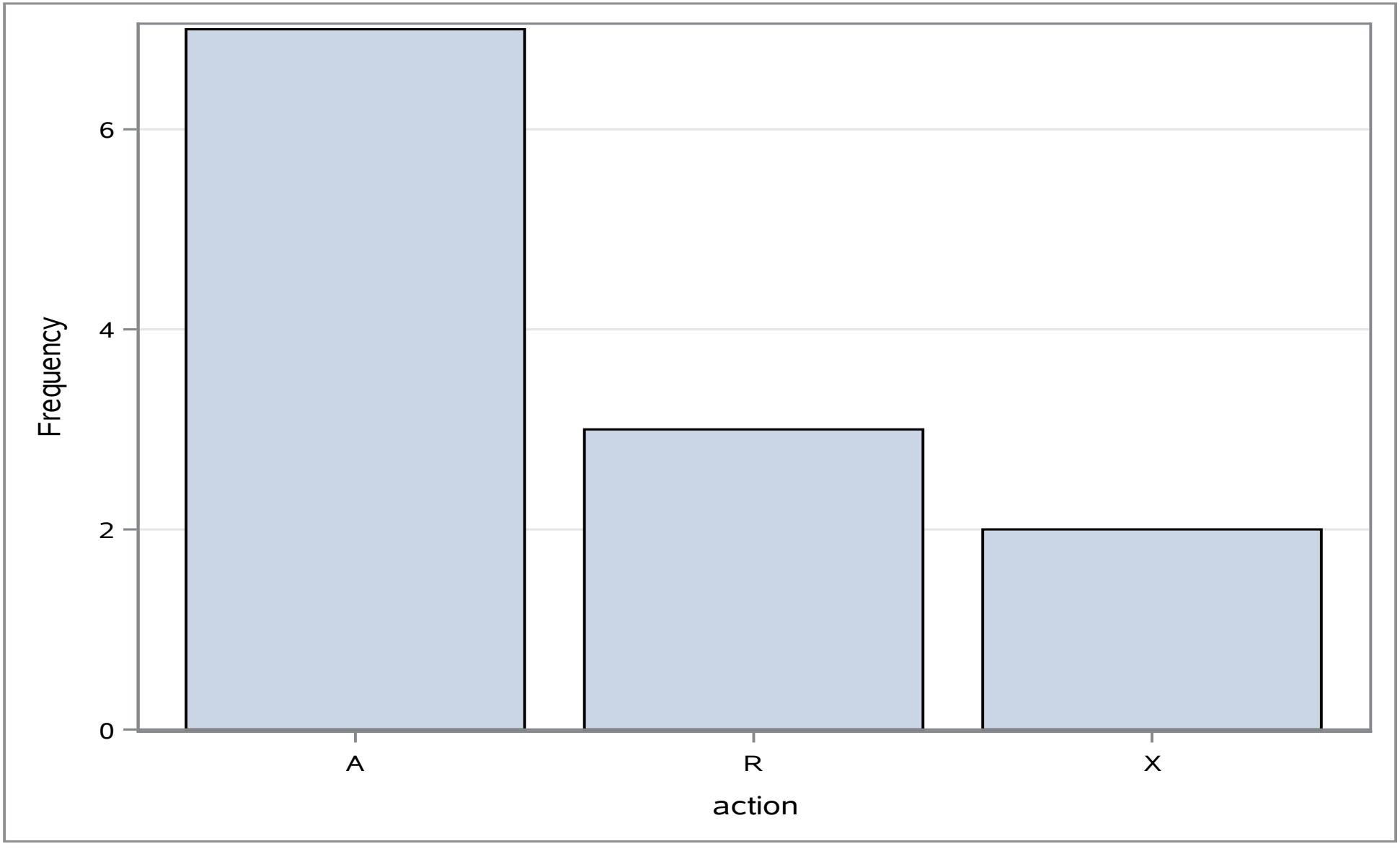


Figure: 1: Bar chart of action.

Displaying Data Distributions

You can use a **Needle Plot** for a small number (say 20 observations or fewer) of quantitative data. Figure 2 shows a needle plot of the diameters of the 12 knobs. (SAS code here)

$\approx$

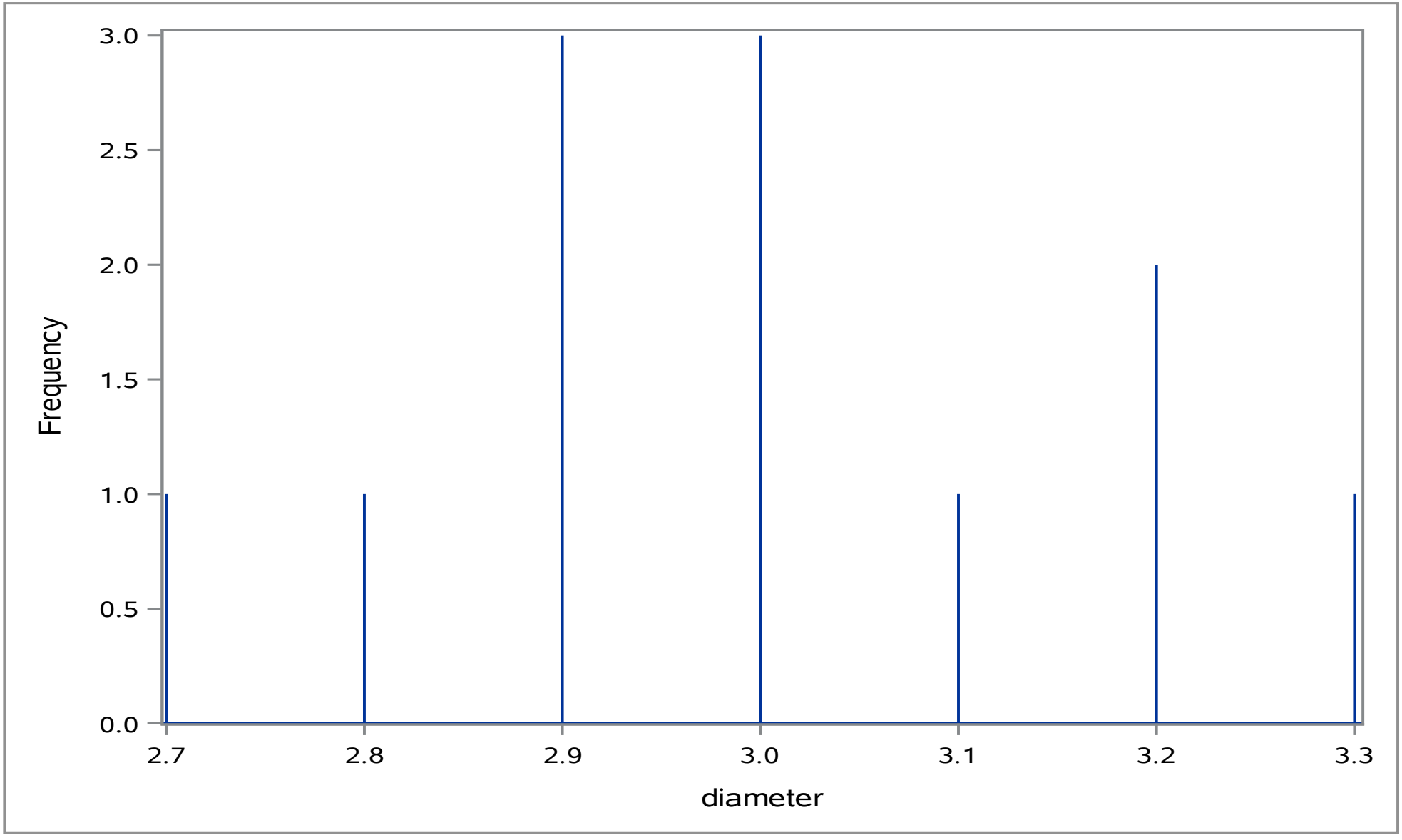


Figure: 2: Needle plot of knob spindle diameters.

Displaying Data Distributions

Use a **Frequency Histogram** for a larger number of quantitative data. Figure 3 shows a frequency histogram of the heights in cm of 105 high school students (data found in sasdata.armspan, SAS code here).

$\approx$

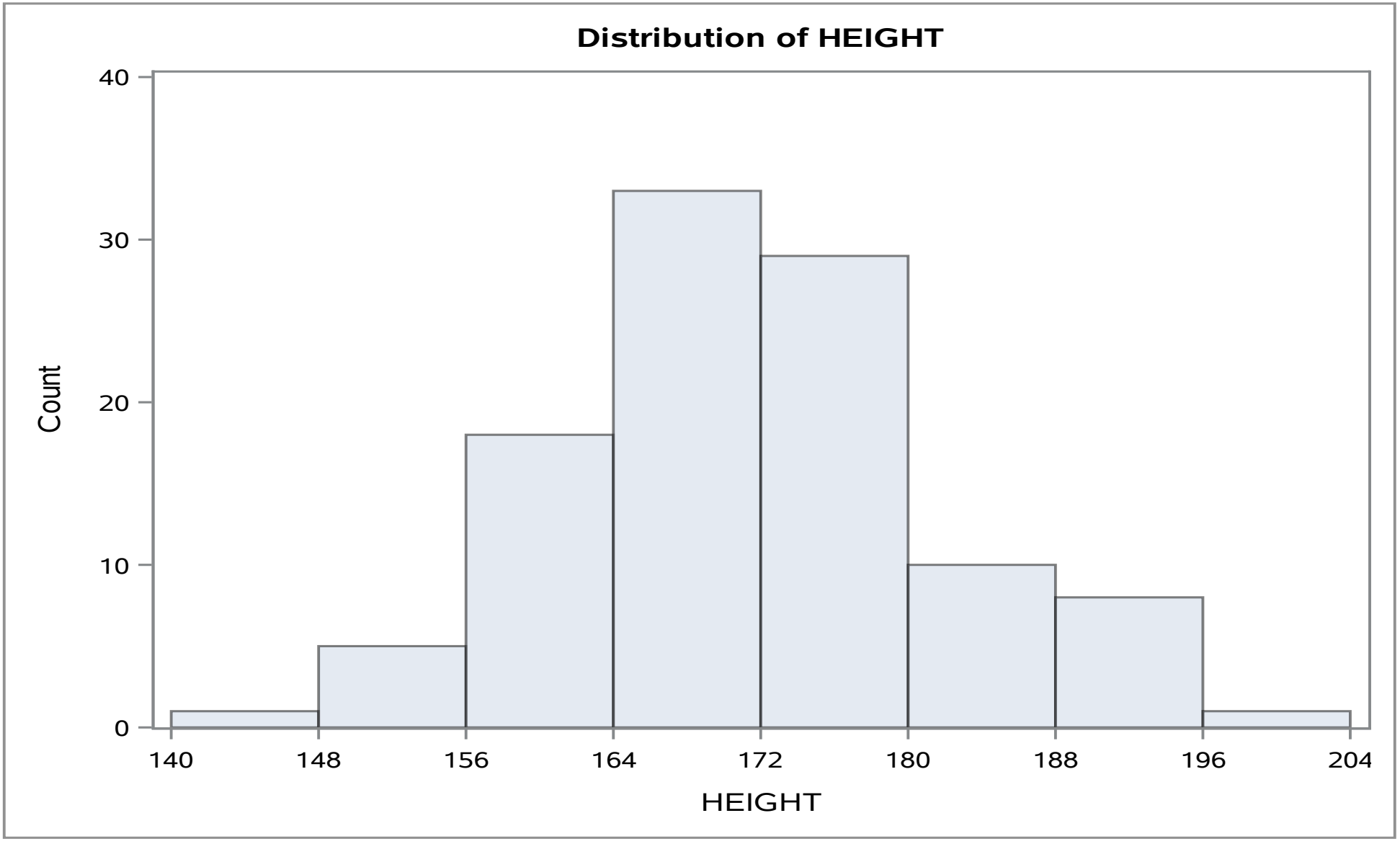


Figure: 3: Frequency histogram of the heights in cm of 105 high school students.

Analyzing Frequency Histograms

Different choices of interval locations and widths can make frequency histograms for the same set of data look very different. Have a look at the following two histograms based on the same data (data found in sasdata.geyser1, SAS code here):

$\approx$

Analyzing Frequency Histograms

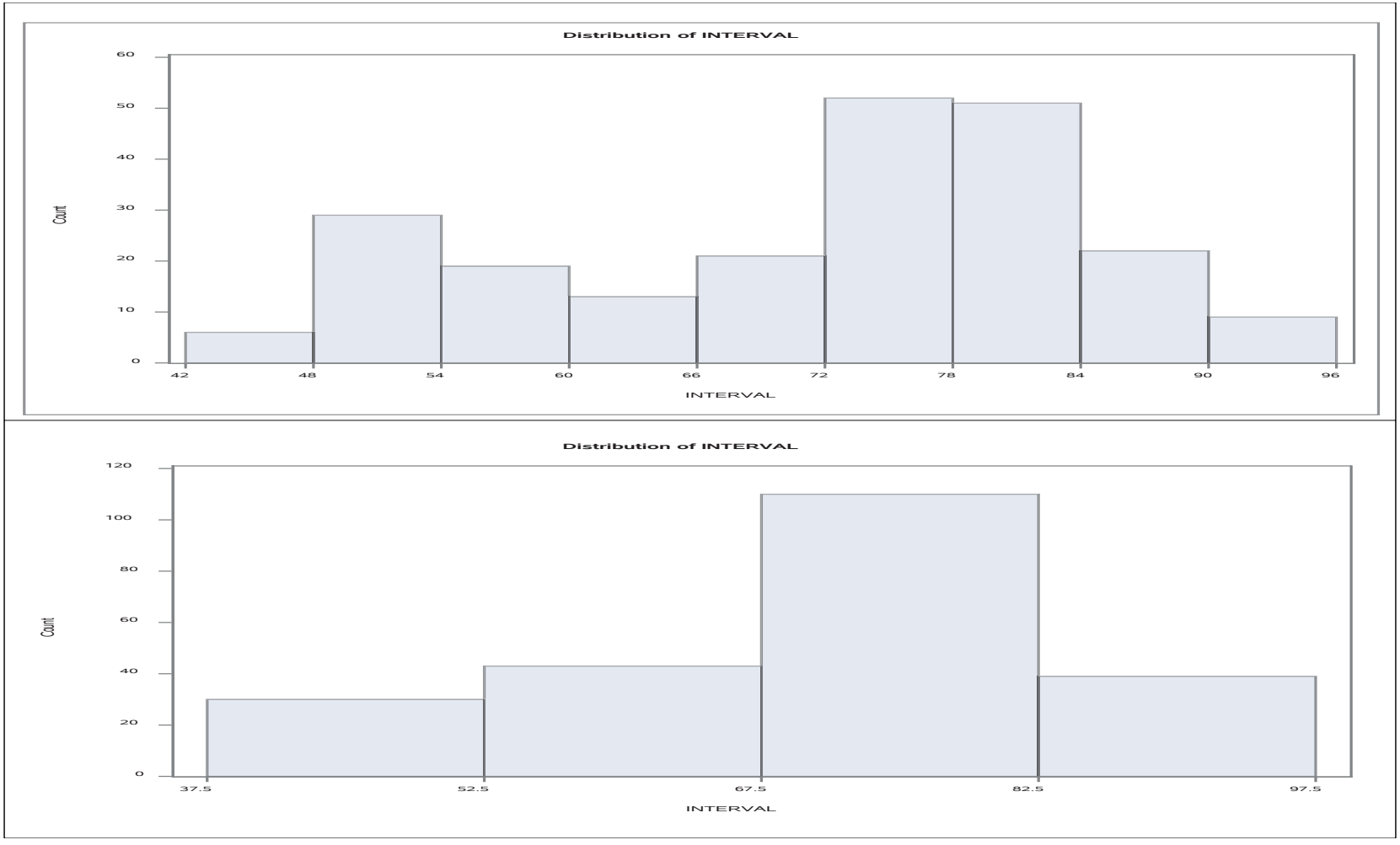


Figure: 4: *Two frequency histograms of the same data.*

Analyzing Frequency Histograms

Therefore, before analyzing a frequency histogram it is important to make sure it represents the true pattern of variation in the data. A good strategy is to create several histograms and choose one for analysis that displays features common to most of them.

$\approx$

Analyzing Frequency Histograms

What to look for:

- Modality: How many peaks?
- Symmetry: Mirrored about some vertical line?
- Center: Can you find one? Where is it?
- Spread: How spread out are the data?
- Pattern and deviations: What are the main patterns? What points don't follow these patterns?

$\approx$

Analyzing Frequency Histograms

Example: Figure 5 shows a frequency histogram of the lifetimes of 157 electrical transformers (data in sasdata.transform).

$\approx$

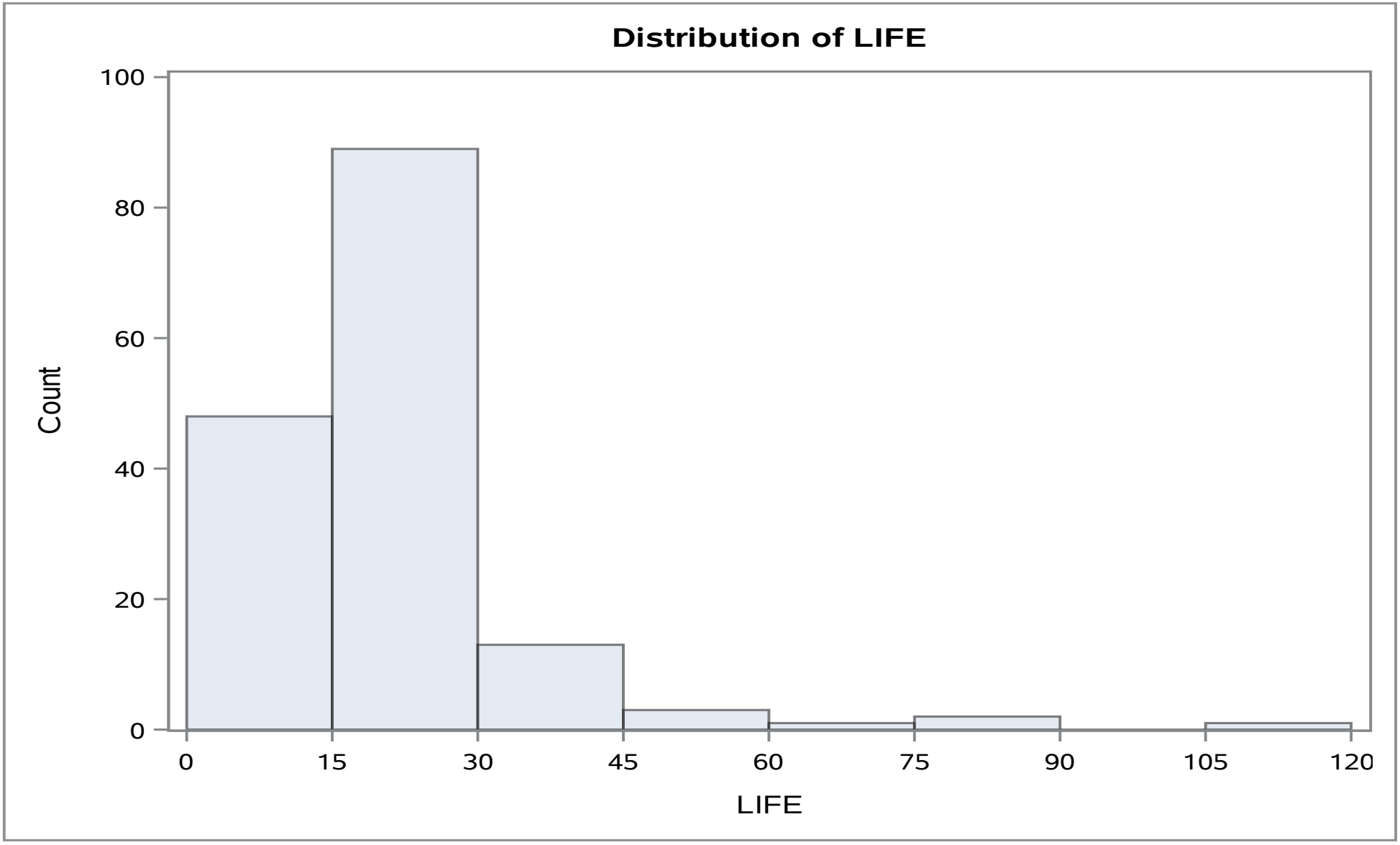
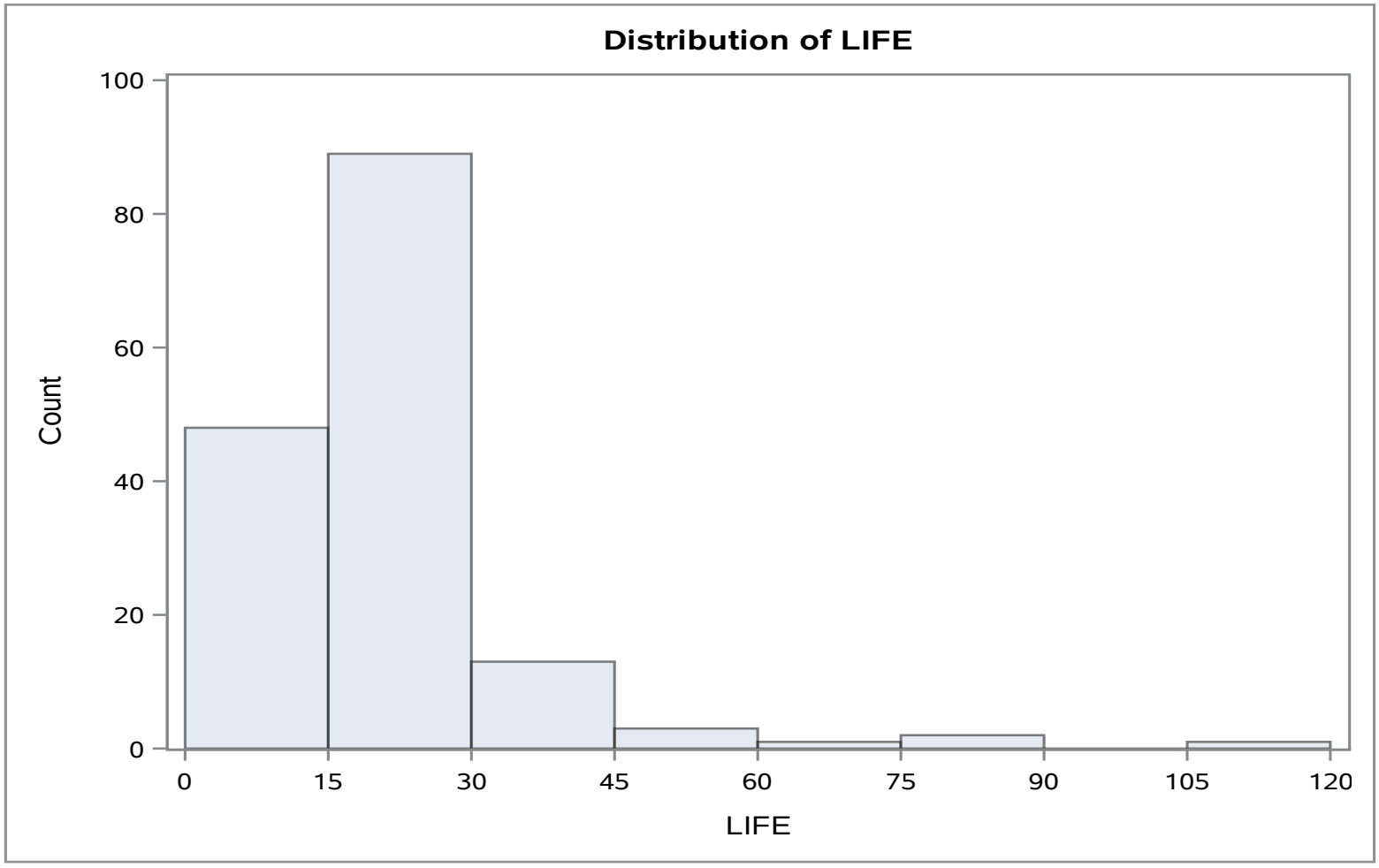
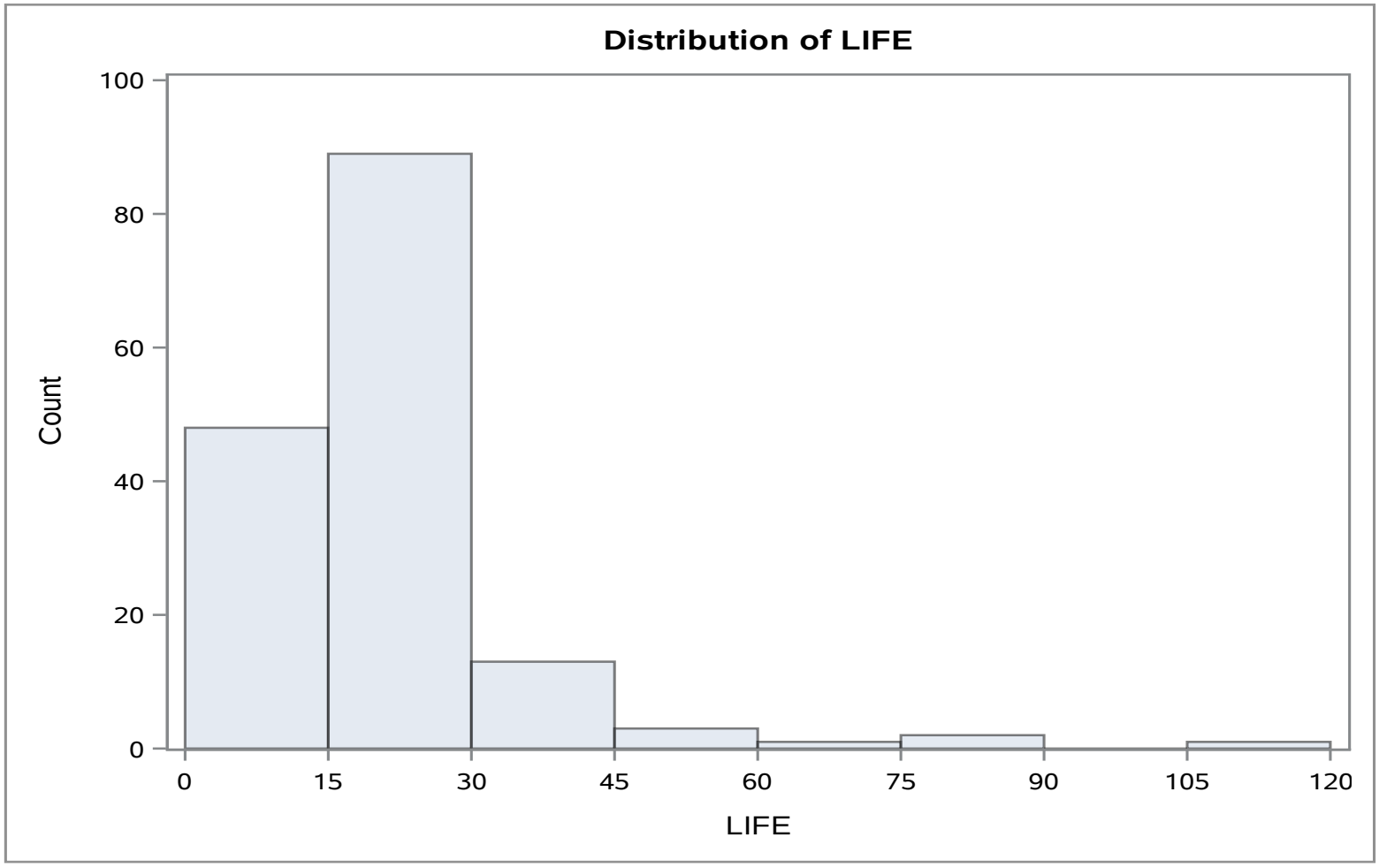


Figure: 5: Frequency histogram of the lifetimes of 157 electrical transformers.



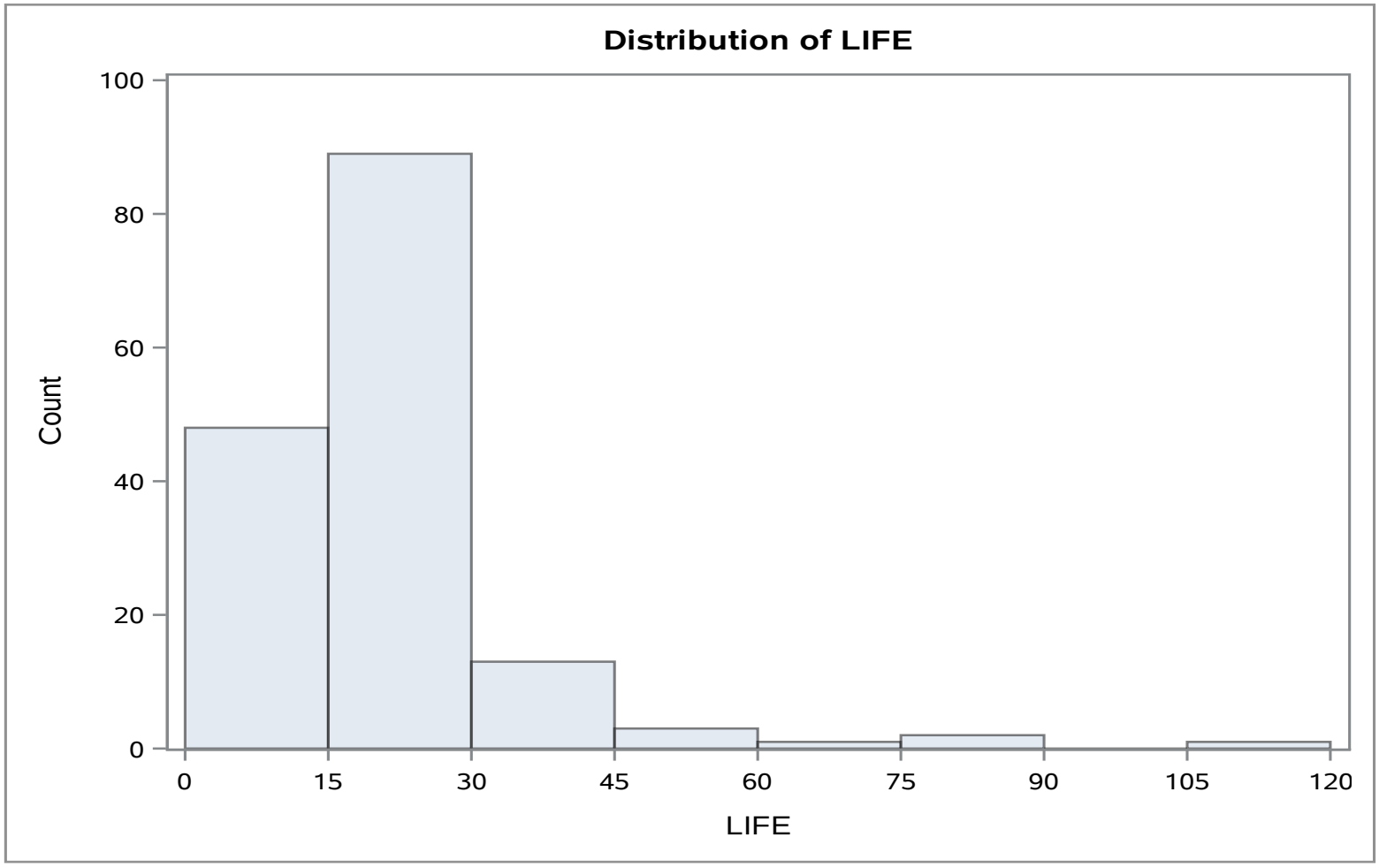
Modality: How many peaks? One.

≈



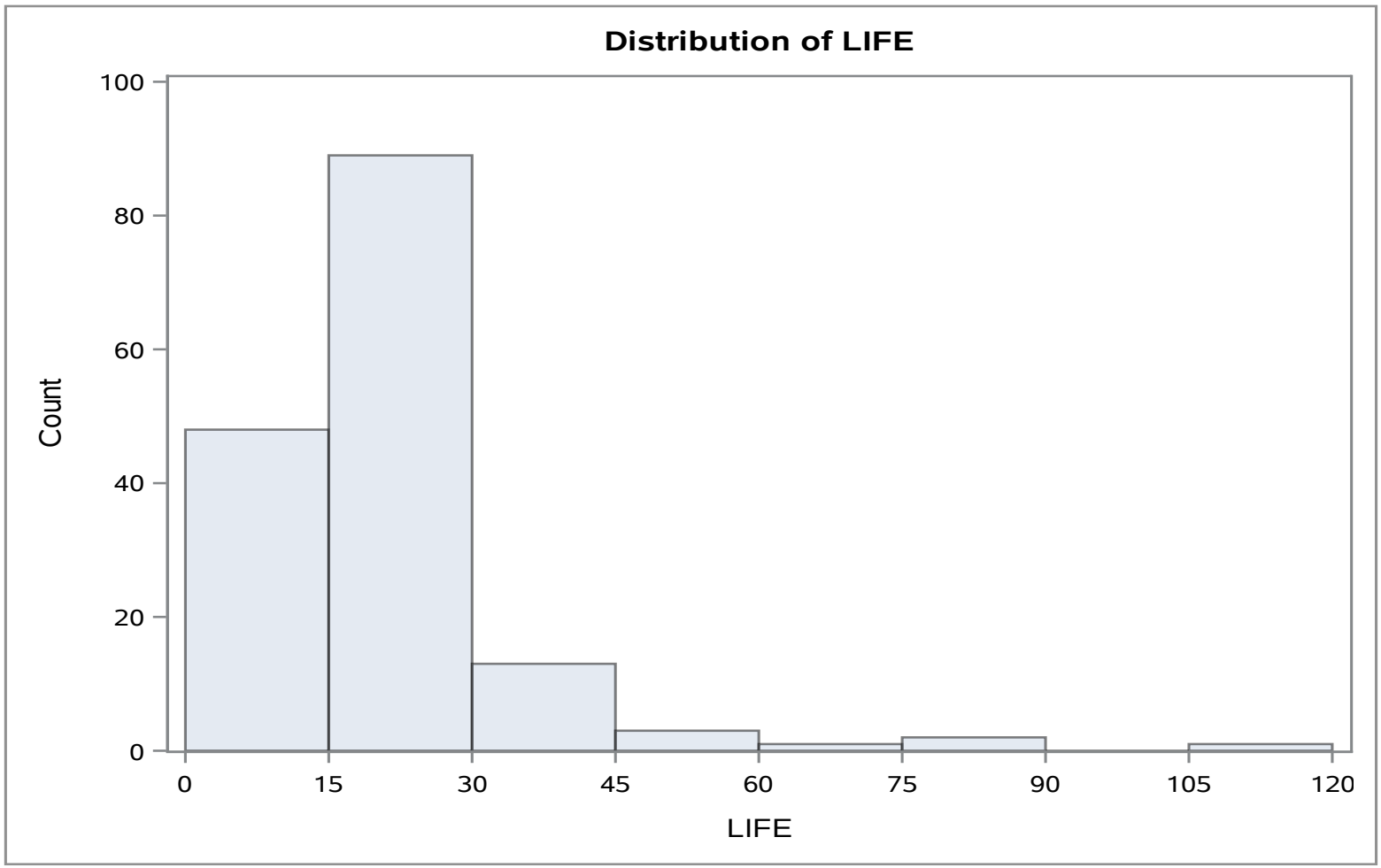
Symmetry: Mirrored about some vertical line? No. Right skewed.

≈



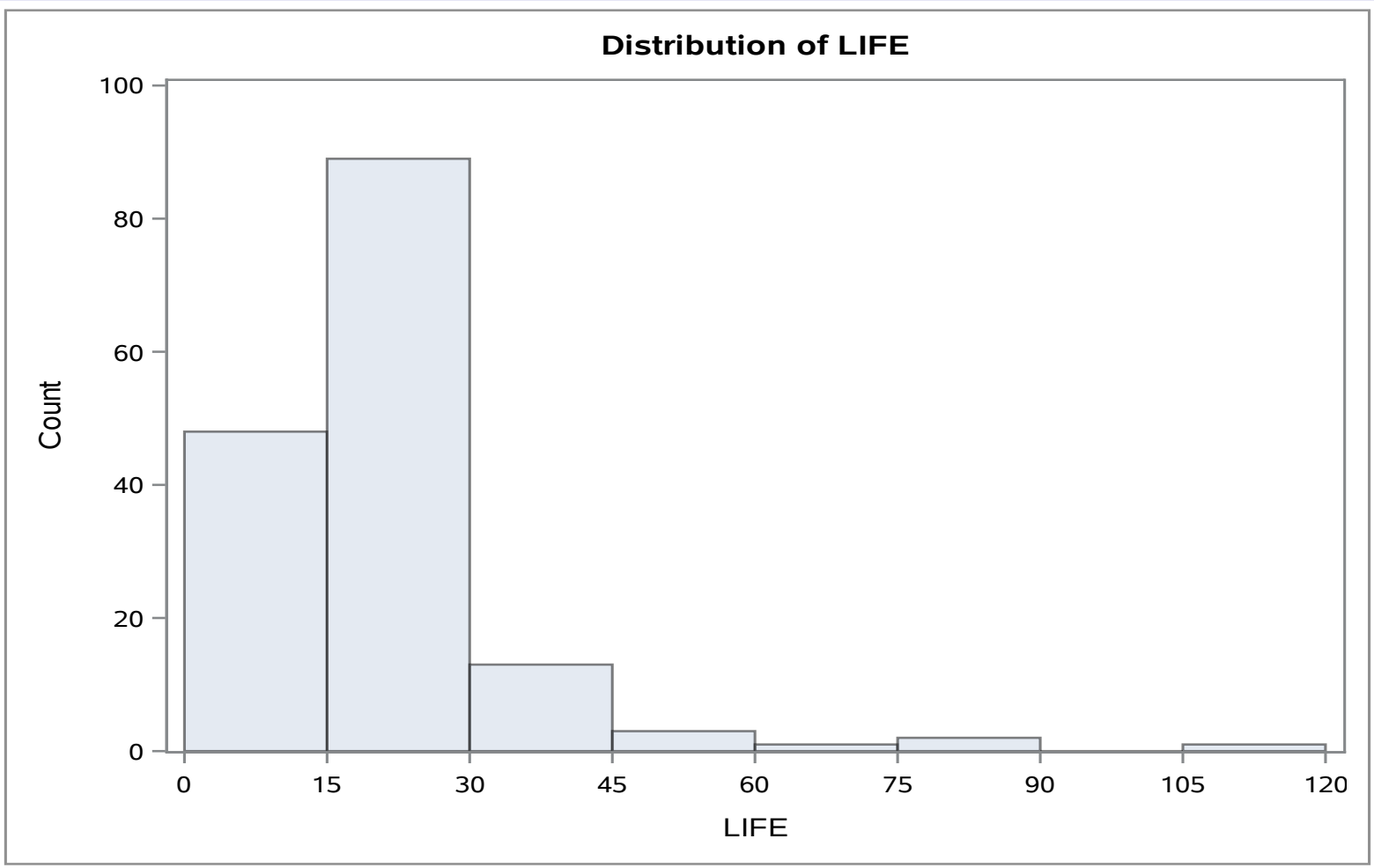
Center: Can you find one? Where is it? No clear center.

≈



Spread: How spread out are the data? A great majority lie between 0 and 30. Virtually all between 0 and 45.

≈



Pattern and deviations: What are the main patterns? What points don't follow these patterns? Unimodal and heavily right skewed. Virtually all between 0 and 45. A few extreme values in the 105-120 range.

Analyzing Frequency Histograms

The histogram of the transformer lifetimes is uni-modal and right-skewed. Transforming (no pun intended!) skewed data by taking a function of the values rather than the values themselves, can sometimes make it symmetric. Figure 6 shows a frequency histogram of the natural logs of the transformer lifetimes.

$\approx$

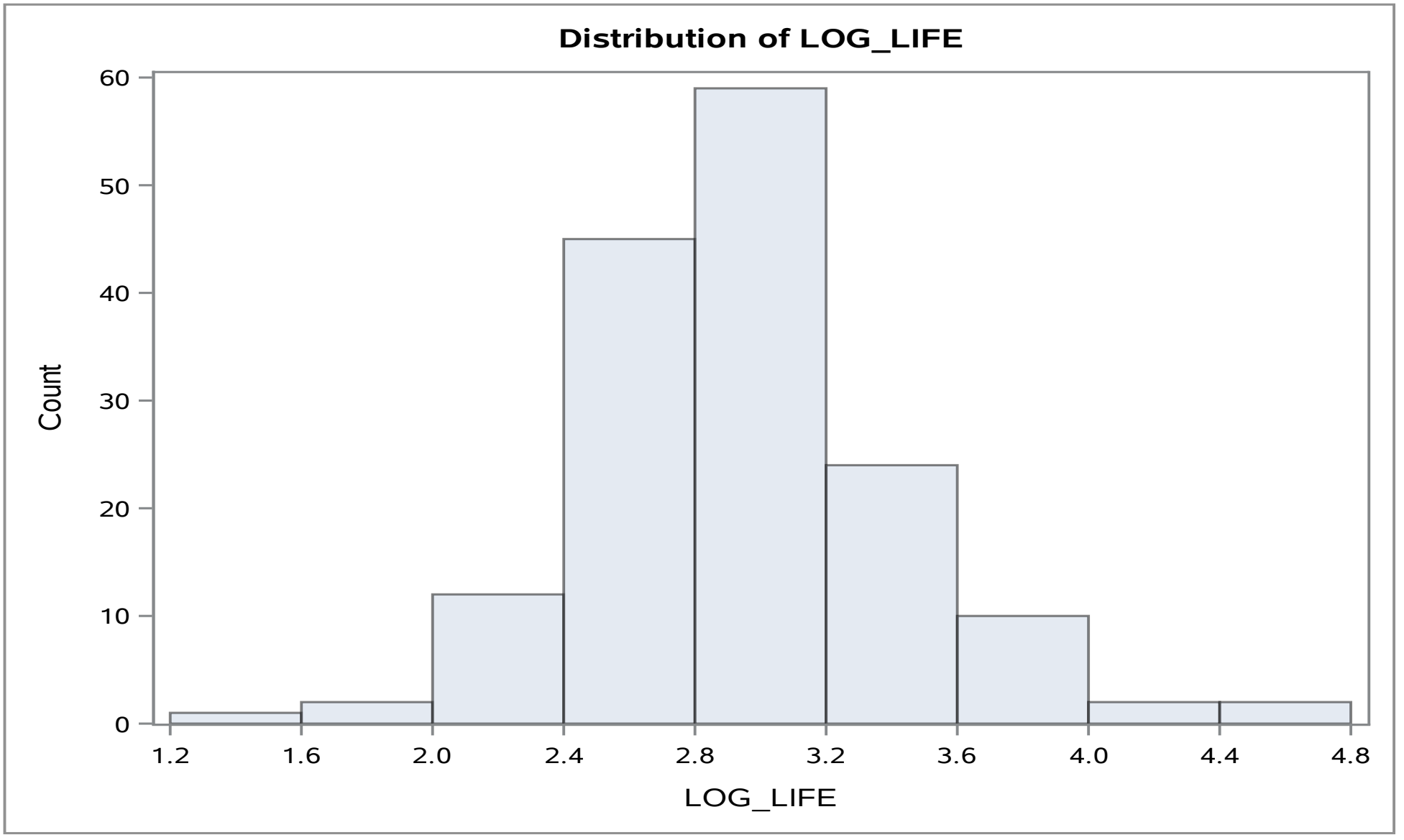
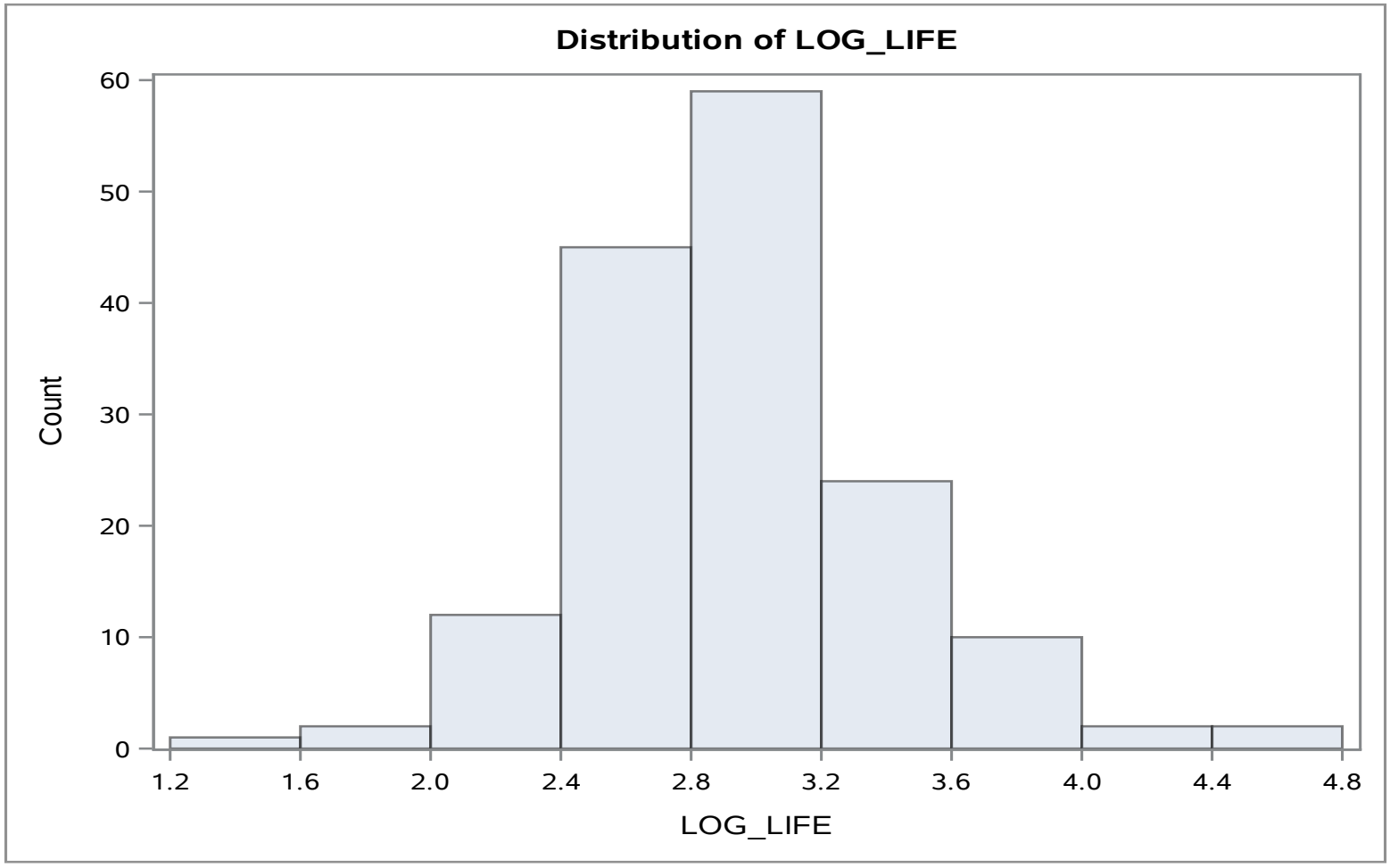
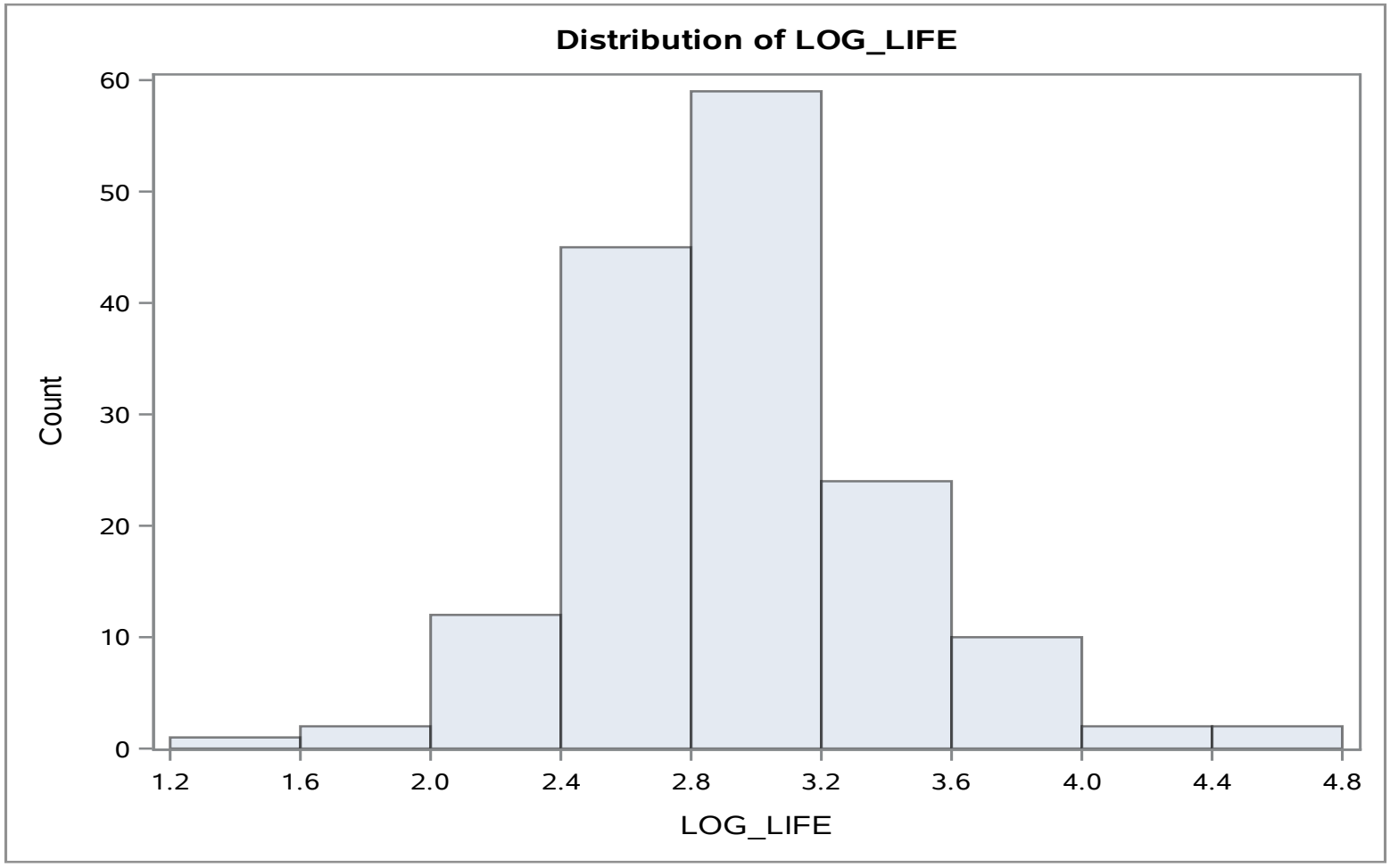


Figure: 6: *Frequency histogram of the natural logs of the lifetimes of 157 electrical transformers.*



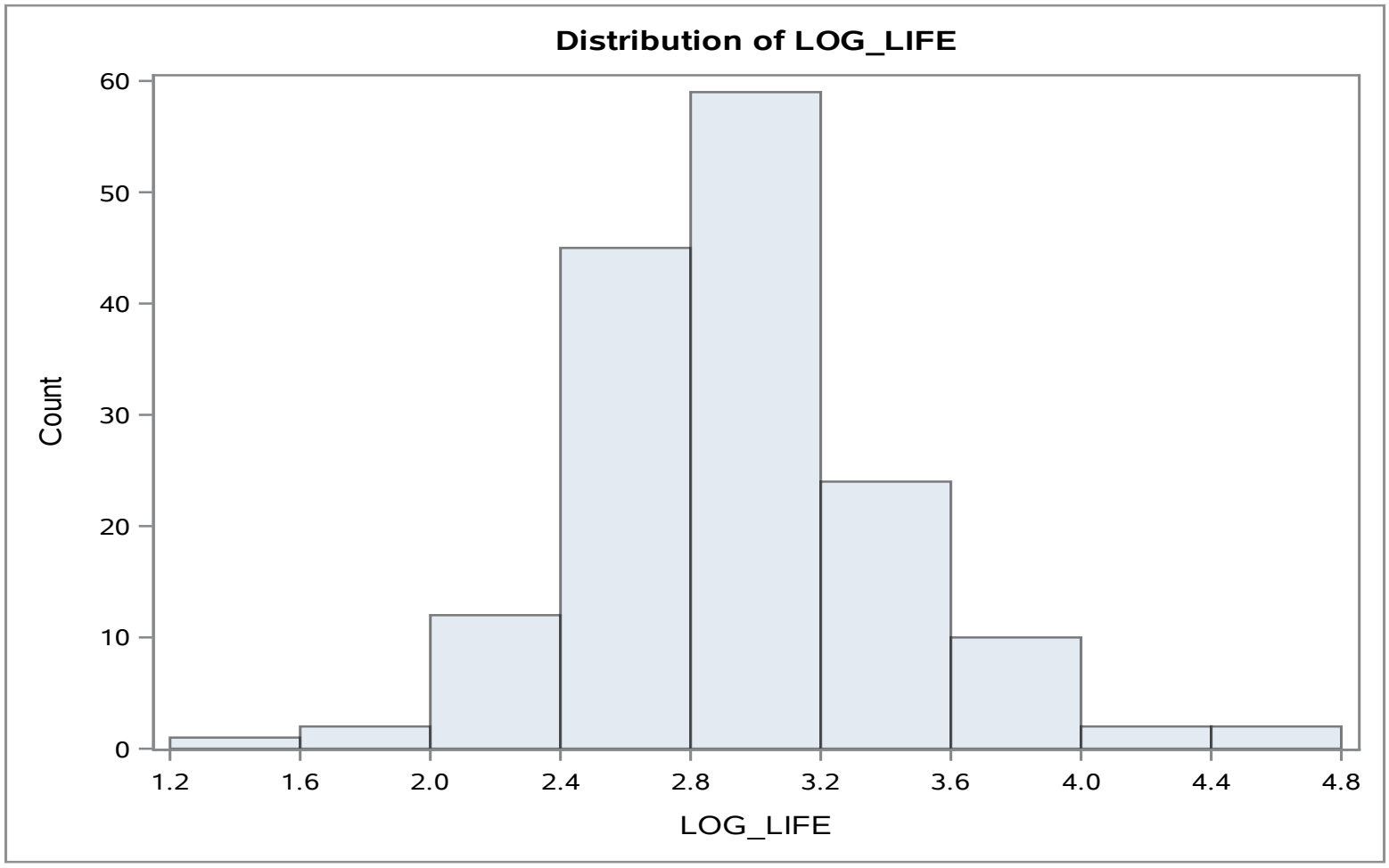
Modality: How many peaks? One.

≈



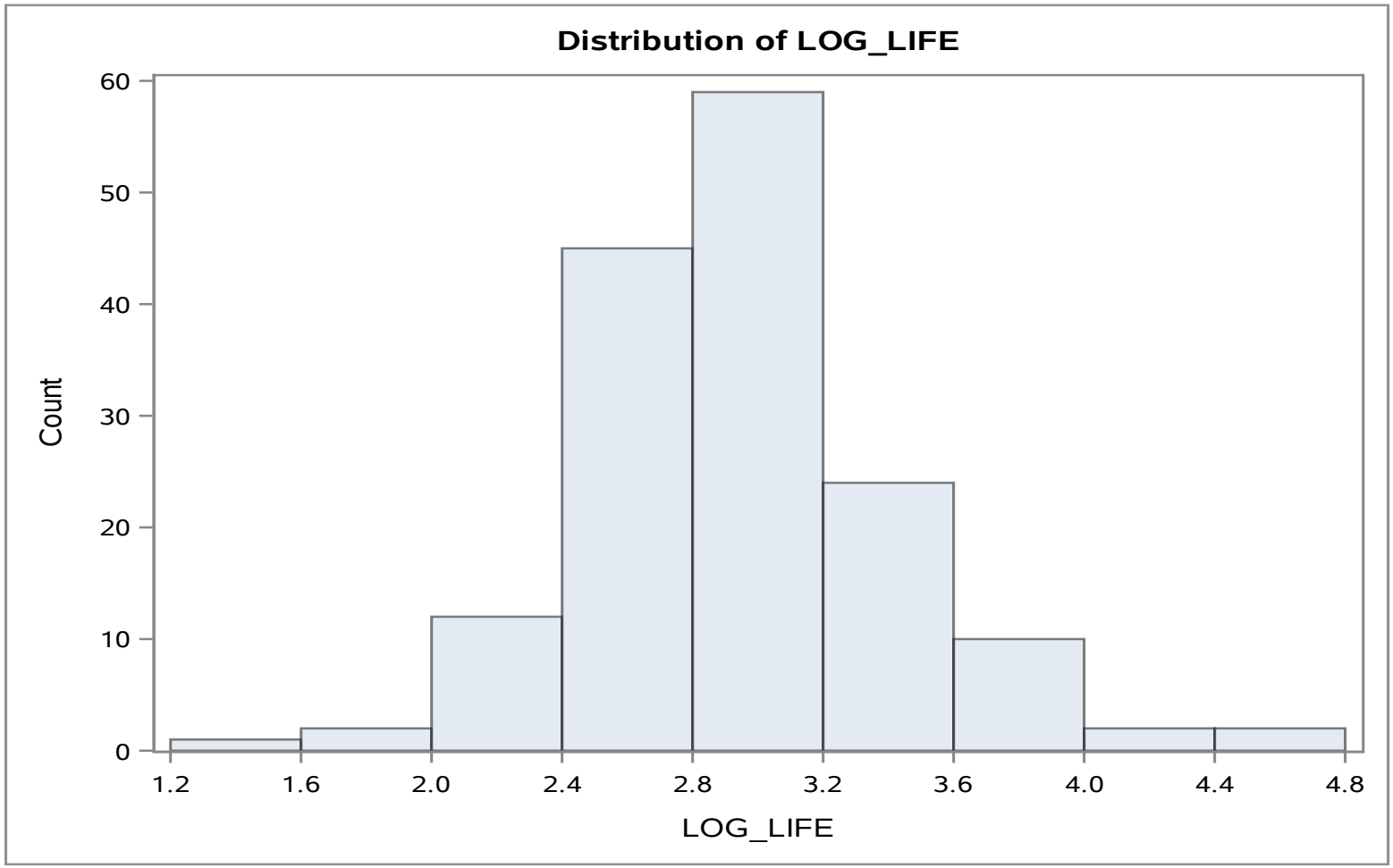
Symmetry: Mirrored about some vertical line? Yes.

≈



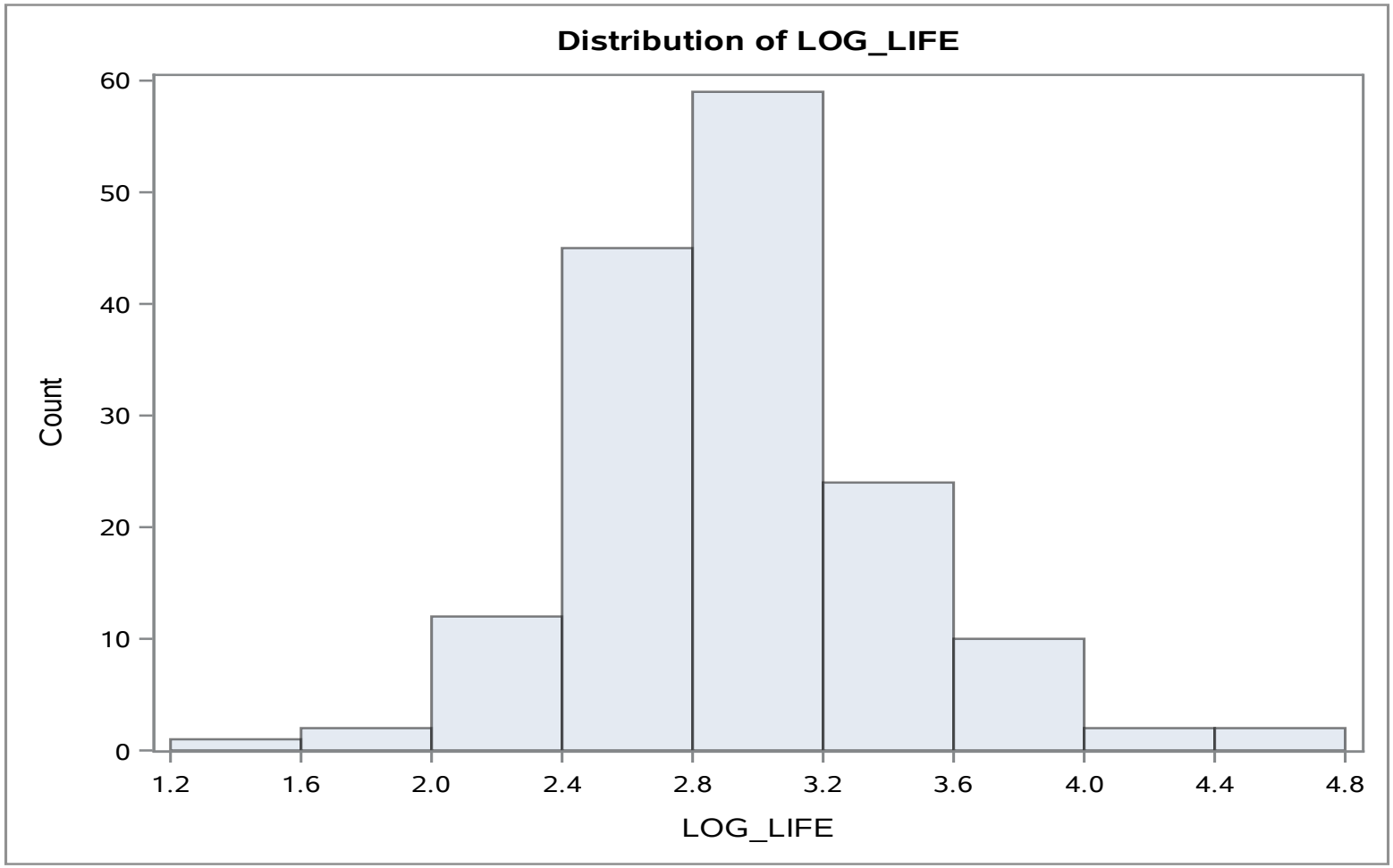
Center: Can you find one? Where is it? Yes. Around 3.0

≈



Spread: How spread out are the data? Most data between 2.0 and 4.0. Range is from 1.2 to 4.8.

≈



Pattern and deviations: What are the main patterns? What points don't follow these patterns? Unimodal and roughly symmetric. No rogue points evident.

≈

Analyzing Frequency Histograms

Example: Figure 7 shows a frequency histogram of a set of 222 intervals between eruptions of the Old Faithful geyser in Yellowstone National Park (data in sasdata.geyser1).

≈

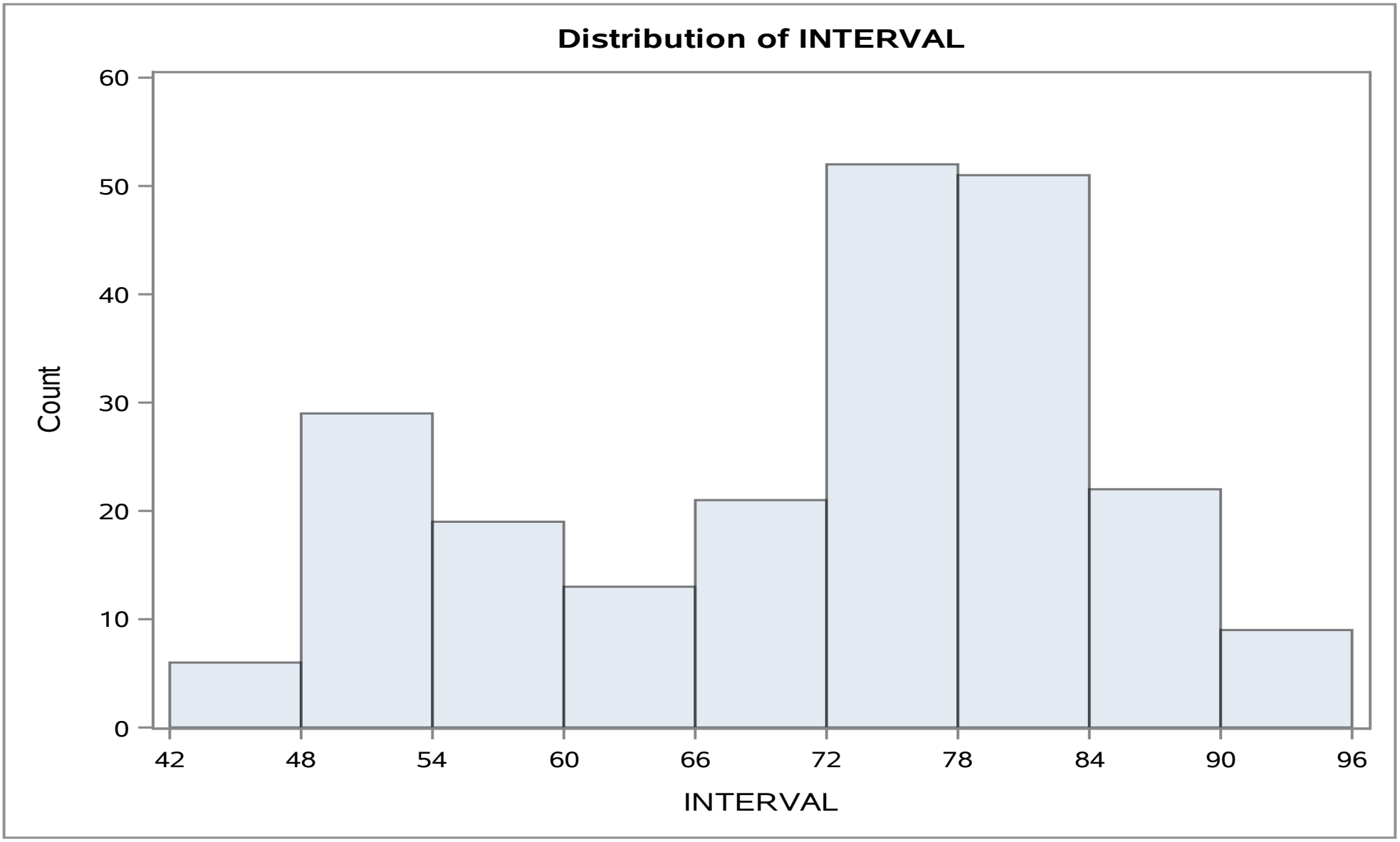
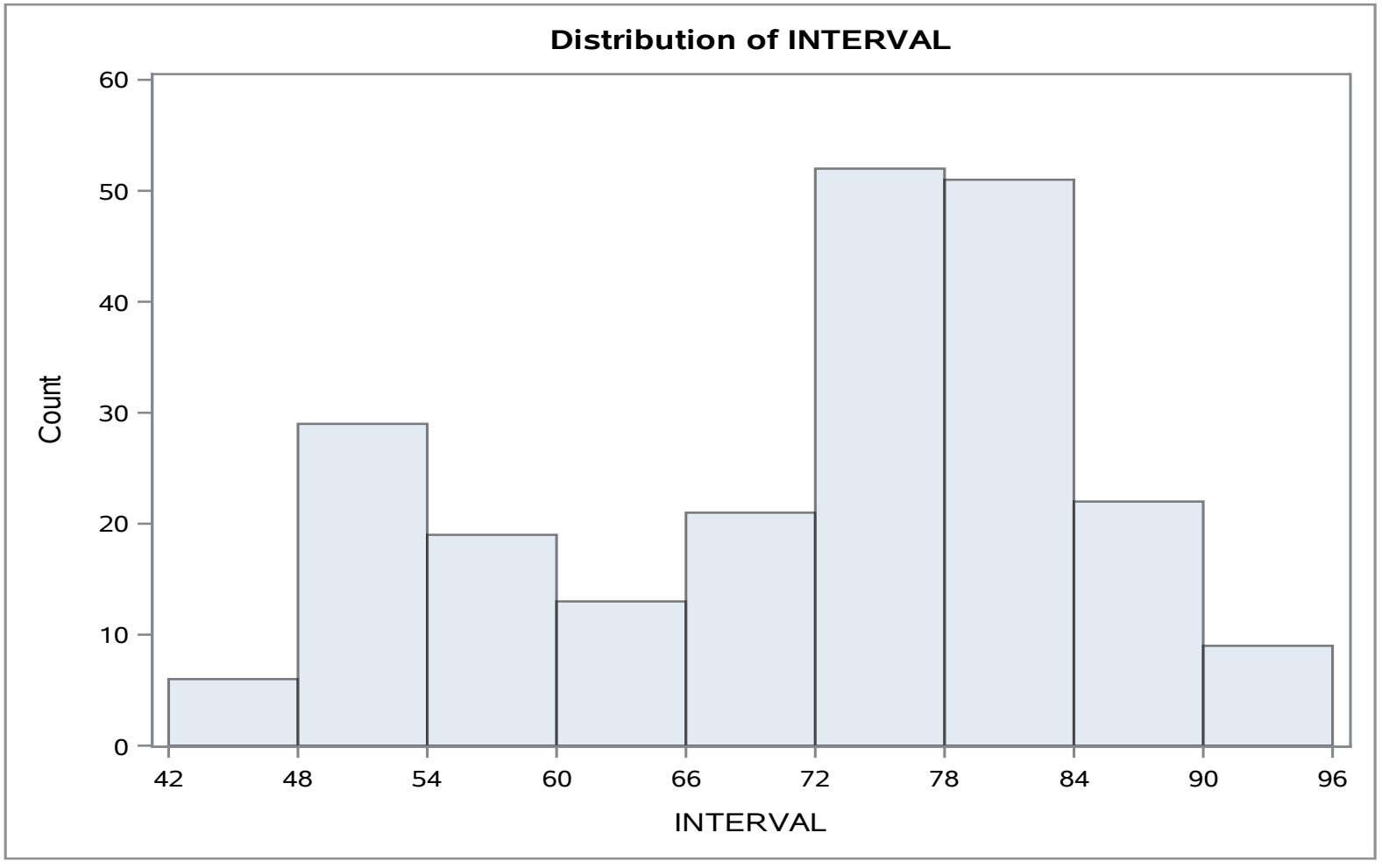
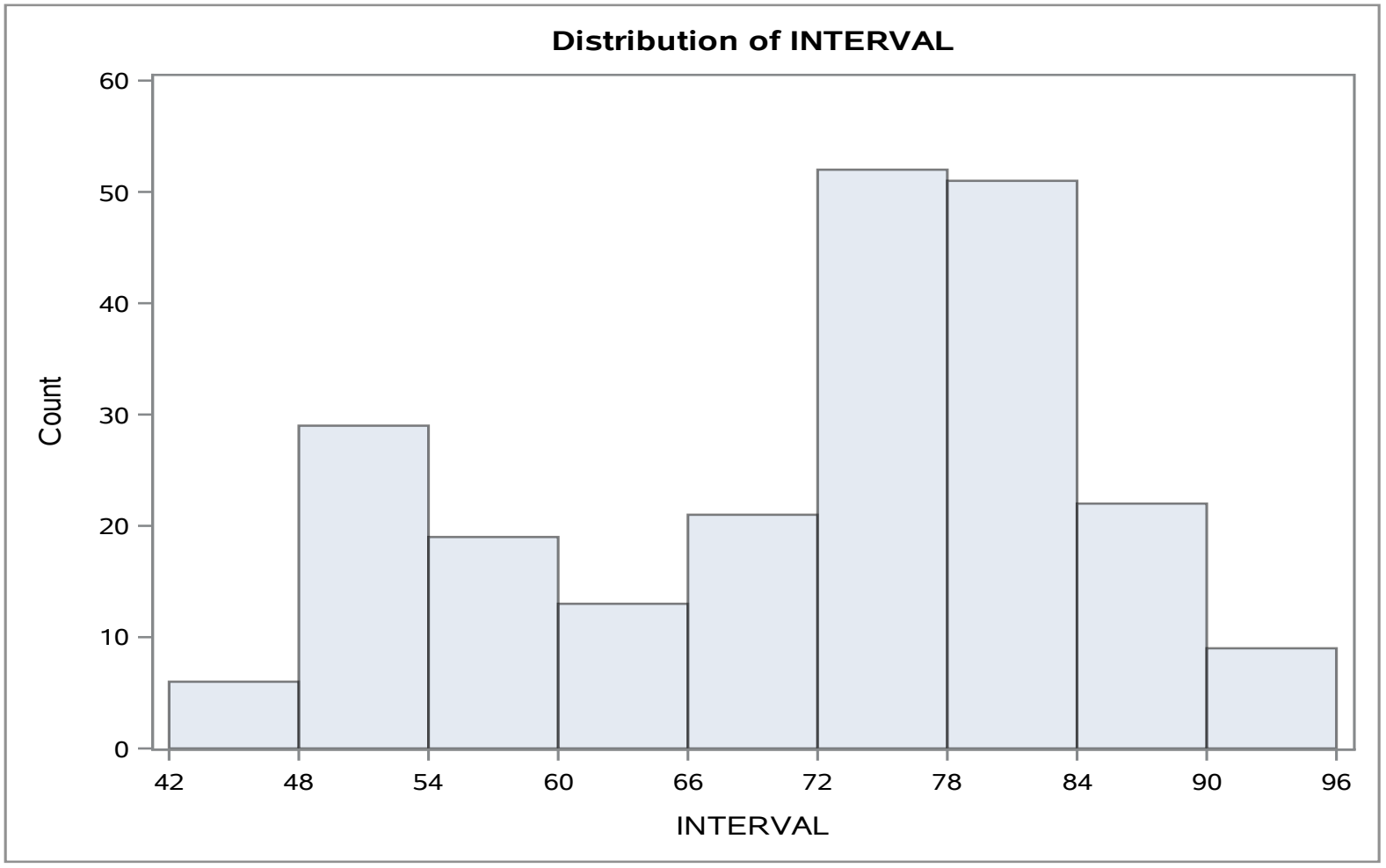


Figure: 7: *Frequency histogram of a set of 222 intervals between eruptions of the Old Faithful geyser in Yellowstone National Park.*



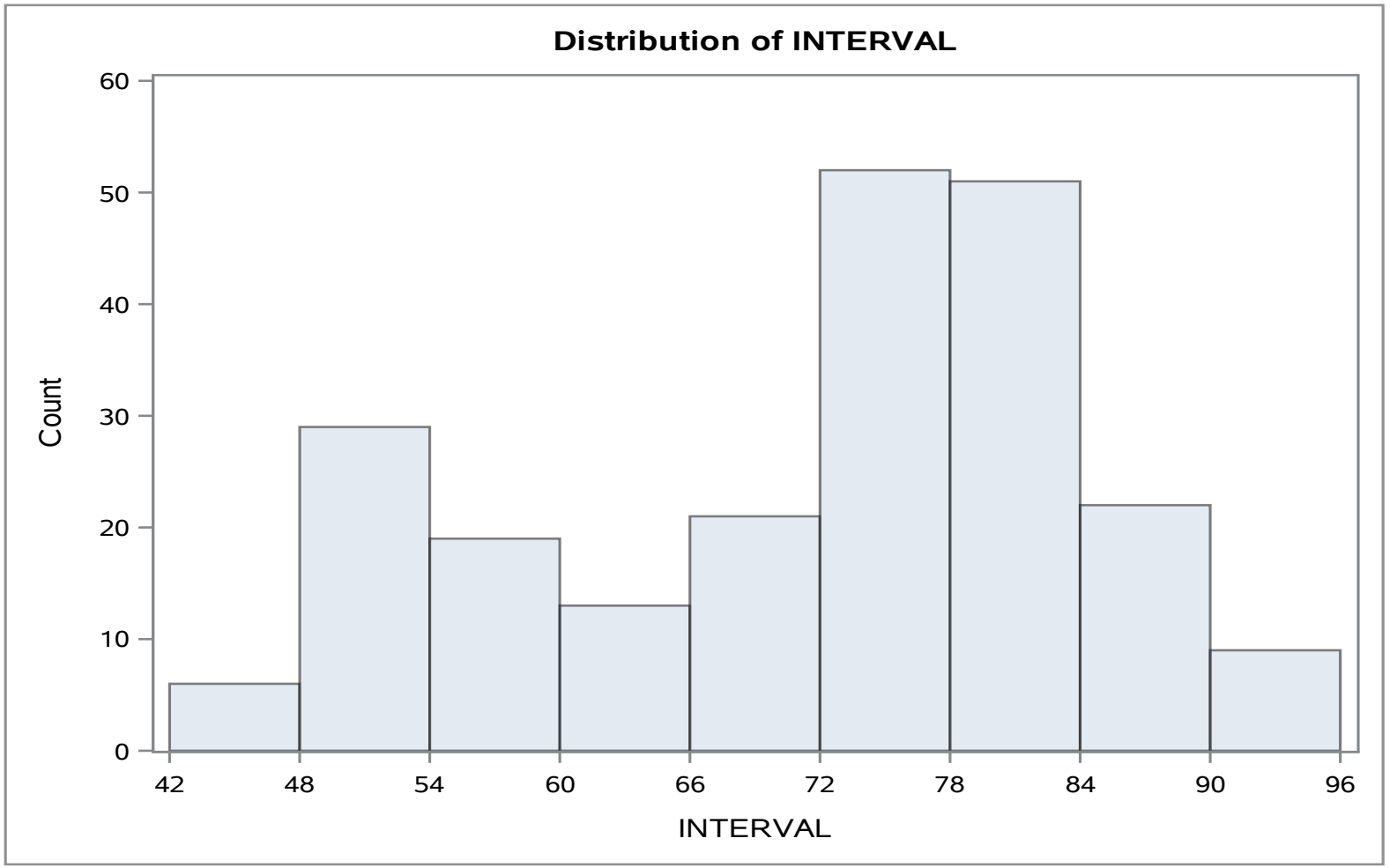
Modality: How many peaks? Two.

≈



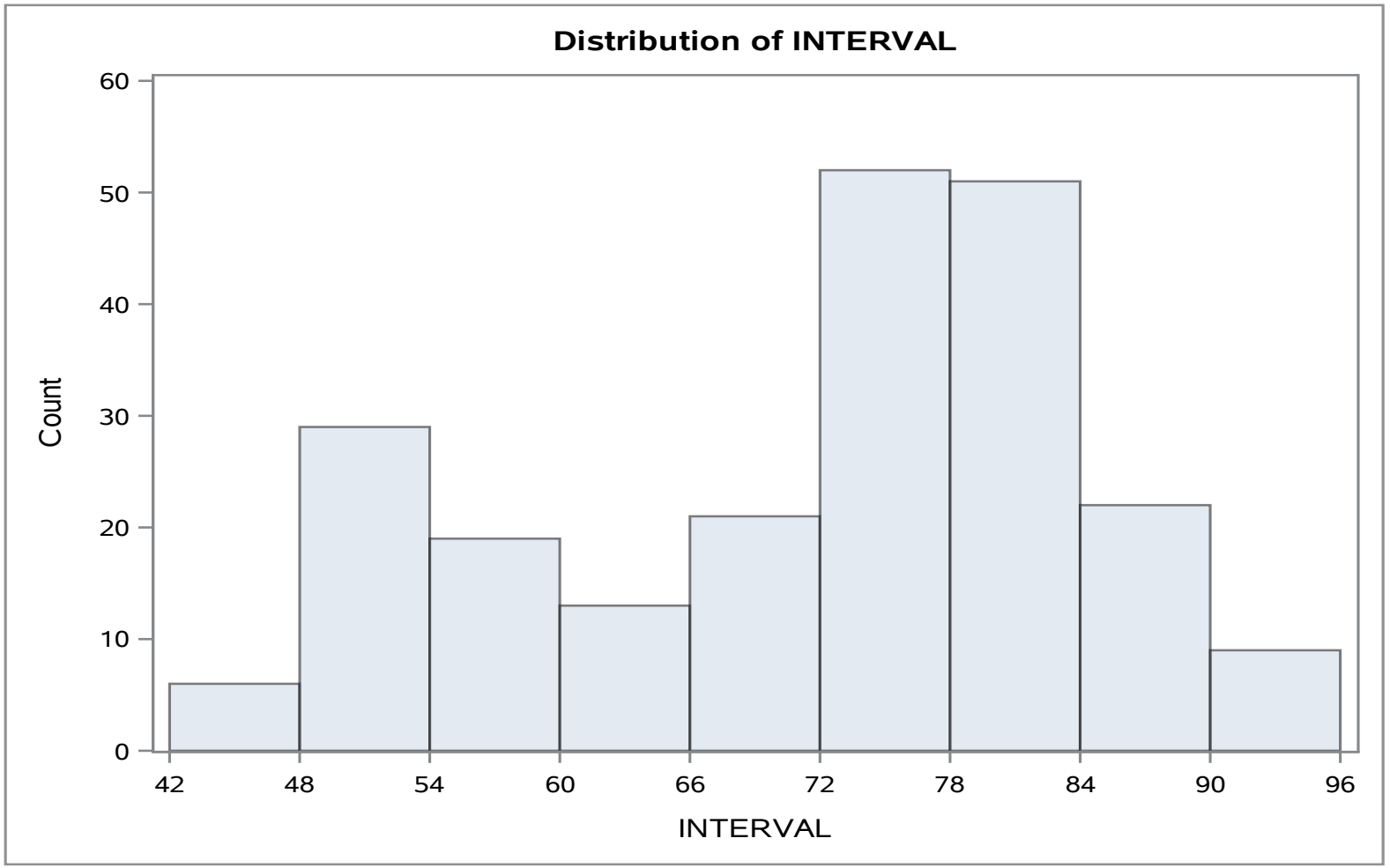
Symmetry: Mirrored about some vertical line? No. Left peak is smaller than the right one.

≈



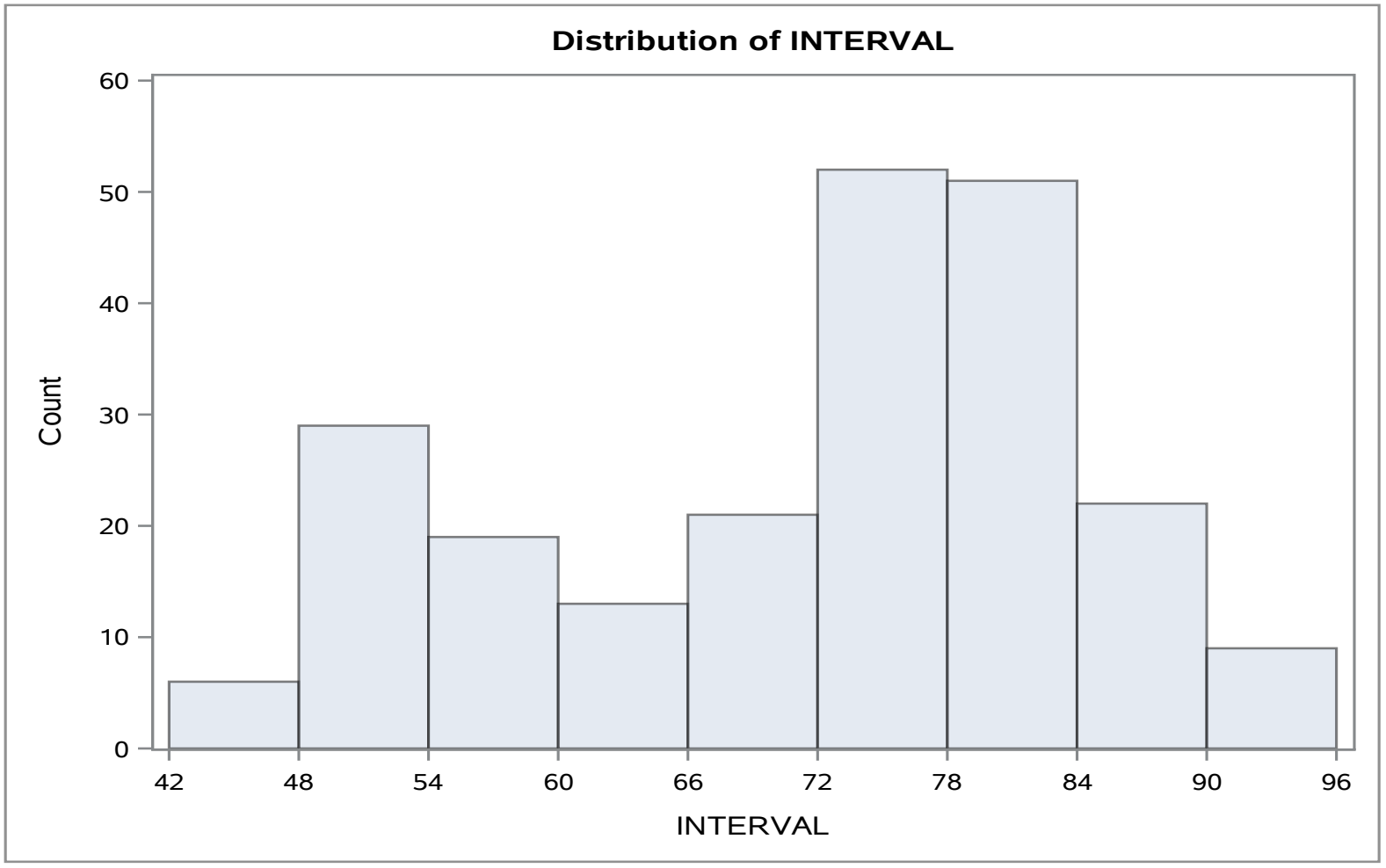
Center: Can you find one? Where is it? No, there are two centers. For the left peak around 48-54 and for the right 72-84.

≈



Spread: How spread out are the data? The left peak ranges from 42 to around 63 and the right one from 63 to 96.

≈



Pattern and deviations: What are the main patterns? What points don't follow these patterns? Bimodal. No rogue points evident.

≈

Analyzing Frequency Histograms

The histogram is **bi-modal**, with modes at approximately 48-54 minutes and 72-84 minutes. This shows that the intervals between eruptions tend to be in the neighborhood of 50 minutes or 80 minutes. The size of the modal peak around 80 minutes is larger, which tells us intervals around 80 minutes are more likely than those around 50 minutes. It turns out that the time until the next eruption is related to the duration of the present eruption, but that's a story for another day.

$\approx$

Analyzing Frequency Histograms

What Might Cause These Histogram Shapes?

- Symmetric, unimodal: measurement errors; data from homogeneous populations
- Skewness: lower bound, no upper bound, or vice-versa
- Multi-modal: data from nonhomogeneous populations
- Short-tailed: mixture of process streams (picture, next slide)

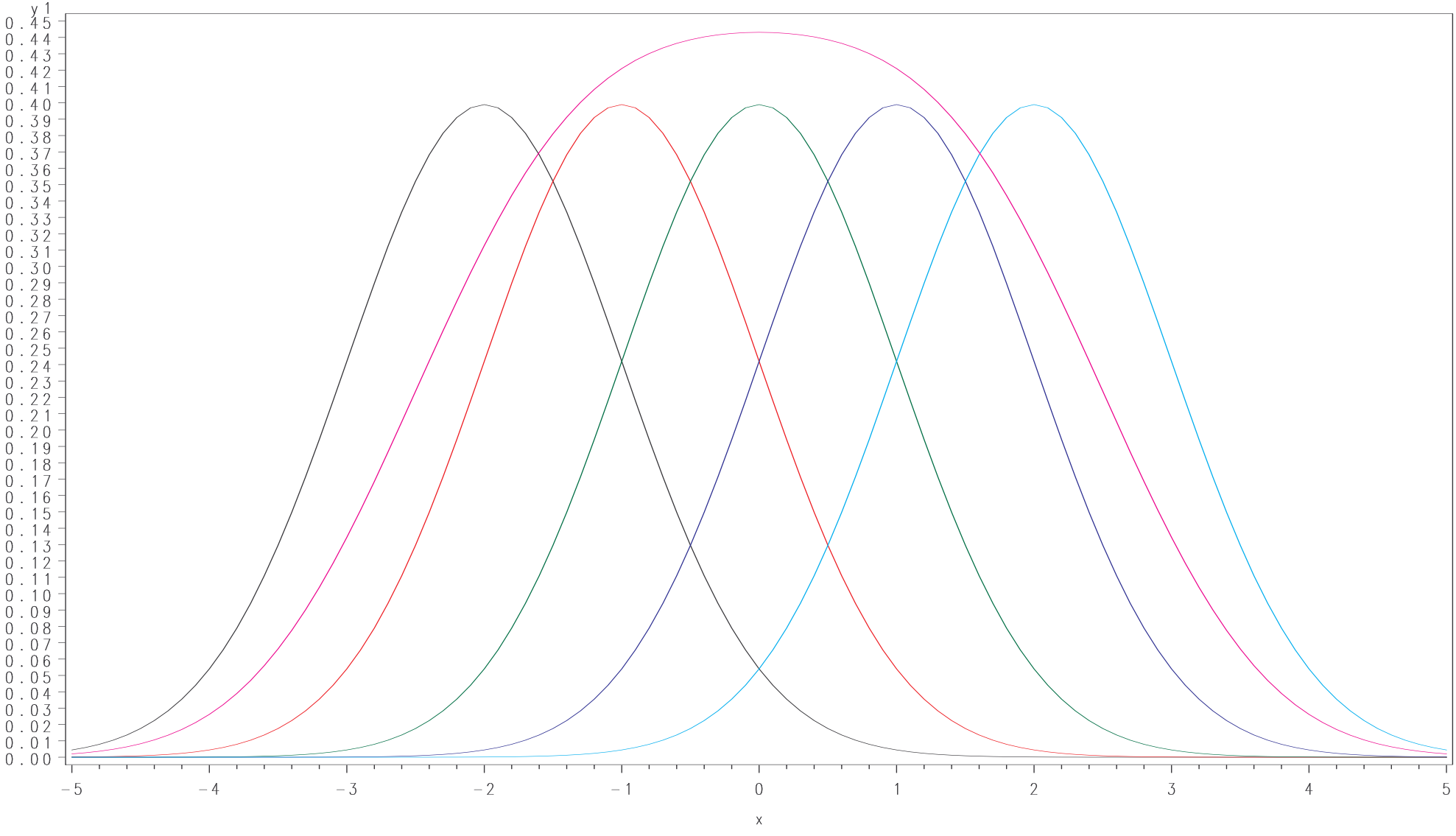


Figure: 8: Short-tailed distribution.

Summary measures of quantitative data

While data should always be graphed as a first step, numerical summaries can be a valuable addition to any analysis.

However, it is important to choose summary measures appropriate to the pattern of variation in the data.

The most commonly used measures fall into two categories: measures of location, and measures of spread.

In the formulas that follow, we assume there are n data values denoted $y_1, y_2, \ldots, y_n$.

$\approx$

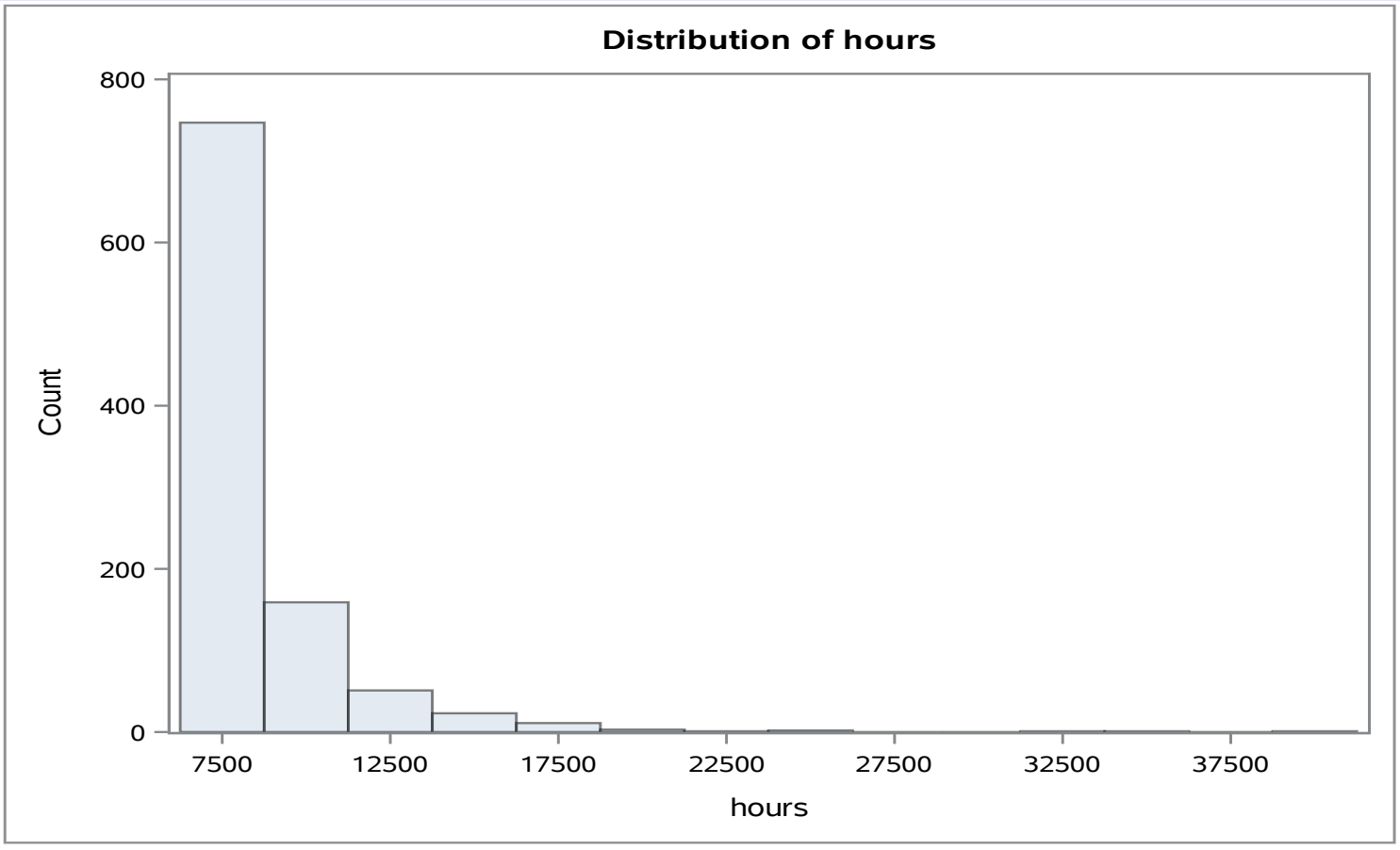
Summary measures of quantitative data: location

- Mean: Average; denoted $\overline{y}$; Formula: $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Median: Halfway point (point below which at least half the data values lie, and above which at least half the data values lie); denoted Q_2.
- Mode: Location of a modal bar on a frequency histogram.
- Quantiles: For a number q between 0 and 1, the q^{th} quantile is a value at or below which a proportion at least q of the data lies and at or above which a proportion at least $1 - q$ of the data lies.
- Quartiles: These are the .25, .5 and .75 quantiles; denoted Q_1, Q_2, and Q_3.

Why define the mode this way?

The mode is most often defined as the value that occurs most frequently in the data set. Unfortunately, this is not always related to the data distribution in a sensible way. The histogram on the next slide plots lifetimes in hours of 1000 led lightbulbs (found in the data set sasdata.whats_wrong_with_the_mode).

$\approx$



The mode computed in the standard way is 17682.49, which is not at all representative of location in the displayed pattern.
The mode computed as the center of the modal bar is 7500, a much better summary of location.

Summary measures of quantitative data: spread

- Mean absolute deviation: "Average" distance from mean:

$$\frac{1}{n-1}\sum_{i=1}^{n}|y_i-\overline{y}|$$

- Standard deviation (RMS): Square root of "average" squared distance from mean:

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

  The square of the standard deviaion, called the **variance** is also used to measure spread.

- Interquartile range (IQR): $Q_3 - Q_1$; the range of middle 50% of data.

What you should know about choosing summary measures

- Some patterns of variation, such as unimodal/symmetric, skewed, short tails and multi-modal, occur often in practice, and are often associated with specific population characteristics or methods of data generation.
- The summary measures most appropriate for a set of data depend on its pattern of variation. For this reason, a graph of the pattern of variation should always be viewed before choosing summary measures, and should accompany summary measures in presenting data.

What you should know about choosing summary measures

Some good rules of thumb for choosing summary measures are:

Pattern	Location	Spread
unimodal and symmetric	mean	standard deviation
unimodal and skewed	median (or mode)	interquartile range
multi-modal	modes	range of modal peaks

What you should know about choosing summary measures

Figures 9, 10 and 11 show a frequency histogram and summary measures for the transformer data, logged transformer data, and geyser data, respectively. Which measures are appropriate for each? (SAS code for Figure 9 here)

$\approx$

Moments			
N	157	Sum Weights	157
Mean	21.3441583	Sum Observations	3351.03286
Std Deviation	13.519659	Variance	182.78118
Skewness	3.35201013	Kurtosis	16.3104457
Uncorrected SS	100038.84	Corrected SS	28513.8641
Coeff Variation	63.3412609	Std Error Mean	1.07898625

Basic Statistical Measures			
Location		Variability	
Mean	21.34416	Std Deviation	13.51966
Median	18.05260	Variance	182.78118
Mode	.	Range	107.25961
		Interquartile Range	10.32325

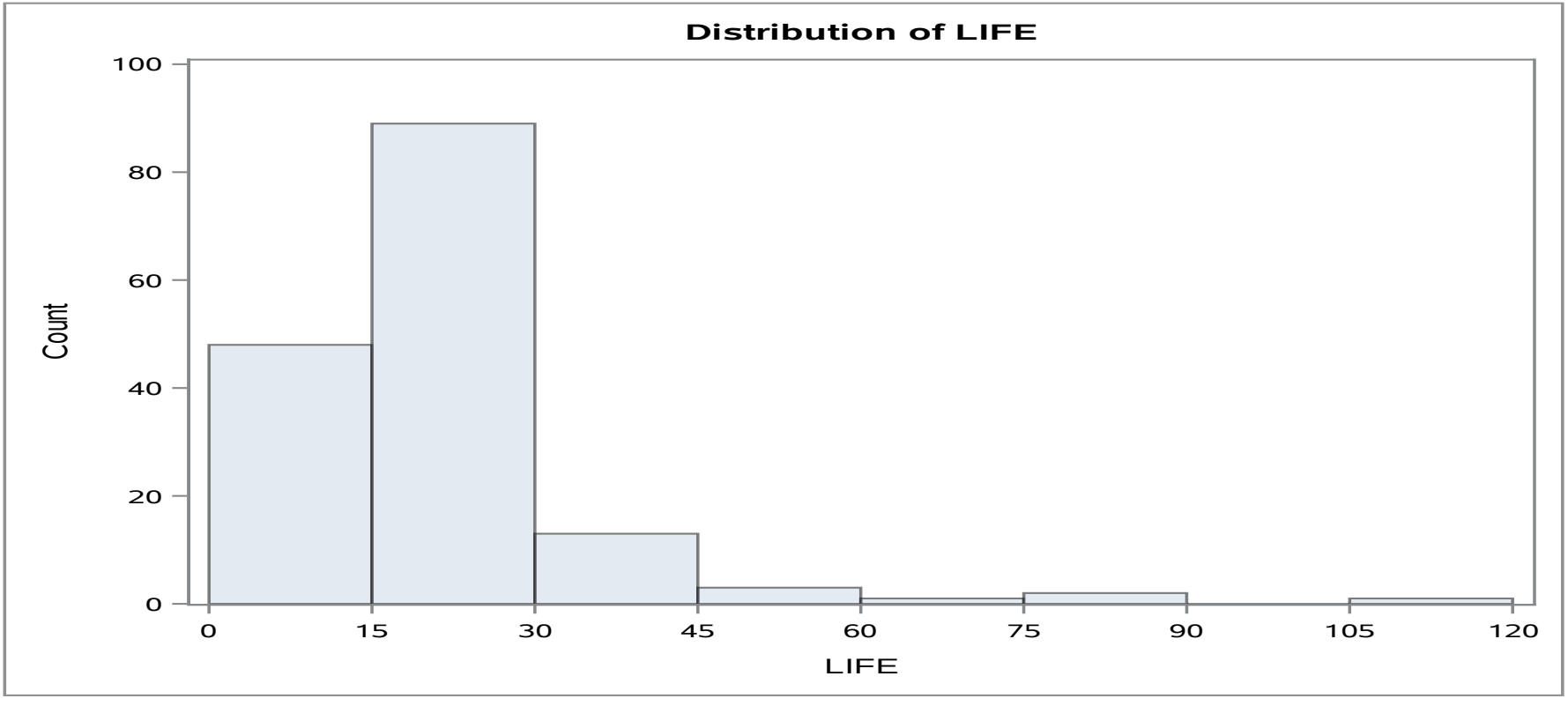


Figure: 9: *Histogram and summary measures, transformer lifetimes.*

Moments			
N	157	Sum Weights	157
Mean	2.93056479	Sum Observations	460.098671
Std Deviation	0.48630474	Variance	0.2364923
Skewness	0.48206193	Kurtosis	1.6400187
Uncorrected SS	1385.24176	Corrected SS	36.8927988
Coeff Variation	16.5942327	Std Error Mean	0.03881134

Basic Statistical Measures			
Location		Variability	
Mean	2.930565	Std Deviation	0.48630
Median	2.893290	Variance	0.23649
Mode	.	Range	3.27137
		Interquartile Range	0.56135

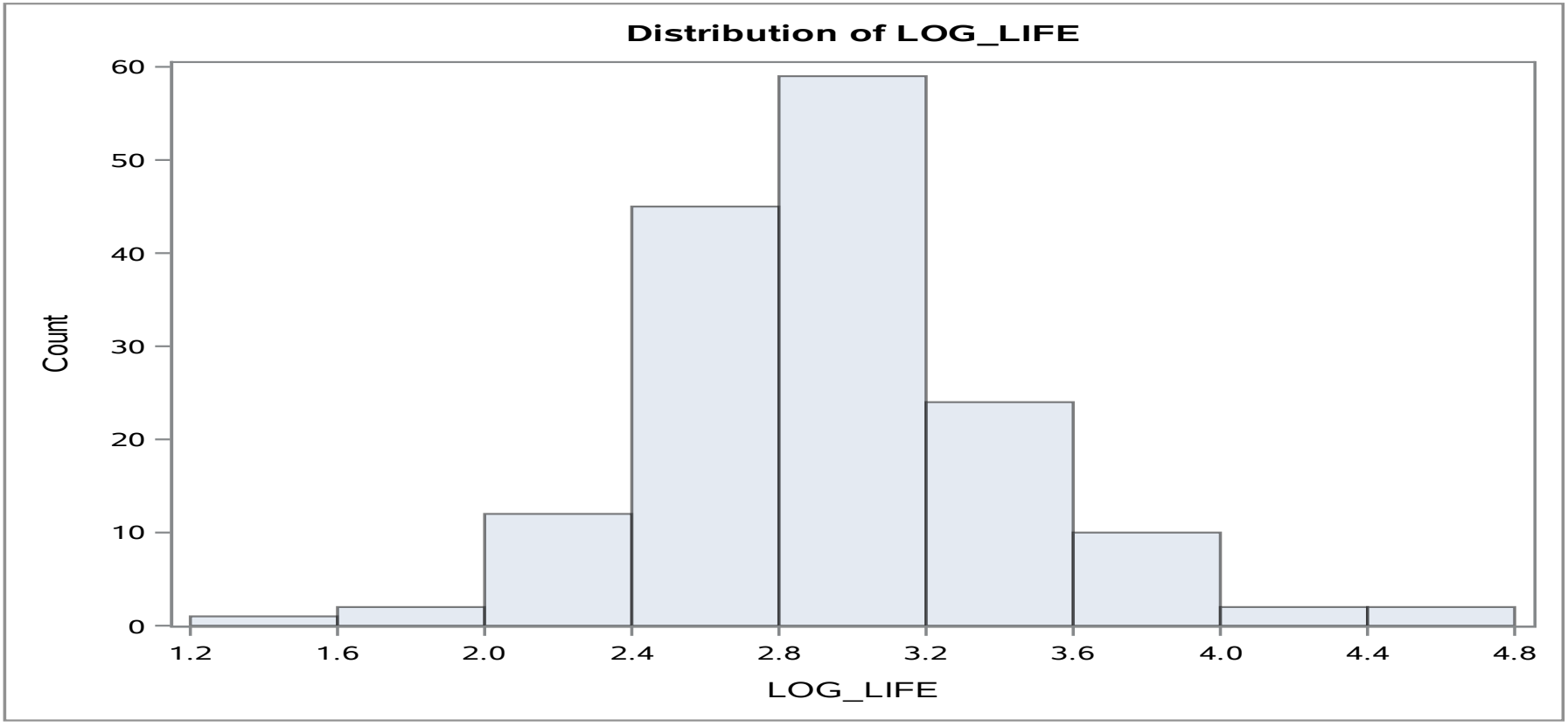


Figure: 10: Histogram and summary measures, logged transformer lifetimes.

Moments			
N	222	Sum Weights	222
Mean	71.009009	Sum Observations	15764
Std Deviation	12.7991767	Variance	163.818923
Skewness	-0.4855168	Kurtosis	-0.8856534
Uncorrected SS	1155590	Corrected SS	36203.982
Coeff Variation	18.0247223	Std Error Mean	0.85902449

Basic Statistical Measures			
Location		Variability	
Mean	71.00901	Std Deviation	12.79918
Median	75.00000	Variance	163.81892
Mode	75.00000	Range	53.00000
		Interquartile Range	21.00000

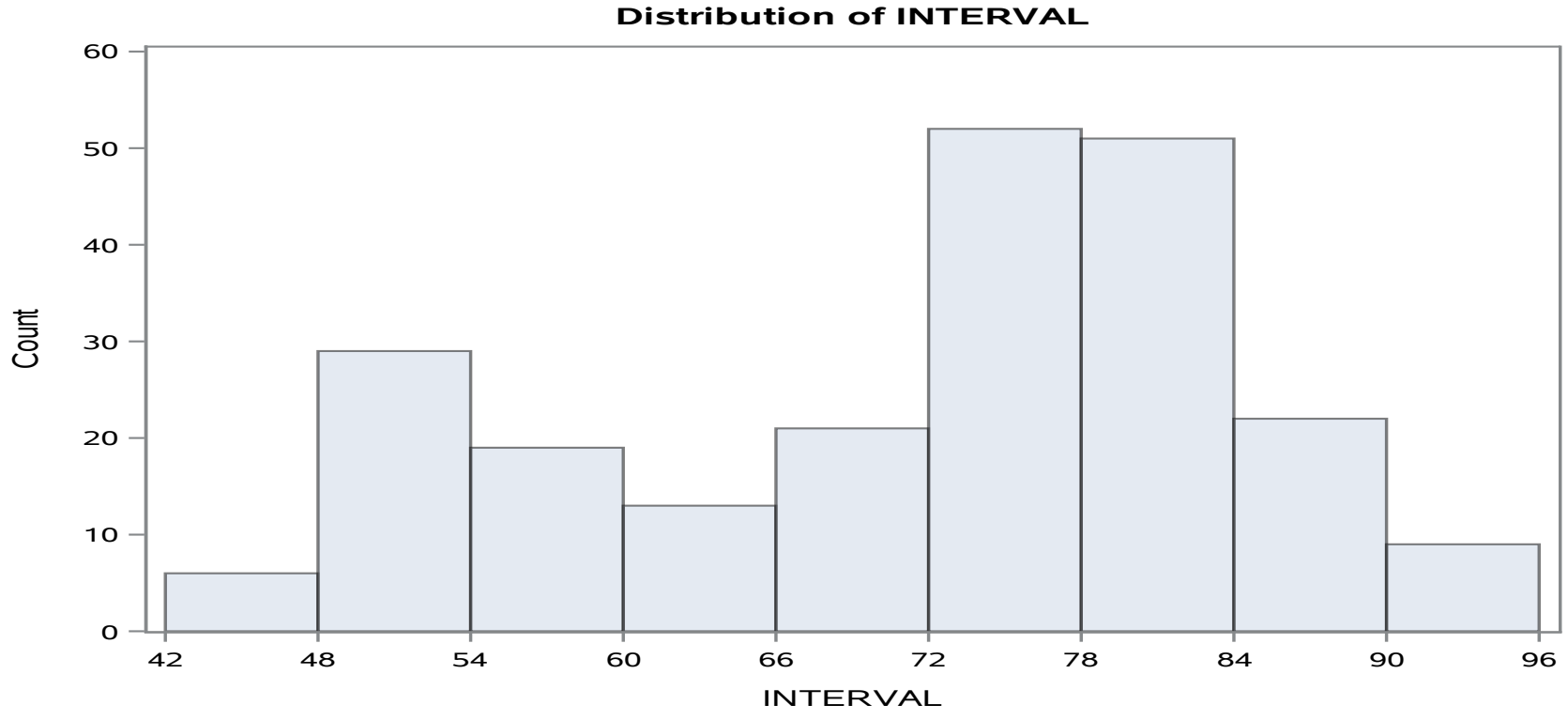


Figure: 11: Histogram and summary measures, geyser data.

Graphical Interpretations

We have already seen the graphical interpretation (indeed, definition) of a mode: the center of a modal bar on a frequency histogram.

There are nice conceptual interpretations of the mean and median as well as the following illustrates.

$\approx$

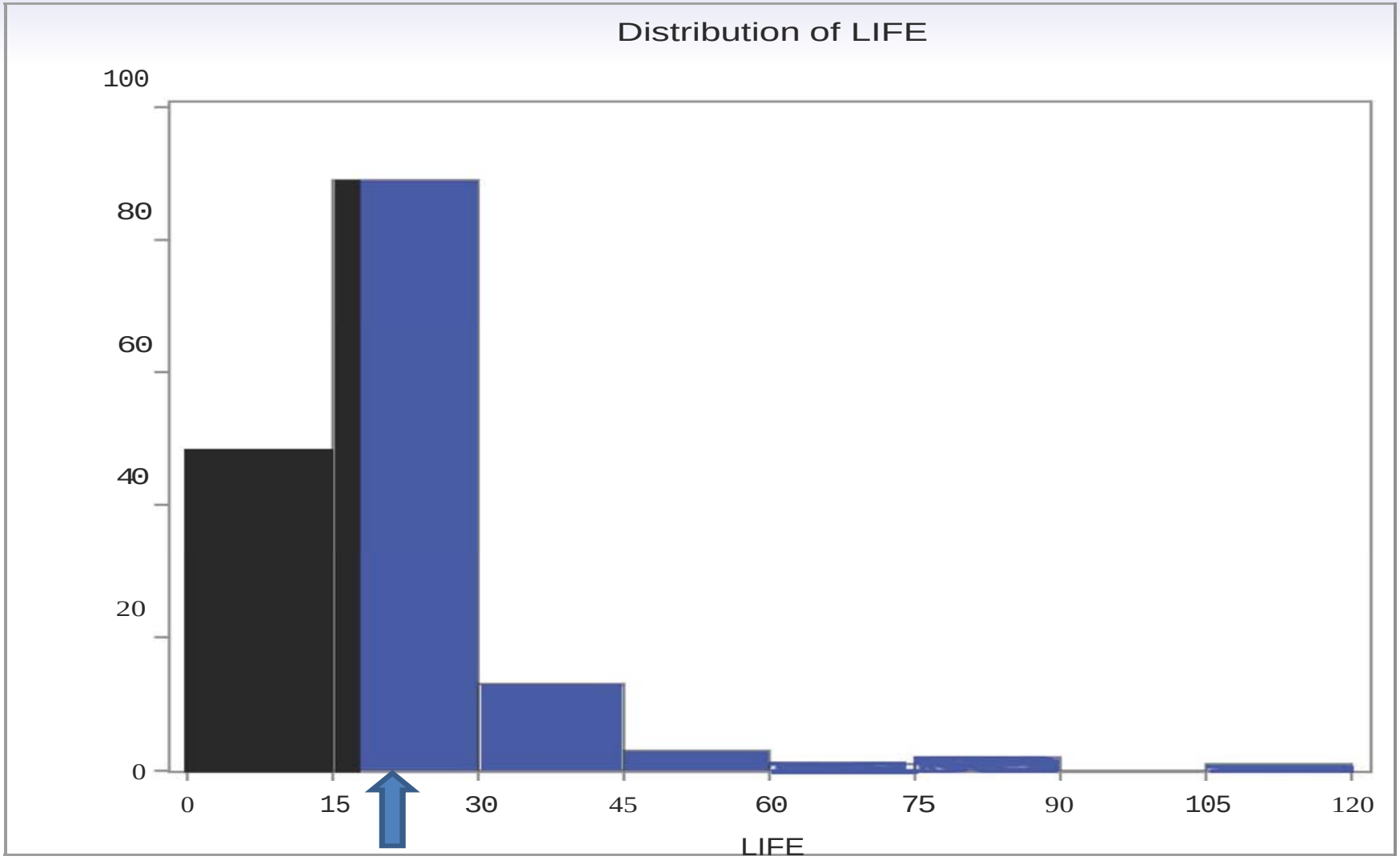


Figure: 11a: Conceptually, the median is the value that separates the histogram into equal areas (colored black and blue). Here, for the transformer lifetimes, it is 18.05. The mean is where the histogram "balances". Here, the arrow shows the location, $\overline{y} = 21.34$.

Outliers

- Outliers are extremely unrepresentative data.
- Outliers may be "bad" data values. Or, they may be perfectly good, but unexpected, data values.
- Either way, they should be identified and checked.
- Box-and whisker plots, based on the **five number summary**, can help detect outliers.

Outliers

The five number summary consists of the quartiles Q_1, Q_2, and Q_3, and the upper and lower adjacent values, A_-, and A_+.

The lower adjacent value, A_-, is the smallest data value greater than $Q_1 - (1.5)(IQR)$. The upper adjacent value, A_+, is the largest data value smaller than $Q_3 + (1.5)(IQR)$.

$\approx$

Outliers

Example: Calculating the Five Number Summary.
The data are times I obtained (in 1/1000 seconds) from 6 clicks of a digital stopwatch (my finger slipped on the first one):
240, 144, 167, 172, 143, 133

Never mind that we would never compute a five number summary for six data points; this is just for illustration.

$\approx$

Outliers

We first calculate the quartiles Q_1, Q_2, and Q_3, or equivalently, the .25, .5 and .75 quantiles. We could use the general quantile algorithm described in the text, but instead we will present a simplified version that works when we only need to compute Q_1, Q_2, and Q_3:

$\approx$

Calculating Quartiles

1. Sort the data from smallest to largest.
2. If there are an odd number of data values, the median, Q_2, is the middle one.
3. If there are an even number of data values, the median, Q_2. is the average of the middle two.
4. Q_1 is the median of all data values $<$ or $\leq Q_2$, whichever is easier to compute.
5. Q_3 is the median of all data values $>$ or $\geq Q_2$, whichever is easier to compute.

Back to the Example

1. Begin by ordering the data: 133,143,144,167,172,240
2. Now compute the median. Since the number of observations, 6, is even, the median is the average of the middle two values: $Q_2 = [144 + 167]/2 = 155.5$.
3. The values $< Q_2$ are the same as those $\leq Q_2$: 133, 143, 144. Q_1 is the median of these, or 143. Similarly, the values $> Q_2$ are the same as those $\geq Q_2$: 167, 172, 240. Q_3 is the median of these, or 172.

$\approx$

Outliers

Recall that the sorted data are: 133,143,144,167,172,240
And that the interquartile range is
$IQR = Q_3 - Q_1 = 172 - 143 = 29$.

We calculate the lower adjacent value, A_-, as the smallest data value greater than $Q_1 - (1.5)(\text{IQR}) = 143 - (1.5)(29) = 99.5$. Looking at the data, this is the smallest data value, 133.

The upper adjacent value, A_+, is the largest data value smaller than $Q_3 + (1.5)(\text{IQR}) = 172 + (1.5)(29) = 215.5$. Looking at the data, this is the value 172, which just happens to correspond to Q_3.

$\approx$

Outliers

A **boxplot** (actually, a **box and whiskers plot**) is generated by forming a box with edges at Q_1 and Q_3 and a line at Q_2. Whiskers are extended from the end of the box to the corresponding adjacent value. Any observations outside the whiskers are considered possible outliers and are displayed individually.

The completed boxplot can be displayed either horizontally (Figure 12a) or vertically (Figure 12b). SAS code to produce these plots is found here. Notice that the time I recorded when my finger slipped is identified as a possible outlier.

$\approx$

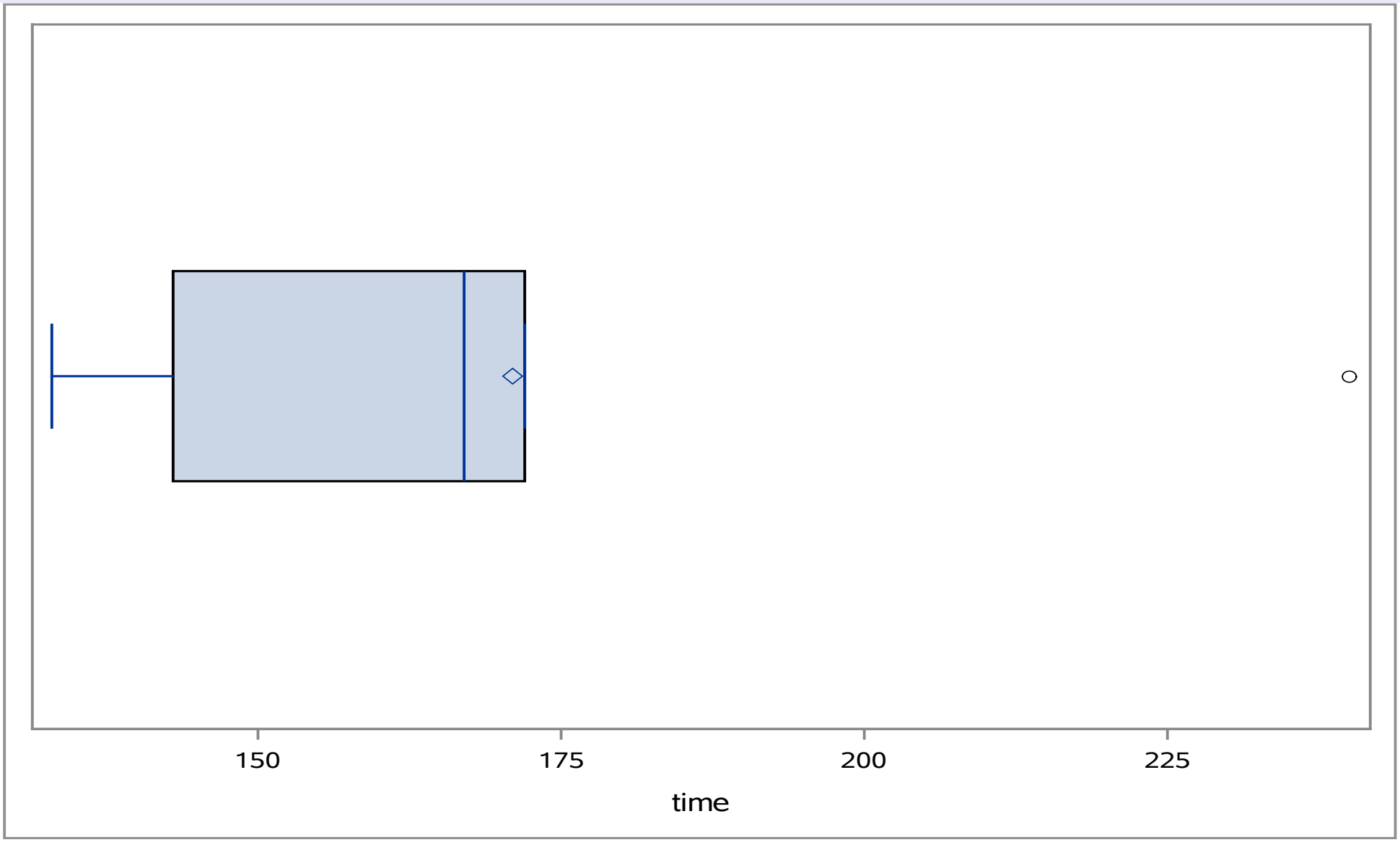


Figure: 12a: Horizontal boxplot for the stopwatch data.

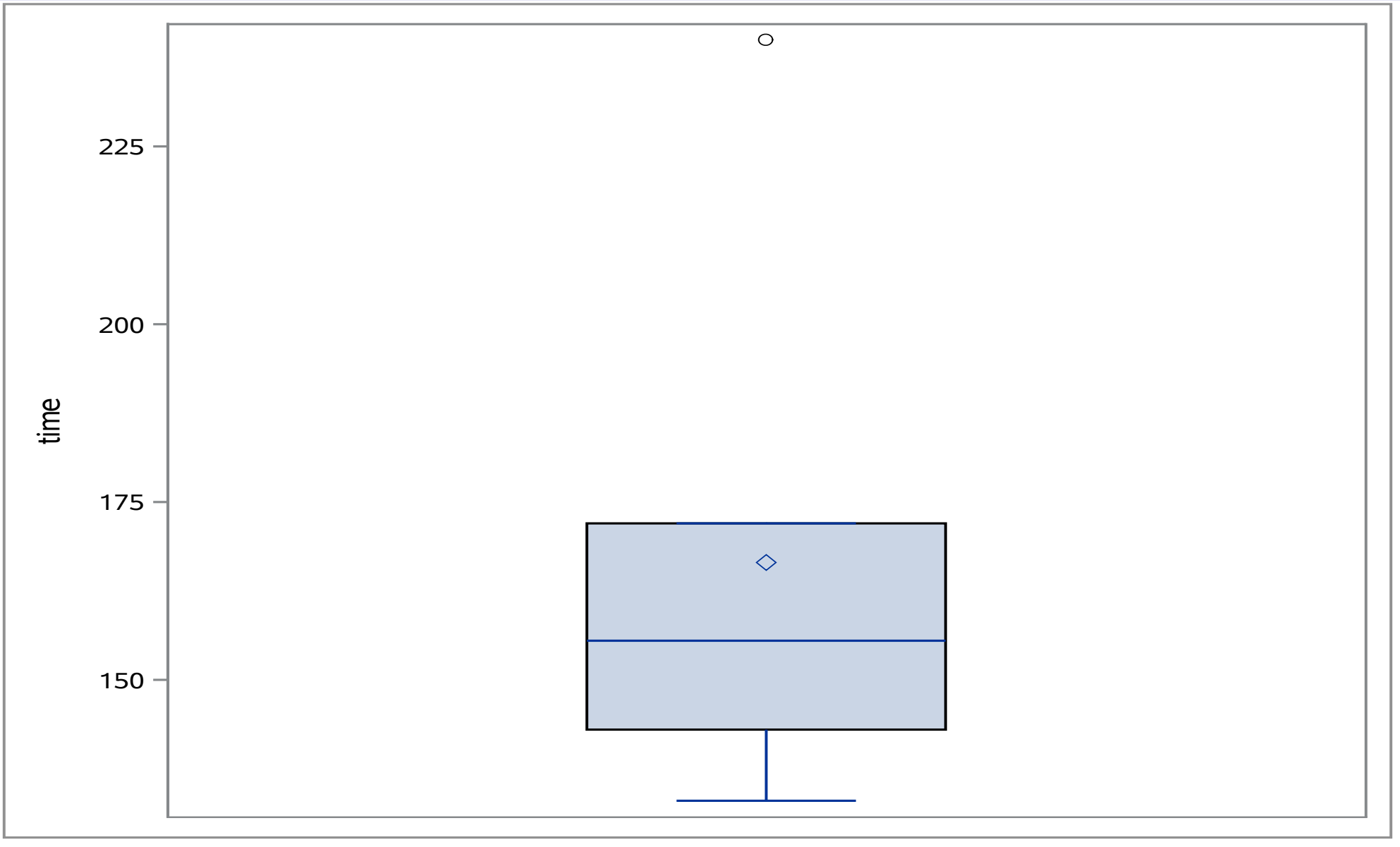


Figure: 12b: Vertical boxplot for the stopwatch data.

Outliers

Note that a boxplot is a good summary for some data sets (e.g., unimodal) but not for others (e.g., multimodal).

$\approx$

Outliers

Here are the boxplots for some of the data sets we've considered. First, the transformer lifetimes (SAS code here):

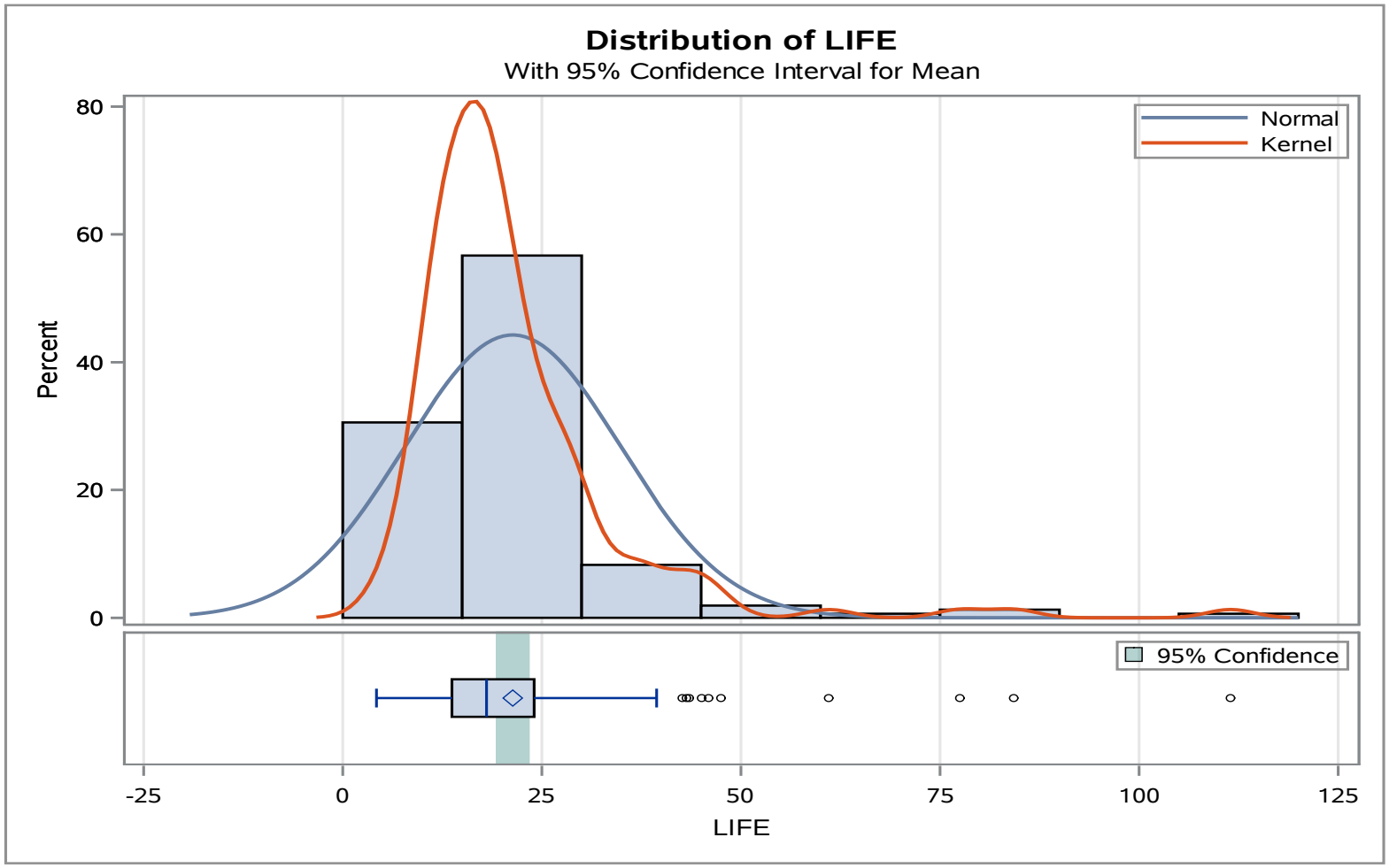


Figure: 13: Boxplot and histogram for the transformer lifetime data.

Outliers

Next, the logged transformer lifetimes:

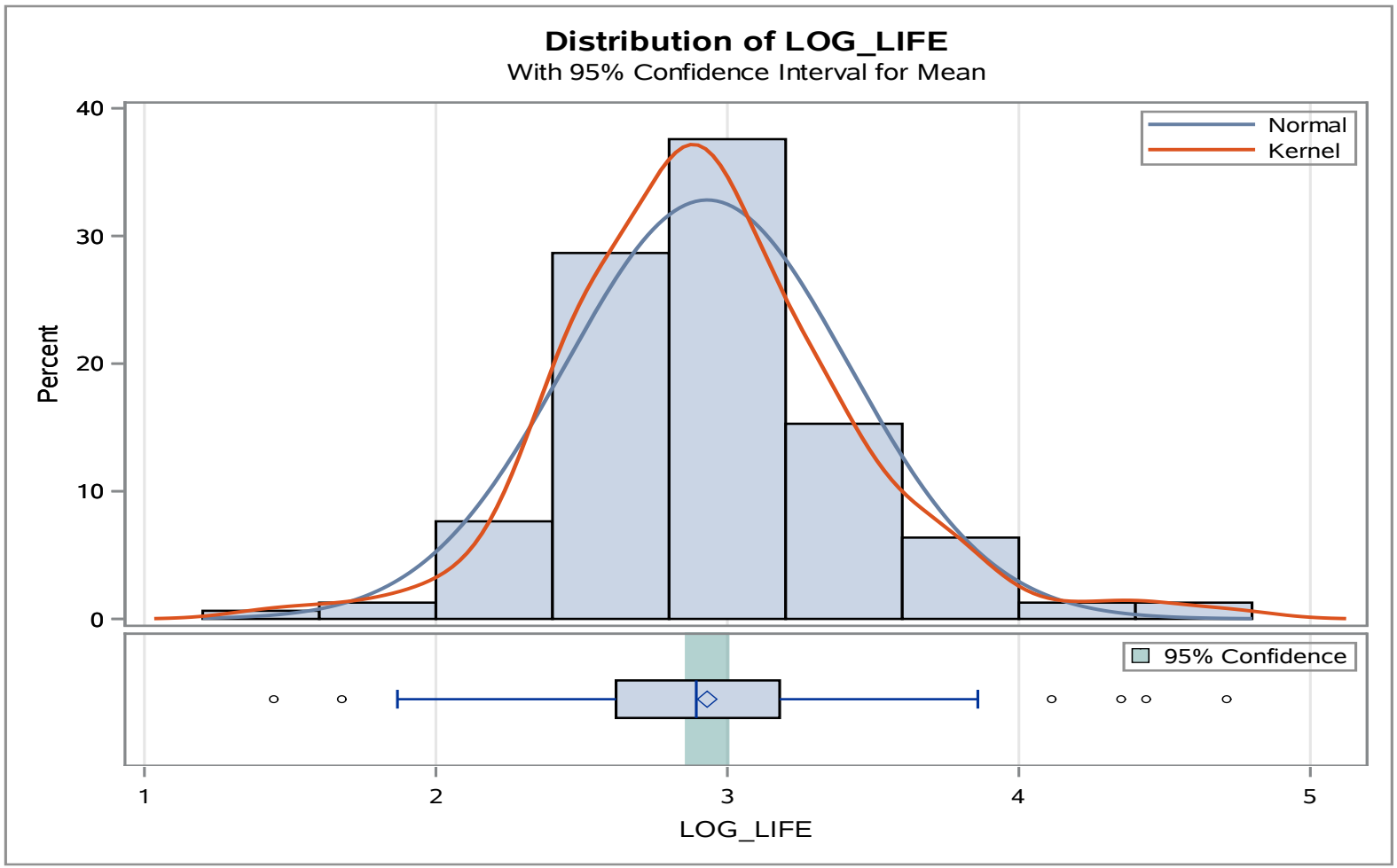


Figure: 14: Boxplot and histogram for the logged transformer lifetime data.

Outliers

Finally, the geyser data:

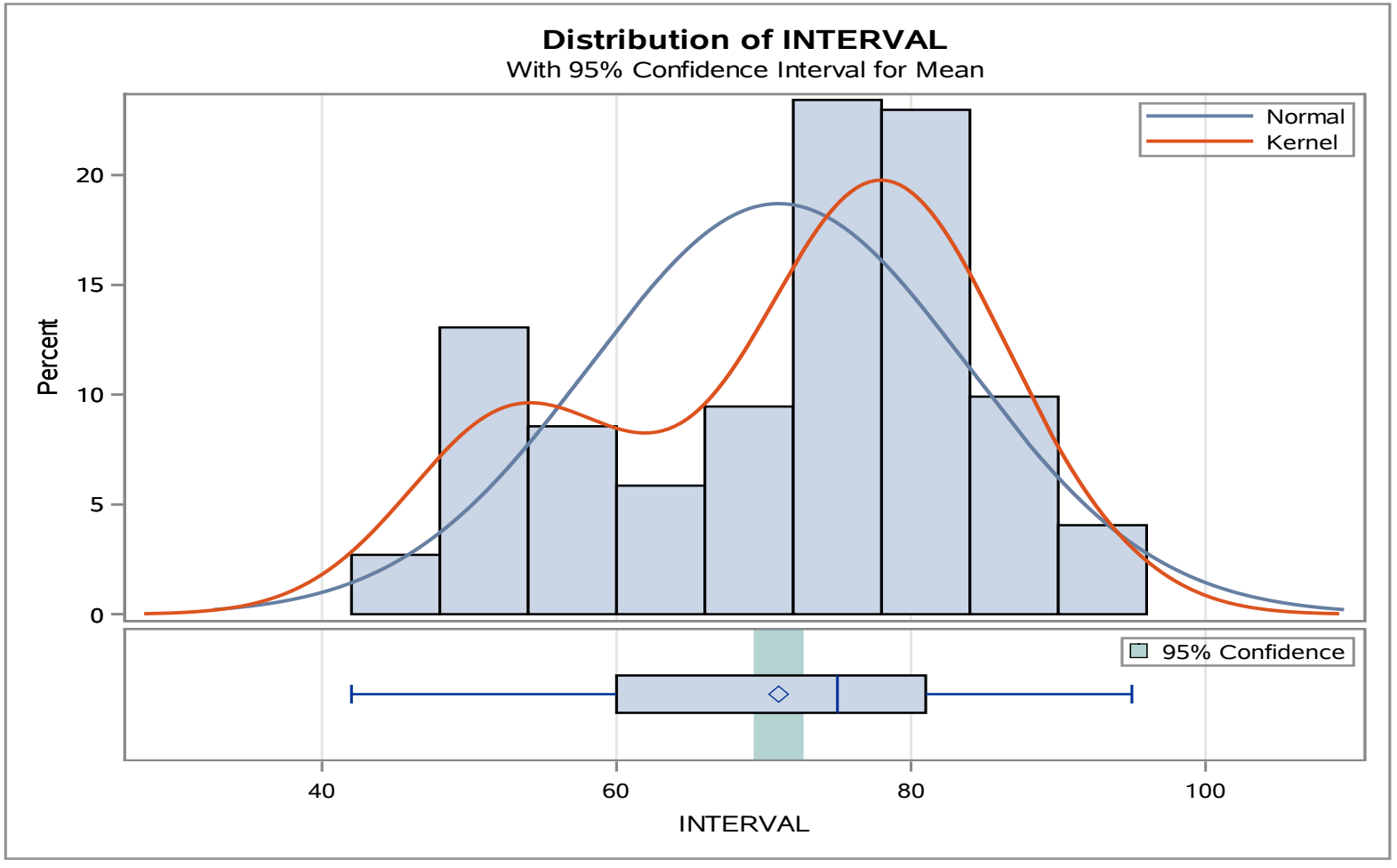


Figure: 15: Boxplot and histogram for the geyser data.

Calculating Quartiles: More Examples

Let's look at more examples of calculating quartiles. First, let's revisit the rules we use to calculate them:

≈

Calculating Quartiles

1. Sort the data from smallest to largest.
2. If there are an odd number of data values, Q_2 is the middle one.
3. If there are an even number of data values, Q_2 is the average of the middle two.
4. Q_1 is the median of all data values $<$ or $\leq Q_2$, whichever is easier to compute.
5. Q_3 is the median of all data values $>$ or $\geq Q_2$, whichever is easier to compute.

Calculating Quartiles: More Examples

Applying the above rules to the following data sets, we obtain:

Ordered data	Q_2	Q_1	Q_3
121,133,143,144,167,172,240	144.0	133.0	172.0
121,133,143,144,167,172,240,250	155.5	138.0	206.0
121,133,143,144,167,172,240,250,260	167.0	143.0	240.0

Side-by-Side Boxplots

Boxplots can also be useful in comparing different groups of data. The next plot shows side-by-side boxplots comparing the weights of bread from different ovens (SAS code here). Compare this plot with the stratified plot used for the same purpose in the Chapter 1 lecture notes.

$\approx$

Side-by-Side Boxplots

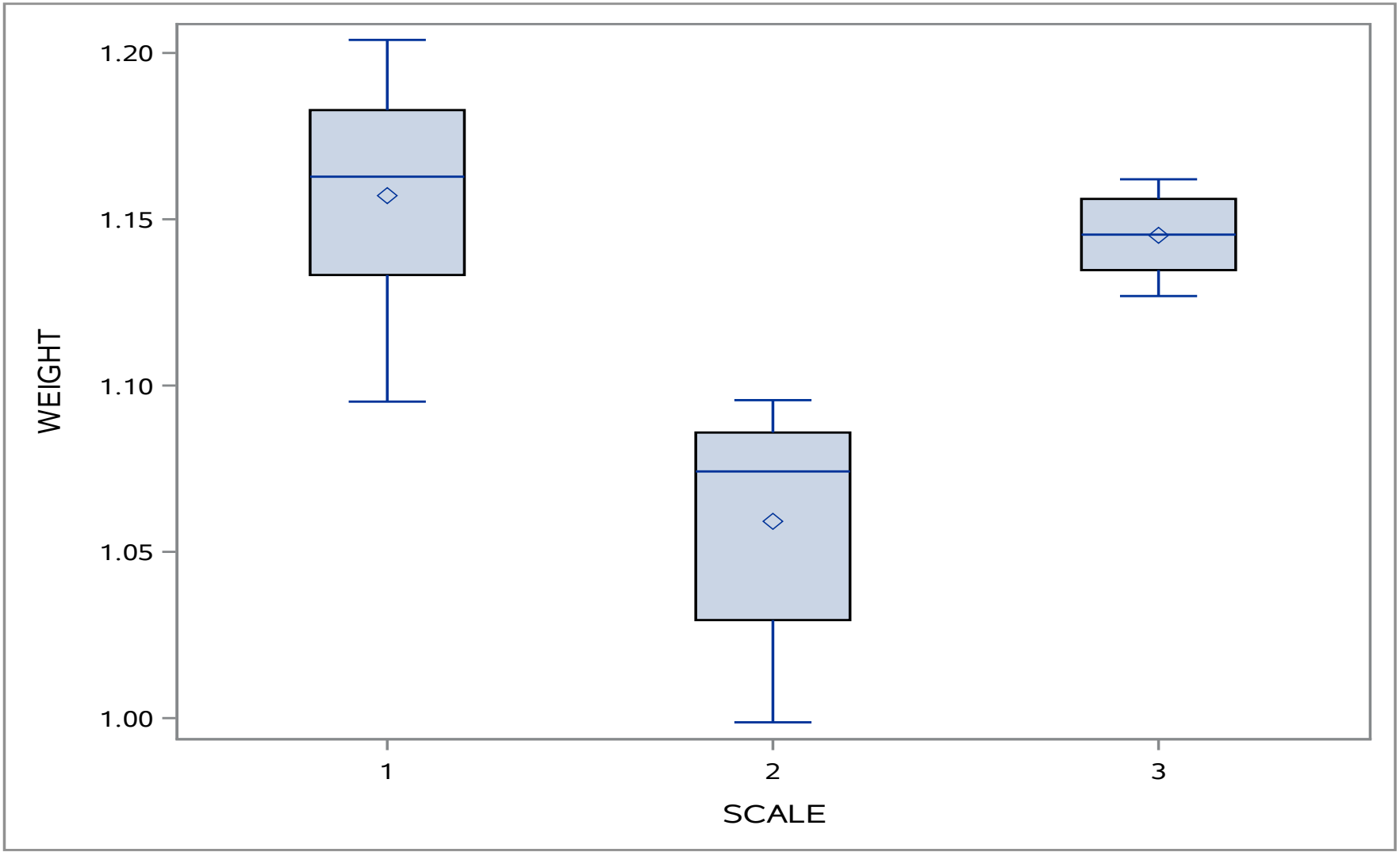


Figure: 16: Side-by-Side Boxplots for the bread data.

A More Topical Example

Recall the data set, from lab 1, of *US News and World Report* rankings of the 50 states in terms of how well they perform for their citizens in seven categories as well as overall. The next two graphs (code here) compare the overall rankings for those states carried by Donald Trump with those carried by Hillary Clinton.

$\approx$

A More Topical Example

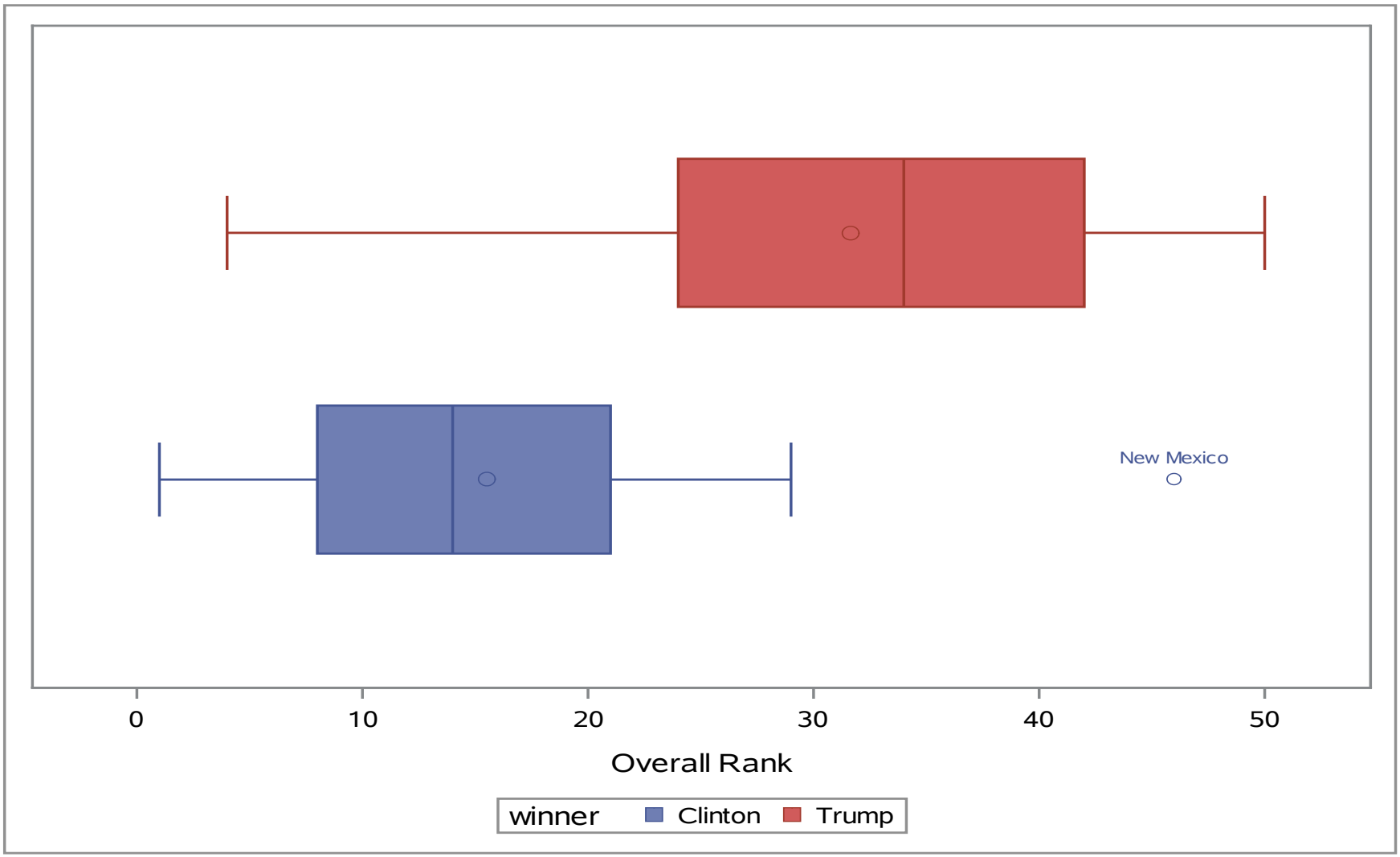


Figure: 17: Comparison Boxplots for the US News and World Report Rankings.

A More Topical Example

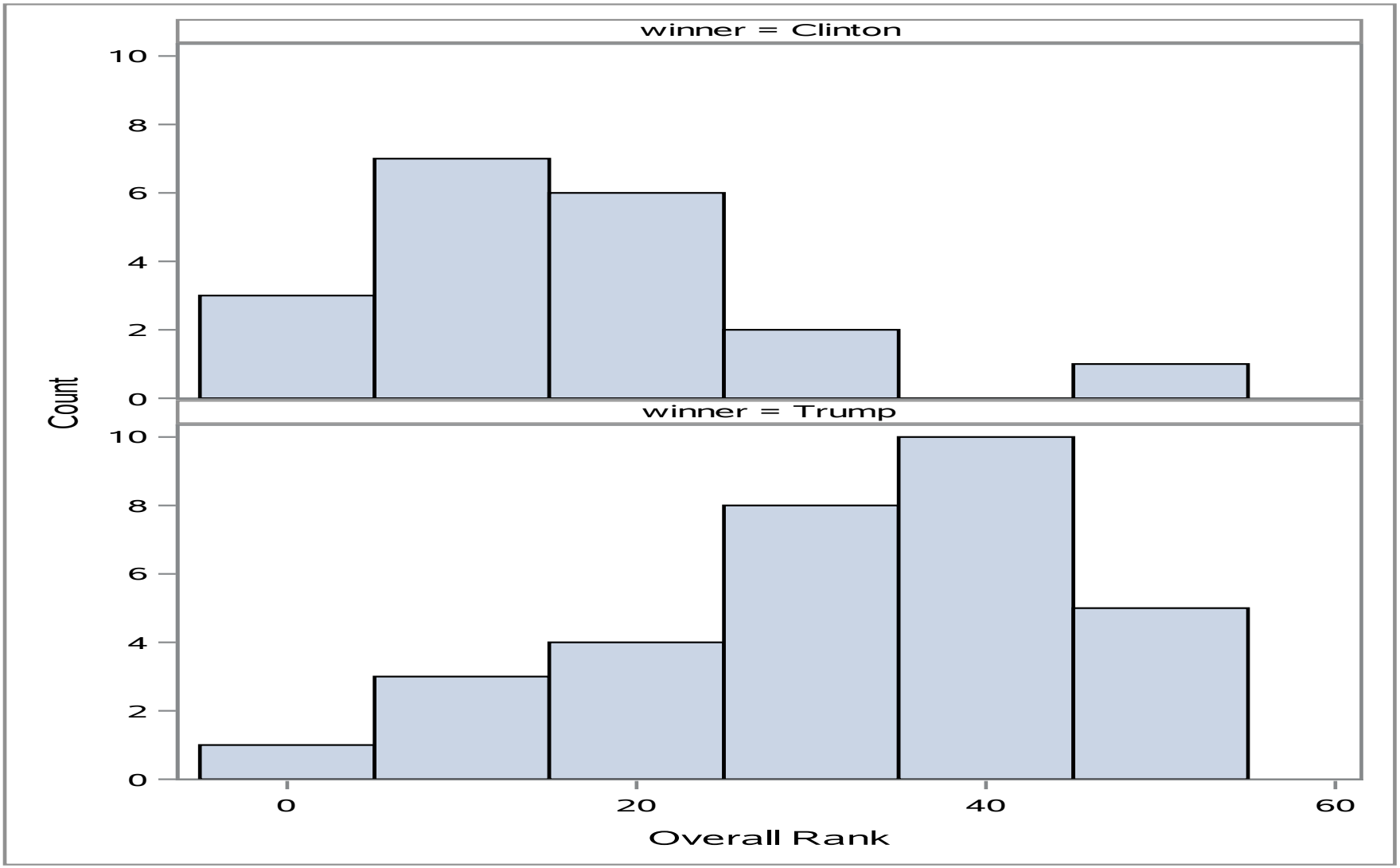


Figure: 18: Comparison Histograms for the US News and World Report Rankings.

What Tukey Meant

Recall the quote from John Tukey which began this chapter:

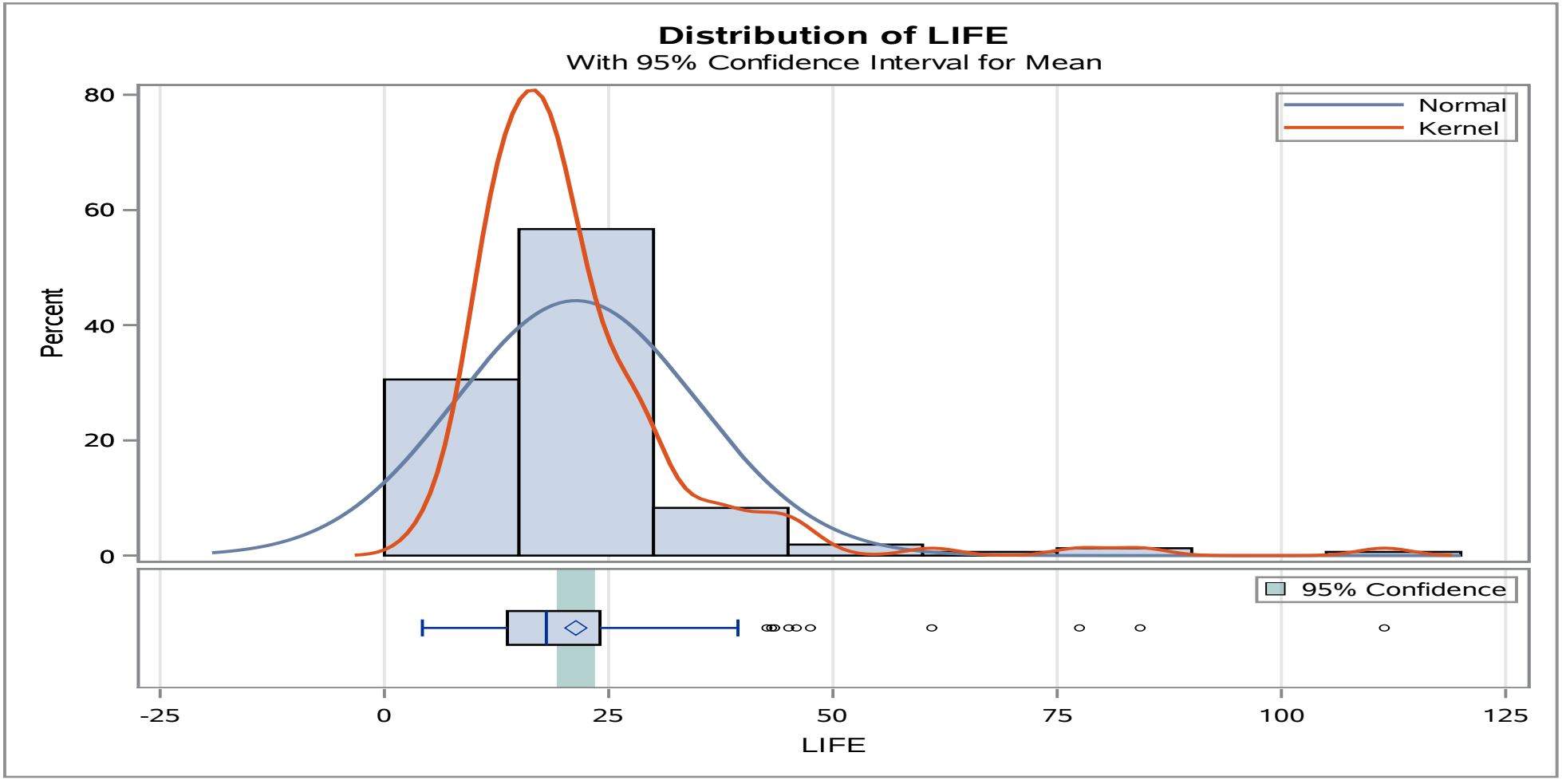
"Numerical quantities focus on expected values, graphical summaries on unexpected values."

With what you have learned, you should be able to appreciate it more, as the following graph will show.

$\approx$

What Tukey Meant

The numerical summaries tell us about most of the data; for the transformer data the median is around 18. On the graphs, however, our eyes are drawn to the exceptions, here the outliers.



Resistant Summary Measures

- Summary measures are **resistant** if they are not seriously affected by outliers.
- The median and IQR are resistant measures of location and spread.
- The mean and standard deviation are not resistant.

Resistant Summary Measures

Though the mean is not resistant, for many data sets without outliers it is a better measure of location than the median. Two measures which attempt to add some resistance to the mean, while retaining its good properties when there are no outliers, are the **trimmed mean** and the **Winsorized mean**.

Resistant Summary Measures

- The k-times trimmed mean omits the k largest and k smallest data values and takes the mean of the remaining ones.
- To compute the k-times Winsorized mean, first create a new data set by replacing the k smallest data values with the value of the $k+1$st smallest, and the k largest data values with the value of the $k+1$st largest, while leaving the other data values untouched. The k-times Winsorized mean is the mean of all values in this new data set.

Example: Trimmed and Winsorized Means

As an example, consider again the last data set used to illustrate computation of quartiles. In ascending order the values are: 121, 133, 143, 144, 167, 172, 240, 250, 260.

To compute the 2-times trimmed mean, discard the two largest values (250, 260) and the two smallest values (121, 133), and take the average of the remaining values. The value of the two times trimmed mean is therefore

$(143 + 144 + 167 + 172 + 240)/5 = 173.2.$

$\approx$

Resistant Summary Measures

To compute the 2-times Winsorized mean for these same data (121, 133, 143, 144, 167, 172, 240, 250, 260), set the two largest values to the value of the third largest value, 240, and set the two smallest values to the value of the third smallest value, 143, giving a modified data set: 143, 143, 143, 144, 167, 172, 240, 240, 240. Then take the average of these values. The answer is $181.\overline{3}$.

$\approx$

Recap, Chapter 2: Summarizing Data:

- Displaying stationary data distributions
  - Bar charts, frequency histograms
  - Analysis of same
  - Causes of common patterns
- Summary measures for stationary data distributions and when each is appropriate
- Boxplots and outliers
- Resistant summary measures

$\approx$