Chapter 2: Summarizing Data

- Preview:
  - 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

- **What's the IDEA?**

  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 should always be checked for stationarity before using them.

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

  Example: In the PREZHITE data, the variable YEAR is quantitative and discrete, HEIGHT is quantitative and continuous, and WINPARTY is categorical and discrete. In fact, I can't think of an example of a categorical variable that is continuous.

- Displaying data distributions:
  - Bar chart for categorical data. Figure 1 shows a bar chart of WINPARTY produced by SAS/INSIGHT.
  - Needle Plot for a small amount (say 20 observations or less) quantitative data. Figure 2 shows the numbers of moons among the nine planets.
  - Frequency Histogram for a larger amount of quantitative data. You have already seen the frequency histogram of presidential winner heights shown in Figure 3.
- Analyzing Frequency Histograms
  - Modality
  - Symmetry
  - Center
  - Spread
  - Pattern and deviations

- **What's the IDEA?**

  Different choices of interval locations and widths can make frequency histograms for the same set of data look very different. 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.

  Example: The data set SASDATA.TRANSFORM contains the lifetimes of 157 electrical transformers. Figure 4 shows a frequency histogram of these lifetimes.

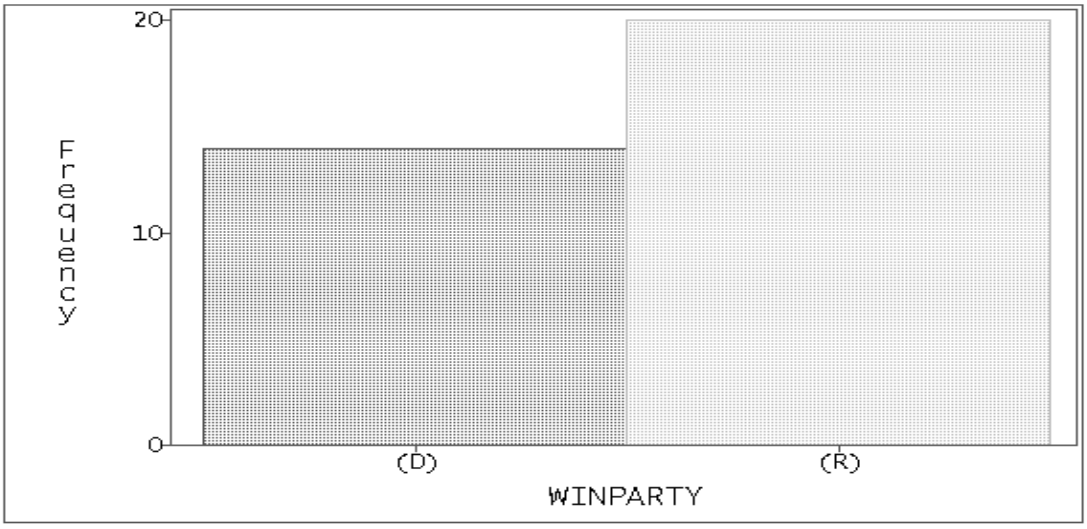


Figure 1: *Bar chart of election winner's party.*

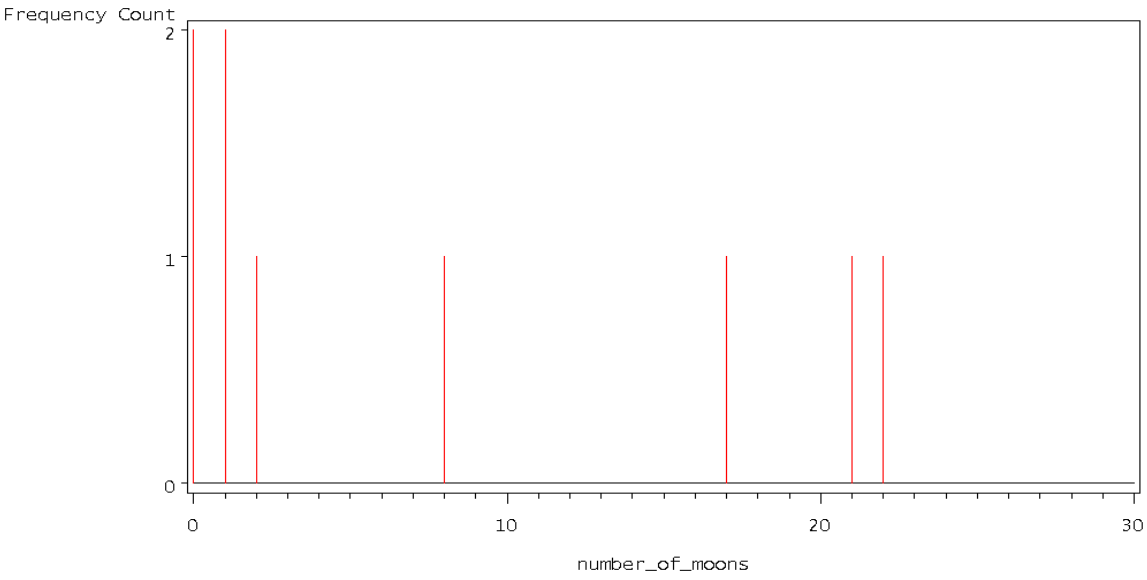


Figure 2: *Needle plot of the numbers of moons among the nine planets.*

This histogram is uni-modal and right-skewed. Transforming skewed data can sometimes make it symmetric. Figure 5 shows a frequency histogram of the natural logs of the transformer lifetimes. The distributional pattern is unimodal and roughly symmetric.

The data set SASDATA.GEYSER1 contains the durations of and intervals between eruptions of the Old Faithful geyser in Yellowstone National Park. The intervals between eruptions were judged to be stationary, so the pattern of variation is represented by the frequency histogram in Figure 6.

The histogram is **bi-modal**, meaning the intervals between eruptions tend to be in the neighborhood of 50 minutes or 80 minutes. Can we predict which? 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.

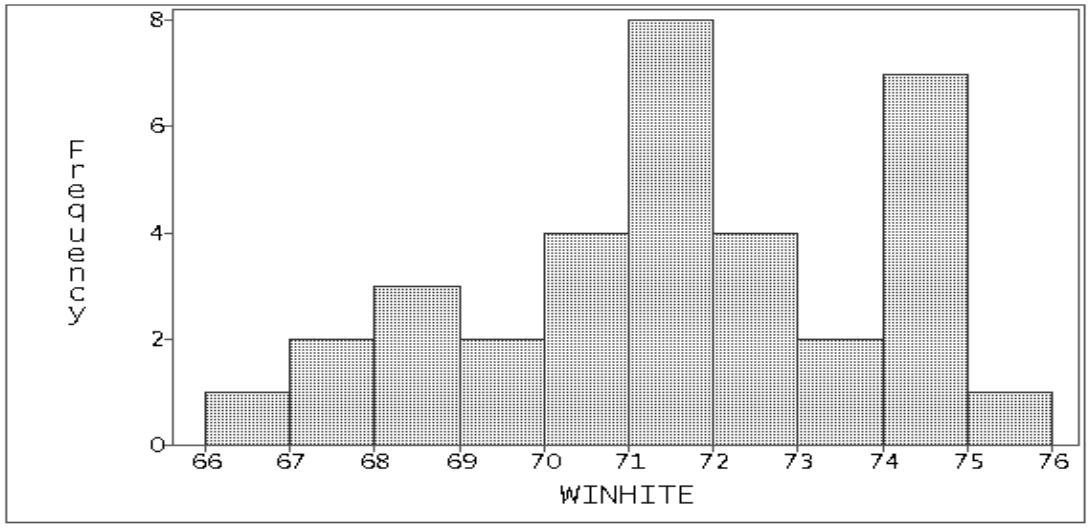


Figure 3: *Frequency histogram of election winner's heights.*

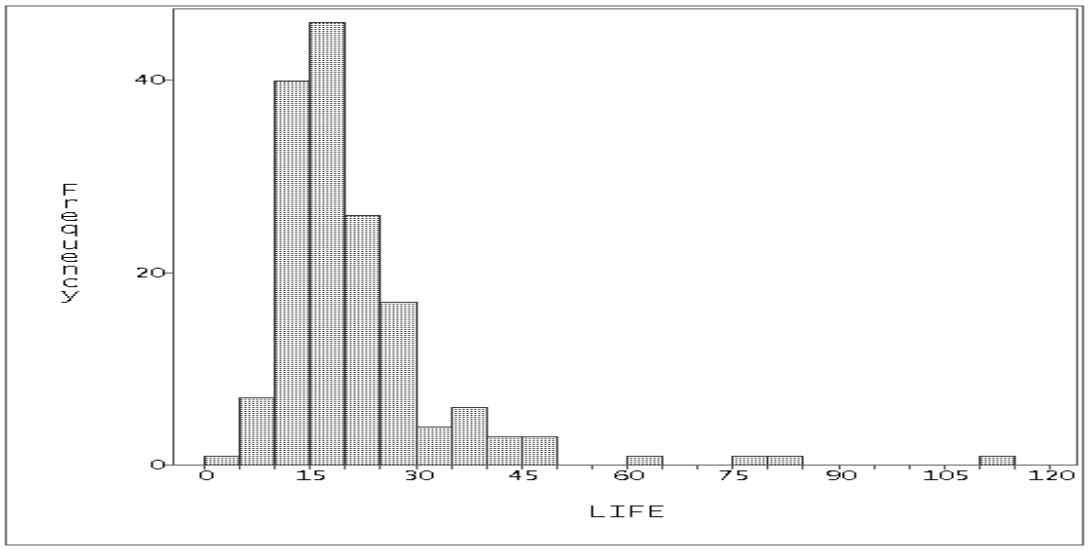


Figure 4: *Frequency histogram of lifetimes of 157 electrical transformers.*

- What Might Cause These Shapes?
  - o Symmetric, unimodal: measurement errors; homogeneous populations
  - o Skewness: upper bound, no lower bound, or vice-versa
  - o Short-tailed: mixture of process streams
  - o Multi-modal: nonhomogeneous populations
- Summary measures of quantitative data: location
  - o Mean: Average

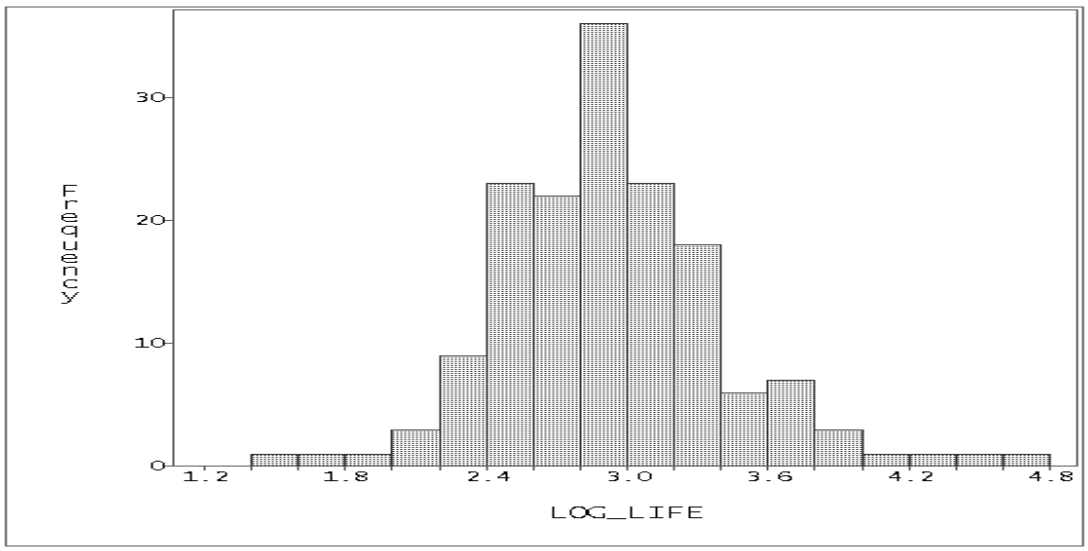


Figure 5: *Frequency histogram of the natural logs of lifetimes of 157 electrical transformers.*

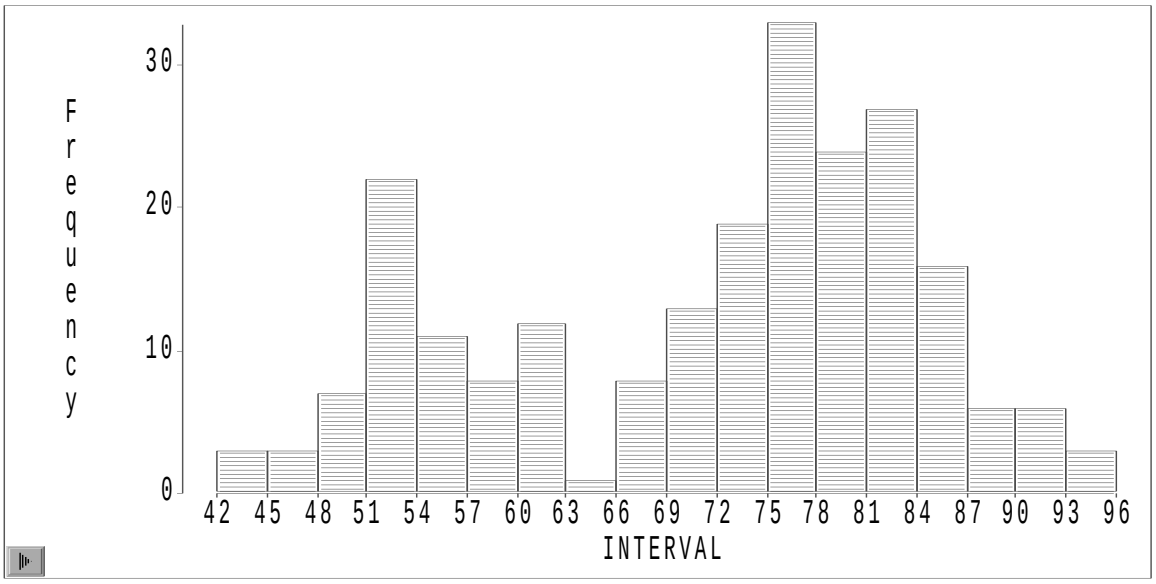


Figure 6: *Frequency histogram of the intervals between eruptions of the Old Faithful geyser.*

- Median: Halfway point
- Mode: Location of the modal bar on a frequency histogram.
- Quartiles
- Quantiles

• Summary measures of quantitative data: spread

- Mean absolute deviation: Average distance from mean
- Standard deviation (RMS): Square root of average squared distance from mean

- o Interquartile range (IQR): Range of middle 50% of data.
- o Quartiles
- o Quantiles

Figure 7 is output from a SAS/INSIGHT distribution analysis of the transformer data, which includes a frequency histogram and the above summary measures (**except for the mode**). Figures 8 and 9 show similar output for the logged lifetimes and the geyser data.

• What's the IDEA?

- o Statistical summary measures are used to parsimoniously summarize the pattern of variation of data from a stationary process. Summary measures tend of focus on two aspects of a data distribution: location and spread.
- o 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.
- o 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. Generally, the mean and standard deviation are appropriate for a unimodal/symmetric pattern, median (or mode) and interquartile range are appropriate for a skewed pattern, and modes and the range of modal peaks for a multi-modal pattern.

• Outliers

- o Outliers are extremely unrepresentative data.
- o Box-and whisker plots, based on the five number summary, can help detect outliers.

Example: Calculating the Five Number Summary

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 data are the times I obtained just now from 6 clicks of the digital stopwatch (my finger slipped on the first one):

240, 144, 167, 172, 143, 133

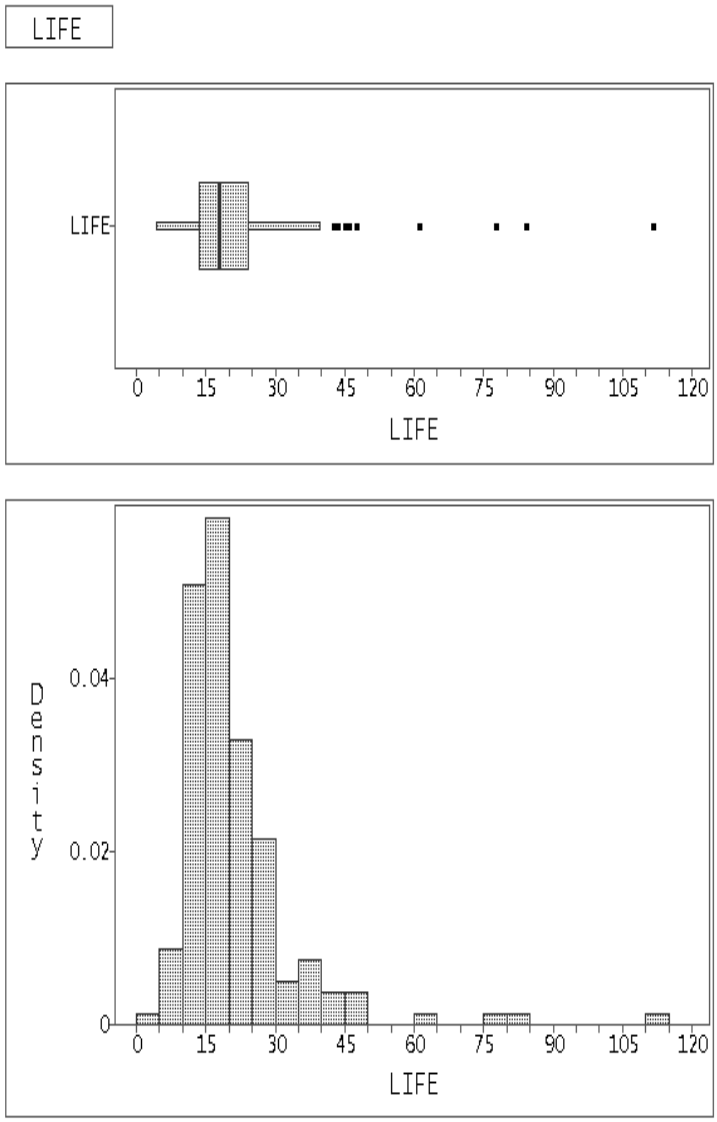
We first calculate the quartiles Q_1, Q_2, and Q_3, or equivalently, the .25, .5 and .75 quantile. To do this, we use the algorithm on pp. 69-70 of the text.

1. Begin by ordering the data $y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq y_{(4)} \leq y_{(5)} \leq y_{(6)}$. So $y_{(1)} = 133$, $y_{(2)} = 143$, $y_{(3)} = 144$, $y_{(4)} = 167$, $y_{(5)} = 172$, $y_{(6)} = 240$.
2. Start with quantile $q = .25$, which is the first quartile, Q_2. Then, since $m/n < .25 < (m+1)/n$, where n is the number of observations, 6, and $m = 1$, the .25 quantile is $y_{(m+1)} = y_{(2)} = 143$. A similar computation shows that the $q = .75$ quantile, or third quartile, Q_3, equals 172.
3. Next find the quantile $q = .5$, which is the second quartile, or median, Q_2. Since $.5 = k/n = 3/6$, the .5 quantile is

$$(y_{(k)} + y_{(k+1)})/2 = (y_{(3)} + y_{(4)})/2 = (144 + 167)/2 = 155.5$$

From this, we see 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.

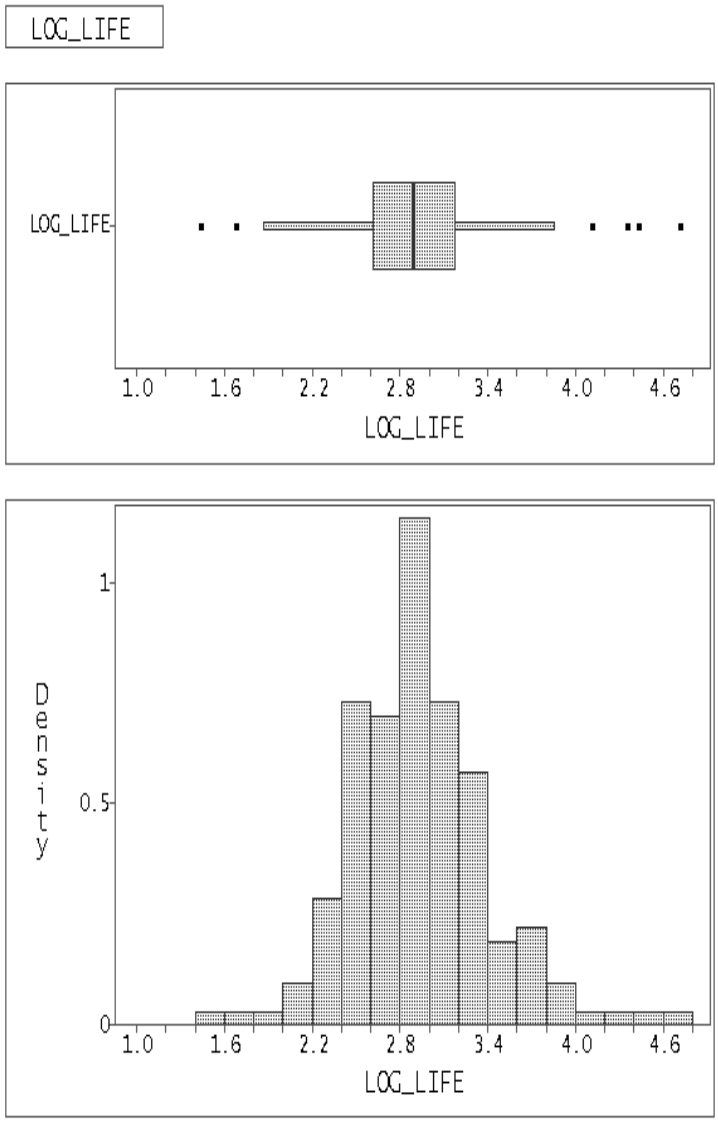


Moments			
N	157.0000	Sum Wgts	157.0000
Mean	21.3442	Sum	3351.0329
Std Dev	13.5197	Variance	182.7812
Skewness	3.3520	Kurtosis	16.3104
USS	100038.840	CSS	28513.8641
CV	63.3413	Std Mean	1.0790

Quantiles			
100% Max	111.4912	99.0%	84.2795
75% Q3	24.0319	97.5%	61.0434
50% Med	18.0526	95.0%	43.4375
25% Q1	13.7087	90.0%	35.7246
0% Min	4.2316	10.0%	11.0943
Range	107.2596	5.0%	9.8025
Q3–Q1	10.3232	2.5%	7.4034
Mode	.	1.0%	5.3448

Figure 7: *Distribution analysis of lifetimes of 157 electrical transformers.*

The upper adjacent value, A_+, is the largest data value smaller than $Q_3 + (1.5)(\text{IQR}) = 172 +$

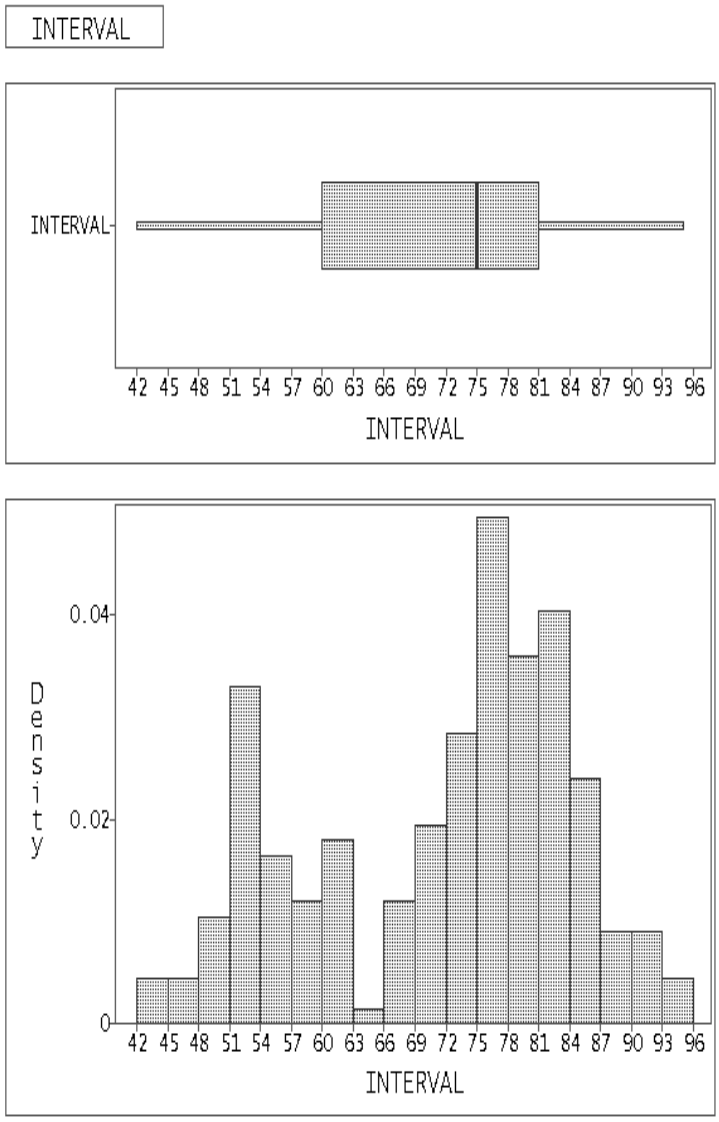


Moments			
N	157.0000	Sum Wgts	157.0000
Mean	2.9306	Sum	460.0987
Std Dev	0.4863	Variance	0.2365
Skewness	0.4821	Kurtosis	1.6400
USS	1385.2418	CSS	36.8928
CV	16.5942	Std Mean	0.0388

Quantiles			
100% Max	4.7139	99.0%	4.4341
75% Q3	3.1794	97.5%	4.1116
50% Med	2.8933	95.0%	3.7713
25% Q1	2.6180	90.0%	3.5758
0% Min	1.4426	10.0%	2.4064
Range	3.2714	5.0%	2.2826
Q3–Q1	0.5614	2.5%	2.0019
Mode	.	1.0%	1.6761

Figure 8: *Distribution analysis of the natural logs of lifetimes of 157 electrical transformers.*

$(1.5)(29) = 215.5$. Looking at the data, this is the value 172, which just happens to correspond to



Moments			
N	222.0000	Sum Wgts	222.0000
Mean	71.0090	Sum	15764.0000
Std Dev	12.7992	Variance	163.8189
Skewness	-0.4855	Kurtosis	-0.8857
USS	1155590.00	CSS	36203.9820
CV	18.0247	Std Mean	0.8590

Quantiles			
100% Max	95.0000	99.0%	93.0000
75% Q3	81.0000	97.5%	91.0000
50% Med	75.0000	95.0%	88.0000
25% Q1	60.0000	90.0%	84.0000
0% Min	42.0000	10.0%	51.0000
Range	53.0000	5.0%	50.0000
Q3-Q1	21.0000	2.5%	47.0000
Mode	75.0000	1.0%	44.0000

Figure 9: *Distribution analysis of the intervals between eruptions of natural logs of the Old Faithful geyser.*

Q_3.

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 outliers and are displayed individually.

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

- **What's the IDEA?**

  An outlier is an extremely unrepresentative data point. A box-and whisker plot, based on the five number summary, can help detect 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.
  - The mean is not resistant, but for many data sets has better behavior than the median if there are no outliers. Two measures which attempt to add some resistance to the mean are the trimmed mean and Winsorized mean.
    * 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 six stopwatch measurements I took, which are, in ascending order: 133, 143, 144, 167, 172, 240.

  To compute the 2-times trimmed mean, discard the two largest values (172, 240) and the two smallest values (133, 143), and take the average of the remaining values: 144 and 167. The value of the two times trimmed mean is therefore $(144+167)/2 = 155.5$.

  To compute the 2-times Winsorized mean, set the two largest values to the value of the third largest value, 167, and set the two smallest values to the value of the third smallest value, 144, giving a modified data set: 144, 144, 144, 167, 167, 167. Then take the average of these values. The answer, in this case is the same as the 2-times trimmed mean: 155.5.

- **What's the IDEA?** Summary measures are called "resistant" if they are not seriously affected by outliers. The median and IQR are resistant measures, the mean and standard deviation are not. Trimmed and Winsorized means are more resistant variations of the mean.