Summarizing & Organizing Univariate Data (i.e.measurements of just 1 variable)

(a) Frequency distributions

(b) Numerical summaries

(c) Graphical summaries

Organizing & Summarizing

- Now that you have chosen a sample & collected data (i.e. made measurements) on the units in that sample, it's time to *organize* & *summarize*
- Recall that the data can be *qualitative* (nominal, ordinal) or *quantitative* (interval-type, ratio-type discrete or ratio-type continuous)
- Organizing via a *frequency distribution*: An ordinary frequency dist. (OK for nominal, ordinal & discrete quantitative data with many repetition) is a list of distinct data-points & how many times each is repeated.

Ordinary Frequency Distribution

- The following is a sample of 30 U of Akron students F,F,So,F,J, Se, F, J, J, F, So, Se, F, Se, So, F, F, J, So, So, J, Se, F,F,F, J, F, J, Se, So

F	12	F	0.4	40%
So	6	So	0.2	20%
J	7	J	0.233	23.3%
Se	5	Se	0.167	16.7%

The one in red is an ordinary relative frequency distribution

- Below is a sample of 25 families in the Akron area who were asked how many times they ate out in August

2, 2, 4, 1, 1, 1, 4, 2, 3, 1,1, 2, 3, 1, 1, 5, 1, 2, 2, 3, 4, 1, 1, 2, 1

1	11	1	0.44	44%
2	7	2	0.28	28%
3	3	3	0.12	12%
4	3	4	0.12	12%
5	1	5	0.04	4%

Class Frequency Distributions

- Good for ordinal, discrete quantitative data with few repetitions and continuous quantitative data
- Below are responses of 25 neuralgia patients about the pain severity on a 1-10 scale

6,8, 3,6, 5,4,3,8,5,5,7, 5, 5, 1,3,1,1,7,2,1

Class	Freq	Rel. Freq.
1-2	5	0.25 =25%
3-4	4	0.2 = 20%
5-6	7	0.35 = 35%
7-8	4	0.2 = 20%

- Below are the widths of metal rods made by a company

5.7, 3.8,4.0,3.1,2.5, 3.8, 1.9, 2.7, 1.8, 5.0, 8.2, 9.1,4.5,4.9, 0.8,1.1,2.8, 1.4, 0.9, 9.7 (in cm)

Class	Freq	Rel. Fre
0.0-<2.0	6	0.3=30%
2.0-<4.0	6	0.3=30%
4.0-<6.0	5	0.25 = 25%
6.0-<8.0	0	0 = 0%
8.0-<10.0	3	0.15=15%

Numerical Summaries

- Numbers computed from the data that try to capture key features of the dataset (e.g. where is its center, how much variability is there around the center, is it asymmetric, what kind of asymmetry, is (are) there 1 or many peaks, are there any unusually large/small data-point)
- Center: Measures of *Central Tendency* (mean, median, mode, quartiles, percentiles)
- Variability around the center: Measures of *dispersion* or *spread* (Std. deviation, range, IQR, MAD)
- Shape & exceptions: Fences, skewness, kurtosis

Measures of Central Tendency

- Sample mean: The simple average of all data (just add them up & divide by the sample size)
- Sample median: Arrange the data-points in the increasing order & pick the middle one (or, if the sample size n is *even*, take the average of the two middle data-points
- Sample mode: Data-point with the highest frequency
- Sample quartiles: Numbers that divide the data set into quarters (as closely as possible)
- Sample percentiles: The p-th percentile (approx) has p % of the data below it & (100-p)% data above it.

Examples

Dataset: {1,3,7,4,1,5,5,9,2,5, 8,2,3,3,6,7,10,6}

Mean: 87/18 = 4.833

Median: The middle number of {1,1,2,2,3,3,3,4, 5 ,5,5, 6,6,7,7,8,9,10} = average of 5 & 5 = 5

Mode= 3 or 5

1st quartile (Q1): 0.25(n+1) = 0.25(19) = 4.75 (not a whole number). So Q1= average of the 4th & 5th data-points = (2+3)/2 = 2.5

3rd quartile (Q3): 0.75(19) = 14.25(not whole no.) So Q3= avg of the 14th & 15th values = (7+7)/2

Measures of Central Tendency

- The 60th percentile: 0.6(n+1) = 0.6(19)=11.4 (not a whole number). So P_60 = average of the 11th and 12th data points = (5+6)/2 = 5.5
- Note: The median is the 2nd quartile (Q2) as well as the 50th percentile (P_50)
- The mode is not a good description of the center (specially when the dataset has many peaks)
- The mean is a good description of the center only when there isn’t any outlier (exceptional data value). Otherwise it tends to get pulled toward such outliers.
- The median is the best (most robust) center descriptio

Measures of Dispersion

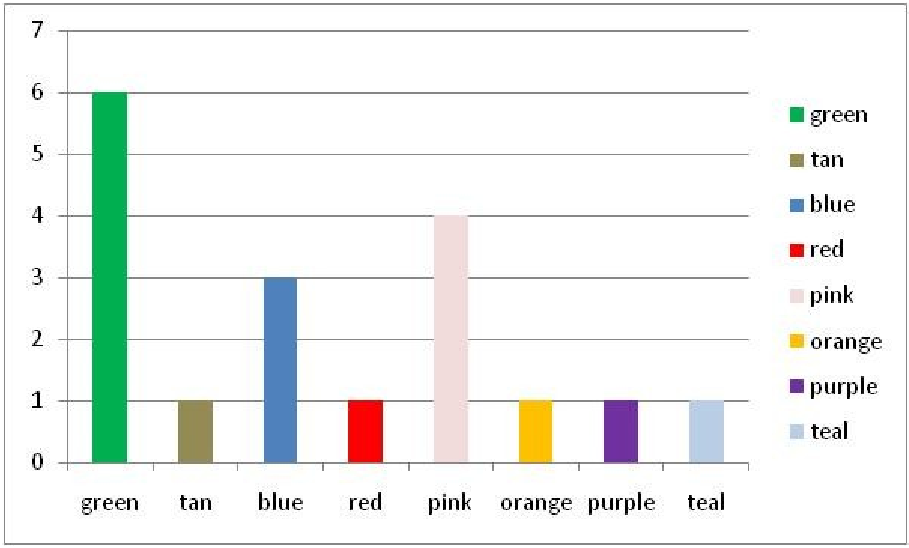
- Range = max – min
- Inter-quartile range (IQR) = Q3 – Q1
- Standard deviation (square-root of variance) = basically the average squared distance of all the data-points from the mean. So, for our ‘toy’ dataset from 2 slides ago,variance={(1-4.442)^2 + (3-4.442)^2+(7-4.442)+…+(6-4.442)^2}/(18-1)
- Median absolute deviation (MAD)= median of the absolute deviations of all data-points from the median of the dataset. So for our dataset,it’s med{ |1-5|, |1-5|, |2-5|,…,|10-5|}=med{4,4,3,…,5}= 2

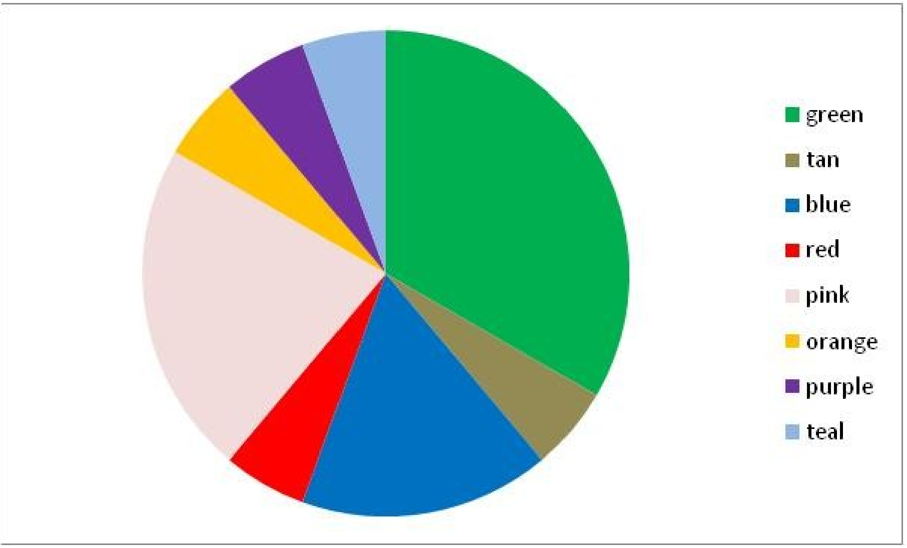
Graphical Summaries of datasets

- More visually and intuitively appealing than numerical summaries; easier and quicker to understand the key features
- However, pictures can be interpreted slightly differently by different people, so there's room for some subjectivity
- First we will see what pictures can be drawn for a *qualitative* (or *categorical*) dataset. Next we'll consider *quantitative* data.

Barplots and Pie Charts

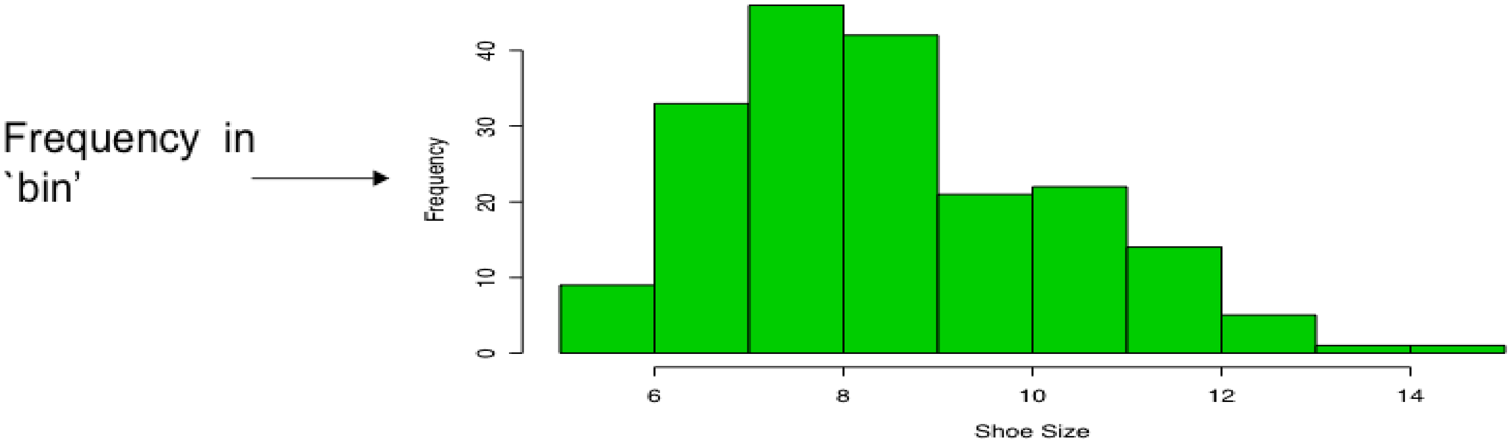
- For categorical variables, we can graph the distribution using bar plots and pie charts





Histograms

- Histograms emphasize **frequency** of different values in the distribution

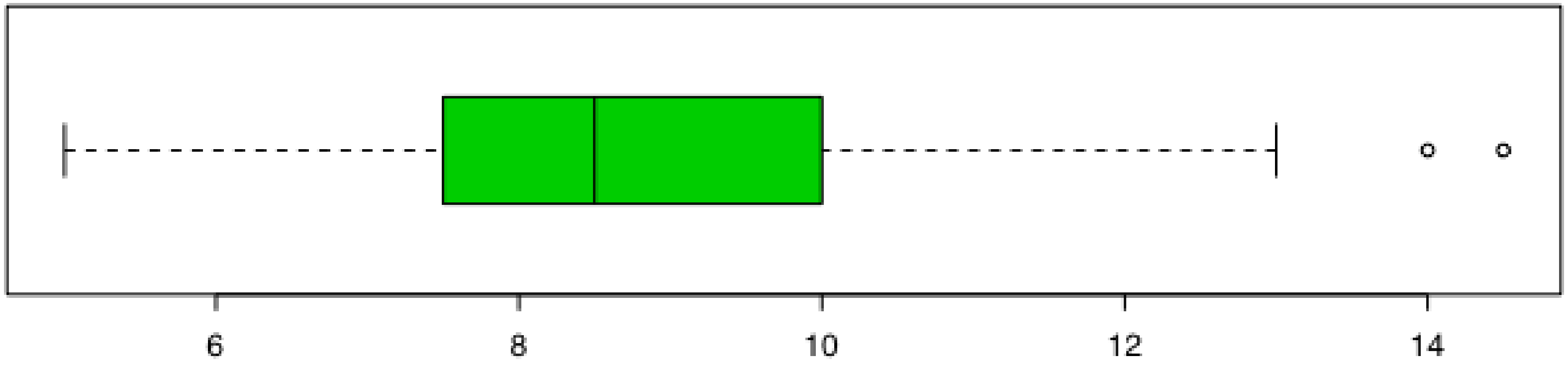


Frequency in `bin'

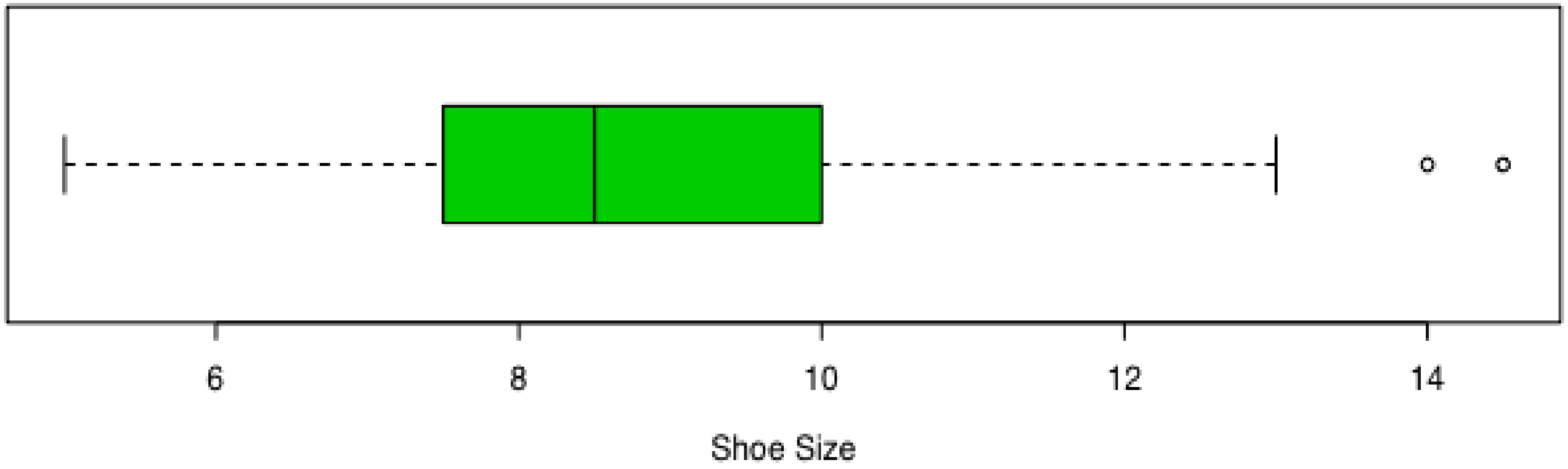
- X-axis: Values are divided into bins
- Y-axis: Height of each bin is the frequency that values from that bin appear in dataset

Boxplots

- Box plots are an effective tool for conveying information of continuous variables
- **Box** contains the central 50% of the data, with a line indicating the median
- **Median** is the value with 50% of data on either side
- **Whiskers** contain most of the rest of the data, except for suspected outliers
- **Outliers** are suspiciously large or small values



Boxplot: Shoe Size of a Stat-401 class



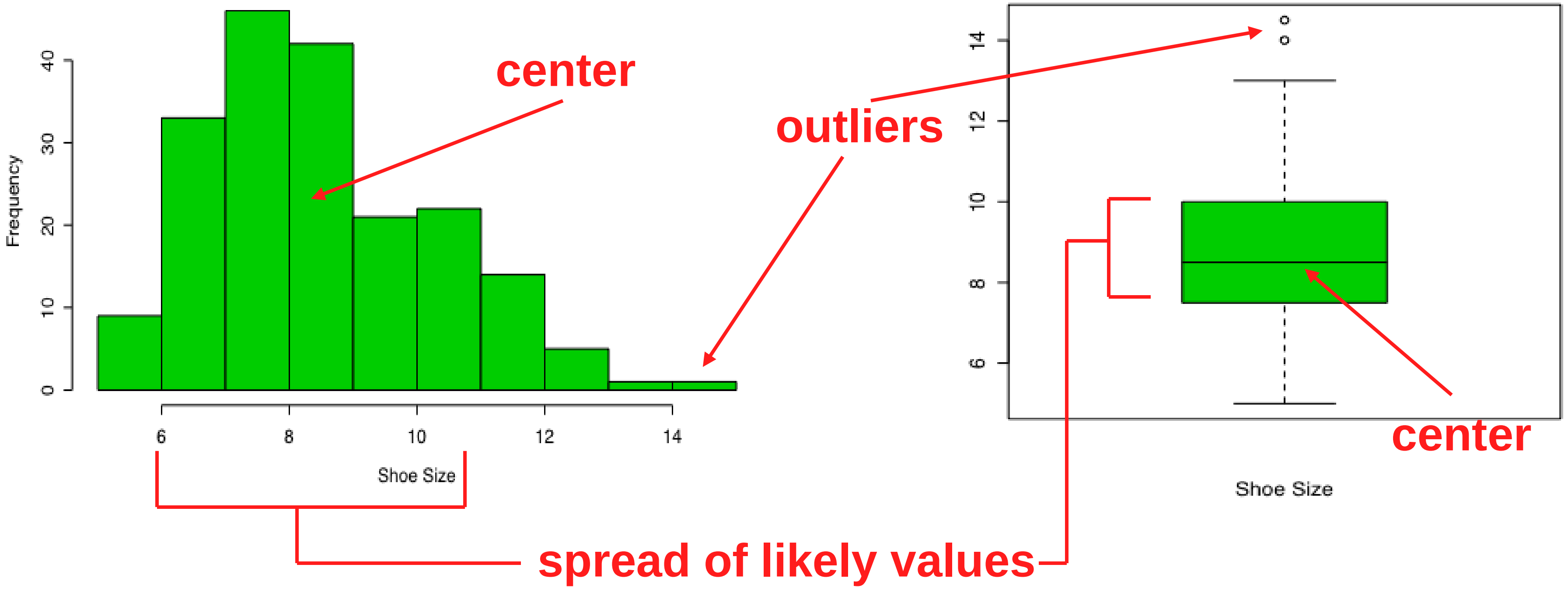
- Almost all values are between 5 and 13
- 50% of values are between 7.5 and 10
- Center (Median) is around 8.5
- Couple of suspected outliers: 14 and 14.5

The 5-point summary & Outliers

- The boxplot provides a ‘5-point summary’ of the dataset (a description of the center, two descriptions of dispersion or spread, another two measures of location)
- What are they??
- L.O.S.S. (location, outliers, spread and shape)
- An outlier is an (exceptional) data-point that falls outside the LEFT & RIGHT FENCES in a boxplot (LEFT FENCE = Q1 – 1.5*IQR, RIGHT FENCE = Q3 + 1.5*IQR)

Histograms versus Boxplots

- Both graphs give a good idea of the **spread**
- Boxplots may be a little clearer in terms of the **center** and **outliers** in a distribution



center

outliers

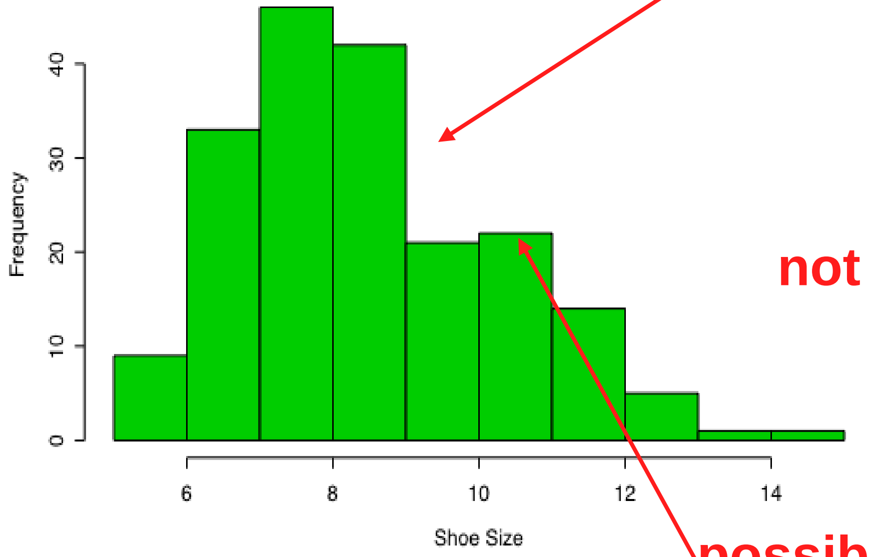
center

spread of likely values

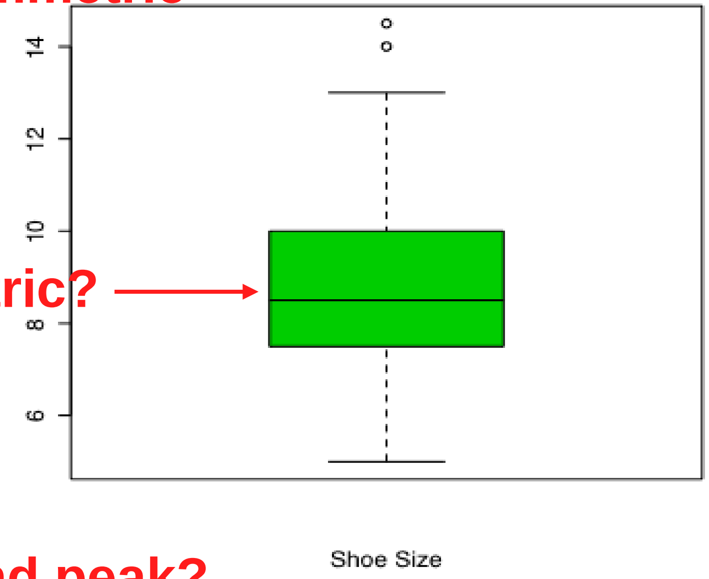
Histograms versus Boxplots

- Histograms much more effective at displaying the **shape** of a distribution
  - **Skewness**: departure from left-right symmetry
  - **Multi-modality**: presence of multiple high frequency values

clearly not symmetric

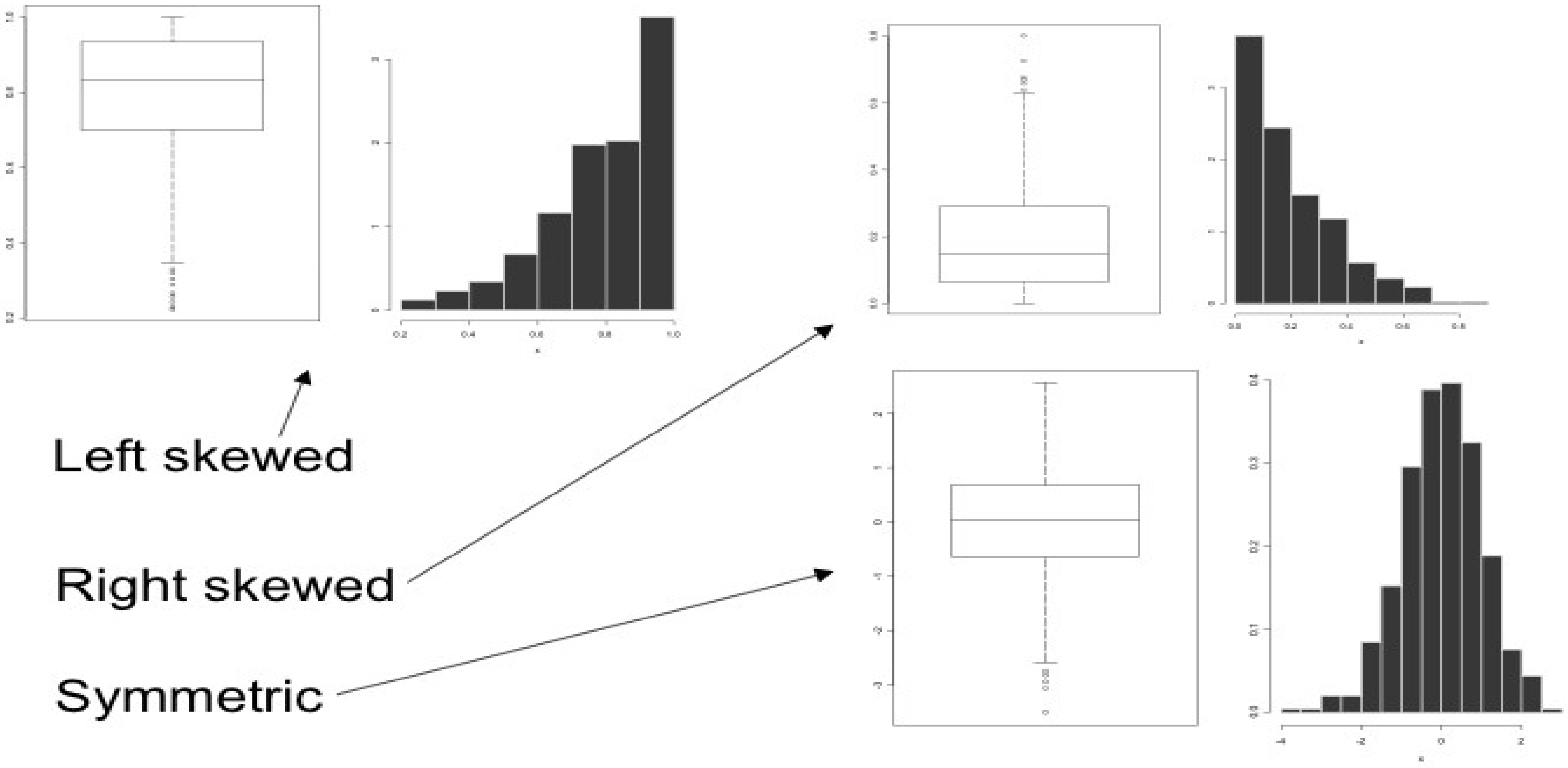


not symmetric?



possible second peak?

Symmetry - Histograms vs. Boxplots



Left skewed

Right skewed

Symmetric

Stem-and-Leaf Plots

Dataset: {20.6 21.9 25.1 20.9 23.5 22.7 24.2 20.6 25.1 29.9 22.3 24.9 24.4 25.7 25.2 25.5 28.7 27.1 28.4 26.1 26.6 26.9 21.3 21.7}

- **Stem-and-Leaf Display:**
-
- Stem-and-leaf of C1 N = 24
- Leaf Unit = 0.10
-

```
20  669
21  379
22  37
23  5
24  249
25  11257
26  169
27  1
28  47
29  9
```

A stem-and-leaf plot is another picture of the dataset that's kind of like the histogram but is more informative than it, as you see the data-values themselves.

Steps in constructing it:

- For each data-value, call the last or the last two digits (even if they are digits after the decimal point) the *leaf* & the remaining digits the *stem* (example: for the data-value 20.9, the leaf is 9 and the stem is 20; for the data-value 150.65, the leaf can be 5 and the stem can be 150.6)
- Make a vertical list of all the distinct stems present in the data (ordered from smallest to largest)
- Next to each stem, list all the leaves in the dataset (smallest to largest)