CP Statistics Chapter 1 Notes - Exploring Data

1.1: Displaying Distributions with Graphs

Individuals and Variables

- **Individuals** are objects described by a set of data. Individuals may be people, but they may also be animals or things.
- A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Categorical and Quantitative Variables

- A **categorical variable** places an individual into one of several groups or categories.
- A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

Distribution

The **distribution** of a variable tells us what values the variable takes and how often it takes these variables.

Describing the Overall Pattern of a Distribution – Remember your SOCS

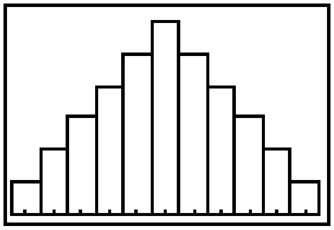
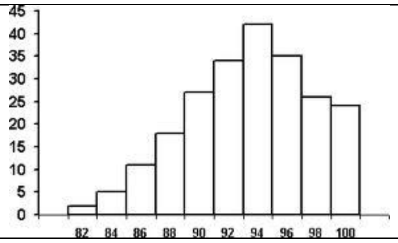
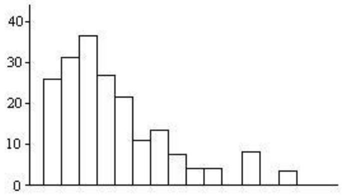
To describe the overall pattern of a distribution, address all of the following:

- **S**pread – give the lowest and highest value in the data set
- **O**utliers – are there any values that stand out as unusual?
- **C**enter – what is the approximate average value of the data (only an estimation)
- **S**hape – does the graph show symmetry, or is it skewed in one direction (see below)

Outliers

An outlier in any graph of data is an individual observation that falls outside the overall pattern of the graph.

Describing the SHAPE of a distribution – Symmetric and Skewed Distributions

Symmetric	Skewed Left	Skewed Right
		
Mean = Median	**Mean < Median**	**Mean > Median**

Time Plot

- A **time plot** of a variable plots each observation against the time at which it was measured.
- Always mark the time scale on the horizontal axis and the variable of interest on the vertical axis. If there are not too many points, connecting the points by lines helps show the pattern of changes over time.

1.2: Describing Distributions with Numbers

The Mean ($\overline{x}$)

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are x1, x2, …, xn, their mean is:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \qquad \text{or simply,} \qquad \overline{x} = \sum_{i=1}^{n} x_i$$

The Median (M)

- The median M is the midpoint of distribution, the number such that half the observations are smaller and the other half are larger. To find the median of distribution:
- Arrange all observation in order of size, from smallest to largest.
- If the number of observations n is odd, the median M is the center observation in the ordered list. The position of the center observation can be found at (n + 1) / 2
- If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The position of the two middle values are n/2 and n/2 + 1

The Five-Number Summary

The five-number summary of a data set consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is:

Minimum – Q_1 – M – Q_3 – Maximum

The Quartiles (Q_1 and Q_3)

- To calculate the quartiles, arrange the observations in increasing order and locate the median M in the ordered list of observations.
- The 1st quartile (Q1) is middle number of the values that are less than the median.
- The 3rd quartile (Q3) is the middle number of the values that are greater than the median.

Example

2	14	28	29	30	32	33	34	40	42	52
Min		**Q1**			**Med**			**Q3**		**Max**

The Interquartile Range (IQR)

The IQR is the distance between the first and third quartiles, IQR = Q3 - Q1

Outliers: The 1.5 x IQR Criterion

Call an observation an outlier if it falls more than 1.5 x IQR below the first quartile or above the third quartile. Using the 5-number summary from above as an example (IQR = 40-28=12)

- Low outlier cutoff: $Q_1 - 1.5 \times IQR$ (example: 28 – 1.5(12) = 28 – 18 = 10) Therefore, the 2 is an outlier.
- High outlier cutoff: $Q_3 + 1.5 \times IQR$ (example: 40 + 1.5(12) = 40 + 18 = 58) no outlier

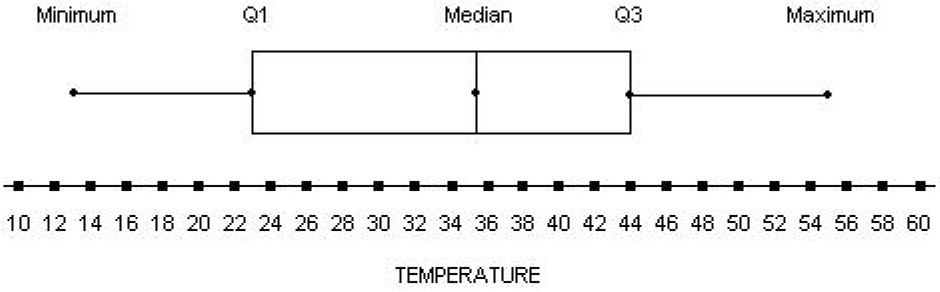
1.2: Describing Distributions with Numbers

Boxplot

A boxplot is a graph of the five-number summary, with outliers plotted individually.

- A central box spans the quartiles.
- A line in the box marks the median.
- Observations more than 1.5 x IQR outside the central box are plotted individually.
- Lines extend from the box out to the smallest and largest observations, not the outliers.

Example:



The Standard Deviation (S or Sx)

The standard deviation of a set of observations is the average of the squares of the deviations of the observations from their mean. The formula for the standard deviation of *n* observations $x_1, x_2, \ldots, x_n$ is:

$$s = \sqrt{\frac{\sum (x_i - x)^2}{n-1}}$$

Calculation of the Standard Deviation

Consider the data below which has a mean of 4.8:

x_i	x_i – mean	$(x_i\text{-mean})^2$
6	6 – 4.8 = 1.2	$(1.2)^2 = 1.44$
3	3 – 4.8 = -1.8	$(-1.8)^2 = 3.24$
8	8 – 4.8 = 3.2	$(3.2)^2 = 10.24$
5	5 – 4.8 = 0.2	$(0.2)^2 = 0.04$
2	2 – 4.8 = -2.8	$(-2.8)^2 = 7.84$
Sum	**0**	**22.8**

So the standard deviation is $\sqrt{22.8/(5-1)} = \sqrt{22.8/4} = \sqrt{5.7} = 2.387$