3 Graphing and Summarizing Data

MTH 3240 Environmental Statistics

Spring 2020

Notes

Objectives

Objectives:

- Interpret histograms.
- Interpret measures of center (mean, median).
- Interpret measures of variation (variance, standard deviation).
- Interpret box plots.

Notes

The Distribution of a Variable

- The ***distribution*** of a variable refers to the set of values that the variable takes and the frequencies with which it takes those values.
- Distributions can be displayed as *histograms*.

Notes

Histograms

- ***Histograms*** show values of the variable on the horizontal axis and bars whose heights indicate the frequencies of the values in the data set (grouped into intervals).

Example

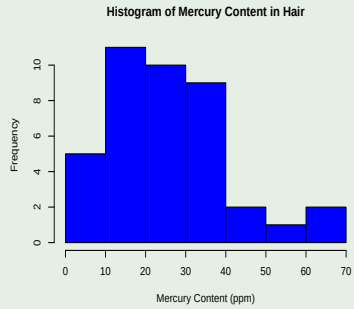
Citizens of the Republic of Seychelles are among those who consume the most fish in the world, much of it predatory species.

The data below are **mercury** contents (ppm) in the hair of $n = 40$ fishermen in the Seychelles. A **histogram** of the data follows.

Notes

13.3	32.4	18.1	58.2	64.0	68.2	35.4	33.9	23.9
18.3	22.1	39.1	31.4	18.5	21.0	5.5	7.9	5.2
28.7	26.3	13.9	25.9	9.8	26.9	16.8	37.7	19.6
21.8	31.6	30.1	42.4	16.5	21.2	33.0	9.8	10.6
29.6	40.7	12.9	13.8					

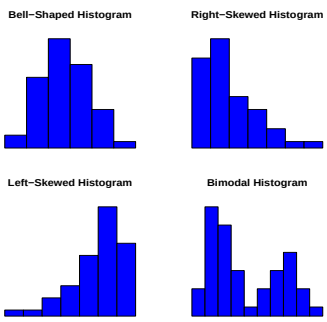
Notes



Notes

- A distribution **shape** is usually one of:
  - ***Bell-shaped***
  - ***Right skewed*** (long "tail" extending to the right)
  - ***Left skewed*** (long "tail" extending to the left)
  - ***Bimodal*** (having two peaks)

Notes



Notes

Measures of Center

- A ***statistic*** is any numerical quantity calculated from a set of random sample data.
- To summarize the **center** of a set of data, we usually use:
  1. The *sample mean*
  2. The *sample median*

Notes

- The ***sample mean*** $\bar{X}$ is the **average value** in the data.

Sample Mean: For data $X_1, X_2, \ldots, X_n$, the sample mean is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \,.$$

Notes

- The ***sample median*** (or ***50th percentile***) $\tilde{X}$ is **middle value** (after ordering the data from smallest to largest).

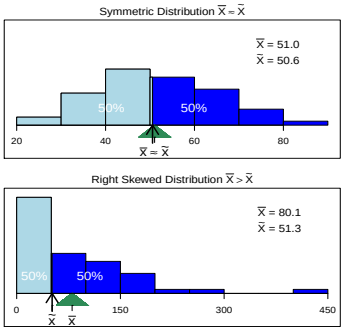
Sample Median: For data $X_1, X_2, \ldots, X_n$, the sample median is

$$\tilde{X} = \begin{cases} \text{The } (\frac{n+1}{2})\text{th ordered value if } n \text{ is odd.} \\ \text{The average of the } (\frac{n}{2})\text{th and the} \\ \quad (\frac{n+2}{2})\text{th ordered values if } n \text{ is even.} \end{cases}$$

Notes

- The **median** is **resistant** to outliers, the **mean isn't**.
- In the figures below, the **mean** is the "balancing point" and the median is the "equal areas point".

Notes



Notes

- As seen in the figures:
  - For data with a **bell-shaped** distribution, the **mean** and **median** will be approximately **equal**, $\bar{X} \approx \tilde{X}$.
  - For data with a **right skewed** distribution, the **mean** will be **larger** than the **median**, $\bar{X} > \tilde{X}$.

Notes

Measures of Variation

- To summarize the **variation** of a set of data, we usually use the **sample standard deviation** (or its square, the **sample variance**).

Notes

- The ***sample standard deviation*** S represents an **"average" deviation** away from the **mean**.

Sample Standard Deviation: For data $X_1, X_2, \ldots, X_n$, the sample standard deviation is

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

- The **standard deviation** is measured in the **same units** as the data.

Notes

- The **square** of the standard deviation is called the ***sample variance*** S^2, and represents the **"average" squared deviation** away from the mean.

Sample Variance: For data $X_1, X_2, \ldots, X_n$, the sample variance is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- The **variance** is measured in the **squared units** of the data.

Notes

- The **variance** and **standard deviation *aren't*** **resistant** to outliers.

  (The ***interquartile range*** is another measure of variation that ***is*** **resistant**.)

Notes

Choosing Measures of Center and Variation

- **Choosing a Measures of Center**:
  - For data with a symmetric distribution and no outliers, the mean is preferred.
  - For data with a right skewed distribution or with outliers, the median is preferred.
- **Choosing a Measure of Variation**:
  - The standard deviation is preferred whenever the mean is used to summarize the center of the data.
  - The interquartile range is preferred when the median is used to summarize the center.

Notes

Boxplots

- A ***boxplot*** is a graphical display of five statistics: Minimum, Q_1, $\tilde{X}$, Q_3, Maximum.

  (Q_1 and Q_3 are the ***1st*** and ***3rd quartiles*** (i.e. the **25th** and **75th sample percentiles**).

Notes

Notes

Example

The data on the next slide are **benzene** concentrations (ppb) at two locations in the South Platte River after a spill of toxic materials from the Suncor Energy oil refinery north of Denver.

Notes

Benzene in South Platte River

Date	Benzene at Confluence with Sand Creek	Benzene Downstream from the Confluence
Dec. 27	640	190
Dec. 28	240	300
Dec. 29	140	130
Dec. 30	190	130
Dec. 31	170	160
Jan. 2	300	240
Jan. 3	730	250
Jan. 4	630	240
Jan. 5	650	240
Jan. 6	190	590
Jan. 7	310	260
Jan. 8	400	260
Jan. 9	720	240

Notes

Here's the sample from the **first location**, sorted smallest to largest:

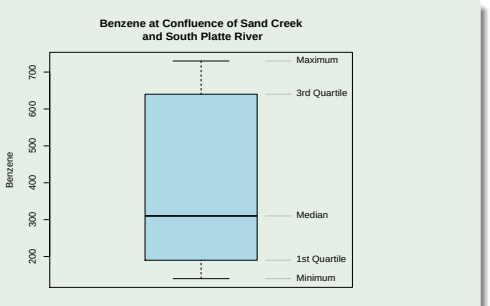
140 170 190 190 240 300 310
400 630 640 650 720 730

The five numbers used in a boxplot of the data are:

Min	Q_1	$\tilde{X}$	Q_3	Max
140	190	310	640	730

The **boxplot** is shown on the next slide.

Notes

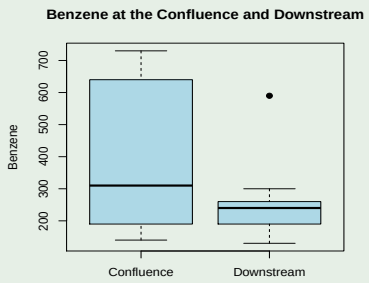


- **Boxplots** are used mainly to **compare** samples taken from **two or more** populations side by side in the same graph.

Example

The side by side boxplots for comparing the **benzene** concentrations for the **two locations** in the South Platte River are on the next slide.

Notes



Notes

Notes

Notes