P8130: Biostatistical Methods I

Lecture 2: Descriptive Statistics

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

Lecture 1: Recap

- Intro to Biostatistics
- Types of Data
- Study Designs

Descriptive Statistics

- The collection and presentation of the data through graphical and numerical displays
- Look for patterns in the data and summarize information
  - Measures of location
  - Measures of dispersion
  - Graphical display

Measures of Location

- Measures of location or *central tendency* indicate the center of the data
- Mean (average)
- Median (the 50^{th} percentile)
- Mode

Measures of Location: Mean

Definition: the arithmetic mean represents the sum of all observations divided by the number of observations

Sample mean for a sample of *n* observations is given by:

$$\overline{x} = \sum_{i=1}^{n} x_i / n$$

Sample mean is used to estimate the population mean μ which is typically unknown

Measures of Location: Mean

- The most common used measure of location
- Overly sensitive to outliers (unusual observations), thus not recommended if the data are skewed
- Not appropriate for nominal or categorical variables

Measures of Location: Median

Definition: The sample median is computed as:

1. If n is odd, median is computed as $\left(\frac{n+1}{2}\right)^{th}$ largest item in the sample
2. If n is even, median is computed as the average between $\left(\frac{n}{2}\right)$ and $\left(\frac{n}{2}+1\right)^{th}$ largest items

Example:

Given n=7 (odd) total sample observations, median is the $\frac{7+1}{2} = 4^{th}$ largest item

Given n=10 (even) total sample observations, median is the average of the

$\frac{10}{2} = 5th$ and $\frac{10}{2} + 1 = 6th$ largest items

Measures of Location: Median

- Compared to the *mean,* the median is not affected by every value in the data set including outliers
- The median is defined as the middle value or the 50^{th} percentile
  - This means that half of the data are less than or equal to it, and at least are greater tan or equal to it
- Median calculation starts by first ordering the data (increasing order)
- Appropriate measure for ordinal data

Other Measures of Location

Percentiles: median is the 50th percentile

- In general: the kth percentile is a value such that most k% of the data are smaller than it and (100-k)% are larger
- Deciles: 10th, 20th, 30th, …
- Quartiles: 25th (Q1), 50th, 75th (Q3)

- Question: what does it mean if your GRE score is in the 90th percentile?

Measures of Location: Mode

Definition: the most frequently occurring value in the data

- You can have multiple modes or none (really?)

- Problematic if there is a large number of possible values with infrequent occurrence

Measures of Dispersion

Describe the spread of the data:

- Range
- Inter-quartile range (IQR)
- Variance/Standard deviation
- Coefficient of variation (CV)

Measures of Dispersion

Range: Max – Min

Inter-quartile range: IQR = 75th (Q3) – 25th (Q1)

Since the range only depends on the minimum and maximum values, it can be influenced by the extremes

Solution? Use the IQR

Measures of Dispersion

Population Variance is the average squared deviation from the mean:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Population Standard Deviation is just the square root of the variance:

$$\sigma = \sqrt{\sigma^2}$$

Values often unknown and then we refer back to sample …

Measures of Dispersion

Sample Variance is the average squared deviation from the mean:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Population Standard Deviation is just the square root of the variance:

$$\mathrm{s} = \sqrt{s^2}$$

Lots of changes in notation and also formula!!

Measures of Dispersion

Mean and standard deviations are the most used measures of location and spread.

Why? It’s all about the …

Property: linear transformations do affect these measures

Let $Y = cX + b$ be a linear transformation a variable X

Mean of $Y = c\overline{X} + b$

Standard Deviation $s_Y = cs_X$

Measures of Dispersion

Coefficient of Variation (CV) is a measure that relates the mean and the standard deviation.

- Sometimes the variance changes with its mean

- Population: $CV = \frac{\sigma}{\mu} \times 100\%$

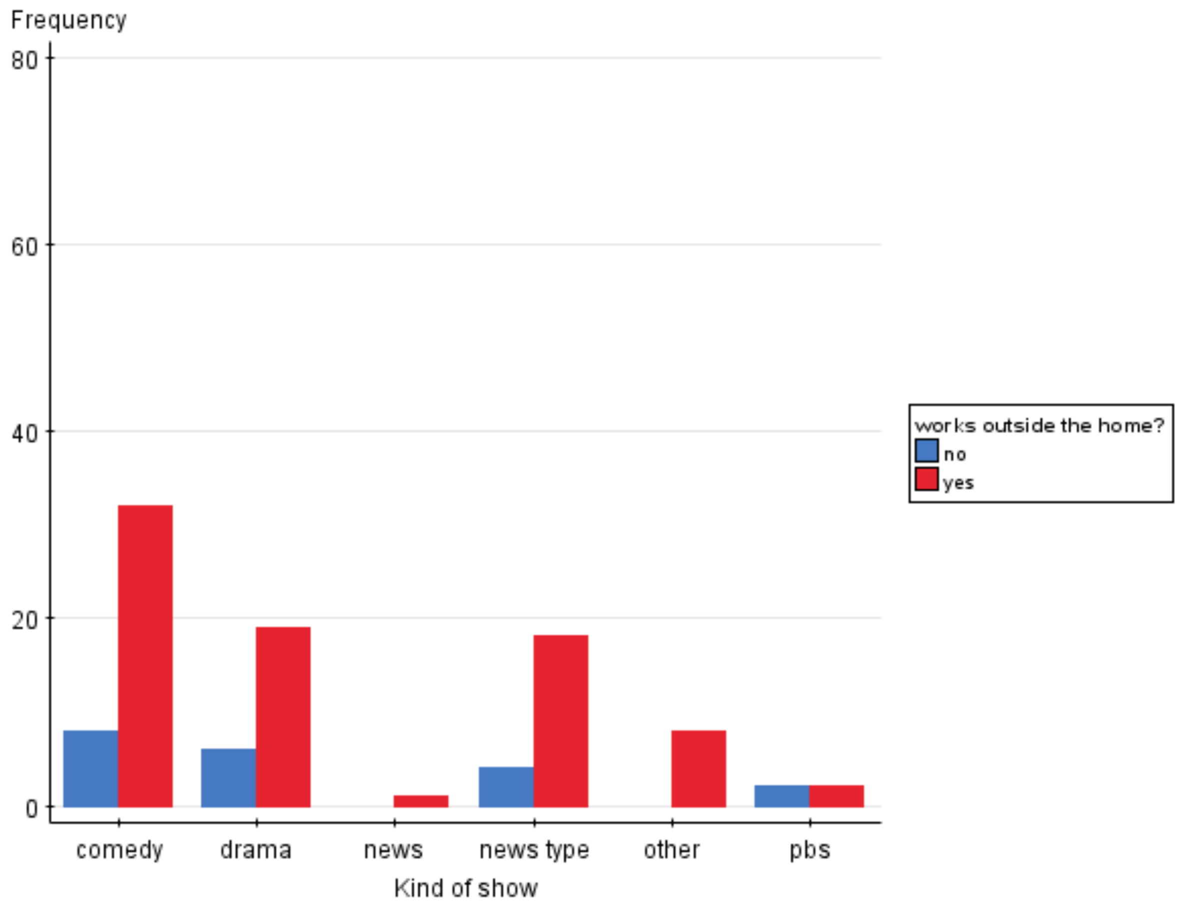
- Sample: $CV = \frac{s}{\bar{x}} \times 100\%$

- CV is unitless and can be interpreted in terms of variability to the average

Graphical Display

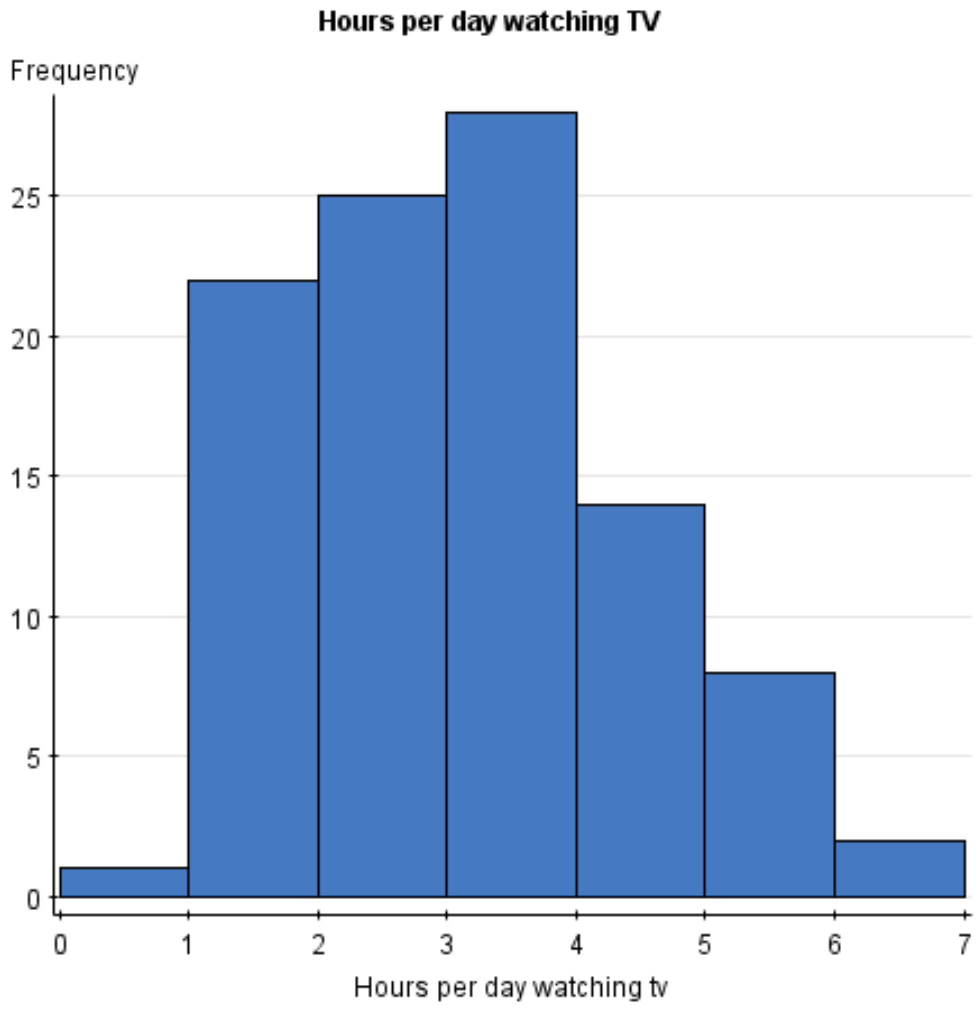
- A picture is worth a thousand words (sometimes)
- Bar graphs
- Histograms
- Box-plots
- Scatter plots (later in linear regression)

Bar Graph



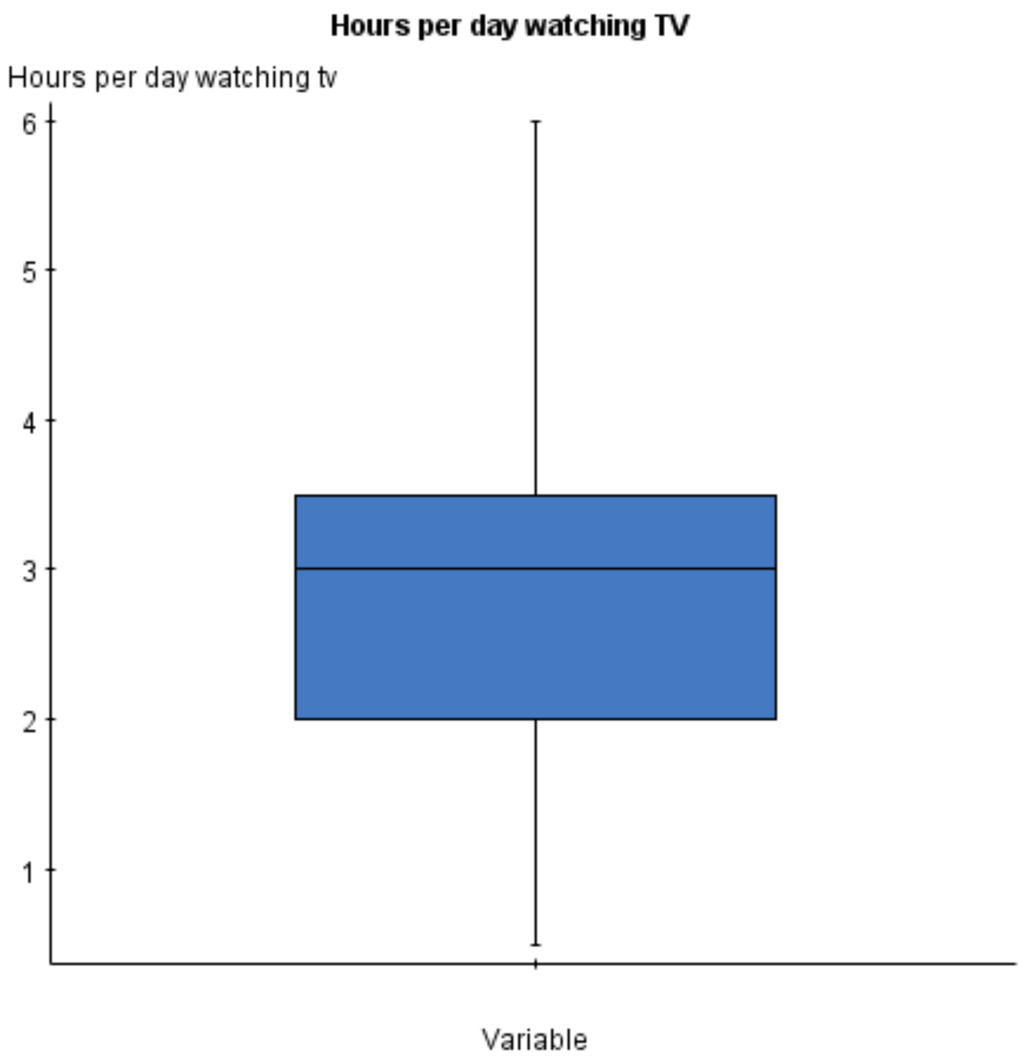
- Data are divided into groups and frequencies are determined for each group
- Rectangles are constructed with the base of constant width and heights proportional to the frequencies

Histogram



- Numerical values are grouped into measurements classes, defined by equal-length intervals along the numerical scale
- Each value belongs to only one class
- Usually 5-12 classes
- Like bar graph, this plot has frequencies on the vertical axis
- If the mean > median: right skew
- If the mean < median: left skew

Box-plot



- Extends from the Q1(25th) to the Q3(75th) quartile – the box
- The ‘whiskers’ extend from the smallest to the largest values
- If one of the whiskers is long, it indicates skewness in that direction
- If a data value is less than Q1 – 1.5(IQR) or greater than Q3 + 1.5(IQR), then it is considered an outlier and given a separate mark on the boxplot

Readings

Rosner, *Fundamentals of Biostatistics*, Chapter 2

- Sections: 2.2 – 2.6
- Sections: 2.9 – 2.10