Summarizing Data

Jerome Dumortier

29 September 2021

Lecture Overview

Measures of central tendency

- Mean (arithmetic and weighted), median, mode

Measures of dispersion

- Range
- Variance and standard deviation
- Interquartile range (IQR)
- Coefficient of variation (CV)

Graphical summaries

- Histograms and skewness
- Empirical cumulative distribution functions
- Boxplots

Covariance and correlation coefficicent

Central Tendency: Arithmetic Mean and Weighted Mean

Arithmetic mean:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Weighted mean:

- Suppose you have a set of observations $x_1, x_2, \ldots, x_n$ and a set of corresponding weights $w_1, w_2, \ldots, w_n$. Then the equation for the weighted mean is written as follows:

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \cdots w_n x_n}{w_1 + w_2 + \cdots w_n}$$

Arithmetic Mean

Entering the data in R:

```
state1 = c(10,10,10,10,10,10,10,10,10,10)
state2 = c(2,3,7,8,9,10,10,14,17,20)
state3 = c(2,2,2,2,2,2,2,2,2,82)
states = data.frame(state1,state2,state3)
rm(state1,state2,state3)
```

Income distribution among citizens in three states. Note that the average income in all three states is 10. Although, there is considerable variation among the citizens.

Weighted Mean

Suppose you take an O'Neill class and the professor bases the grades on homework (10%), midterm exam (20%), final exam (30%), term paper (25%), presentation of term paper (10%), and participation (5%). Each item is based on a 100 point scale and you score 85, 57, 78, 92, 95, and 10 points, respectively.

```r
weights = c(10,20,30,25,10,5)
scores  = c(85,57,78,92,95,10)
weighted.mean(scores,weights)
```

```
## [1] 76.3
```

Mode and Median

Mode

- The mode is the value of the observations that appears the most often.

Median

- The median is the value that divides the data set into two equal part. 50% of the observations are below the value and 50% are above the value.
- Eliminates the influence of very small or very large values.
- There is a unique median for each data set.

What are the mode and the median for the example data with the three different states?

Summarizing Data in R

```r
summary(states)
```

```
##      state1       state2          state3
##  Min.   :10   Min.   : 2.00   Min.   : 2
##  1st Qu.:10   1st Qu.: 7.25   1st Qu.: 2
##  Median :10   Median : 9.50   Median : 2
##  Mean   :10   Mean   :10.00   Mean   :10
##  3rd Qu.:10   3rd Qu.:13.00   3rd Qu.: 2
##  Max.   :10   Max.   :20.00   Max.   :82
```

Range, Variance, and Standard Deviation

Range

- The range is the largest value minus the smallest value.

Population variance:

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

Population standard deviation:

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

The standard deviation is simply the square root of the variance.

Calculation of the Variance

Consider the following score from four homework assignments: 85, 56, 71, 92. The population variance calculation is done as follows:

$$Var = \frac{(85-76)^2 + (56-76)^2 + (71-76)^2 + (92-76)^2}{4} = 190.5$$

Variance and Standard Deviation for a Sample

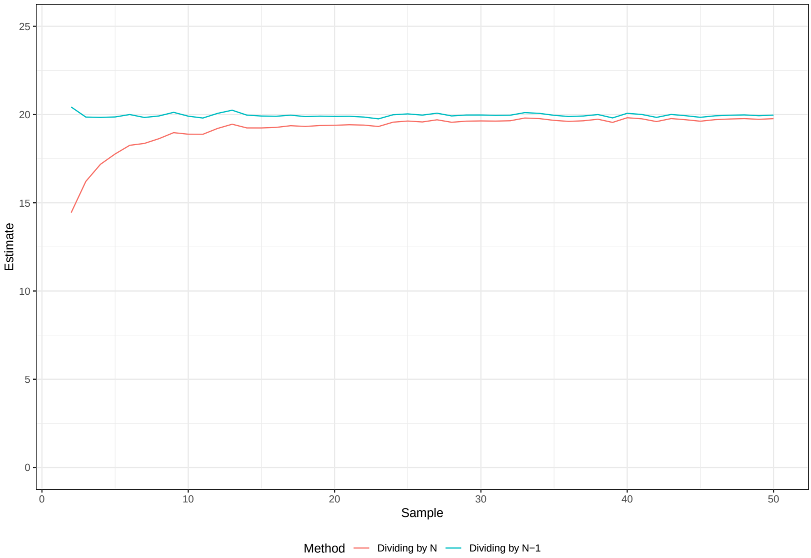
Sample variance

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

Sample standard deviation

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Sample Variance



Summarizing Data with R

Data: `mh1`

- Sample of 101 housing values in the Meridian Hills neighborhood of Indianapolis

Do and answer the following:

- Use the command `summary()` to summarize the data.
- What does the command `range()` do?
- What does the command `diff(range(mh1))` do?

Variance and Standard Deviation in R

Assume that the data set of the Meridian Hills homes is not a sample but that it represents the population of all homes! Calculate the population variance as follows:

```
varp   = function(x) mean((x-mean(x))^2)
stdevp = function(x) sqrt(mean((x-mean(x))^2))
varp(mh1$price)
```

```
## [1] 98220205028
```

```
stdevp(mh1$price)
```

```
## [1] 313401
```

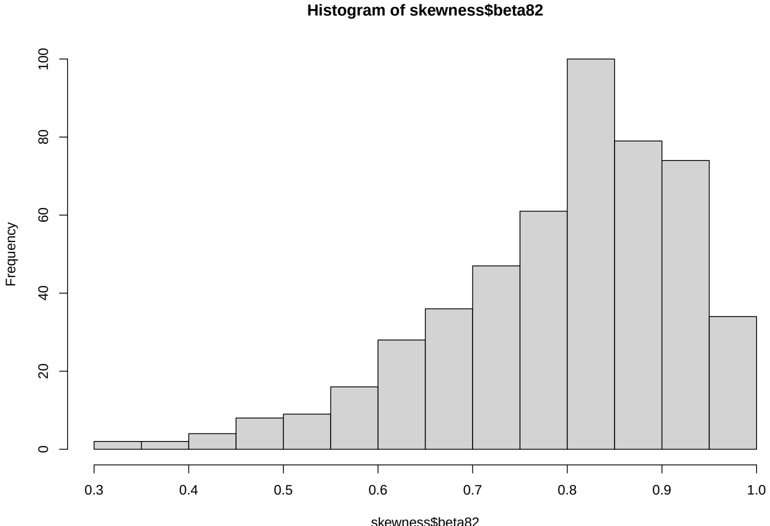
Coefficient of Variation

The coefficient of variation standardizes the standard deviation σ by the mean, i.e., $CV = \sigma/\mu$. Because the magnitude of the standard deviation depends on the mean, it is sometimes necessary to calculate the coefficient of variation to make two or more standard deviations comparable. For example, suppose you are comparing residential home values in California and Indiana. You calculate the mean and standard deviation for California as \$2,000,000 and \$400,000, respectively. The mean and standard deviation for Indiana are \$125,000 and \$50,000, respectively. Calculating the coefficient of variation (CV) for both states leads to $CV_{CA} = 0.2$ and $CV_{IN} = 0.4$. Hence, there is much more variation in the home values for Indiana than there is in California.

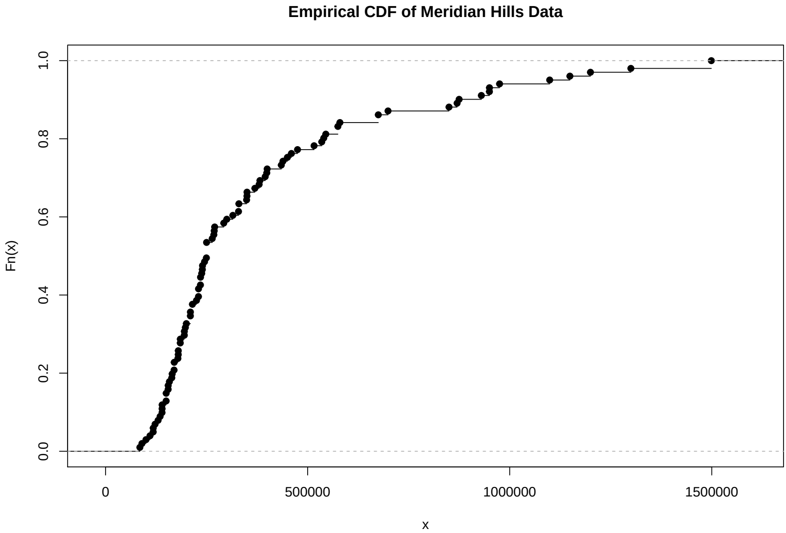
Skewness

Skewed left (negative skewness): Long tail to the left

```
hist(skewness$beta82)
```



Graphical Summaries: Empirical Distribution Function



Empirical Distribution Function: Quantiles

Quantiles are the value that divide the ordered observations into n subsets each containing the same percentage of observations. There are $n - 1$ quantiles.

- Median divides the observation into two subsets each containing 50% of the observations.

Specific quantiles:

- Quartiles: 25%
- Quintiles: 20%
- Percentiles: 1%

Interquartile Range: Norway Murder Data

```r
norway = subset(eucrime,geo=="NO")
quantile(norway$values,c(0.25,0.5,0.75))
```

```
##   25%   50%   75% 
## 27.00 28.50 30.25 
```

```r
summary(norway$values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   24.00   27.00   28.50   36.42   30.25  111.00 
```

Interquartile Range: Norway Murder Data

IQR

- $30.25 - 27 = 3.25$

Lower and upper bounds

- Lower bound: $27 - 1.5 \cdot 3.25 = 22.125$. Thus, it is simply the minimum value of 24. There are no outliers.
- Upper bound: $30.25 + 1.5 \cdot 3.25 = 35.125$. Thus, it is the value of 35.125. There are two outliers.

Interquartile Range: Meridian Hills Data

```r
quantile(mh1$price,c(0.25,0.5,0.75))
```

```
##    25%    50%    75%
## 179998 249900 450000
```

IQR

- $450,000 - 179,998 = 270,002$

Lower and upper bounds

- Lower bound: $179,998 - 1.5 \cdot 270,002 = -225,005$. Thus, it is simply the minimum value of $84,900.
- Upper bound: $450,000 + 1.5 \cdot 270,002 = 855,003$. Note that R displays the last observation within that bound, i.e., $849,900. Every observation above that value are outliers.

Covariance and Correlation

Data in `mh2` also contains information about the square footage of the homes. The covariance (and later the correlation coefficient) measures joint variability and movement of two random variables. Two definitions of covariance:

$$Cov(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(X))(y_i - E(Y)) = E(X \cdot Y) - E(X)E(Y)$$

Sign of the covariance:

- Positive: X and Y move in the same direction
- Negative: X and Y move in the opposite direction
- Zero: X and Y are uncorrelated

If two random variables X and Y are independent, then X and Y are uncorrelated.

Covariance between Price and Square Footage

Calculating covariance in R:

```
cov(mh2$price,mh2$sqft)
```

```
## [1] 1076268259
```

Issues associated with covariance:

- Useful to determine the direction of change but not the magnitude
- Transform the square footage in square meters (i.e., divide the square footage by 10.764)
- Using different units changes the covariance although nothing has changed in terms of relationship

The correlation coefficient overcomes those issues.

Correlation Coefficient

Correlation does not mean causation!

- Causation requires a strong theoretical believe that one variable is the cause of another variable, e.g., influence of education on income

Correlation coefficient (sometimes called Pearson's r)

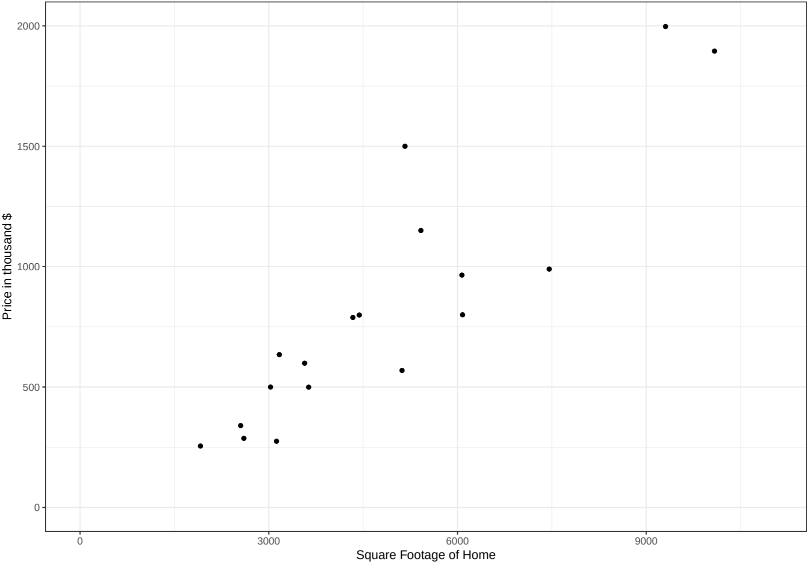
$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

Properties

- Varies between -1 and 1
- Sign provides the direction
- Value provides the magnitude

Note that the correlation coefficient has no dimensions!

Scatter Plot between Price and Square Footage



Simpson's Paradox: Baseball Example

Definitions:

- AB: At bats
- H: Hits, hits allowed

Derek Jeter

- 1995: AB = 48, H = 12
- 1996: AB = 582, H = 183

David Justice

- 1995: AB = 411, H = 104
- 1996: AB = 140, H = 45