Descriptive statistics

Petr Ocelík

ESS401 Social Science Methodology

13th October 2015

Outline

- Measures of central tendency, position, and variability.
- Graphic displays of descriptive statistics.
- R introduction: cont’d.

Descriptive statistics

- The purpose is to **summarize data**.
- Quantitative variables have two key features:
  - The **center** of the data – a typical observation.
  - The **variability** of the data – the spread around the center.

Notation

	Mean	Standard Deviation	Variance
Population	μ	σ	σ^2
Sample	$\bar{x}$	s	s^2

$\sum$ = "the sum of ..."

n = number of pieces of data (population)

$n-1$ = number of pieces of data (sample)

$\bar{x}$ = mean (average) of data

x_i = each of the values in the data

$x_1, x_2, x_3, x_4, \ldots x_n$ (as i goes from 1 to n)

Central tendency

- The statistics that describe **the center of a frequency** distribution for a quantitative variable.
- Shows a **typical** observation/case.
- Most common measures: mean, mode, and median.

Central tendency: mean

- **Arithmetic mean**

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

- **Properties:**
  - Center of gravity of a distribution.
  - Can be used **only for metric scales.**
  - Strongly influenced by outliers.

Central tendency: mode

- Value that **occurs most frequently** in the sample.
- Applicable at **all levels of measurement**.
- Used mainly for highly discrete variables such as **categorical data**.

- {“catholic”, “Muslim”, “Hindu”, “catholic”, “catholic”, “Muslim”, “catholic”, “catholic”}
- {1, 2, 3, 1, 1, 2, 1, 1}
- {“agree”, “agree”, “disagree”, “agree”, “neutral”, “disagree”, “disagree”, “disagree”, “agree”}
- {1, 1, -1, 1, 0, -1, -1, -1, 1}
- Years of education.

Central tendency: median

- Observation that is in **the middle of the ordered sample** (between 50th bottom and 50th upper percentile).
- Splits data into **two parts with equal # of observations**.
- For even sized samples: average value of the two middle observations.
- Applicable **at least at ordinal level**.

Central tendency: median

- Identification of median: **(n + 1) / 2 ;**

  n = # of observations in the data

- **Odd** numbered *n*: {1, 1, 2, 2, 3, 3, **5**, 6, 6, 6, 7, 10, 39}
- Median = (13 + 1)/2 = 7^{th} position = **5**

- **Even** numbered *n*: {1, 1, 2, 2, 3, **3, 5**, 6, 6, 6, 7, 10}
- Median = (12 + 1)/2 = 6.5^{th} position
  = (6^{th} + 7^{th} position)/2 = (3 + 5)/2 = **4**

Central tendency: median

Set 1	8	9	**10**	11	12
Set 2	8	9	**10**	11	100
Set 3	0	9	**10**	10	10
Set 4	8	9	**10**	100	100

Finlan & Agresti 2009: 43

Central tendency

- Mode
- Median
- Mean
- {1, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, 39}

Central tendency

- Mode
- Median
- Mean
- {1, 1, 2, 2, 3, 3, **5,** **6, 6, 6,** **7,** 10, 39}

Position

- The measures of central tendency are not sufficient for description of data for a quantitative variable.
- Does not describe the **spread of the data**.

- **Position measures:** describe the point at which a given percentage of the data fall below or above that point.

Position: percentile

- **Percentile.** The *pth* percentile is the point such that *p%* of the observations fall below that point and (and 100 - p)% fall above it.

  - E.g. 89th percentile = indicates a point where 89% of observations lie below and 11% lie above it.
  - **Median is a 50th percentile**.
  - "Standard" percentiles: (25, 50, 75), or (10, 25, 50, 75, 90).

Position: IQR

- **Interquartile range**
  - Difference between the values of observations at **75%** (upper quartile) and **25%** (lower quartile).
  - Shows spread of middle half of the observations.

{1, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, 39}

Median = (13 + 1)/2 = 7^{th} observation = 5

Q1 = (6 + 1)/2 = 3.5^{th} observation = (2 + 2)/2 = 2

Q2 = (6 + 1)/2 = 3.5^{th} observation = (6 + 7)/2 = 6.5

IQR = Q3 – Q1

IQR = 6.5 – 2 = **4.5**

Position: quartile

- **Quartile**
  - Values of observations at 25% (Q1), 50% (Q2), and 75% (Q3) of a distribution.

{1, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, 39}

Q1 (25 %) = 2

Q2 (50 %) = 5

Q3 (75 %) = 6.5

Measures of center and position: R commands

mean(data)

mode does not have standard R function

median(data)

range(data)

IQR(data, na.rm=F)

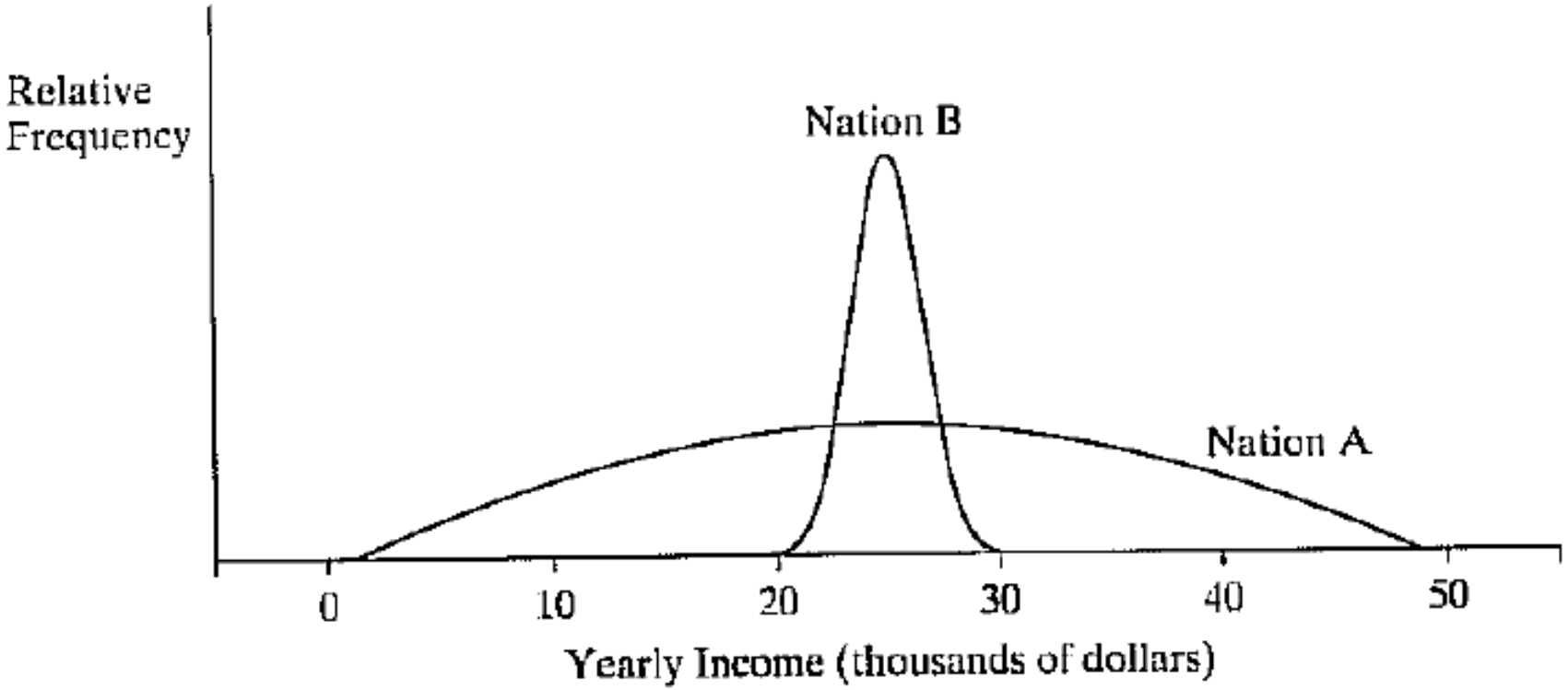
quantile(data, c(0.25, 0.5, 0.75))

Variability

- The measures of central tendency are not sufficient for description of data for a quantitative variable.
- Does not describe the **spread of the data**.

- **Variability measures:** describe the deviations of the data from a measure of center (such as mean).
  - With exception of a **range**.

Variability



Finlan & Agresti 2009: 46

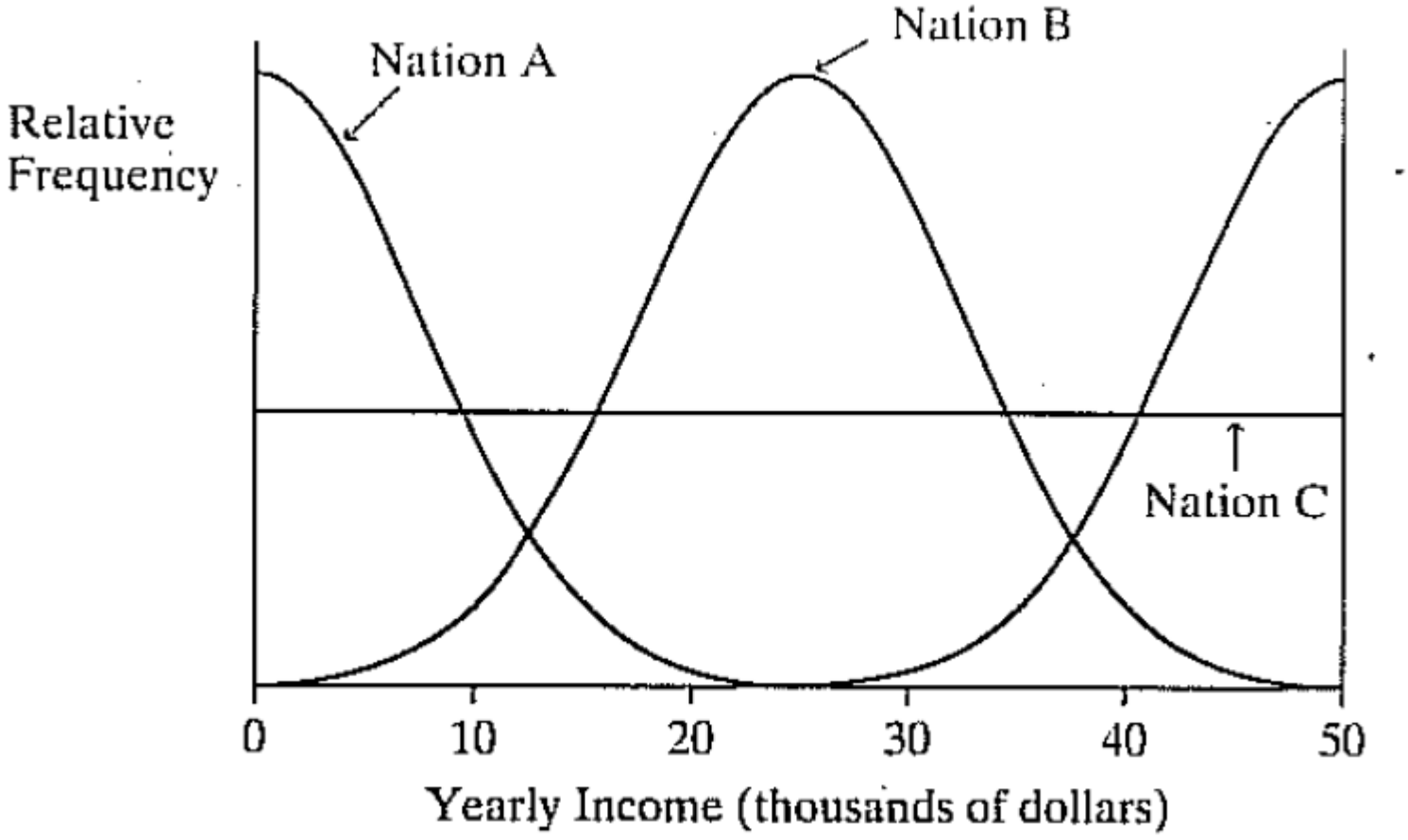
Variability: range

- **Range:** difference between largest and smallest value.
- The simplest measure of variability.
- Does not describe deviations from the mean.

{**1**, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, **39**}

Range = 39 – 1 = 38

Variability



Finlan & Agresti 2009: 47

Variability: deviation

- **Deviation**
  - Difference between value of observation and mean.

$$(x_i - \mu)$$

$$(x_i - \bar{x})$$

{1, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, 39}

(1 - 7), (1 - 7), (2 - 7), … , (39 - 7)

-6, -6, -5, -5, -4, -4, -2, -1, -1, -1, 0, 3, 32

Variability: deviation

- **Deviation**
  - Difference between value of observation and mean.
  - **Positive** deviation: observation value > mean
  - **Negative** deviation: observation value < mean
  - **Zero** deviation: observation value = mean.
  - Since **sum of deviations = 0**, the absolute values or the squares are used in measures that use deviations.

Variability: variance

- Mean is usually not very indicative for data dispersion:

  {4, 4, 6, 6}; mean = 5; s^2 = 1.33

  {0, 0, 10, 10}; mean = 5; s^2 = 33.33

- Therefore we need other measures such as **variance (s^2)**.

Variability: variance

- **Variance**
  - Squared **mean deviation** from mean.

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

population = {1, 3, 6, 10}

¼ * ((1 - 5)^2 + (3 - 5)^2 + (6 - 5)^2 + (10 - 5)^2)

¼ * ((-4)^2 + (-2)^2 + 1^2 + 5^2)

¼ * (16 + 4 + 1 + 25) = ¼ * 46 = **11.5**

Variability: variance

- **Variance**
  - Squared **approximate mean deviation** from mean.

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

sample = {1, 3, 6, 10}

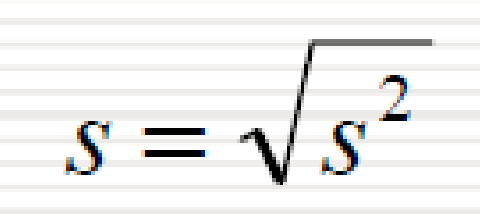
1/3 * ((1 - 5)^2 + (3 - 5)^2 + (6 - 5)^2 + (10 - 5)^2)

1/3 * ((-4)^2 + (-2)^2 + 1^2 + 5^2)

1/3 * (16 + 4 + 1 + 25) = 1/3 * 46 = **15.33**

Variability: standard deviation

- **Standard deviation**
  - Measure of average deviation.
  - Typical distance of observation from the mean.
  - Sensitive to outliers.



sample = {1, 3, 6, 10}

s^2 = 15.33

s = sqrt(15.33) = 3.92

Variability: standard deviation

- **Properties**
  - $s >= 0$
  - $s = 0$ only when all observations have same value.
  - The greater variability about mean, the larger *s*.
  - If data are rescaled, the s is rescaled as well.
  - E.g. if we rescale s of annual income in $ = 34,000 to thousands of $ = 34, the *s* also changes by factor of 100 from 11,800 to 11.8.

Variability: standard deviation

- **Interpretation**
  - Scale dependent.
  - E.g. assume that average amount of points received in this course is 50 points graded on a scale 0 to 60.
  - s = 0 extremely unlikely (no differences in performance).
  - As well as s > 20 (huge differences in performance).

Variability: dimensionless measures

- **Coefficient of variability**
  - Allows comparisons across different distributions (units, means, …).
  - Applicable only to ratio scale.

$$c_v = \frac{s}{\overline{x}}$$

- **Z-score**
  - Standardized measure of variability.
  - Express variation in standard deviations instead of original metric.

$$z_i = \frac{\left(x_i - \overline{x}\right)}{s}$$

Variability: dimensionless measures

- **Coefficient of variability**

$$c_v = \frac{s}{\bar{x}}$$

  - Proportion of std. dev. on the mean value.
  - Allows to compare variability of different data sets.

  - mean = 80, std. dev. = 12, **CV = 12 / 80 = 0.15**
  - mean = 50, std. dev. = 20, **CV = 20 / 50 = 0.40**

Variability: dimensionless measures

- **Z-score**

$$z_i = \frac{(x_i - \bar{x})}{s}$$

  - Shows a distance of an observation in # of standard deviations from the mean.
  - For bell-shaped distributions very unlikely to have values larger than 3 std. deviations from the mean.

  - Data = {1, **3**, 6, 10} ; mean = 5 ; s = 3.92
  - Z-score for 2nd case: (3 - 5) / 3.92 = **- 0.51**
  - 0.51 * 3.92 = 1.99 ; 1.99 + 3 ~ 5
  - Z-scores = (-1.02, -0.51, 0.26, 1.28)

Measures of variability: R commands

range(data)

var(data)

sd(data)

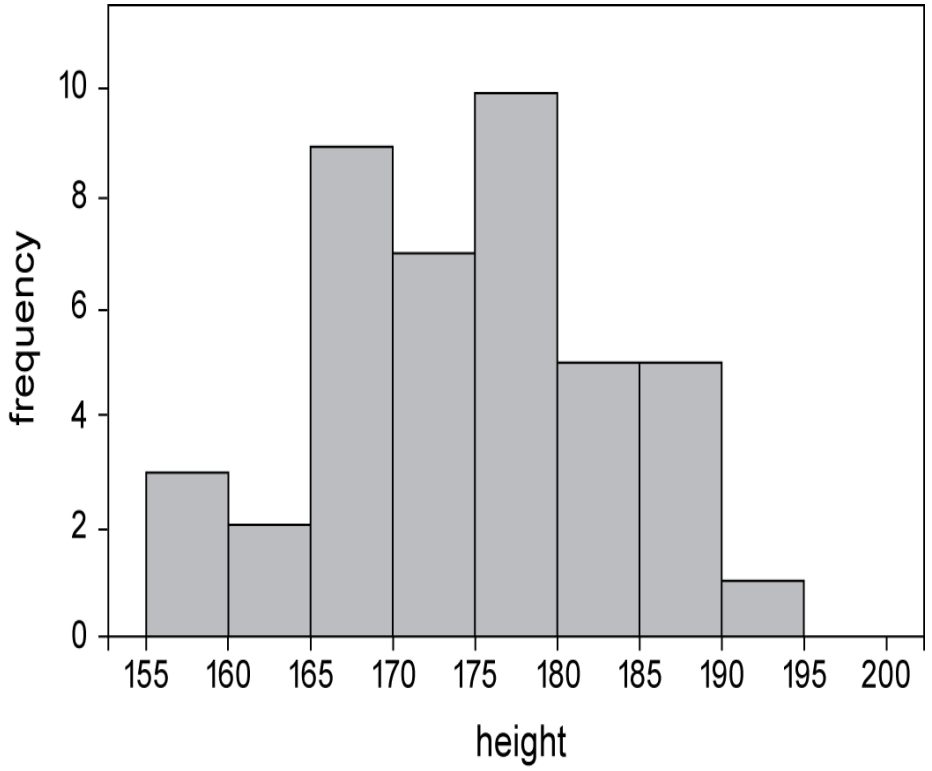
scale(data) = z-scores

sd(data) / mean(data) = coefficient of variability

Frequency distribution

- Frequency distribution: table or visual display of the **frequency** of variable values.

155-160	3
160-165	2
165-170	9
170-175	7
175-180	10
180-185	5
185-190	5
190-195	1
195-200	0



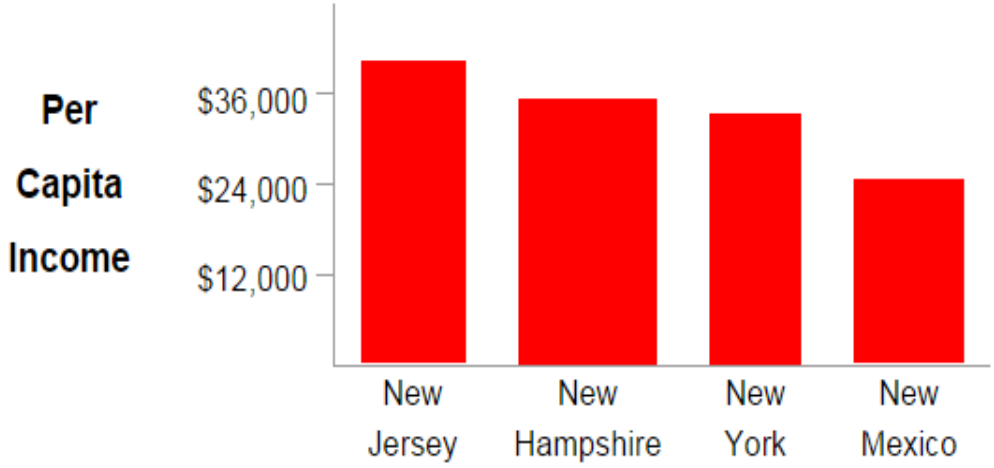
Frequency distribution

- **Absolute frequency:** # of the observations of a category.
- **Relative frequency:** proportion of the observations of a category over total # of observations.
- **Percentage:** proportion multiplied by 100.

155-160	3	0.07	7%
160-165	2	0.05	5%
165-170	9	0.21	21%
170-175	7	0.17	17%
175-180	10	0.24	24%
180-185	5	0.12	12%
185-190	5	0.12	12%
190-195	1	0.02	2%
195-200	0	0	0%

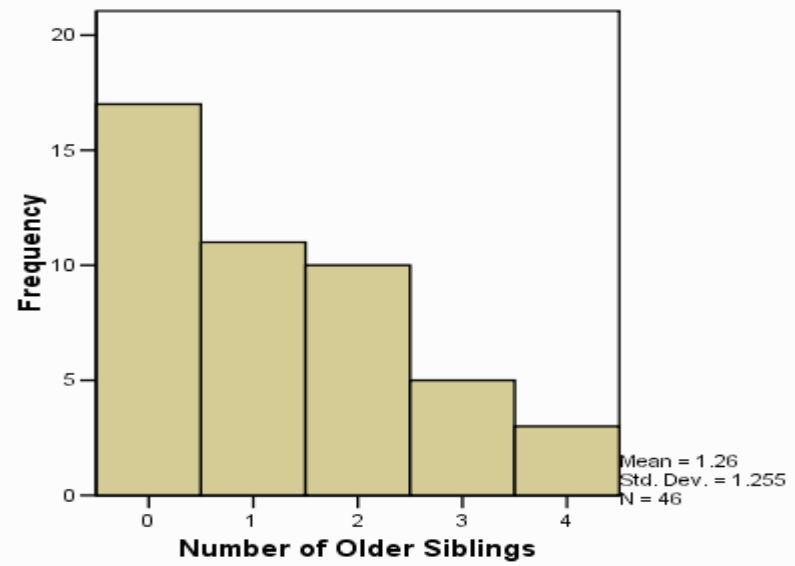
Bar chart

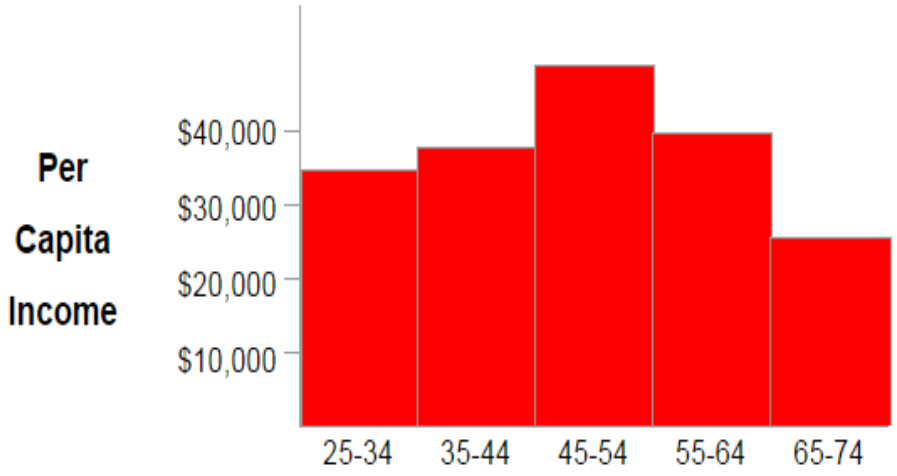
- The columns are positioned over values of **categorical variable** (U.S. states).
- The height of the column indicates the value of the variable (per capita income).



Histogram

- The columns are positioned over a values of **quantitative variable.**
- The column label can be single value or range of values.
- The height of the column indicates the value of the variable.





Boxplot

- Splits data into quartiles (position measure).
- Box: from Q1 to Q3.
- Median (Q2): line within the box.
- Whiskers: indicate the range from:
  - Q1 to smallest non-outlier.
  - Q3 to largest non-outlier.
- Outlier > 1.5 * (Q3 – Q1) from Q1 or Q3
- Outliers are represented separately.