An example: The CPU times given in the table are the amounts of time (in seconds) 25 jobs were in control of a large mainframe computer's central processing unit (CPU).

0.02	0.15	0.19	0.47	0.71
0.75	0.82	0.92	0.96	1.16
1.17	1.23	1.38	1.40	1.59
1.61	1.94	2.01	2.16	2.41
2.59	3.07	3.53	3.76	4.75

Describing Quantitative Data

Graphical Methods

- Stem-and-leaf (small date set, no loss of information)
- Histogram (large data set)
- Time series plot

Numerical Measures

- Central Tendency: mean, median
- Variation: range, variance, standard deviation
- Relative Standing: percentile, z-score

Stem-and-Leaf Display

Steps to construct a stem-and-leaf display:

- Divide each observation into two parts: *stem* and *leaf*.
- List the stems in order in a column
- Place the leaf for each observation in the appropriate stem row. Arrange the leaves in each row in ascending order.

Stem-and-leaf Display for CPU time:

Histogram

Steps to construct histogram

1. Calculate **range** of the data: min(data) - max(data)

2. Divide the range into **classes** (**bins, intervals**) of equal width. For the number of classes, refer to the following rule

Number of observations	Number of classes
Less than 25	5 or 6
25-50	7-14
More than 50	15-20

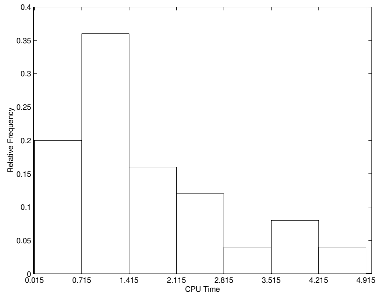
3. For each class, calculate the **class frequency**, which is equal to the number of observations falling in that class.

4. Similarly, **relative class frequency** can be defined as

$$\frac{\text{class frequency}}{\text{total number of observations}}.$$

5. The **histogram** is a bar graph in which the categories are classes and the heights of the bars are determined by class frequency or relative class frequency.

Histogram for CPU data

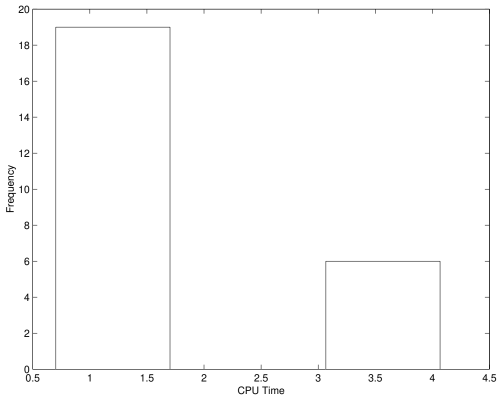


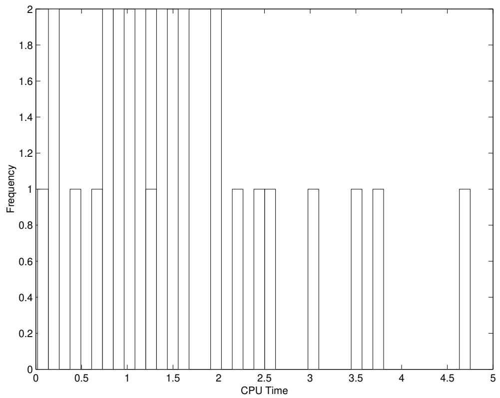
- Because the bars are of *equal* width,

$$\begin{aligned}\text{Area of a bar} &\sim \text{Class frequency}\\ &\sim \text{Relative class frequency.}\end{aligned}$$

- The percentage of number of observations falling within a particular interval is proportional to the area of the bar above the interval.

- If we select one observation from the data, the likelihood (or *probability*) of that observation falling in a particular class (or interval) is proportional to the area of the bar above the interval.





Shapes of Distribution

Single Mode (mound-shaped)

Symmetric

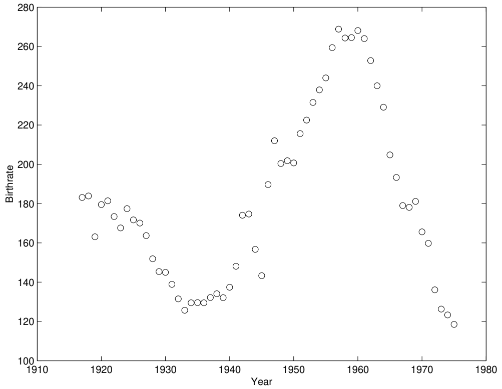
Right skewed (long right tail)

Left skewed (long left tail)

Time Series Plot

Some data sets are **time series**, that is, measurements taken at regular intervals over time. Plots against time can reveal some important features of the data.

Time series plot of number of live births per 10,000 23-year-old women in the United States between 1917 and 1975:



Numerical Descriptive Measures

- Central Tendency
  - mean : average value
  - median : middle value
- Variation
  - range , IQR
  - variance (sample and population)
  - standard deviation
- Relative Standing
  - 100pth percentile : lower (or upper) quartile
  - z-score

Term and Notation

A **statistics** is a numerical descriptive measure computed from sample data.

A **parameter** is a numerical descriptive measure of a population.

Notation:

$$\begin{aligned} \text{Sample :} \quad & y_1, y_2, \ldots, y_n \\ \text{Ordered Sample :} \quad & y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)} \end{aligned}$$

Location

- sample mean $\bar{y} = (\sum_{i=1}^{n} y_i)/n$ (population mean μ)

- sample median

$$m(y) = \begin{cases} y_{((n+1)/2)} & \text{n odd} \\ \frac{y_{(n/2)}+y_{(n/2+1)}}{2} & \text{n even} \end{cases}$$

(population median τ)

- For symmetric distribution, the two are the same; For skewed distribution, the **mean** is located farther toward the long tail than the **median**.

- The **mean** is sensitive to large or small observations, while the **median** is resistant to the influence of extreme observations (i.e. *robust*).

Variation

- range $= y_{(n)} - y_{(1)}$

- sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n}{n-1}.$$

population variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}.$$

- standard deviation $= \sqrt{\text{variance}}$.

$$s^2 = \sum (y_i - \bar{y})^2/(n-1).$$

Why $n-1$? Why not average in the usual way by dividing by n?

- If we *knew* the population mean μ, then we'd define the sample variance as

$$s^2 = (\sum_{i=1}^{n} (y_i - \mu)^2)/n \quad (*).$$

- But we don't know μ, so we estimate it by $\bar{y}$.

- The total deviation $\sum(y_i - \bar{y}))^2$ is smaller than $\sum(y_i - \mu)^2$ (why?), so just plugging in $\bar{y}$ in the formular (*) will result in underestimating of the population variance

- So we divide by $n-1$ instead of n. The reason why $n-1$ is used instead of any other value less than n, follows from the fact that there is a constraint on the n quantities $y_i - \bar{y}$.

- Given any $n-1$ of the $y_i - \bar{y}$, one can solve for the omitted data value ($n-1$ degree of freedom).

Effect of Linear Transformation

- Multiplying each observation by a number b multiplies measures of center (mean and median) by b, and measures of variance (range and standard deviation) by $|b|$.

- Adding the same number a to each observation adds a to location measure, but does not change measures of variation.

Relative Standing

Percentile

- The $100p$th percentile: a value so that $100p\%$ of data are smaller than it.

- the lower quartile Q_L, the mid-quartile and the upper quartile Q_U.

  $$\text{Interquartile Range IQR} = Q_U - Q_L$$

- Finding percentiles with small data set

  Calculate $l = p(n+1)/100$ and round to the nearest integer (if l falls halfway between two integers, round up). Then $y_{(l)}$ is the pth percentile.

- Note: *By definition, percentiles are not unique and the outputs from different software may not be the same*

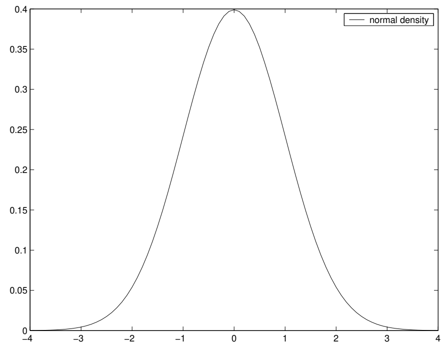
Standardizing Observations

The **z-score** for a value y of a data set is the distance that y lies above or below the mean, measured in units of the standard deviation:

$$\text{Sample } z-score: \quad z = (y - \bar{y})/s$$
$$\text{Population } z-score: \quad z = (y - \mu)/\sigma$$

Let $z_i =$ z-score of y, then $z_1, \ldots, z_n$ have mean 0, variance 1.

Normal Distribution



The Empirical Rule (68-95-99.7 Rule)

For mound-shaped distribution

(1) Approximately 68% of the data will lie within 1 standard deviation of their mean

(2) Approximately 95% of the data will lie within 2 standard deviation of their mean

(3) almost all the data (99.7%) will lie within 3 standard deviation of their mean

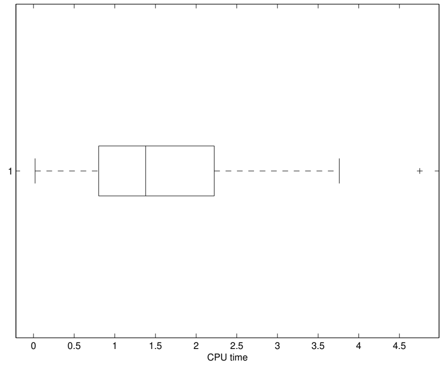
Detecting Outliers

An observation y that is unusually large or small relative to the other values in a data set is called an **outlier**.

Outlier $\neq$ Bad observation

Rule of Thumb for detecting outliers

- z-score greater than 3
- Boxplot: 1.5 IQL away from the median



Boxplot

Five-number summary: min, Q_L, median, Q_U and max

Boxplot is a graph of five-number summary

(1) A box spans the quartile

(2) A line in the box marks the median

(3) Suspected outliers (more than 1.5IQR above Q_L or below Q_U) are plotted individually

(4) Lines extend from the box to the smallest and largest observations that are not suspected outliers

A fun example of using boxplot to compare distributions : The data set is from

McClave, J. T. and Dietrich II, F. H. (1991). *Statistics*. Dellen Publishing , San Francisco.

The data is collected to answer the question: *"Do blonds and brunettes have a different pain threshold?"*.

hair color	pain threshold score				
light blonde	62	60	71	55	48
dark blonde	63	57	52	41	43
light brunette	42	50	41	37	
dark brunette	32	39	51	30	35

Blondes are tougher!