

# Describing univariate distributions

Paul M. Magwene

Department of Biology

# Overview

- Terminology for describing univariate distributions
- Measures of location (centrality)
- Measures of dispersion (spread)

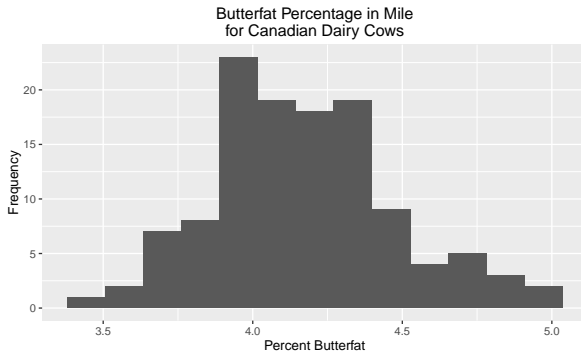
## Example data set: butterfat data

We'll use a data set that records the butter fat percentage in milk from 120 Canadian dairy cows (Sokal and Rohlf, Biometry, 4th ed)

- See the link on the course wiki for `butterfat.csv`
- Load `butterfat.csv` using the `read.csv` function

# Generate a histogram

Using the `ggplot2` library, generate a histogram for the butterfat data set.



**Figure:** Histogram of butter fat percentage from 120 Canadian cows.

# Mean

- Most common measure of location
- Measure of location that minimizes the sum of the squared deviations around it
- Statistical measure of location that has the smallest standard error (to be defined later)
- Physical analogy: If we think of observations as points of mass on a line, the mean is the center of mass (balance point)

Let  $X = \{x_1, x_2, \dots, x_n\}$ . The mean of  $x$  is:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

# Median

- The middle point of a frequency distribution
- The value of the variable that has an equal number of items on either side of items

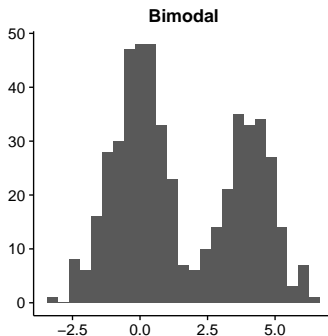
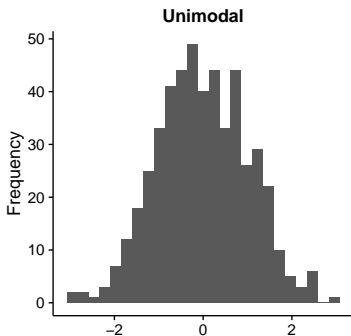
The median is a robust estimator of location. Robust statistics are those that are not strongly affected by outliers or violations of model assumptions.

## Robustness of median: Example

Changes in estimates of location when three outlier values (8, 10, 15) are added to butterfat data.

# Mode

- The most common value (or interval) in a distribution
- Unimodal, bimodal, multi-modal





## Some other “means”

**Weighted mean** – useful when there is some a priori notion of weight or importance for different observations

$$\bar{X}_w = \frac{1}{(\sum^n w_i)} \sum^n w_i x_i$$

where the  $w_i$  represent the weights attached to each observation.

**Geometric mean** – most often used to study proportional growth (populations, tissues, organs, etc)

$$GM_X = \sqrt[n]{\prod x_i}$$

**Harmonic mean** – rarely used in biology.

$$HM_X = \frac{1}{n} \sum^n \frac{1}{x_i}$$

# Range

- The difference between the largest and smallest items in a sample

$$\max(x) - \min(x)$$

# Deviates

**Deviate** – the difference between an observation and the mean; can be negative or positive. Units same as the  $x_i$ .

$$x_i - \bar{X}$$

**Squared deviate** – the square of a deviate; always  $\geq 0$  (units<sup>2</sup>).

$$(x_i - \bar{X})^2$$

**Sum of squared deviations** – the sum of all the squared deviations in a sample (units<sup>2</sup>).

$$\sum_{i=1}^n (x_i - \bar{X})^2$$

# Variance and standard deviation

**Variance** – the mean squared deviation (units<sup>2</sup>).

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

**Standard deviation** – the square root of the variance (units same as the  $x_i$ ).

$$\sigma_X = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2}$$

The above are the population variance and standard deviation.

# Sample estimators of variance and standard deviation

The *unbiased* sample estimators of the variance and standard deviation are given by:

$$\text{Variance: } s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$$

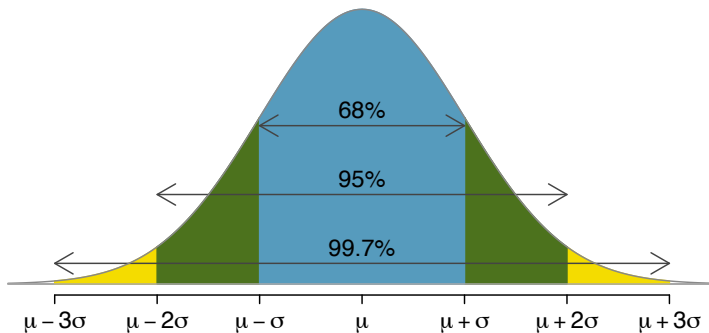
$$\text{Standard deviation: } s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2}$$

You almost always want to use the sample estimators of variance and standard deviation.

# Standard deviation rules of thumb

If data are normally distributed:

- Approximately 68% of observations fall within 1 standard deviation about the mean
- Approximately 95% of observations fall within 2 standard deviations about the mean
- Approximately 99.7% of observations fall within 3 standard deviations about the mean



# Coefficient of variation

- Standard deviation expressed as percentage of mean
- Unitless measure

$$V = \frac{s_X \times 100}{\bar{X}}$$

# Quantiles, quartiles, interquartile range

- **Quantiles** – points that will divide a frequency distribution into equal sized groups
  - quartiles – points dividing a distribution into 4 equal groups
  - deciles – points dividing a distribution into 10 equal groups
  - percentiles – points dividing a distribution into 100 equal groups
- **Interquartile range (IQR)**– range of values that captures the central 50% of the distribution
  - Q1 = lower quartile, Q3 = upper quartile



## Boxplots typically depict information about quartiles

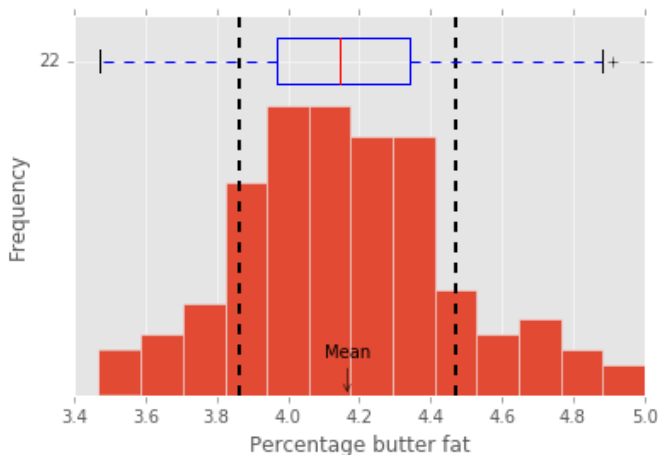


Figure: Histogram of butterfat data set, with superimposed boxplot.

# Median absolute deviation (MAD)

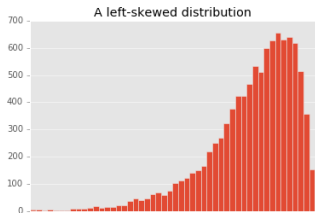
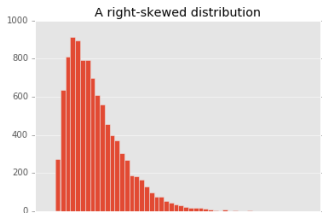
- A robust estimator of dispersion

$$\text{MAD}(X) = \text{median}(|x_i - \text{median}(X)|)$$

For normal distribution,  $\sigma_X \approx 1.486 \times \text{MAD}(X)$ .

# Skewness

- Skewness describes asymmetry of distributions



Common measure of skewness:

$$\text{skewness} = E \left[ \left( \frac{x - \mu}{\sigma} \right)^3 \right]$$