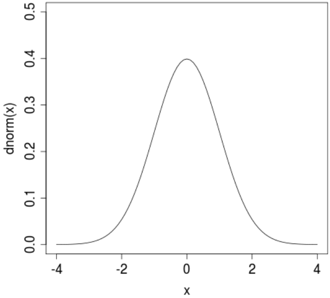
Math 207: Statistics

The Average and the Standard Deviation



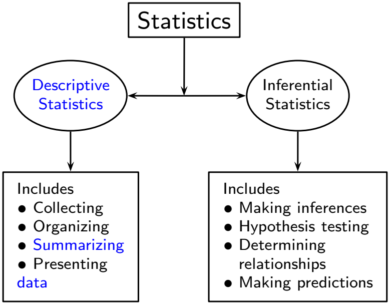
Dr. Ralph Wojtowicz

Mathematics Department

SHENANDOAH UNIVERSITY

Introduction

- We have studied tools such as histograms for summarizing and gaining insights into data.
- The **center** (average or mean and median) and **spread** (standard deviation or sd) are numerical tools of descriptive statistics.
- Range (maximum value – minimum value), interquartile range, and quantiles are other descriptive statistics we will meet.



The Average (Mean)

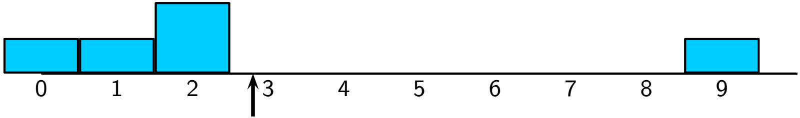
- The **average** (**mean**) of a list of numbers equals their sum, divided by how many there are.

$$\text{average (mean) of a list of numbers} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Example: The list 9, 1, 2, 2, 0 has $n = 5$ entries. Its average is

$$\frac{9+1+2+2+0}{5} = \frac{14}{5} = 2.8$$

- The histogram *balances* when supported at the average.

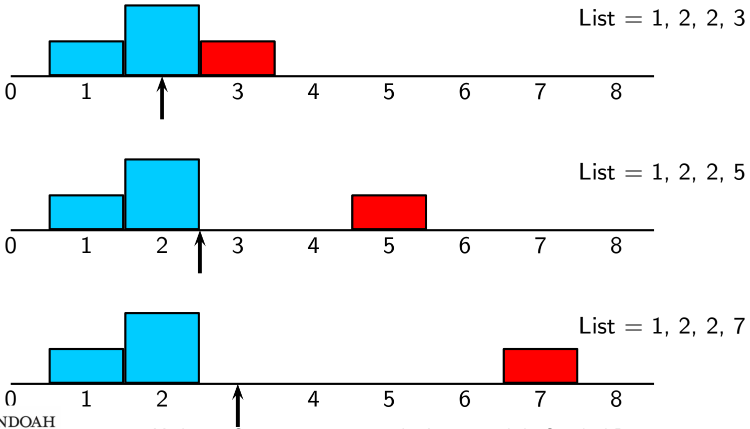


- In R we can compute the mean of a list of numbers as follows:

```
> x <- c(9, 1, 2, 2, 0)
> mean(x)
[1] 2.8
```

The Average and the Histogram

- The histogram balances when supported at the mean.
- The first histogram below is **symmetric** about its mean. Half the data is to the left of the mean and half is to the right.



The Median

- The **median** of a list of numbers is the value with half the area to the left and half to the right.
- Examples:
  - For the list 1, 2, 2, 3, the median is 2 (as is the mean).
  - For the list 1, 2, 2, 5, the median is 2 (but the mean is 2.5).
  - For the list 1, 2, 2, 1000, the median is 2 (but the mean is 251.25).
  - For the list 1, 2, 3, 8, the median is any number (such as 2.5) that is greater than 2 but less than 3 (but the mean is 3.5).

```
> x <- c(1,2,3,8)
> median(x)
[1] 2.5
```

- We typically use the median rather than the mean to describe the center of a histogram with a long tail (e.g., incomes or home prices).
- The **mode** of a list of numbers is the most frequent value. We will not use the mode.

Quantiles and Percentiles

- The **median** of the data set is also called the 50th **percentile** because 50% of the data is less than or equal to it.
- The pth percentile is a number that is greater than or equal to p% of the data.
- For example, the 25th percentile is the median of the first half of the data. The 75th percentile is the median of the second half of the data.
- To compute the median or other percentiles, you first have to put the data in order.
- Here is how to compute these in `R`:

```
> x = c(2, 3, 5, 5, 6, 10, 12, 17, 19, 19, 20)
> summary(x)
  Min.  1st Qu.  Median    Mean   3rd Qu.    Max.
  2.00     5.00   10.00   10.73     18.00   20.00
> quantile(x, probs=c(0.0, 0.25, 0.35, 0.50, 0.75, 1.00))
  0%   25%   35%   50%   75%   100%
 2.0   5.0   5.5  10.0  18.0   20.0
```

The Root-Mean-Square

- The **root-mean-square** (or RMS) of a list of numbers measures the average magnitude (ignoring signs) of the numbers in the list.

$$\text{RMS size of a list of numbers} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

- Example: The list 0, 5, -8, 7, -3 has $n = 5$ entries. Its RMS size is

$$\sqrt{\frac{0^2 + 5^2 + (-8)^2 + 7^2 + (-3)^2}{5}} = \sqrt{\frac{0 + 25 + 64 + 49 + 9}{5}} = \sqrt{\frac{147}{5}} = \sqrt{29.4} \approx 5.4$$

- The calculation steps are: (1) **square** the entries of the list, (2) take the **mean** of this new list, and (3) take the square **root** of this mean.
- RMS is used to compute the sd (or spread) of a list of numbers.

The Standard Deviation

- **Standard deviation (SD)** measures the **spread** of the data.
  - Roughly 68% of the data falls within one SD of the average.
  - Roughly 95% of the data falls within two SDs of the average.
- Average (mean) and median are measures of the center of the data.
- Units of SD and average are the same as those of the data.
- **Variance** is SD^2.

Computing the Population Standard Deviation (SD)

- The **standard deviation** (**SD**) of a list of numbers equals the RMS deviation from average. SD is **population standard deviation**.

$$\text{SD of a list of numbers} = \sqrt{\frac{(x_1 - \text{mean})^2 + (x_2 - \text{mean})^2 + \cdots + (x_n - \text{mean})^2}{n}}$$
$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \text{mean})^2}$$

- Example: The list 20, 10, 15, 15 has $n = 4$ entries. To compute the SD, we first need the mean of the list.

$$\text{mean} = \frac{1}{4}(20 + 10 + 15 + 15) = 60/4 = 15$$

We then calculate the mean of the square deviations from this average.

$$\frac{(20-15)^2 + (10-15)^2 + (15-15)^2 + (15-15)^2}{4} = \frac{5^2 + 5^2}{4} = \frac{50}{4} = 12.5$$

We then take the square root: $\text{SD} = \sqrt{12.5} \approx 3.5$.

Computing Sample Standard Deviation (SD+)

- Most calculators (and statistical software such as R) compute the **sample standard deviation**:

$$\text{SD+} = \sqrt{\frac{(x_1 - \text{mean})^2 + (x_2 - \text{mean})^2 + \cdots + (x_n - \text{mean})^2}{n-1}}$$
$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \text{mean})^2}$$

- To compute the population SD using R, we must correct the denominator:

```
> x <- c(20,10,15,15)
> sd(x)                          # This is SD+
[1] 4.082483
> n = length(x)
> sd(x) * sqrt((n - 1)/n)        # This is SD
[1] 3.535534
```