3: Summary Statistics

Notation

Consider these 10 ages (in years):

```
21    42    5    11    30    50    28    27    24    52
```

- The symbol ***n*** represents the sample size ($n = 10$).
- The **capital letter** X denotes the **variable**.
- x_i represents the i^{th} value of variable X. For the data, $x_1 = 21$, $x_2 = 42$, and so on.
- The symbol Σ ("capital sigma") denotes the summation function. For the data, $\Sigma x_i = 21 + 42 + \ldots + 52 = 290$.

Measures of Central Location

Mean (arithmetic average)

The three main measures that summarize the center of a distribution are the mean, median, and mode. While there are several different types of **mean**, we will focus on the **arithmetic average**.

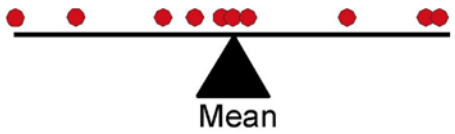
To calculate the arithmetic mean, sum all the values and divide by n (equivalently, multiple $1/n$):

$$\bar{x} = \frac{1}{n}\sum x_i$$

For the dataset, $\bar{x} = \frac{1}{10}(290) = 29$ years.

Interpretation of the mean. The mean tells you:

- The expected value of an individual drawn at random from the sample.
- The expected value of an individual drawn at random from the population.
- The expected value of the population mean.
- The **gravitational center** of a distribution. This is where the distribution would balance:



The mean is like a seesaw. A small child can sit farther from the center of a seesaw in order to balance an adult sitting closer to the center. Similarly, a single outlier sitting far off-center can have a profound effect on the mean. The mean is sensitive to the effects of outliers.

The distinction between the **sample mean (denoted $\bar{x}$)** and **population mean (denoted μ)** is critical to your future understanding of statistical inferences.

Reporting the mean: The mean should be **rounded** to three (or at most four) significant digits before **reporting** in order to avoid an appearance of pseudo-precision. In addition, units of measure should be included when reporting the mean. For example, the mean of our data is 29.0 *years*.

Median

The median was covered in the previous chapter. Briefly, this is the value with a *depth* of $\frac{n+1}{2}$ in the **ordered array**. When *n* is odds, this will fall on a value in the dataset. When *n* is even, this will fall between two values. Here are the sizes of 5 benign tumors (cubic centimeters):

```
4     7     8     11     12
            ↑
          median
```

The median has a depth of (5 + 1) / 2 = 3 and a value of 8.

The median is slightly more *resistant* to outliers. This means it resists the pull of outliers and is more apt to stay put. For instance, if the biggest value in the above set had been mistakenly entered 120 instead of 12, the mean would shift from 8.4 to 30 but the median would stay put at 8.

Mode

The mode is the most frequently occurring value in a data set.

```
AGE   |  Freq     %
------+---------------
 3    |     2    0.3%
 4    |     9    1.4%
 5    |    28    4.3%
 6    |    37    5.7%
 7    |    54    8.3%
 8    |    85   13.0%
 9    |    94   14.4%   ← Mode (most frequent value)
10    |    81   12.4%
11    |    90   13.8%
12    |    57    8.7%
13    |    43    6.6%
14    |    25    3.8%
15    |    19    2.9%
16    |    13    2.0%
17    |     8    1.2%
18    |     6    0.9%
19    |     3    0.5%
------+---------------
Total |   654  100.0%
```

The mode is unreliable in all but very large data sets.

Comparison of the Mean, Median, and Mode

The mean, median, and mode are equivalent when the distribution is unimodal and symmetrical. However, with asymmetry, the median is approximately one-third the distance between the mean and mode:

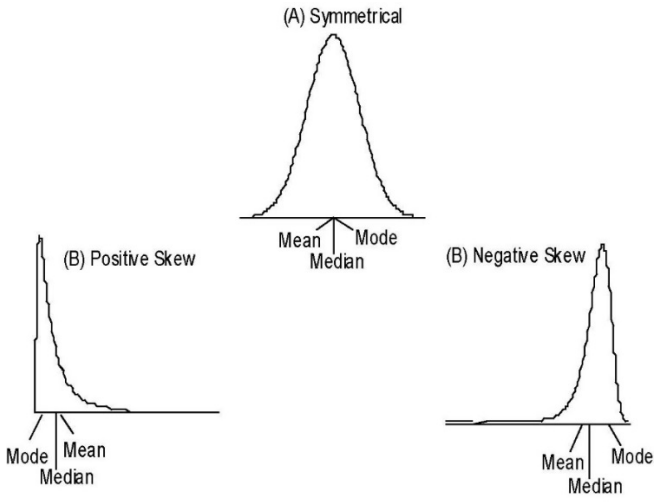


Fig: sumstats2.ai

The mean, median, and mode offer different advantages and disadvantages. The mean offers the advantages of familiarity and efficiency. It also has advantages when making inferences about a population mean. On the downside, the mean is markedly influenced by extreme skewness and outliers. Under circumstances of extreme skewness, the median is a more "stable." An often cited example of this advantage come when considering the salary of employees, where the salary of highly paid executives skews the average income toward a misleadingly high value. Another example is the average price of homes, in which case high priced homes skew the data in a positive direction. In such circumstances, the median is less likely to be misinterpreted, and is therefore the preferred measure of central location.

You can judge the asymmetry of a distribution by comparing its mean and median. When the mean is greater than the median, the distribution has a positive skew. When the mean is about equal to the median, the distribution is symmetrical. When the mean is less than the median, the distribution has a negative skew:

$$\text{mean} > \text{median} : \text{positive skew}$$
$$\text{mean} \cong \text{median} : \text{symmetry}$$
$$\text{mean} < \text{median} : \text{negative skew}$$

In general, the mean is preferred when data are symmetrical and do not have outliers. In other instances, the median may be the preferred measure of central location.

Measures of Spread

Range

Simply noting the minimum and maximum values can be useful when describing the spread of a distribution. However, calculating the sample range (maximum – minimum) is not an acceptable measure of spread because it will consistently underestimate the population range.

5-point summaries and quartiles

The **5-point summary** consists of:

Q0 ≡ Quartile 0 (the *minimum*)
Q1 ≡ Quartile 1 (bigger than 25% of the data points)
Q2 ≡ Quartile 2 (the median)
Q3 ≡ Quartile 3 (bigger than 75% of the data points)
Q4 ≡ Quartile 4 (the *maximum*)

The median is called Q2 because it is equal to or greater than 2-quarters of the data points. The minimum is Q0 because it is greater than or equal to zero-quarters of the data points. You get it.

When the data set is large ($n >= 100$), it is easy to find Q1 and Q3. With small data sets, the exact location of quartiles must be **interpolated**. The two most common methods of interpolation for this purpose are weighted averages and Tukey's hinges. To find **Tukey's hinges:**

- Put the data in rank order
- Locate the median of the data set.
- Divide the data set into two groups: a low group and a high group. When *n* is odd, the median should be placed in both groups.
- Find the 'median' of the low group. This is Q1.
- Find the 'median' of the high group. This is Q3.

Example #1. Consider this ordered array ($n = 10$)

```
5     11     21     24     27     28     30     42     50     52
            low group          M            high group
```

The low group has a 'median' of 21. This is Q1. The high group has a 'median' of 42. This is Q3. The five-point summary (5, 21, 27.5, 42, 52)

Example #2. Consider this new ordered array ($n = 7$) .

```
1.47    2.06    2.36    3.43    3.74    3.78   3.94
      low group          M        high group
```

Recall that you must include the median in both the low group and high group for Tukey's hinges. Therefore the low group consists of {1.47, 2.06, 2.36, 3.43}. The 'median' of this low group (Q1) is the average of 2.06 and 2.36, or 2.21. The 'median' of the high group (Q3) is 3.76. The five point summary is (1.47, 2.21, 3.43, 3.76, 3.95).

Inter-quartile range (IQR)

An good measure of spread (esp. for asymmetrical data) is the **inter-quartile range (IQR):**

$$IQR = Q3 - Q1$$

The IQR for Example #1 (prior page) = 42 – 21 = 21. For Example #2 (prior page), IQR = 3.76 – 2.21 = 1.55. Our IQRs are also called **hinge-spreads** because they quantify the spread from the lower hinge to the upper hinge.

Box-and-whiskers plots (boxplots)

The Tukey boxplot consists of a box showing Q1, Q2, and Q3, whiskers and, occasionally *outside values*.

After determining the 5 point summary and IQR for a dataset, then calculate (but do not draw) **fences** as follows:

- $Fence_{Lower} = Q1 - (1.5)(IQR)$
- $Fence_{Upper} = Q3 + (1.5)(IQR)$

Note that the fences are 1.5 hinge-spreads above below the hinges. **Do *not* plot these fences.** Any value that is above the upper fence is an **upper outside value**. Any values below the lower fence is a **lower outside value**. Plot these points, if any, on the graph.

The largest value still inside the upper fence is the **upper inside value**. The smallest value still inside the lower fence is the **lower inside value**. Drawn **whiskers** from the upper hinge to the upper inside value and from the bottom hinge to the lower inside value.

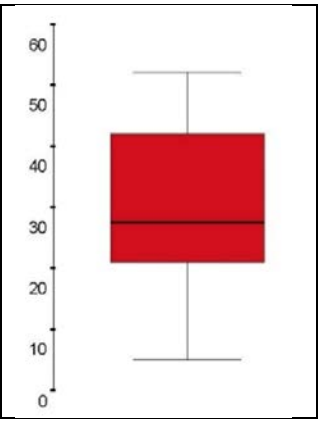
Boxplot example #1.

5 11 21 24 27 28 30 42 50 52

The 5-point summary (determine on prior page) is (5, 21, 27.5, 42, 52).

The box extends from 21 to 42 and has a line in its midst to identify the median at 27.5.

The $F_L = 21 - (1.5)(21) = -10.5$ and $F_U = 42 + (1.5)(21) = 73.5$. No values in the data set are above 73.5 or below −10.5.Therefore, there are no outside values. The upper inside value is 52 and the lower inside is the 5. Whiskers are drawn from the hinges to the inside values. The boxplot is shown on the next page.



Interpretation of boxplots. Think shape, location and spread.

- Shape is easiest to consider in terms of symmetry and potential outliers. Is the median about halfway between the hinges? Is the box about half-way between the whiskers? Outside values are potential outliers.
- The central location is be summarized by the median and box, which is the middle 50% of values.
- Spread is quantified in terms of the hinge-spread and whisker-spread.

The boxplot for example #1 is fairly symmetrical and has no outside values. The a median that is a little less than 30 and a hinge-spread is from 21 to 42.

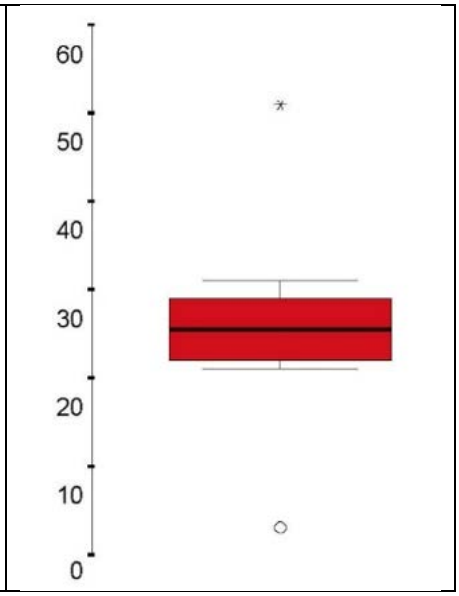
New example: Here's a new set of values:

3 21 22 24 25 26 28 29 31 51

The five-point summary is (3, 22, 25.5, 29, 51). IQR = 29 – 22 = 7.

$F_U = 29 + (1.5)(7) = 39.5$. $F_L = 22 - (1.5)(7) = 11.5$. There is one value above the upper fence (51). There is one value below the lower fence (3). The largest value still inside the upper fences (upper inside value) is 31.

The smallest value still inside the lower fence (lower inside value) is 21.



Standard Deviation and Variance

Both the standard deviation and variance are based on **deviations (d_i),** defined as

$$d_i = x_i - \bar{x}$$

Square each deviation in the dataset and sum them to derive the **sum of squares (*SS*)**:

$$SS = \sum (x_i - \bar{x})^2$$

The **sample variance** is the "mean" sum of squares: $s^2 = \frac{1}{n-1} \cdot SS$

Note: The denominator in this formula is by $n - 1$, not n. This is necessary to derive an unbiased estimate of the population variance. The number $n - 1$ is called the **degree of freedom**.

The **sample standard deviation** is simply the square root of the variance: $s = \sqrt{s^2}$

Illustrative example. Recall these 10 ages: `21, 42, 5, 11, 30, 50, 28, 27, 24, 52.` The sample mean $\bar{x}$ is 29.0 (page 1). The variance is calculated:

i	Values (x_i)	Deviations (d_i)	*Sqyared* d^2
1	21	21 – 29 = –8	$-8^2 = 64$
2	42	42 – 29 = 13	$13^2 = 169$
3	5	5 – 29 = –24	$-24^2 = 576$
4	11	11 – 29 = –18	$-18^2 = 324$
5	30	30 – 29 = 1	$1^2 = 1$
6	50	50 – 29 = 21	$21^2 = 441$
7	28	28 – 29 = –1	$-1^2 = 1$
8	27	27 – 29 = –2	$-2^2 = 4$
9	24	24 – 29 = –5	$-5^2 = 25$
10	52	52 – 29 = 23	$23^2 = 529$
Sums→	290	0	Sum of Squares → 2134

The sample variance is $s^2 = \frac{SS}{n-1} = \frac{2134}{10-1} = 237.1111$ years^2. [Note the squared units.]

The standard deviation $s = \sqrt{s^2} = \sqrt{237.1111} = 15.398 =$ 15.4 years.

Report the standard deviation in conjunction with the mean, round accordingly, and include units of measure. "The mean age of the participants was 29.0 years with a standard deviation of 15.4 years."

Calculating a few variance and standard deviations by hand is instructive. However, most of the time we calculate the standard deviation with a computer or calculator. Do not need to calculate variances and standard deviations by hand unless the instructions specifically request a step-by-step calculation.