Basic Statistical Descriptions of Data


STATISTICS
DATA MINING
ARTIFICIAL INTELLIGENCE
MACHINE LEARNING

Basic Statistical Descriptions of Data

The entire subject of statistics is based around the idea that we have this big set of data, and we want to analyse that set in terms of the relationships between the individual points in that data set.

- ***Motivation***
  - To better understand the data: central tendency, variation and spread
- ***Data dispersion characteristics***
  - median, max, min, quantiles, outliers, variance, etc.
- ***Numerical dimensions*** correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- ***Dispersion analysis on computed measures***
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

Mean (algebraic measure) (sample vs. population):

- The most common and effective numeric measure of the "center" of a set of data is the ***(arithmetic) mean***. Let $x_1, x_2, \ldots\ldots, x_N$ be a set of N values or ***observations***, such as for some numeric attribute X, like salary.

The **mean** of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Note: *N* is population size.

- Sometimes, each value x_i in a set may be associated with a **weight** w_i for $i = 1, \ldots, N$. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

This is called the ***weighted arithmetic mean*** *or the* ***weighted average.***

Measuring the Central Tendency

- A **trimmed mean** (sometimes called a ***truncated mean***) is similar to a mean, but it trims any outliers. Outliers can affect the mean (especially if there are just one or two very large values), so a trimmed mean can often be a better fit for data sets with erratic high or low values or for extremely skewed distributions. Even a small number of extreme values can corrupt the mean.
- For example, the mean salary at a company may be substantially pushed up by that of a few highly paid managers. Similarly, the mean score of a class in an exam could be pulled down quite a bit by a few very low scores.
- Which is the mean obtained after chopping off values at the high and low extremes.
- Example: Find the trimmed 20% mean for the following test scores: 60, 81, 83, 91, 99.
  - Step 1: Trim the top and bottom 20% from the data. That leaves us with the middle three values: 60, 81, 83, 91, 99.
  - Step 2: Find the mean with the remaining values. The mean is (81 + 83 + 91) / 3) = 85.

Measuring the Central Tendency

- **Median**:
  - The median of a set of data is the middlemost number in the set. The median is also the number that is halfway into the set.
  - To find the median, the data should first be arranged in order from least to greatest.
  - Middle value if odd number of values, or average of the middle two values otherwise

  - **What will be the median estimated by interpolation (for *grouped data*)?**

Measuring the Central Tendency

- **Mode**
  - The mode for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes.
  - It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.
  - Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal,** and **trimodal**.
  - In general, a data set with two or more modes is **multimodal**.
  - At the other extreme, if each data value occurs only once, then there is no mode.
  - Example:
    - 53, 55, 56, 56, 58, 58, 59, 59, 60, 61, 61, **62, 62, 62**, 64, 65, 65, 67, 68, 68, 70
    - 62 appears three times, more often than the other values, so **Mode = 62**

Mean, Median and Mode from Grouped Frequencies

The Race.....This starts with some raw data (not a grouped frequency yet) ...

Alex timed 21 people in the sprint race, to the nearest second:

59, 65, 61, 62, 53, 55, 60, 70, 64, 56, 58, 58, 62, 62, 68, 65, 56, 59, 68, 61, 67

To find the Mean Alex adds up all the numbers, then divides by how many numbers:

Mean = (59+65+61+62+53+55+60+70+64+56+58+58+62+62+68+65+56+59+68+61+67)/21

= 61.38095...

Mean, Median and Mode from Grouped Frequencies

- To find the **Median** Alex places the numbers in value order and finds the middle number.
- In this case the median is the 11th number:

53, 55, 56, 56, 58, 58, 59, 59, 60, 61, **61**, 62, 62, 62, 64, 65, 65, 67, 68, 68, 70

- **Median** = 61

- To find the **Mode**, or modal value, Alex places the numbers in value order then counts how many of each number. The Mode is the number which appears most often (there can be more than one mode):

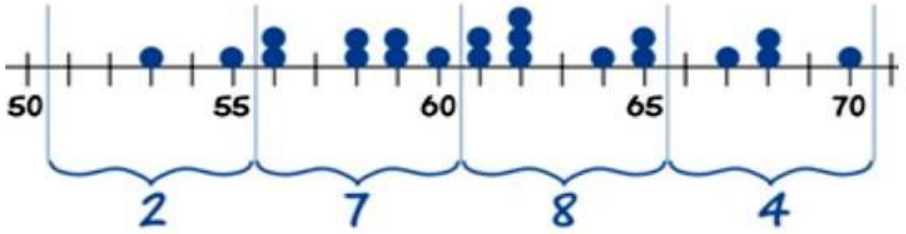
53, 55, 56, 56, 58, 58, 59, 59, 60, 61, 61, **62, 62, 62**, 64, 65, 65, 67, 68, 68, 70

- 62 appears three times, more often than the other values, so **Mode** = 62

Mean, Median and Mode from Grouped Frequencies

- Grouped Frequency Table
- Alex then makes a Grouped Frequency Table:
- *So 2 runners took between 51 and 55 seconds, 7 took between 56 and 60 seconds, etc*

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4



Oh No!

Suddenly all the original data gets lost (naughty pup!)

Only the Grouped Frequency Table survived ...

... can we help Alex calculate the Mean, Median and Mode from just that table?

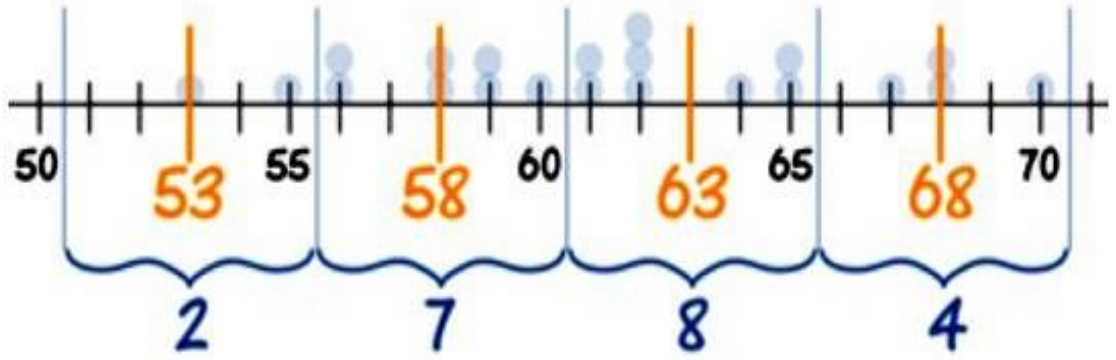
The answer is ... no we can't. Not accurately anyway. But, we can make **estimates**.

Estimating the Mean from Grouped Data

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

The groups (51-55, 56-60, etc), also called **class intervals**, are of **width** 5

The **midpoints** are in the middle of each class: 53, 58, 63 and 68



We can estimate the Mean by using the ***midpoints.***

Let's now make the table using midpoints:

Midpoint	Frequency
53	2
58	7
63	8
68	4

Estimating the Mean from Grouped Data

- Our thinking is: "2 people took 53 sec, 7 people took 58 sec, 8 people took 63 sec and 4 took 68 sec". In other words we imagine the data looks like this:

53, 53, 58, 58, 58, 58, 58, 58, 58, 63, 63, 63, 63, 63, 63, 63, 63, 68, 68, 68, 68

- Then we add them all up and divide by 21. The quick way to do it is to multiply each midpoint by each frequency:

Midpoint	Frequency
53	2
58	7
63	8
68	4

And then our estimate of the mean time to complete the race is:

$$\text{Estimated Mean} = \frac{1288}{21} = 61.333\ldots$$

Midpoint x	Frequency f	Midpoint × Frequency fx
53	2	106
58	7	406
63	8	504
68	4	272
Totals:	**21**	**1288**

Estimating the Median from Grouped Data

- Let's look at our data again:
- The median is the middle value, which in our case is the 11th one, which is in the 61 - 65 group:
- We can say "the **median group** is 61 - 65"
- But if we want an estimated **Median value** we need to look more closely at the 61 - 65 group.

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

We call it "61 - 65", but it really includes values from 60.5 up to (but not including) 65.5.

Why? Well, the values are in whole seconds, so a real time of 60.5 is measured as 61. Likewise 65.4 is measured as 65.

Estimating the Median from Grouped Data

- At 60.5 we already have 9 runners, and by the next boundary at 65.5 we have 17 runners.
- By drawing a straight line in between we can pick out where the median frequency of n/2 runners is:

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

And this handy formula does the calculation:

$$\textbf{Estimated Median} = L + \frac{(n/2) - B}{G} \times w$$

where:

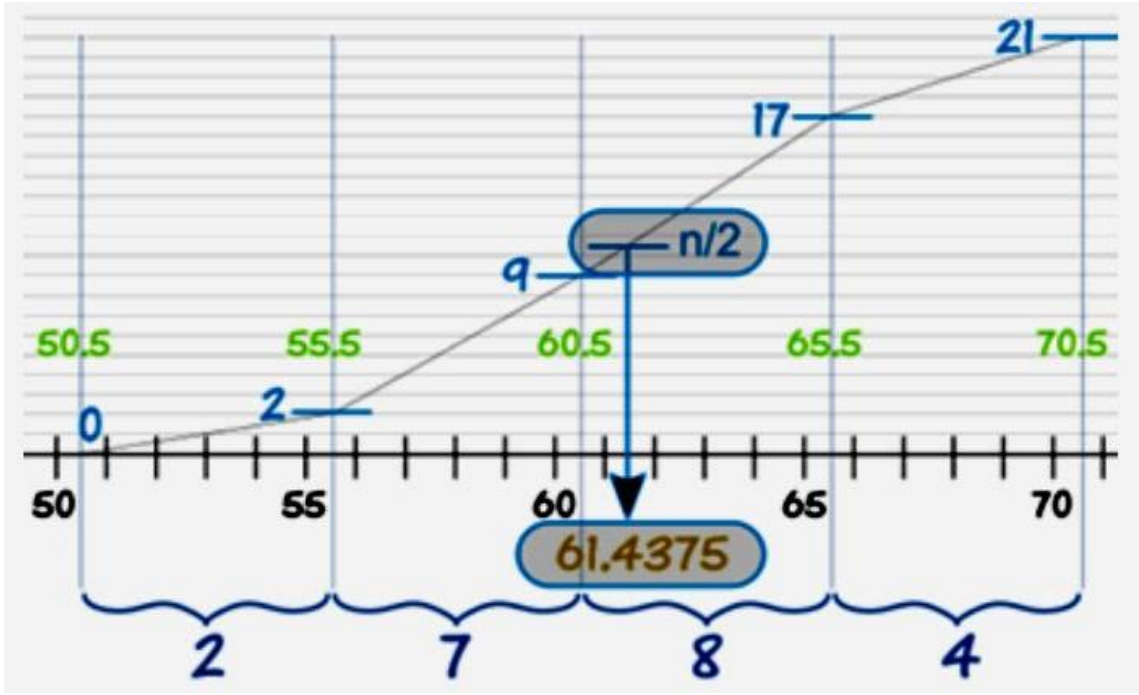
L is the lower class boundary of the group containing the median

n is the total number of values

B is the cumulative frequency of the groups before the median group

G is the frequency of the median group

w is the group width



For our example:

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

L = 60.5

n = 21

B = 2 + 7 = 9

G = 8

w = 5

$$median = L_1 + (\frac{n/2 - (\sum freq)l}{freq_{median}})width$$

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Estimated Median $= 60.5 + \frac{(21/2) - 9}{8} \times 5$

$= 60.5 + 0.9375$

$= \mathbf{61.4375}$

Estimating the Mode from Grouped Data

- Again, looking at our data:
- We can easily find the **modal group** (***the group with the highest frequency***), which is 61 - 65
- We can say **"the modal group is 61 - 65"**
- But the actual Mode may not even be in that group! Or there may be more than one mode. Without the raw data we don't really know. But, we can estimate the Mode using the following formula:

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

$$\text{Estimated Mode} = L + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times w$$

where:

L is the lower class boundary of the modal group

$\mathbf{f_{m-1}}$ is the frequency of the group before the modal group

$\mathbf{f_{m+1}}$ is the frequency of the group after the modal group

$\mathbf{f_m}$ is the frequency of the modal group

w is the group width

For our example:

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

$L = 60.5$

$f_{m-1} = 7$

$f_m = 8$

$f_{m+1} = 4$

$w = 5$

Estimated Mode $= 60.5 + \frac{8 - 7}{(8 - 7) + (8 - 4)} \times 5$

$= 60.5 + (1/5) \times 5$

$= \mathbf{61.5}$

Our final result is:

- Estimated Mean: **61.333...**
- Estimated Median: **61.4375**
- Estimated Mode: **61.5**

(Compare that with the true Mean, Median and Mode of **61.38..., 61 and 62** that we got at the very start.)

Baby Carrots Example

Example: You grew fifty baby carrots using special soil. You dig them up and measure their lengths (to the nearest mm) and group the results:

Length (mm)	Frequency
150 - 154	5
155 - 159	2
160 - 164	6
165 - 169	8
170 - 174	9
175 - 179	11
180 - 184	6
185 - 189	3

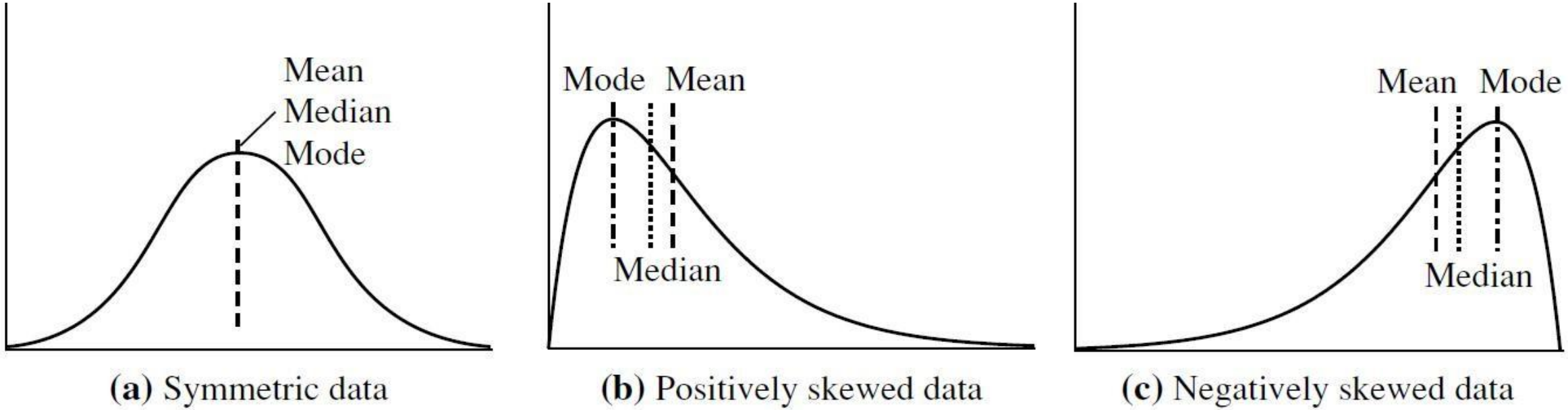
Age Example

The ages of the 112 people who live on a tropical island are grouped as follows:

Age	Number
0 - 9	20
10 - 19	21
20 - 29	23
30 - 39	16
40 - 49	11
50 - 59	10
60 - 69	7
70 - 79	3
80 - 89	1

Symmetric vs. Skewed Data

- Median, mean and mode of ***symmetric, positively*** *and* ***negatively skewed*** data
- Data in most real applications are not **symmetric (a)**. They may instead be either ***positively skewed (b)***, where the mode occurs at a value that is smaller than the median or ***negatively skewed (c)***, where the mode occurs at a value greater than the median.



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Measuring the Dispersion of Data

- **Range, Quartiles, Variance, Standard Deviation, and Interquartile Range**
- We now look at measures to assess the dispersion or spread of numeric data. The measures include **range, quartiles, percentiles,** and **the interquartile range**.
- The five-number summary, which can be displayed as a boxplot, is useful in identifying outliers.
- **Variance** and **standard deviation** also indicate the spread of a data distribution.

Measuring the Dispersion of Data (cont..)

Range

- Let $x_1, x_2, \ldots, x_N$ be a set of observations for some numeric attribute, X.
- The range of the set is the difference between the ***largest (max())*** and ***smallest (min())*** values.

Standard Deviation

- The standard deviation (usually abbreviated SD, sd, or just s) of a bunch of numbers tells you how much the individual numbers tend to differ (in either direction) from the mean. It's calculated as follows:

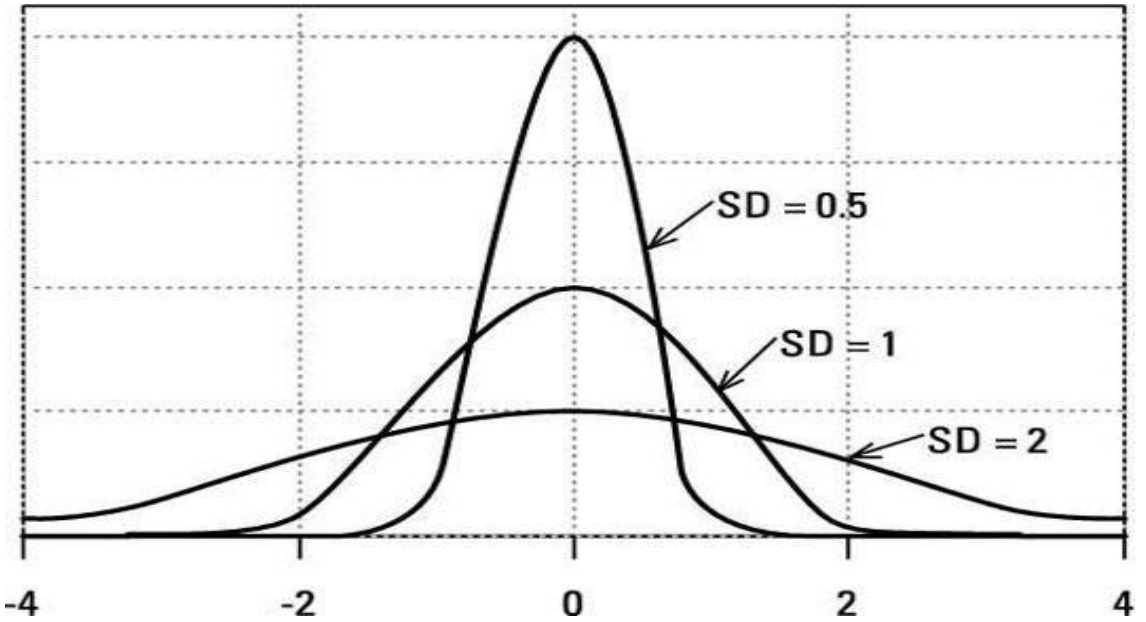
$$SD = sd = s = \sqrt{\frac{\sum_i (d_i)^2}{N-1}}, \text{ where } d_i = X_i - \overline{X}$$

This formula is saying that you calculate the standard deviation of a set of N numbers (X_i) by subtracting the mean from each value to get the deviation (d_i) of each value from the mean, squaring each of these deviations, adding up the $(d_i)^2$ terms, dividing by $N - 1$, and then taking the square root.

Measuring the Dispersion of Data (cont..)

Standard Deviation (cont.)

- This is almost identical to the formula for the root-mean-square deviation of the points from the mean, except that it has ***N – 1*** in the denominator instead of ***N***.
- This difference occurs because the sample mean is used as an approximation of the true population mean (which you don't know). If the true mean were available to use, the denominator would be ***N***.
- When talking about population distributions, the SD describes the width of the distribution curve. The figure shows three normal distributions. They all have a mean of zero, but they have different standard deviations and, therefore, different widths. Each distribution curve has a total area of exactly 1.0, so the peak height is smaller when the SD is larger.



For an IQ example (84, 84, 89, 91, 110, 114, and 116) where the mean is 98.3, you calculate the SD as follows:

$$SD = \sqrt{\frac{(84-98.3)^2 + (84-98.3)^2 + \ldots + (116-98.3)^2}{7-1}} = 14.4$$

Standard deviations are very sensitive to extreme values (outliers) in the data. For example, if the highest value in the IQ dataset had been 150 instead of 116, the SD would have gone up from 14.4 to 23.9.

Why n-1 in Standard Deviation?

Bessel's correction

Why divide by n-1 rather than n?
You compute the difference between each value and the mean of those values. You don't know the true mean of the population; all you know is the mean of your sample. Except for the rare cases where the sample mean happens to equal the population mean, the data will be closer to the sample mean than it will be to the true population mean.
So the value you compute will probably be a bit smaller (and can't be larger) than what it would be if you used the true population mean.
To make up for this, divide by n-1 rather than n. This is called **Bessel's correction.**

But why n-1? If you knew the sample mean, and all but one of the values, you could calculate what that last value must be. Statisticians say there are n-1 degrees of freedom.

Measuring the Dispersion of Data (cont..)

- Several other useful measures of dispersion are related to the SD:
- **Variance:** Variance of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = (\frac{1}{N}\sum_{i=1}^{N}x_i^2) - \bar{x}^2$$

- The variance is just the square of the SD. For the IQ example, the variance = 14.42 = 207.36.
- **Coefficient of variation:** The coefficient of variation (CV) is the SD divided by the mean. For the IQ example, CV = 14.4/98.3 = 0.1465, or 14.65 percent.

Measuring the Dispersion of Data (cont..)

Quartiles

It divide an ordered data set into four equal parts.

The values which divide each part are called the first, second, and third ***quartiles***; they are denoted by Q1, Q2, and Q3, respectively.

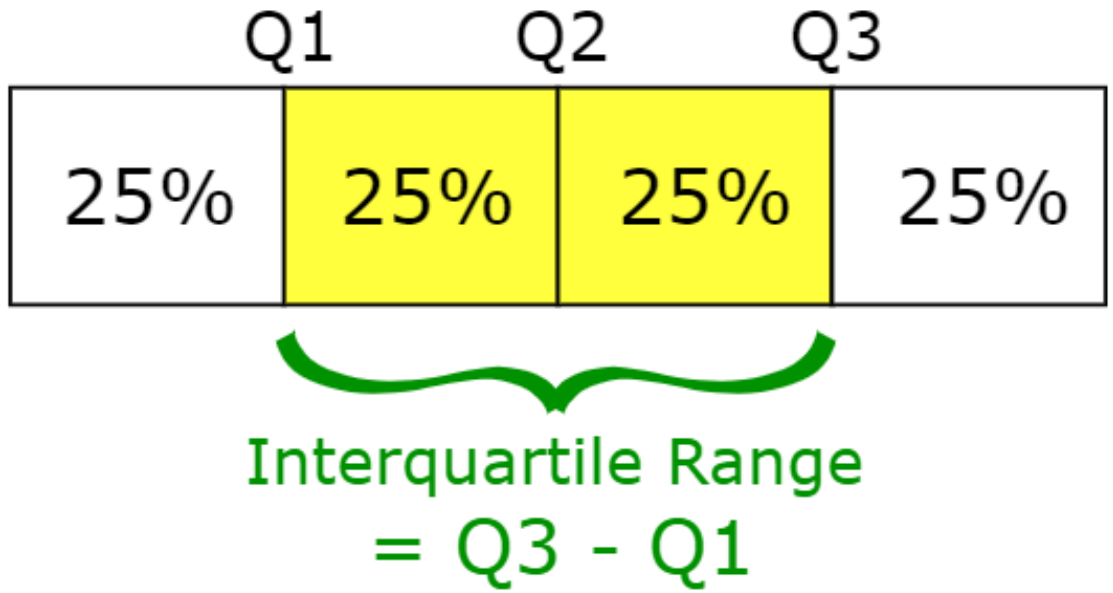
Q1 is the middle value of the first half of the ordered data set, Q2 is the median value in the set, and Q3 is the middle value in the second half of the ordered data set.

Interquartile range (IQR)

A good measure of the spread of data is the ***interquartile range (IQR)*** or the difference between Q3 and Q1. This gives us the width of the box, as well. A small width means more consistent data values since it indicates less variation in the data or that data values are closer together. So, **IQR = Q3 - Q1**

Interquartile Range

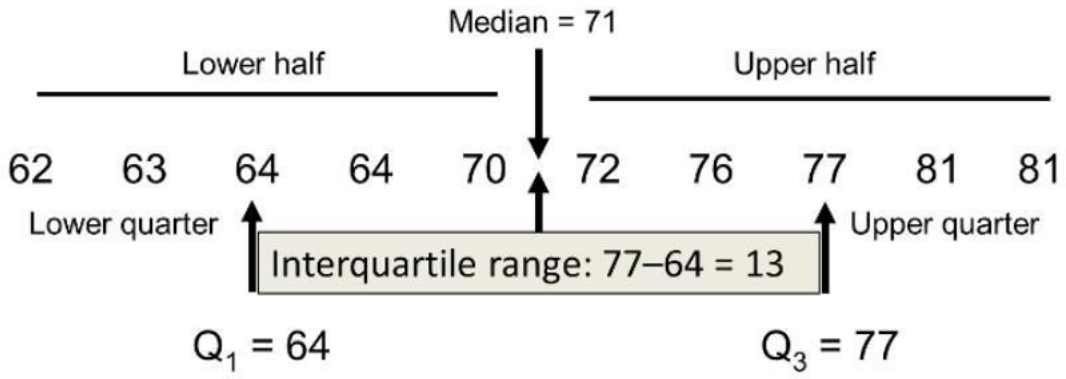
The "Interquartile Range" is from Q1 to Q3:



Interquartile Range = Q3-Q1

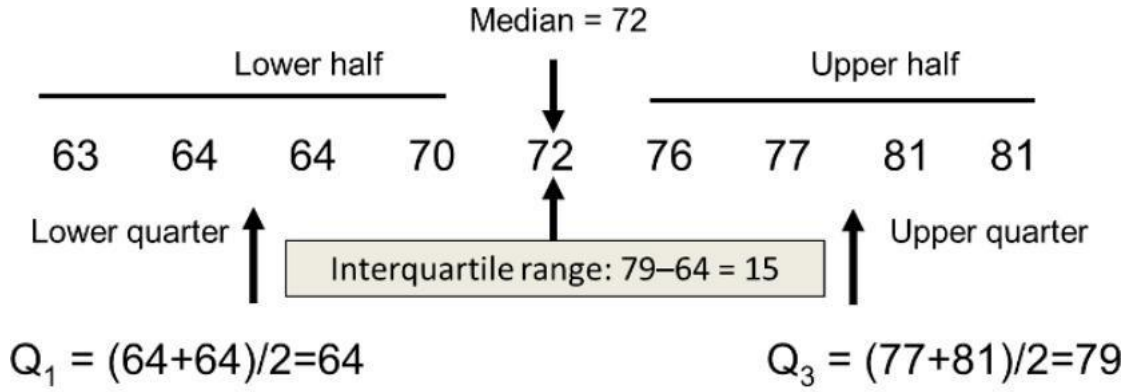
- **With an Even Sample Size:**

- For the sample (n=10) the median diastolic blood pressure is 71 (50% of the values are above 71, and 50% are below).
- The quartiles can be determined in the same way we determined the median, except we consider each half of the data set separately.



- **With an Odd Sample Size:**

- For the sample (n=10) the median diastolic blood pressure is 72 (50% of the values are above 72, and 50% are below).
- When the sample size is odd, the median and quartiles are determined in the same way.
- Suppose in the previous example, the lowest value (62) were excluded, and the sample size was n=9. The median and quartiles are indicated below.



Outliers and Tukey Fences:

- **Tukey Fences**
- When there are no outliers in a sample, the mean and standard deviation are used to summarize a typical value and the variability in the sample, respectively.
- When there are outliers in a sample, the median and interquartile range are used to summarize a typical value and the variability in the sample, respectively.
- Outliers are values **below Q1-1.5(Q3-Q1)** or **above Q3+1.5(Q3-Q1)** or equivalently, values below **Q1-1.5 IQR** or above **Q3+1.5 IQR**.
- In previous example, for the diastolic blood pressures, the lower limit is 64 - 1.5(77-64) = 44.5 and the upper limit is 77 + 1.5(77-64) = 96.5. The diastolic blood pressures range from 62 to 81. Therefore there are no outliers.

Example : The Full Framingham Cohort Data

- The Framingham Heart Study is a long-term, ongoing cardiovascular cohort study on residents of the city of Framingham, Massachusetts. The study began in 1948 with 5,209 adult subjects from Framingham, and is now on its third generation of participants.
- Table 1 displays the means, standard deviations, medians, quartiles and interquartile ranges for each of the continuous variables in the subsample of n=10 participants who attended the seventh examination of the Framingham Offspring Study.

Table 1 - Summary Statistics on n=10 Participants

Characteristic	Mean	Standard Deviation	Median	Q1	Q3	IQR
Systolic Blood Pressure	121.2	11.1	122.5	113.0	127.0	14.0
Diastolic Blood Pressure	71.3	7.2	71.0	64.0	77.0	13.0
Total Serum Cholesterol	202.3	37.7	206.5	163.0	227.0	64.0
Weight	176.0	33.0	169.5	151.0	206.0	55.0
Height	67.175	4.205	69.375	63.0	70.0	7.0
Body Mass Index	27.26	3.10	26.60	24.9	29.6	4.7

- Table 2 displays the observed minimum and maximum values along with the limits to determine outliers using the quartile rule for each of the variables in the subsample of n=10 participants.
- Are there outliers in any of the variables? Which statistics are most appropriate to summarize the average or typical value and the dispersion?

Table 2 - Limits for Assessing Outliers in Characteristics Measured in the n=10 Participants

Characteristic	Minimum	Maximum	Lower Limit[1]	Upper Limit[2]
Systolic Blood Pressure	105	141	92	148
Diastolic Blood Pressure	62	81	44.5	96.5
Total Serum Cholesterol	150	275	67	323
Weight	138	235	68.5	288.5
Height	60.75	72.00	52.5	80.5
Body Mass Index	22.8	31.9	17.85	36.65

[1] Determined by $Q_1-1.5(Q_3-Q_1)$

[2] Determined by $Q_3+1.5(Q_3-Q_1)$

Since there are no suspected outliers in the subsample of n=10 participants, the mean and standard deviation are the most appropriate statistics to summarize average values and dispersion, respectively, of each of these characteristics.

Continue.....

- For clarity, we have so far used a very small subset of the Framingham Offspring Cohort to illustrate calculations of summary statistics and determination of outliers. For your interest, Table 3 displays the means, standard deviations, medians, quartiles and interquartile ranges for each of the continuous variable displayed in Table 1 in the full sample (n=3,539) of participants who attended the seventh examination of the Framingham Offspring Study.

Table 3-Summary Statistics on Sample of (n=3,539) Participants

Characteristic	Mean $\overline{X}$	Standard Deviation (s)	Median	Q1	Q3	IQR
Systolic Blood Pressure	127.3	19.0	125.0	114.0	138.0	24.0
Diastolic Blood Pressure	74.0	9.9	74.0	67.0	80.0	13.0
Total Serum Cholesterol	200.3	36.8	198.0	175.0	223.0	48.0
Weight	174.4	38.7	170.0	146.0	198.0	52.0
Height	65.957	3.749	65.750	63.000	68.750	5.75
Body Mass Index	28.15	5.32	27.40	24.5	30.8	6.3

Continue.....

- Table 4 displays the observed minimum and maximum values along with the limits to determine outliers using the quartile rule for each of the variables in the full sample (n=3,539).

Table 4 - Limits for Assessing Outliers in Characteristics Presented in Table 3

			Tukey Fences	
Characteristic	**Minimum**	**Maximum**	**Lower Limit[1]**	**Upper Limit[2]**
Systolic Blood Pressure	81.0	216.0	78	174
Diastolic Blood Pressure	41.0	114.0	47.5	99.5
Total Serum Cholesterol	83.0	357.0	103	295
Weight	90.0	375.0	68.0	276.0
Height	55.00	78.75	54.4	77.4
Body Mass Index	15.8	64.0	15.05	40.25

[1] Determined by $Q_1-1.5(Q_3-Q_1)$

[2] Determined by $Q_3+1.5(Q_3-Q_1)$

Are there outliers in any of the variables?

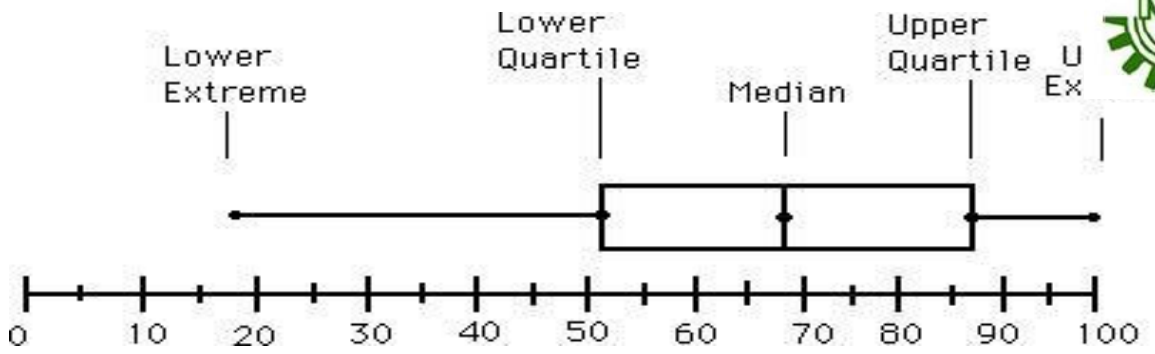
Which statistics are most appropriate to summarize the average or typical values and the dispersion for each variable?

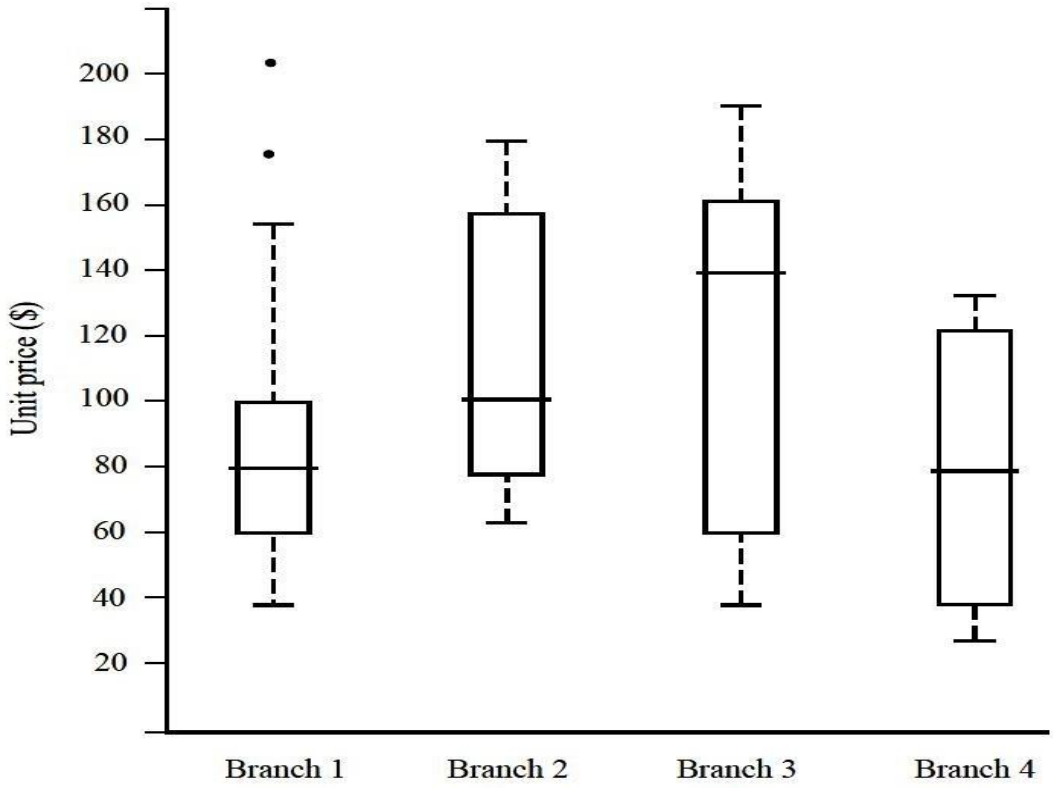
Observations on example.....

- In the full sample, each of the characteristics has outliers on the upper end of the distribution as the maximum values exceed the upper limits in each case. There are also outliers on the low end for diastolic blood pressure and total cholesterol, since the minimums are below the lower limits.
- For some of these characteristics, the difference between the upper limit and the maximum (or the lower limit and the minimum) is small (e.g., height, systolic and diastolic blood pressures), while for others (e.g., total cholesterol, weight and body mass index) the difference is much larger. This method for determining outliers is a popular one but not generally applied as a hard and fast rule. In this application it would be reasonable to present means and standard deviations for height, systolic and diastolic blood pressures and medians and interquartile ranges for total cholesterol, weight and body mass index.

Boxplot Analysis

- **Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:**
- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Two lines (called **whiskers**) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.
  - **Outliers:** points beyond a specified outlier threshold, plotted individually

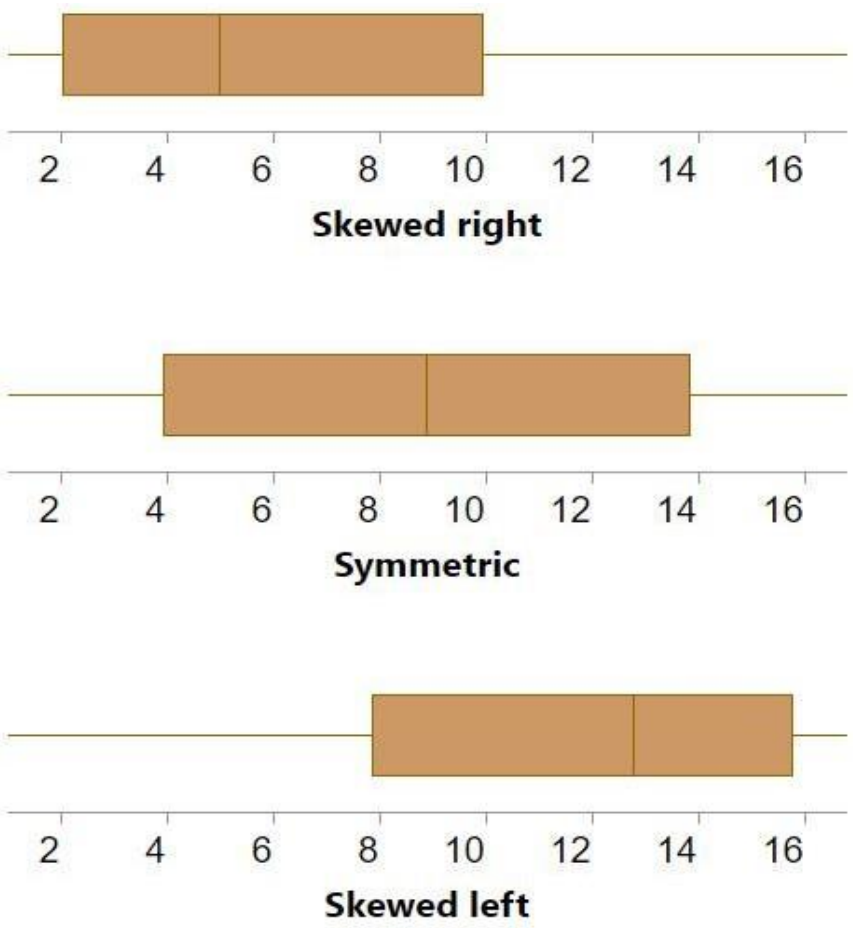




Boxplot for the unit price data for items sold at four branches of AllElectronics during a given time period.

Boxplot Analysis (cont..)

Finally, boxplots often provide information about the shape of a data set. The examples below show some common patterns.



Each of the boxplots illustrates a different **skewness pattern.**

If most of the observations are concentrated on the low end of the scale, the distribution is skewed right; and vice versa.

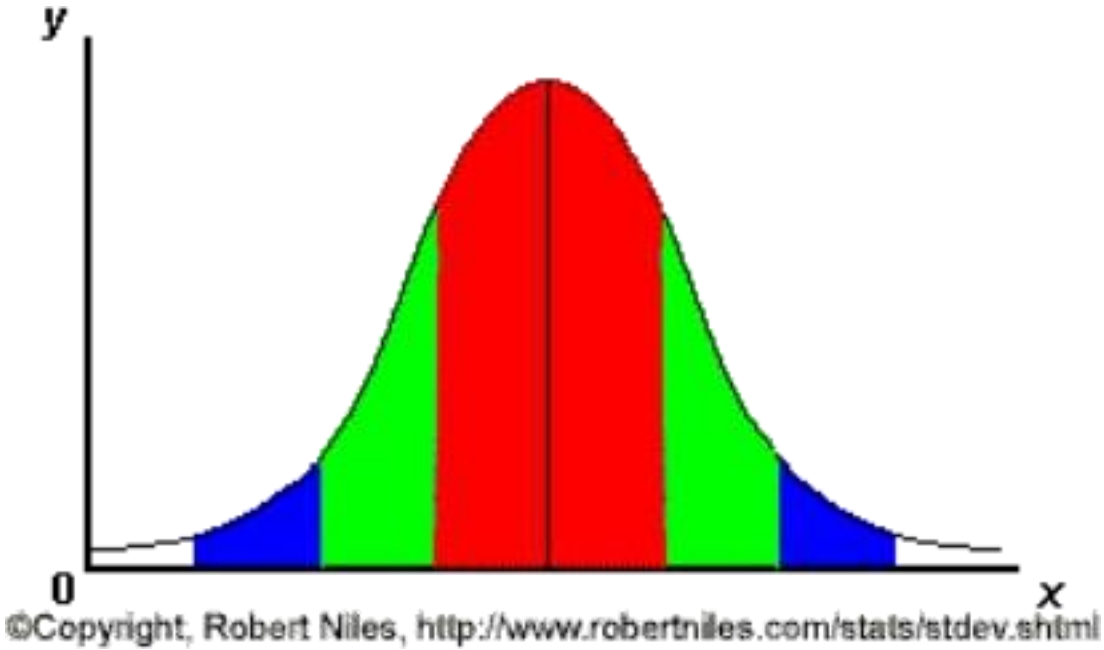
If a distribution is symmetric, the observations will be evenly split at the median, as shown in the middle figure.

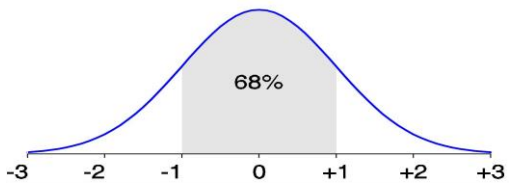
The beauty of the normal curve:

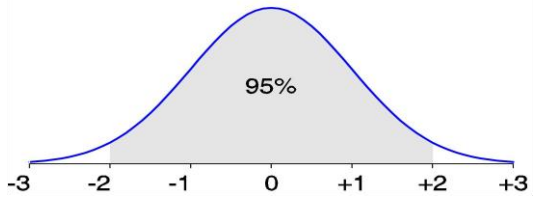
68-95-99.7 Rule

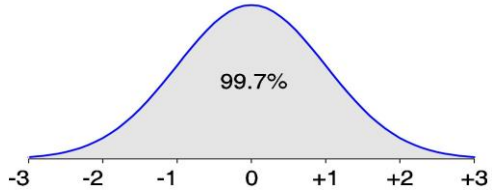
No matter what μ and σ are,

- the area between μ-σ and μ+σ is about 68%; the area between μ-2σ and μ+2σ is about 95%;
- and the area between μ-3σ and μ+3σ is about 99.7%.
- Almost all values fall within 3 standard deviations. (μ: mean, σ: standard deviation)









Are my data “normal”?

- Not all continuous random variables are normally distributed!!
- It is important to evaluate how well the data are approximated by a normal distribution