L2. Numerical Measures of Data

Example 1: Neopalpa Donaldtrumpi

Neopalpa Donaldtrumpi is a small species of moth found in Mexico and in parts of Southern California. The scientist who discovered it, named it after Donald Trump, because the moth's blondish-white head scales and small genitals reminded him of the President.
The wingspans in millimetres for ten N.Donaldtrumpis are shown below:

$$33 \quad 33 \quad 33 \quad 36 \quad 36 \quad 36 \quad 40 \quad 40 \quad 42 \quad 44$$

a. Determine the mean, median, and mode for the given data.

b. Add 5 to each data value, and recalculate the mean. How does adding 5 to each data value change the mean?

c. Multiply each data value by 2 and recalculate the mean. How does multiplying each value by 2 change the mean?

Solution

a. Mode: 33 and 36 Median: $\frac{36+36}{2} = 36$

Mean:

$$\begin{aligned}\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + \cdots + x_{10}}{10} &= \frac{33+33+33+36+36+36+40+40+42+44}{10}\\ &= \frac{373}{10}\\ &= 37.3\,cm\end{aligned}$$

b. Updated dataset: 38 38 38 41 41 41 45 45 47 49

Mean:

$$\begin{aligned}\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + \cdots + x_{10}}{10} &= \frac{38+38+38+41+41+41+45+45+47+49}{10}\\ &= \frac{473}{10}\\ &= 47.3\,cm\end{aligned}$$

Comment: increasing each data value by 5 increases the mean by 5.

c. Updated dataset: 66 66 66 72 72 72 80 80 84 88

Mean:

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + \cdots + x_{10}}{10} = \frac{66+66+66+72+72+72+80+80+84+88}{10}$$
$$= \frac{746}{10}$$
$$= 74.6\,cm$$

Comment: multiplying each data value by 2 increases the mean by a factor of 2.

Example 2: Call 1-855-48-VOICE

The term ”alien” is used in legal contexts to denote people living in the United States who are not citizens. As part of his crack down on illegal immigrants, President Trump and his administration set up a 1-800 number to ”assist victims of crimes committed by criminal aliens”. But instead of receiving calls about crimes perpetrated by illegals, the hotline was inundated with reports of UFO sightings, alien activity, and people claiming that they were victimized by extraterrestrials.

a. For callers using the hotline to report a UFO sighting, the majority encountered a busy signal. The table below displays the number of attempts callers made before successfully reaching an agent.

Number of Attempts	Number of Callers
5	16
7	24
10	35
15	23
18	12

Calculate the average number of attempts that this group of people had to make before their call was connected with an agent.

b. The table below shows the number of minutes callers were placed on hold when reporting that they had been victimized by extraterrestrials.

Time on Hold (min)	Number of Callers
$[0, 10)$	17
$[10, 20)$	14
$[20, 30)$	22
$[30, 40)$	16
$[40, 50)$	11

Calculate the average time that these callers were placed on hold.

Solution

a. Here is the table with a column added to facilitate the $f_i x_i$ calculation

Number of Attempts x_i	Number of Callers f_i	$f_i x_i$
5	16	80
7	24	168
10	35	350
15	23	345
18	12	216
	110	1159

Average number of attempts:

$$\bar{x} = \frac{\sum fx}{n} = \frac{1159}{110} = 10.5364 \text{ attempts}$$

b. Here is the table with a column added to facilitate the $f_i m_i$ calculation

Time on Hold (min)	Number of Callers f_i	m_i	$f_i m_i$
$[0, 10)$	17	5	85
$[10, 20)$	14	15	210
$[20, 30)$	22	25	550
$[30, 40)$	16	35	560
$[40, 50)$	11	45	495
	80		1900

Average time on hold:

$$\bar{x} = \frac{\sum fm}{n} = \frac{1900}{80} = 23.75 \text{ minutes}$$

Example 3: Starbucks

Starbucks has its own grammar for ordering drinks - you should say the drink's size before which syrup you want, and your choice of milk before the type of drink.

a. At a local Starbucks, customers were asked how often they visited a Starbucks in a week. The table below shows the number of customers and their corresponding weekly visit frequency:

Number of Visits	Number of Customers
1	10
5	25
10	70
15	20
20	25

Calculate the average number of visits per week.

b. What is the probability that a randomly selected customer visits a Starbucks 10 times per week?

c. What is an appropriate graphical representation for the data: bar chart or histogram? Why?

d. Describe the shape of the distribution for the number of visits.

e. The individuals who were asked how often they visited a Starbucks on a weekly basis, were also asked how old they were. Below is a table showing their responses:

Age of Customer ($years$)	Number of Customers
$[10, 20)$	30
$[20, 30)$	55
$[30, 40)$	25
$[40, 50)$	30
$[50, 60)$	10

Calculate the average age of a client at Starbucks.

Solution

a. Here is the table with a column added to facilitate the $f_i x_i$ calculation.

Number of Visits	Number of Customers	
x_i	f_i	$f_i x_i$
1	10	10
5	25	125
10	70	700
15	20	300
20	25	500
	150	1635

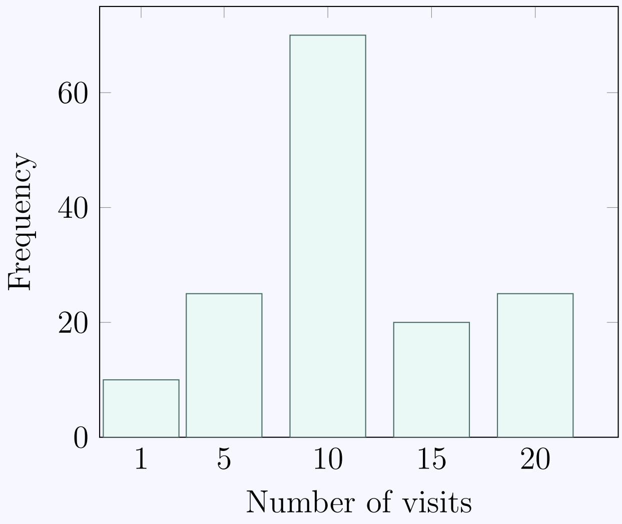
Average number of visits/week

$$\bar{x} = \frac{\sum fx}{n} = \frac{1635}{150} = 10.9 \text{ visits/week}$$

b. Let X = the number of weekly visits to a Starbucks.

$$P(X = 10) = \frac{70}{150} = \frac{7}{15} = 0.4667$$

c. Since the number of visits is a discrete random variable, the appropriate graphical representation is a bar chart.



d. This is a unimodal distribution.

e. Here is the table with a column added to facilitate the $f_i m_i$ calculation.

Age of Customer ($years$)	Number of Customers		
	f_i	m_i	$f_i m_i$
$[10, 20)$	30	15	450
$[20, 30)$	55	25	1375
$[30, 40)$	25	35	875
$[40, 50)$	30	45	1350
$[50, 60)$	10	55	550
	150		4600

Mean age of customer:

$$\bar{x} = \frac{\sum fm}{n} = \frac{4600}{150} = 30.6667 \text{ years old}$$

Example 4: Coffee

In the 1700 s, there was a popular London coffee house called Nando's.
At Nando's they sell three types of coffee beans:

- Arabica costs \$15 per kilogram and makes up 40% of their total coffee inventory.
- Robusta costs \$10 per kilogram and accounts for 35% of their inventory.
- Liberica costs \$25 per kilogram and represents the remaining 25% of their inventory.

a. What is the weighted average cost per kilogram of the coffee beans in the shop?

b. If the shop has 100 kilograms of coffee beans in total, how much would the entire inventory cost based on the weighted average?

Solution

a. Weighted average cost per kilogram of coffee beans:

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1x_1 + w_2x_2 + w_3x_3}{w_1 + w_2 + w_3} = \frac{40 \cdot 15 + 35 \cdot 10 + 25 \cdot 25}{40 + 35 + 25} = \$\,15.75/kg$$

b. Total cost of the inventory:

$$\begin{aligned}\text{Total Cost of Inventory} &= \text{Total Weight} \times \text{Weighted Average} \\ &= 1000\,kg \times 15.75\$/kg \\ &= \$\,1575\end{aligned}$$

Example 5: Made With Love

Rolled oats, brown sugar, and maybe some nuts. But no feelings. In 2017, the Food and Drug Administration went after a small Massachusetts bakery for listing ”love” as an ingredient on its packaging for granola. In a letter posted online, the FDA accused the bakery of mislabelling and misbranding their product. It also inadvertently revealed that the US government doesn't appreciate manufacturers telling consumers that their products are made with love.

Five bags of granola were taken off the shelf a local grocery, and the number of ingredients listed on their packaging were counted and recorded below:

$$6 \quad 8 \quad 10 \quad 12 \quad 14$$

a. For the given data, calculate the sample variance and sample standard deviation.

b. Add 10 to each data value, and recalculate the sample variance and sample standard deviation. How does adding 10 change the value of the variance?

Solution

a. Calculation of the sample variance using $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

Average weight of a package of granola:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{6+8+10+12+14}{5} = 10\, units$$

$Weight$ x_i	$(x_i - \bar{x})^2$
6	$(6-10)^2 = 16$
8	$(8-10)^2 = 4$
10	$(10-10)^2 = 0$
12	$(12-10)^2 = 4$
14	$(14-10)^2 = 16$
50	40

Sample variance:

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{40}{5-1} = 10\, units^2$$

Sample standard deviation:

$$s = \sqrt{s^2} = \sqrt{10} = 3.1623\, units$$

Calculation of the sample variance using $s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]$

$Weight$ x_i	x_i^2
6	36
8	64
10	100
12	144
14	196
50	540

Sample variance:

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] = \frac{1}{5-1}\left[540 - \frac{(50)^2}{5}\right] = \frac{1}{4}[540 - 500] = 10\, units^2$$

Sample standard deviation:

$$s = \sqrt{s^2} = \sqrt{10} = 3.1623\, units$$

b. Updated dataset:

$$16 \quad 18 \quad 20 \quad 22 \quad 24$$

The mean is now $\bar{x} = 10 + 10 = 20$. Therefore,

Weight x_i	$(x_i - \bar{x})^2$
16	$(16 - 20)^2 = 16$
18	$(18 - 20)^2 = 4$
20	$(20 - 20)^2 = 0$
22	$(22 - 20)^2 = 4$
24	$(24 - 20)^2 = 16$
	40

Sample variance:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{40}{5 - 1} = 10\, units^2$$

Comment: Adding 10 to each value in the dataset does not change the value of the variance.

Example 6: Batman

Batman is a province in Turkey, and the name of the caped crusader from Gotham City. Except for the labels, the two have nothing else in common. But that might change if one Batman fan gets his way. Kemal Karic is petitioning to have the borders of Batman province reshaped to look like the Batman logo. So far, some 23 000 people have signed a petition to support the changes

This is not the first time that Batman province made the news for its connection to the Dark Knight. In 2008, the Mayor of Batman announced that he was suing Christopher Nolan for using the province's name without permission. Batman province (pronounced "baht-man") reportedly got its name from an old Turkish unit for weights. Ben Affleck weighs approximately 9.8 batmans.

The weights of other Batmans are shown in the table below

Batman	Weight (*batmans*)
Comic Book	9.5
Adam West	9.1
Michael Keaton	7.2
Val Kilmer	9.3
George Clooney	7.8
Christian Bale	8.2

a. Using the data in the table, calculate the average weight of Batman.

b. Calculate the sample variance and standard deviation for the weight of the Batmans.

c. Lego Batman weighs 0.00004 batmans. If this data point was included in your calculations in parts (a) and (b),

fill in the blanks with one of the following: increases, decreases, stays the same.

i. the mean would ________

ii. the variance would ________

iii. the standard deviation would ________

Solution

a. Mean weight of the Batmans

$$\bar{x} = \frac{\sum x}{n} = \frac{9.5 + 9.1 + 7.2 + 9.3 + 7.8 + 8.2}{6} = \frac{51.5}{6} = 8.5167\, batmans$$

b. Sample Variance

Weight	
x_i	x_i^2
9.5	90.25
9.1	82.81
7.2	51.84
9.3	86.49
7.8	60.89
8.2	67.24
51.1	439.47

$$s^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] = \frac{1}{6-1}\left[439.47 - \frac{(51.1)^2}{6}\right] = \frac{1}{5}[4.2683] = 0.8537\, batmans^2$$

∴ the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{0.8537} = 0.9239\ batmans$$

c. Lego Batman weighs 0.00004 batmans. If this data point was included in your calculations

i. the mean would **decrease**

ii. the variance would **increase**

iii. the standard deviation would **increase**

Example 7: High There?

Scientists at the University of Chicago have developed a proto-app for cannabis users to determine if they are high or not. ”Am I Stoned?” allows users to assess the effects of the drug in their system by testing their level of impairment through a series of tasks designed to measure reaction times, coordination, and recall.
Ratings for the Potbot app in the iTunes store are shown below.

Rating $(in\, stars)$	Number of Customers
5	885
4	12
3	8
2	0
1	95

a. Calculate the average rating of the app.

b. Calculate the variance and standard deviation for the app’s rating.

Solution

a. Augmenting the given table with the necessary columns to compute the mean and variance, we have:

Ratings	Number of Customers		
x_i	f_i	$f_i x_i$	$f_i x_i^2$
5	885	4425	22125
4	12	48	192
3	8	24	72
2	0	0	0
1	95	95	95
	1000	4592	22484

Average rating of the app:

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{4592}{1000} = 4.592\, stars$$

b. Sample variance

$$s^2 = \frac{1}{n-1}\left[\sum fx^2 - \frac{(\sum fx)^2}{n}\right] = \frac{1}{100-1}\left[22484 - \frac{(4592)^2}{1000}\right] = \frac{1}{999}[1397.536] = 1.3989\, stars^2$$

∴ the standard deviation is

$$s = \sqrt{s^2} = \sqrt{1.3989} = 1.1827\, stars$$

Example 8: ER Data

Queensland Health, oversees 16 hospitals in Queensland, Australia. Last year, they released a list of bizarre ER visits. Highlights include 13 hiccup sufferers, 2,089 splinter victims, and two people haunted by nightmares.
The number of minutes that people at the Royal Brisbane Hospital waited in the waiting room before seeing a doctor is shown below.

Time ($mins$)	Number of People
$60 \le x < 120$	10
$120 \le x < 180$	22
$180 \le x < 240$	25
$240 \le x < 360$	15
$360 \le x < 480$	8

a. Calculate the average rating of the app.

b. Calculate the variance and standard deviation for the app's rating.

c. The Empirical Rule states that at least 68% of the data can be found within one standard deviation of the mean (i.e $\bar{x} \pm s$). Calculate the interval for which we would expect that 68% of people visiting the ER would have to wait before seeing a doctor.

d. The Empirical Rule states that at least 95% of the data can be found within two standard deviation of the mean (i.e $\bar{x} \pm 2s$). Calculate the interval for which we would expect that at least 95% of people visiting the ER would have to wait before seeing a doctor.

e. Augment the table with an LTCF column. Use this to estimate the value of the median.

Solution

a. Augmenting the table with the necessary columns we have:

Time ($mins$)	Number of People			
	f_i	m_i	$f_i m_i$	$f_i m_i^2$
$60 \le x < 120$	10	90	900	81000
$120 \le x < 180$	22	150	3300	495000
$180 \le x < 240$	25	210	5250	1102500
$240 \le x < 360$	15	300	4500	1350000
$360 \le x < 480$	8	420	330	1411200
	80		17310	4439700

Average wait time:

$$\bar{x} = \frac{\sum fm}{n} = \frac{17310}{80} = 216.375\, minutes$$

b. Sample variance:

$$s^2 = \frac{1}{n-1}\left[\sum fm^2 - \frac{(\sum fm)^2}{n}\right] = \frac{1}{80-1}\left[4439700 - \frac{(17310)^2}{80}\right] = 8787.9589\, minutes^2$$

∴ the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{8787.9589} = 93.744\, minutes$$

c. 68%

$$\bar{x} \pm s$$
$$216.375 \pm 93.744$$
$$[122.631, 310.119]$$

Interpretation: We expect that at least 68% of patients who visit the ER will need to wait between 122.631 minutes and 310.119 minutes before seeing a doctor.

d. 95%

$$\bar{x} \pm 2s$$
$$216.375 \pm 2(93.744)$$
$$[28.887, 403.863]$$

Interpretation: We expect that at least 95% of patients who visit the ER will need to wait between 28.887 minutes and 403.863 minutes before seeing a doctor.

e. Here is the table with the LCTF column added to it.

Time ($mins$)	Number of People	m_i	$LCTF$
$60 \leq x < 120$	10	90	10
$120 \leq x < 180$	22	150	32
$180 \leq x < 240$	25	210	57
$240 \leq x < 360$	15	300	72
$360 \leq x < 480$	8	420	80
	80		

∵ There are 80 pieces of data, the median would correspond to the 50^{th} data point.

Examining the LCTF's we observe that the 50^{th} data point occurs on the third line of the table and belongs in the category $180 \leq x < 240$. Since we do not know what the actual data point are, we will use the midpoint as a proxy for the median.

$\Rightarrow$ Median = 210