NOTES FOR 23.5
FALL 2024

NIKOS APOSTOLAKIS

Contents

Disclaimer

These notes are a supplement to our textbook, they are not meant to replace it. They reflect the material as I cover it in class.

You should also read the textbook.

Date: Spring 2024.

1. Descriptive statistics

We have some *data* and we want to have effective ways in describing them. Usually our dataset consists of a bunch of numbers.

For example, here is a rather small dataset:

DATA A: 5 0 3 0 0 5 5 3 4 5 0 5 3 5

Size, Frequency, Frequency tables

The first thing we want to know is the /size/ of our dataset, that is how many numbers we have. We count and we see that we have 14 numbers. We use the variable n or N[1] So for our dataset we have

$$N = 14.$$

Notice that some of the values repeat. The /frequency/ of a value x is the number of times that the value appears on the dataset. For example, we have four zeros in our dataset, and we say: 0 appears with frequency 4.

One way of describing a dataset is by its /frequency table/. For example here is the frequency table of the dataset **A**:

x	0	3	4	5
f	4	3	1	6

Discuss: How to find N, the /size/ of our data.

If we add all the frequencies we will get the size of the dataset. We write this using the symbol Σ, that stands for /sum/.

$$N = \sum f.$$

Relative Frequency

Consider the following situation.

Example: Comparing different groups

Group A, of 500 people, took a an exam and 350 people passed. Group B, of 700 people took the same exam and 450 people passed. Which group did better?

Of course, more people from group B passed, but that doesn't necessarily mean that group B did better because there are more people in that group. We really need to compare the *relative frequencies*.

[1] If the data comes from a /population/ we usually use N, and for data that comes from a /sample/ we use n.

$$p = \frac{f}{N}.$$

Relative frequency of success for group A: $\dfrac{350}{500} = 0.7$

Relative frequency of success for group B: $\dfrac{450}{700} = 0.642857142857$

So, group A has a *success rate* of 70% while group B has a success rate of about 64%. So group A did better.

Discus: fractions, decimals, percentages, decimals, rounding.

Discus: Given the previous frequency table construct the relative frequency table.

Example 1. The relative frequency of the grades in an exam in a class with 80 students is shown:

Grade	p
A	0.15
B	0.1
C	0.4
D	0.3
F	0.05
Σ	1.0

(1) How many students got an A?
(2) What percentage of students got a C?
(3) How many students passed? (Any grade except F is considered passing.)

Solution. The formula that gives the frequency in terms of the relative frequency is

$$f = N\,p.$$

In this example $N = 80$. Therefore:

(1) Since $80 \cdot 0.15 = 12$ we have that 12 students got an A.
(2) Since $80 \cdot 0.4 = 32$ we have that 32 students got a C.
(3) It's easier to calculate how many students failed: $80 \cdot 0.05 = 4$ so only 4 students got an F, and everybody else passed. So $80 - 4 = 76$ students passed.

□

Histograms

Data are often given by a histogram. For example see Figure 1. In the horizontal axis we have the *distinct* values and in the vertical axis the frequencys. There is a vertical bar around each value and we can read its frequency value by how high that bar reaches.

x	-1	0	1	2	3
f	25	5	60	20	35

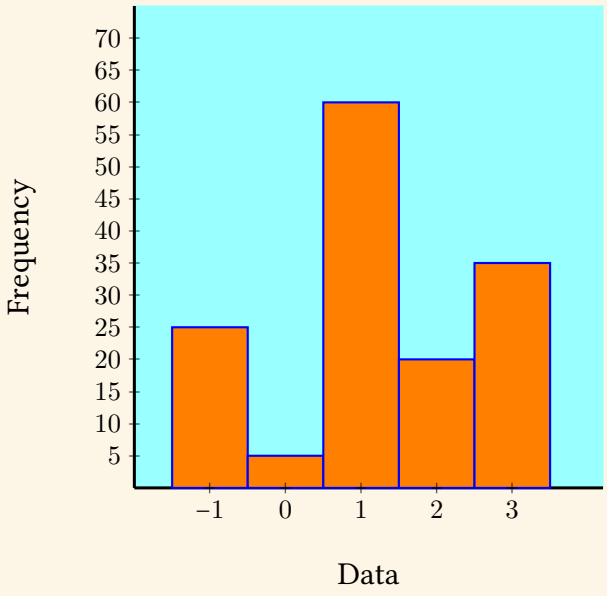


FIGURE 1. Data given as histogram

Often we group the data into classes. For example consider the following dataset:

69	92	57	80	45
62	42	86	96	93
93	70	62	89	65
75	45	79	85	43
53	70	82	98	55
45	73	85	59	77

We can group the data below into classes $40-49$, $50-59$, ..., $90-99$.

4	5	3	5	3	5	
5	7	3	5	9		
6	9	2	2	5		
7	0	5	9	0	3	7
8	0	6	9	5	2	5
9	2	6	3	3	8	

From the *stem and leaf diagram* we can then construct the frequency table for the classes.

Mean and Median

The *mean* (μ for population, $\bar{x}$ for sample) is what we usually refer to as *average*.

The mean

To find the mean

(1) Add all the data.
(2) Divide the sum by the size.

With formulas:

$$\mu = \frac{\sum x}{N}, \qquad \bar{x} = \frac{\sum x}{n}$$

Example 2. Alice has 22$, Bob 35$. Carlos 19$, and Daphne 30$. All together they have

$$\sum x = 22 + 35 + 19 + 30 = 104$$

dollars. There are four people so the size is $N = 4$ we have that the mean amount of money is

$$\mu = \frac{104}{4} = 26.$$

If the data is given by a frequency table, to find the sum we multiply every number with its frequency and then we add the results.

Example 3. Consider the following frequency table of a sample:

x	f
3	10
4	11
5	6
6	9

That means that in our dataset we have ten 3, eleven 4, six 5, and nine 6. To find the sum of all the numbers in our dataset we can first add the threes then all the fours and so on. We can organize these calculation by adding a new column $x \cdot f$ where we multiply each number with its frequency.

	x	f	$x \cdot f$
	3	10	30
	4	11	44
	5	6	30
	6	9	54
$\sum$		36	154

So $n = 36$, and $\sum x = 154$, and we have $\bar{x} = \frac{154}{36} \approx 4.28$

The *median* $\tilde{x}$ is the *middle value*: half of the data are less than the media and half are more.

Example 4. Consider the following data:

$$8 \quad 15 \quad 25 \quad 27 \quad 40.$$

The median is 25 and we have as 2 values less than the median and 2 more than it.

If the size is even then there are two values in the middle. In that case we take the median to be their mean.

Example 5. Consider the following data:

$$1 \quad 6 \quad 11 \quad 15 \quad 17 \quad 21.$$

Now there are two middle numbers. So the median is

$$\tilde{x} = \frac{11 + 15}{2} = 26.$$

Careful: To find the median we have to sort the data if they are not already sorted.

In general, if the size is n, the *position* of the median is

$$\text{Position of the median: } \frac{n+1}{2}.$$

So if we have $n = 5$ then the median is at the third position because $(5 + 1)/2 = 3$. If we have 6 values then the the above formula says that the position of the median is $7/2 = 3.5$, but ofcourse there is no such position. In that case we average the values at the two closest positions namely the third and the fourth.

How to find the median from a frequency table

Consider again the data of Example 3. Since there are $n = 36$ numbers in the dataset the median will be the average of the numbers in the eighteenth and nineteenth positions. But how to find what numbers are in these positions?

Well, the smallest value is 3 with frequency 10, so there are 10 threes and they occupy the first ten positions from position 1 to position 10. The next value in order is 4 and there are 11 of them. So the next 11 positions from position 11 to position 22 are all fours. The next six positions, from position 22 to position 27, are occupied by 5, and the final nine positions, from position 28 to position 36 are occupied by 6.

Positions	1–10	11–21	22–27	28–36
x	3	4	5	6

So the two middle positions, the eighteenth and nineteenth, both have the value 4. Therefore the median is 4.

Variance, Standard deviation

As was the case for the mean, we use different symbols for the variance and the standard deviation depending on whether we consider a population or a sample. For data from a population the *variance* is denoted by σ^2 and the *standard deviation* by σ, while for data from a sample we use s and s^2. But there is also a slight difference in the definitions.

We first define the variance:

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}, \qquad s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}.$$

In both cases the numerator is the sum of the squares of the differences of x from the mean, but for a population the denominator is the size, while for a sample the denominator is *one less* than the size.

In both cases the *standard deviation* is the square root of the variance:

$$\sigma = \sqrt{\sigma^2}, \qquad s = \sqrt{s^2}.$$

It's often easier to do these calculations using a table.

Example 6. Let's find the variance and the standard deviation of the following sample data:

$$46 \quad 37 \quad 40 \quad 33 \quad 42 \quad 36 \quad 40 \quad 47 \quad 34 \quad 45.$$

We make a table with three columns, the first for x, the second for the difference of x and the mean $\bar{x}$, and the third for the squares of the values in the second column.

	x	$x-\bar{x}$	$(x-\bar{x})^2$
	46	6	36
	37	−3	9
	40	0	0
	33	−7	49
	42	2	4
	36	−4	16
	40	0	0
	47	7	49
	34	−6	36
	45	5	25
Σ	400		224

We first added all the values of x in the first column to find that $\sum x = 400$, and since there are $n = 10$ values we have that the mean is

$$\bar{x} = \frac{400}{10} = 40.$$

Then we calculated the second column by subtacting 40 from each value in the first column.

Finally we calculated the third column by squaring each number in the second column. The sum of all the values in the third column is

$$\sum(x-\bar{x})^2 = 224.$$

Since we have sample data to find the variance we divide by one less than the size $n - 1 = 9$:

$$s^2 = \frac{224}{9} \approx 24.888889.$$

To find the standard deviation we take the square root of the variance:

$$s = \sqrt{24.888889} \approx 4.9888765.$$

We usually round to the hundredth place so let's say

$$s = 4.99.$$

2. Thursday, September 12

Example 7. In a population we have $\mu = 36$ and $\sigma = 3$.

(1) Find the z–score for $x = 37, 32, 34, 35, 36.7$.
(2) Find the raw score x for $z = -2.1, 1.6, -05, 0.5, -3, 3$.

Solution. The formulas for calculating the z-score from the raw score x and vice-versa, are:

$$z = \frac{x - \mu}{\sigma}, \qquad x = \mu + z \cdot \sigma.$$

(1) We have:

$$x = 37 \implies z = \frac{37 - 36}{3} = \frac{1}{3} \approx 0.33$$

$$x = 32 \implies z = \frac{32 - 36}{3} = \frac{-4}{3} \approx -1.33$$

$$x = 34 \implies z = \frac{34 - 36}{3} = \frac{-2}{3} \approx -0.67$$

$$x = 36.7 \implies z = \frac{36.7 - 36}{3} = \frac{0.7}{3} \approx 0.23.$$

(2) $2.1 \cdot \sigma = 6.3$, so

$$z = -2.1 \implies x = 36 - 6.3 = 29.7.$$

$1.6 \cdot \sigma = 4.8$, so

$$z = 1.6 \implies x = 36 + 4.8 = 40.8.$$

$0.5 \cdot \sigma = 1.5$, so

$$z = -0.5 \implies x = 36 - 1.5 = 34.5 \qquad z = 0.5 \implies x = 36 + 1.5 = 37.5.$$

$3 \cdot \sigma = 9$, so

$$z = -3 \implies x = 36 - 9 = 27 \qquad z = 3 \implies x = 36 + 9 = 45.$$

□

Example 8. A sample has $\bar{x} = 3.5$ and $s = 0.6$. What range of raw scores does the phrase "Whithin two standard deviations from the mean" describes?

Answer. In general "whithin two standard deviations from the mean" describes the range

$$\bar{x} - 2s \le x \le \bar{x} + 2s.$$

now, $2s = 2 \cdot 0.6 = 1.2$ so the range is

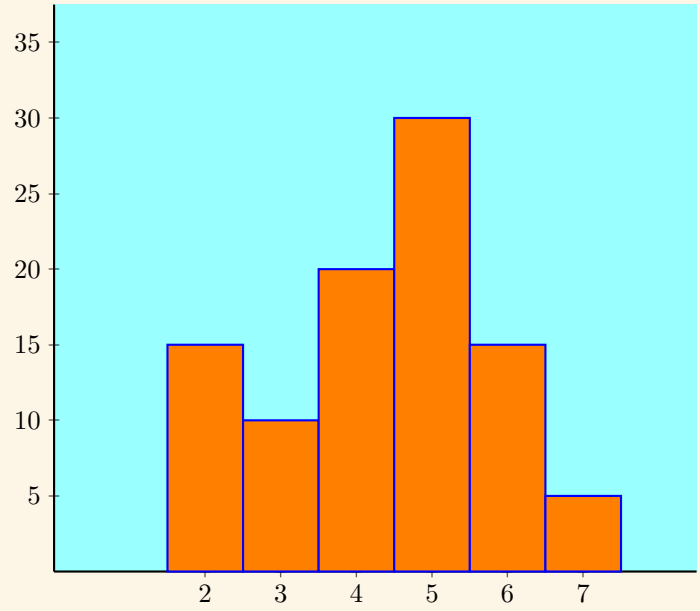
$$3.5 - 1.2 \le x \le 3.5 + 1.2$$

that is

$$2.2 \le x \le 4.7.$$

□

Example 9. Consider the dataset described by the histogram



(1) Calculate the mean μ and the standard deviation σ.
(2) Calculate the z-scores of all distinct values.
(3) Calculate the five-numbers summary and draw the box plot.

We have the following table[2]

	x	f	$x \cdot f$	$x-\mu$	$(x-\mu)^2$	$(x-\mu)^2 \cdot f$	z
	2	15	30	−2.3684211	5.6094185	84.141278	−1.6664153
	3	10	30	−1.3684211	1.8725763	18.725763	−0.96281773
	4	20	80	−0.36842105	0.13573407	2.7146814	−0.25922015
	5	30	150	0.63157895	0.39889197	11.966759	0.44437740
	6	15	90	1.6315789	2.6620497	39.930746	1.1479749
	7	5	35	2.6315789	6.9252075	34.626038	1.8515725
Σ		95	415			192.10527	

We have that $N = 95$ and therefore

$$\mu = \frac{415}{95} \approx 4.37.$$

$$\sigma^2 = \frac{192.10527}{95} \approx 2.02$$

$$\sigma = \sqrt{2.02} \approx 1.42$$

[2] I don't round the intermediate calculations, but it's OK to do so in your homework or in exams.

The z-scores have been computed in the last column in the table above.

Tow compute the five-numbers summary we need to compute the five quartiles Q_0, Q_1, Q_2, Q_3, and Q_4. We have $Q_0 = x_{\min} = 2$ and $Q_4 = x_{\max} = 7$.

Q_2 is the median and since $N = 95$ this will be the 48-th value.

Q_1 is the median of the lower half of the data, that is the 24-th value.

Q_3 is the median of the upper half of the data, that is the 72-nd value.

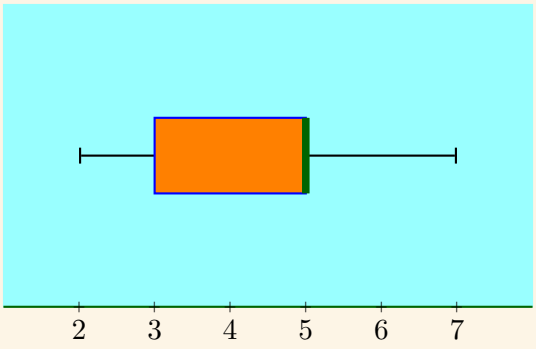
Using the frequency table we have the following correspondence between positions and values:

Position	1–15	16–25	26–45	46–75	76–90	91–95
x	2	3	4	5	6	7

Therefore,

$$Q_1 = 3, \quad Q_2 = 5, \quad Q_3 = 5$$[3]

Using the five-number summary we concstruct the following box plot.



[3]The second and third quartile coincide, this can happen.

3. z–SCORES

If we have a numerical data set, obtained by performing some measurement on a sample or population, we use the variable x to stand for an arbitrary value of the data set.

The mean and standard deviation.

The *mean* or *average* is denoted $\bar{x}$ if our data set comes from a sample, or μ if it comes from the population. The mean is important because it is, in some sense, the *center* of the data set.

The *standard deviation* is denoted s if our data set comes from a sample, or σ if it comes from the population. The standard deviation is important because it measures how *spread* the data set is. In some sense, it tells us how far from the mean the values on our data set are, on average. Small standard deviation means the values are concentrated around the mean, large standard deviation means the values are spread far from the mean.

We often use the standard deviation as a unit of measurement, like *inch*, *foot*, or *meter*. So when you hear (or read) something like "one standard deviation" or "three standard deviations" you should think that it means something like "one foot" or "three feet".

Now, how long our unit of measurement is depends on the sample or the population. So in a sample with standard deviation $s = 3$, "one standard deviation" means 3, "two standard deviations" means 2×3 that is 6, and "half a standard deviation" means 0.5×3 that is 1.5. But in a population with standard deviation $\sigma = 10$, "one standard deviation" means 10, "two standard deviations" means 2×10 that is 20, and "half a standard deviation" means 0.5×10 that is 5.

Since the center of our data is the mean, we often want to know how far from the mean any particular value is. The farthest from the mean the more *exceptional* or *rare* the value is. But, we don't use feet or meters to measure how far from the mean a value is, we use the standard deviation!

Example 10. For a sample we have mean $\bar{x} = 2.4$ and standard deviation $s = 0.2$. What value does the phrase "two standard deviations *above* the mean" indicates? How about "half a standard deviation *below* the mean"?

Answer. Since $s = 0.2$, "two standard deviations" means $2 \times 0.2 = 0.4$. The word "*above*" in this context means "more". In other words, "*above* the mean" means "*added to* the mean" So we have to add "two standard deviations" to the mean.

Now the mean is $\bar{x} = 2.4$ and as we said two standard deviations is 0.4. So the phrase "two standard deviations *above* the mean" indicates the value

$$x = 2.4 + 0.4 = 2.8.$$

In the phrase "half a standard deviation *below* the mean" the word "below" means "less". In other words, "*below* the mean" means "*subtracted from* the mean". Now "half a standard deviation" is $0.5 \times s = 0.5 \times 0.2 = 0.1$. So we have to subtract 0.1 from the mean $\bar{x} = 2.4$. Thus the phrase "half a standard deviation *below* the mean" indicates the value

$$x = 2.4 - 0.1 = 2.3.$$

□

Often we want to know how usual or how rare a value is. For example we may want to know whether a person is tall, short, or about average height. To do that we compare their height with

the average height, if their height is more than the average we'll say that the person is tall, if it is less than the average we'll say that they are short.

Sometimes we may want to know not only if the person is tall or short but also *how tall* or *how short* they are. To do that we check not only if their height is above or below then mean, but also *how much* above or below the mean they are. Again we won't measure this in inches or centimeters but with standard deviations. The reason, of course, is that how tall or short a person in a population is depends not only on their height but on the height of the other people in that population. Here is an example of how we determine how many standard deviations above, or below, the mean a given value in our data set is.

Example 11. The average height of adult men in a certain country is $\mu = 69.92$ inches with a standard deviation of $\sigma = 1.70$ inches. Adolph and Bob are two adult males from that country. Adolph's height is 72.3 inches and Bob's 68.7 inches.

(1) How many standard deviations above the mean is Adolph's height?

Answer. We first find how much more than the average (mean) is Adolph's height by subtracting the mean $\mu = 69.92$ from Adolph's height $x = 72.3$. We find $72.3 - 69.92 = 2.38$, so Adolph is 2.38 inches taller than the average person.

To find out how many standard deviations that difference in height is we simply *divide* 2.38 by the standard deviation. We find that

$$\frac{2.38}{\sigma} = 1.4.$$

So, Adolph's height is 1.4 standard deviations above the mean. □

(2) How many standard deviations below the mean is Bob's height?

Answer. We subtract Bob's height from the mean to find how many inches below the mean his height is. We find $69.92 - 68.7 = 1.22$ inches. So Bob's height is 1.22 inches below the mean. Again to find how many standard deviations that is we divide by the standard deviation:

$$\frac{2.38}{\sigma} \approx 0.72.$$

So, Bob's height is 0.72 inches below the mean. □

Exercises:

(1) A data set coming from a population has mean $\mu = 5$ and standard deviation $\sigma = 6$. Find the values x that are:
   (a) One standard deviation above the mean.
   (b) Two standard deviations below the mean.
   (c) Half a standard deviation above the mean.
   (d) A third of a standard deviation below the mean.

(2) In an exam the average score was $\bar{x} = 76$ and the standard deviation was 8. Jenifer scored 80 in that exam, while Philip scored 68.
   (a) How many standard deviations above the mean was Jenifer's score?
   (b) How many standard deviations below the mean was Philip's score?

The z–score.

The z–score tells us how far from the mean a certain value is, measured in standard deviations. So for any value x in our data set we have its z-score a number that tells us how far above or below the mean that value is, measured in standard deviations. So the mean value has z–score 0, the value 2 standard deviations *above* the mean has z–score 2, the value one standard deviation *below* the mean has z–score -1, and so on. Values above the mean have positive z-score while values below the mean have negative z–score.

Now in Example 11 above we found that Adolph's height was 1.4 standards deviations above the mean, and Bob's was 0.72 standard deviations below the mean. That means that the z–score for Adolph's height is 1.4 and the z–score for Bob's height is -0.72.

If we know a value x from the data set and we want to find its z-score, we subtract the mean from x and then divide by the mean. In formulas

$$z = \frac{x - \bar{x}}{s}, \qquad z = \frac{x - \mu}{\sigma}. \tag{1}$$

Both formulas say the same thing, they just look different because we use different symbols for the mean and standard deviation depending whether we are working with a sample or a population.

Sometimes we want to convert z–scores back to *raw scores* x, (the values of the data set we're working with are called raw scores). In Example 10, we knew the z–score of two values, one had z–score 2, and the other -0.5, and we wanted to find the raw scores. To do that we multiplied the z–score with the standard deviation and added the result to the mean.

To find the raw score x if we know the z–score we use the formulas

$$x = \bar{x} + z \cdot s, \qquad x = \mu + z \cdot \sigma. \tag{2}$$

NOTE: The formulas say to multiply the z–score with the standard deviation and then to *add* the result to the mean. When the z–score is negative when we multiply with the standard deviation we will get a negative number, and so when we add it to the mean we really subtract its absolute value.

Example 12. For a sample we have mean $\bar{x} = 3$ and standard deviation $s = 0.5$.

(1) Find the z–score of $x = 4$.
(2) Find the z–score of $x = 2.75$.
(3) Find the raw score x if $z = 0.25$.
(4) Find the raw score x if $z = -1$.

Solution. (1) We are given the raw score and want to find the z–score. So we use Formula 1. So we first subtract the mean from the value:

$$x - \bar{x} = 4 - 3 = 1,$$

and then divide the result by the standard deviation

$$z = \frac{1}{0.5} = 2.$$

(2) Again we know x and we want to find z. So we use Formula 1, again. We subtract the mean from the given value

$$x - \bar{x} = 2.75 - 3 = -0.75,$$

and then divide the result by the standard deviation

$$z = \frac{-0.75}{0.5} = -1.5.$$

(3) Now we are given z and we want to find x. So now we will use Formula 2. So we first multiply z with s,

$$z \cdot s = 0.25 \cdot 0.5 = 0.125$$

and we add the result to the mean to find x

$$x = 3 + 0.125 = 3.125.$$

(4) We are again given z and we want to find x. So now we use Formula 2 again. So we first multiply z with s,

$$z \cdot s = -1 \cdot 0.5 = -0.5$$

and we add the result to the mean. Please note that we add a negative number -05, so we are really subtracting 0.5:

$$x = 3 + (-0.5) = 2.5.$$

□

We did examples in class and in the homework that used $z-$scores to compare raw scores from different data sets, we compared grades from different exams for example. Do the following exercises:

Exercises:

(1) Carlos and Delilah take two different sections of the same Statistics class, and they both took the midterm exam in their section. Carlos' score was 82 and Delilah's 78. In Carlos' section the average score was 75 and the standard deviation was 8, while in Delilah's section the average score was 70 and the standard deviation was 7.
   (a) Who did better, Carlos or Delilah?
   (b) What score in Delilah's section correspond to Carlos' score?

(2) In Alice's job the average salary is $\$61,000$ with a standard deviation of $\$3,000$, while in Bob's job the average salary is $\$58,000$ with a standard deviation of $\$5,000$. By coincidence the $z-$score of both of their salaries is $z = 2$. Would you rather have Alice's or Bob's salary?

4. The Empirical Rule and Chebyshev's Theorem

The Empirical Rule

This is not a precise statement but rather a *rule of thumb*, and it does not apply to all data sets, it applies only to data sets that have an *approximately bell-shaped histogram*. For these kind of data sets,

(1) Approximately 68% of the data has z-score between -1 and 1.
(2) Approximately 95% of the data has z-score between -2 and 2.
(3) Approximately 99.7% of the data has z-score between -3 and 3.

In Figure 2 we see such a bell-shaped histogram.

Instead of saying that the data set has an approximately bell-shaped histogram, we often just say "the data set has a bell-shaped distribution".

We did many examples in class, here are a few more.

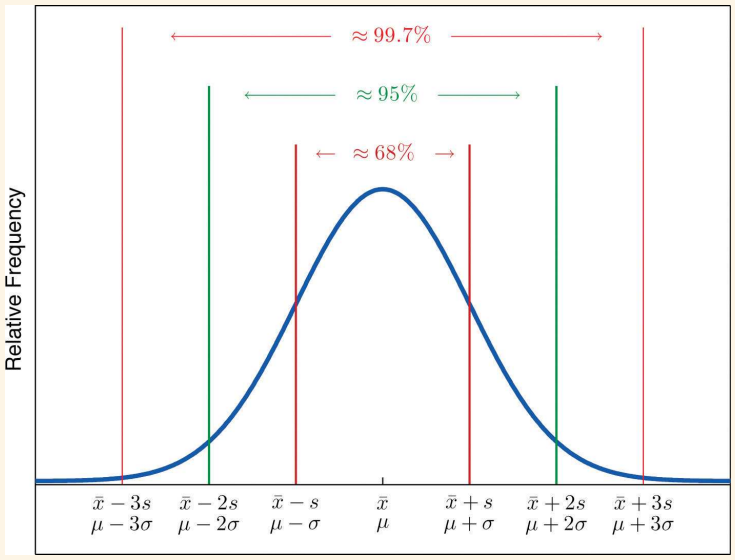


FIGURE 2. The Empirical Rule.

Example 13. The scores in an exam have bell shaped distribution with mean 70 and standard deviation 9. Some typical questions that we can answer:

(1) Approximately what percentage of students scored between 61 and 79?

Answer. We calculate the z–scores of 61 and 79 (look at Formula (1) in the previous section). For $x = 61$ we have

$$z = \frac{61 - 70}{9} = \frac{-9}{9} = -1.$$

For $x = 79$ we have

$$z = \frac{79 - 70}{9} = \frac{9}{9} = 1.$$

So the question asks for the (approximate) percentage of exam scores that have z–score between -1 and 1. According to the Empirical Rule we have that 68% of the data has z-score between -1 and 1. Therefore about 68% of the students scored between 61 and 79. □

(2) About what percentage of students scored below 61?

Answer. We just calculated that the raw score 61 corresponds to $z = -1$. So the question asks what percentage of data has z–score below -1. The Empirical Rule says that the z–scores of approximately 68% of data is within the between -1 and 1. So $100\% - 68\% = 32\%$ is outside that interval. Now, as we see in Figure 2, a bell-shaped histogram is symmetrical about the mean. So of this 32% of data, half of them (that is 16%) have z–score above 1 and half of them have z–score below -1. So we conclude that approximately 16% of students scored below 61. □

(3) About what percentage of students scored above 88?

Answer. We calculate the z–score for $x = 88$:

$$z = \frac{88 - 70}{9} = \frac{18}{9} = 2.$$

So the question asks what percentage of data has z–score above 2. By the Empirical Rule 95% of data has z–score between -2 and 2, and therefore 5% of z–scores is outside that interval. Again by symmetry, we conclude that half of those scores, that is $\frac{1}{2}\,5\% = 2.5\%$, have z–score above 2. So about 2.5% of students scored above 88. □

(4) About what percentage of students scored between 52 and 79?

Answer. We calculate the z–scores. For $x = 52$ we have

$$z = \frac{52 - 70}{9} = \frac{-18}{9} = -2,$$

and we have already calculated above that for $x = 79$ the z–score is $z = 1$. So we want to know the percentage of data that has z-scores between -2 and 1.

Now this is not given directly from the Empirical Rule. However we can calculate the percentage that has z-score above 1 and the percentage that has z-score below -2. Using symmetry as in the previous two question, we have that about 16% of data has z–score above 1 and about 2.5% of data has z–score below -2. So approximately $16\% + 2.5\% = 18.5\%$ of data has z–score outside the interval between -2 and 1. The remaining 81.5% has z–score between -2 and 1.

So, the percentage of students scored between 52 and 79 is approximately 81.5%. □

(5) Find an interval centered around the mean, that contains 99.7% of all the scores.

Answer. From the Empirical Rule the interval that contains 99.7% percent of the data has endpoints with z–scores -3 and 3. So we have to find the raw scores that correspond to these z–scores. We will use Formula (2) from the previous section. We have to multiply the z-scores with the standard deviation and add them to the mean.

$$z = -3 \implies x = 70 + (-3)9 = 70 - 27 = 43$$
$$z = 3 \implies x = 70 + 3 \cdot 9 = 70 + 27 = 97.$$

So the interval that captures approximately 97% of the exam scores is the interval between 43 and 97. □

(6) If we randomly select a student among those that took the test what is the probability that they scored bellow 43?

Answer. From the previous question we know that 43 has z–score $z = -3$. Again using the Empirical Rule and the symmetry of the curve we know that approximately $\frac{1}{2}\,0.3\% = 0.15\%$ of exam scores have z–score below -3. Thus the probability that a randomly selected student scored bellow 43 is 0.15% or, equivalently 0.0015. □

When we actually know the size of the sample or the population we can find not only the percentages but also actual number of data that lies in a given interval.

Example 14. An exam was given to 400 students and the scores had a bell–shaped distribution with mean 75 and standard deviation 5. Approximately how many students scored above 85?

Answer. The z-score for $x = 85$ is

$$z = \frac{85 - 75}{5} = \frac{10}{5} = 2.$$

From the previous example we know that approximately 2.5% of data has z–score above $z = 2$. So approximately 2.5% of students scored above 85. Since we have a total of 400 this means that the actual number of students is approximately

$$2.5\% \text{ of } 400 = 0.025 \cdot 400 = 10.$$

So approximately 10 students scored above 85. □

Chebyshev's Theorem

This is an actual theorem, not a rule. That means it is a *fact*, it is always true. Also notice that it works for all data sets, no matter the shape of their histogram. In full generality it says:

Chebyshev's Theorem At least $\dfrac{k^2-1}{k}$ of the data has z–score between $-k$ and k for all integers k. There are two interesting special cases that we use often:

(1) At least $\dfrac{3}{4}$ of the data, in other words 75% of the data, has z-score between -2 and 2.

(2) At least $\dfrac{8}{9}$ of the data, in other words approximately 88.9% of the data, has z-score between -3 and 3.

Unlike the Empirical Rule, we can be sure that *at least* that proportion of date is within the given interval. We say approximately 88.9% because we converted the fraction $\frac{8}{9}$ to a percentage, we know that for sure eight ninths of the data have z-scores in the interval between -3 and 3.

Example 15. A group of friends went in a fishing expedition and all together they caught 50 fish. One of the friends was taking a statistics class at the time and he decided to calculate the mean and the standard deviation of length of all the fish caught. He found that the average length was 20 inches with a standard deviation of 3 inches. Assuming that his calculations were correct answer the following questions:

(1) What can we say about the number of the caught fish that was between 14 and 26 inches long?

Answer. Since we don't have any information about the shape of the data we will use Chebyshev's Theorem. We start by calculating the z–scores for $x = 14$ and $x = 26$.

$$x = 14 \implies z = \frac{14-20}{3} = \frac{-6}{3} = -2$$
$$x = 26 \implies z = \frac{26-20}{3} = \frac{6}{3} = 2.$$

By Chebyshev's Theorem we know that at least 75% of the data, in our case the length of the fish, has z–score between -2 and 2. So at least 75% of the fish was between 14 and 26 inches long.

Now we calculate what is 75% of 50. We find $0.75 \cdot 50 = 37.5$. So that means *at least* 37.5 fish was between 14 and 26 inches long. Now the number of fish is a whole number so we can say that at least 38 fish was between 14 and 26 inches long. □

(2) After a few months when all the actual date was lost and only the mean and the standard deviation was known, one of the friends was claiming that in that trip he had caught six fish and the were all more than 30 inches long. Can his claim be true?

Answer. We calculate the z-score for $x = 30$

$$x = 30 \implies z = \frac{30 - 20}{3} = \frac{10}{3} \approx 3.33.$$

So that friend claims that the length of all six fish he caught had a z–score above 3.

Now from Chebyshev's Theorem we know that at least 88.9% of the data have z–score between -3 and 3. We calculate and find that 88.9% of 50 is 44.45. This means that at least 45 of the caught fish had length with z–score between -3 and 3.

So it is impossible to have 6 fish with the z–sore of their length more than 3. Therefore that friend's claim cannot possibly be true. □

Example 16. The number of daily sales in a used cars dealership, over a period of a 150 days, was found to have mean $\mu = 5$ and standard deviation $\sigma = 2$. Bob worked at that dealership for 45 of these days, and while on a date night he was overheard to claim that he sold at least ten cars for each of those days. Explain why Bob's claim can't possibly be true.

Solution. 45 is 30% of 150 and the z-score of 10 is 2.5. So if Bob's claim was true at least 30% of the sales during these 150 days would be 2.5 standard deviations above the mean. But, by Chebyshev's Theorem, at least 75% of the sales are within two standard deviations from the mean, and so at most 25% of values could be 10 or more. □