The Empirical Rule and Chebyshev's Theorem: It is the last topic within the Descriptive Statistics.

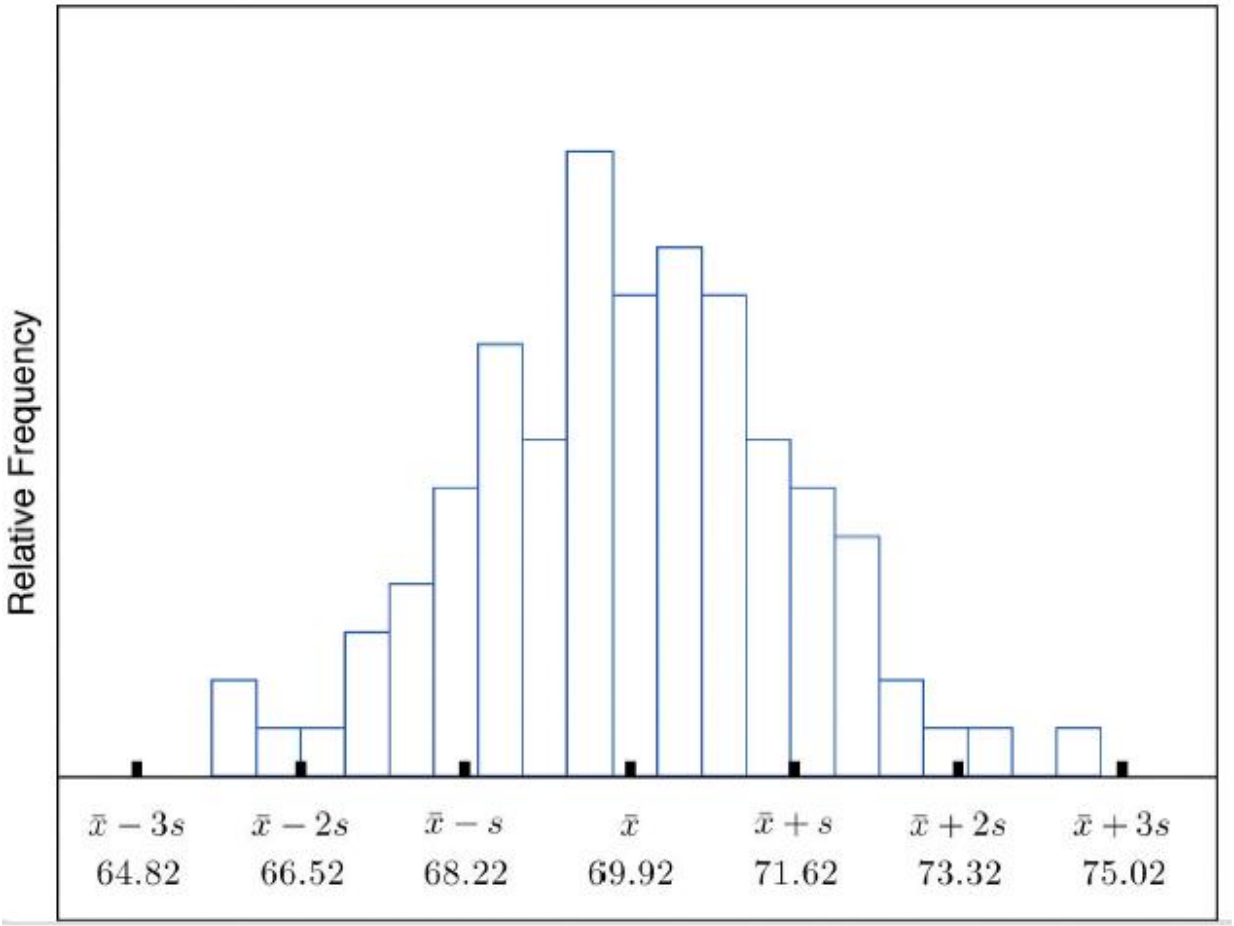
Objectives:

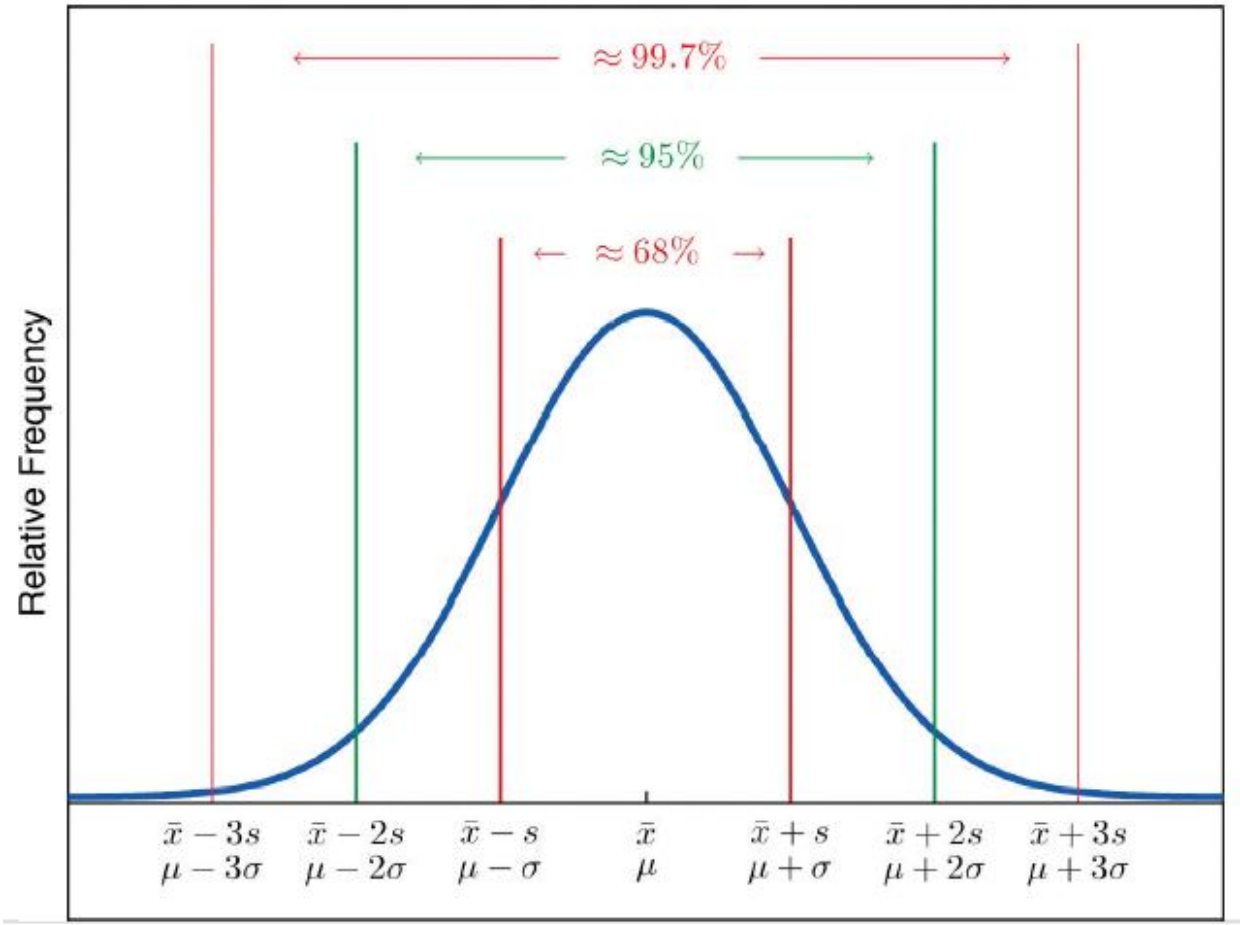
- To learn what the value of the standard deviation of a data set implies about how the data scatter away from the mean as described by the Empirical Rule and Chebyshev's Theorem.
- To use the Empirical Rule and Chebyshev's Theorem to draw conclusions about a data set.

The Empirical Rule:

Suppose the following table represents 100 randomly selected of daily records of relative humidity measurement in percentage during known period for specific meteorological station. Based on the procedures to calculate the mean and standard deviation of the data: $\bar{x}$ = 69.92 and *s* = 1.70, which rounded to two decimal places.

- If we go through the data and count the number of observations that are within one standard deviation of the mean, that is, that are between $\bar{x}$ - *s = 69.92 – 1.70 = 68.22% and* $\bar{x}$ *+ s = 69.92 + 1.70 = 71.62%.* There are 68 of them.
- If we count the number of observations that are within two standard deviations of the mean, that is, that are between $\bar{x}$ - *2(s) = 69.92 – 2(1.70) = 66.52% and* $\bar{x}$ *+ 2(s) = 69.92 + 2(1.70) = 73.32%.* There are 95 of them.
- All of the measurements are within three standard deviations of the mean, that is, between $\bar{x}$ - *3(s) = 69.92 – 3(1.70) = 64.822% and* $\bar{x}$ *+ 3(s) = 69.92 + 3(1.70) = 75.02%.* There are 99.7 of them.





A relative frequency histogram for the data set in the table above

-**Keys:**

- The Empirical Rule can be applied only when data set has approximately bell-shaped relative frequency histogram.
- Approximately 68% of the data lie within one standard deviation of the mean, that is, in the interval with endpoints $\bar{x} \pm s$ for samples and with endpoints $\mu \pm \sigma$ for populations.
- Approximately 95% of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations.
- Approximately 99.7% of the data lies within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations.

68.7	72.3	71.3	72.5	70.6	68.2	70.1	68.4	68.6	70.6
73.7	70.5	71.0	70.9	69.3	69.4	69.7	69.1	71.5	68.6
70.9	70.0	70.4	68.9	69.4	69.4	69.2	70.7	70.5	69.9
69.8	69.8	68.6	69.5	71.6	66.2	72.4	70.7	67.7	69.1
68.8	69.3	68.9	74.8	68.0	71.2	68.3	70.2	71.9	70.4
71.9	72.2	70.0	68.7	67.9	71.1	69.0	70.8	67.3	71.8
70.3	68.8	67.2	73.0	70.4	67.8	70.0	69.5	70.1	72.0
72.2	67.6	67.0	70.3	71.2	65.6	68.1	70.8	71.4	70.2
70.1	67.5	71.3	71.5	71.0	69.1	69.5	71.1	66.8	71.8
69.6	72.7	72.8	69.6	65.9	68.0	69.7	68.7	69.8	69.7

- The Empirical Rule does not apply to data sets with severely asymmetric distributions, and the actual percentage of observations in any of the intervals specified by the rule could be either greater or less than those given in the rule.

Case (1) Consider a specific river discharges have a bell-shaped distribution with mean 69.6m^3/sec and standard deviation 1.4m^3/sec.

a. About what proportion of all such discharges are between 68.2 m^3/sec and 71 m^3/sec?

b. What interval centered on the mean should contain about 95% of all such discharges?

Solution:

A sketch of the distribution of heights is given below:

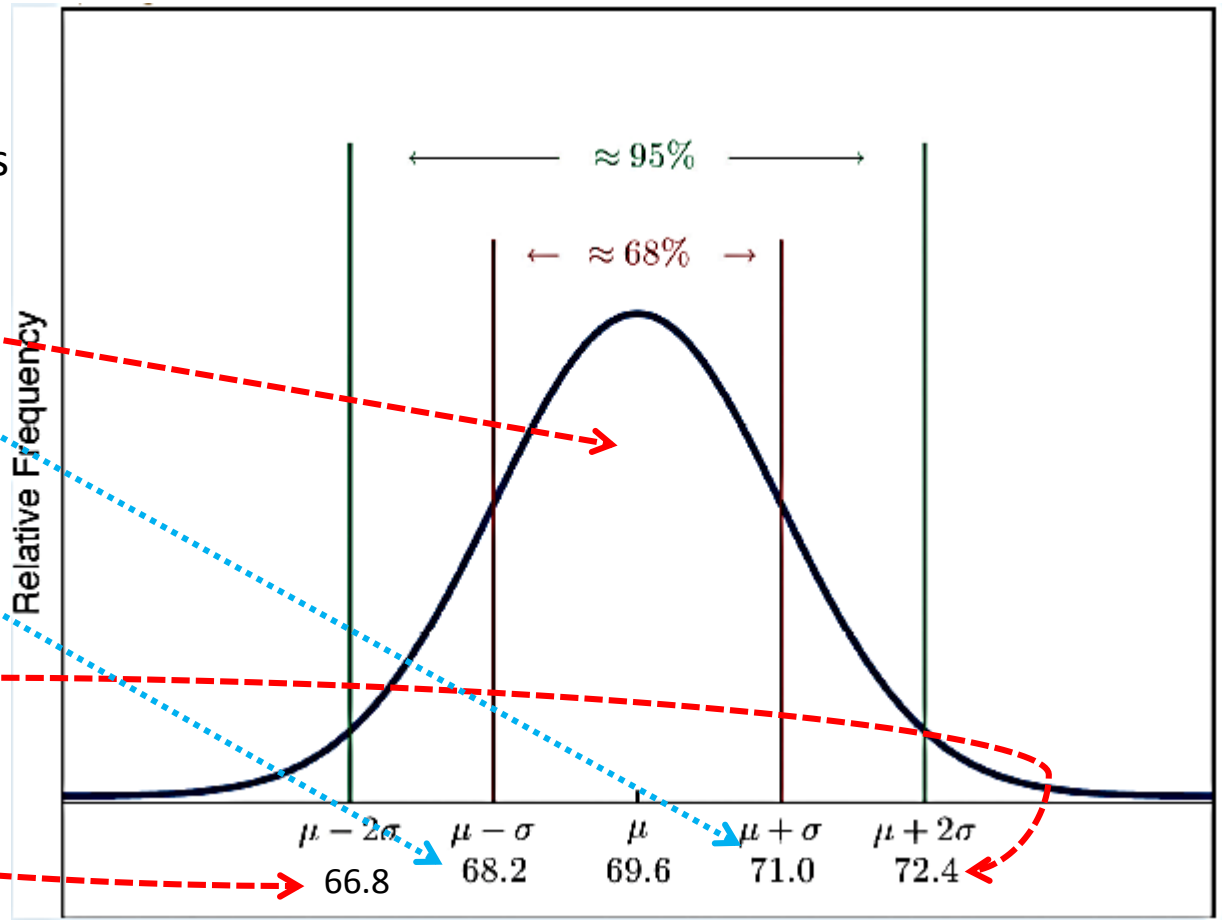
a. Since the interval from 68.2 m^3/sec to 71.0 m^3/sec has endpoints $\bar{x} - s$ and $\bar{x} + s$, by the Empirical Rule about 68% of all recorded observations should have discharges in this range.

b. By the Empirical Rule the endpoints $\bar{x} - 2s$ and $\bar{x} + 2s$

Hence…

$\bar{x} - 2s = 69.6 - 2(1.4) = 66.8$ *and* $\bar{x} + 2s = 69.6 + 2(1.4) = 72.4$

The interval in question is the interval from 66.8 m^3/sec to 72.4 m^3/sec..



Case (2) Suppose the annual averages of precipitation recorded in specific Station have a bell-shaped distribution with mean μ = 100mm and standard deviation σ = 10mm. Discuss what the Empirical Rule implies concerning the measurements with annual averages of 110mm, 120mm, and 130mm.

Solution:

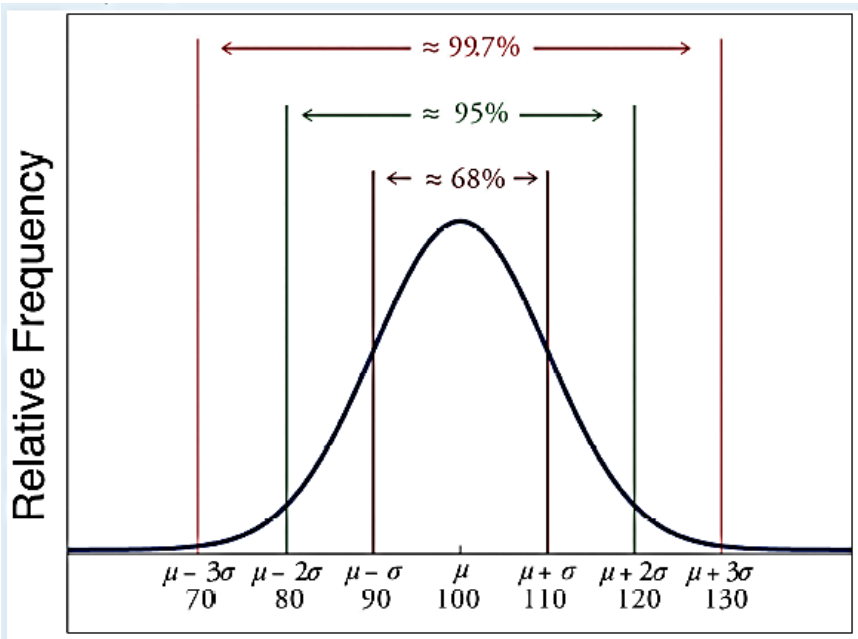
A sketch of annual averages distribution of precipitation is given below:

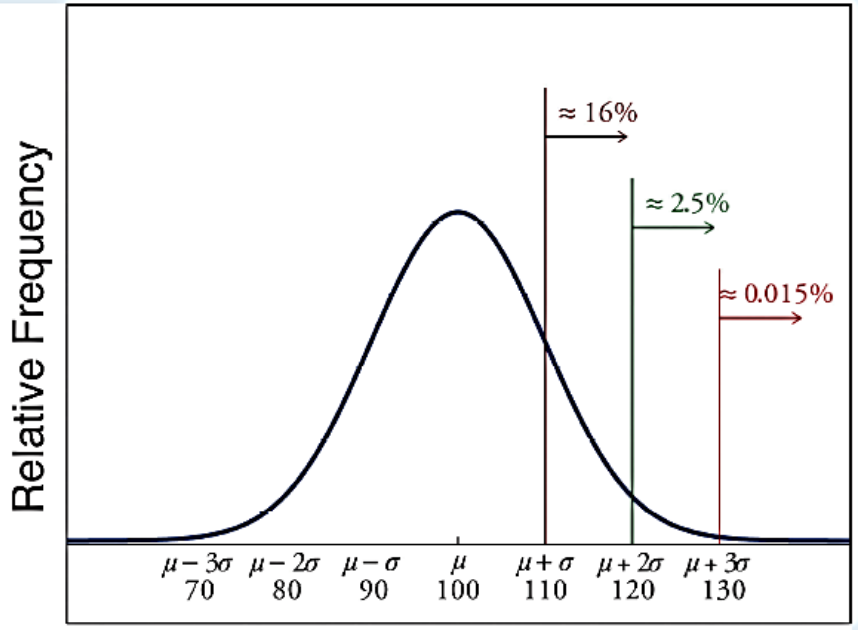
The Empirical Rule states that:

1- Approximately 68% of the annual averages of precipitation in the population lie between 90mm and 110mm.

Since 68% of the measurements lie within the interval from 90mm to 110mm, it must be the case that 32% lie outside that interval.

By symmetry approximately half of that 32%, or 16% of all observations, will lie above 110mm. If 16% lie above 110mm, then 84% lie below. We conclude that the precipitations measurement 110mm represents the 84th percentile.





2- Approximately 95% of the annual averages of precipitation in the population lie between 80mm and 120mm.

The same analysis applies to the observation 120mm. Since approximately 95% of all precipitation averages lie within the interval form 80mm to 120mm, only 5% lie outside it, and half of them, or 2.5% of all measurements, are above 120mm. The measurement 120mm is thus higher than 97.5% of all observations, and is quite a high average.

3- Approximately 99.7% of the annual averages of precipitation in the population lie between 70mm and 130mm.

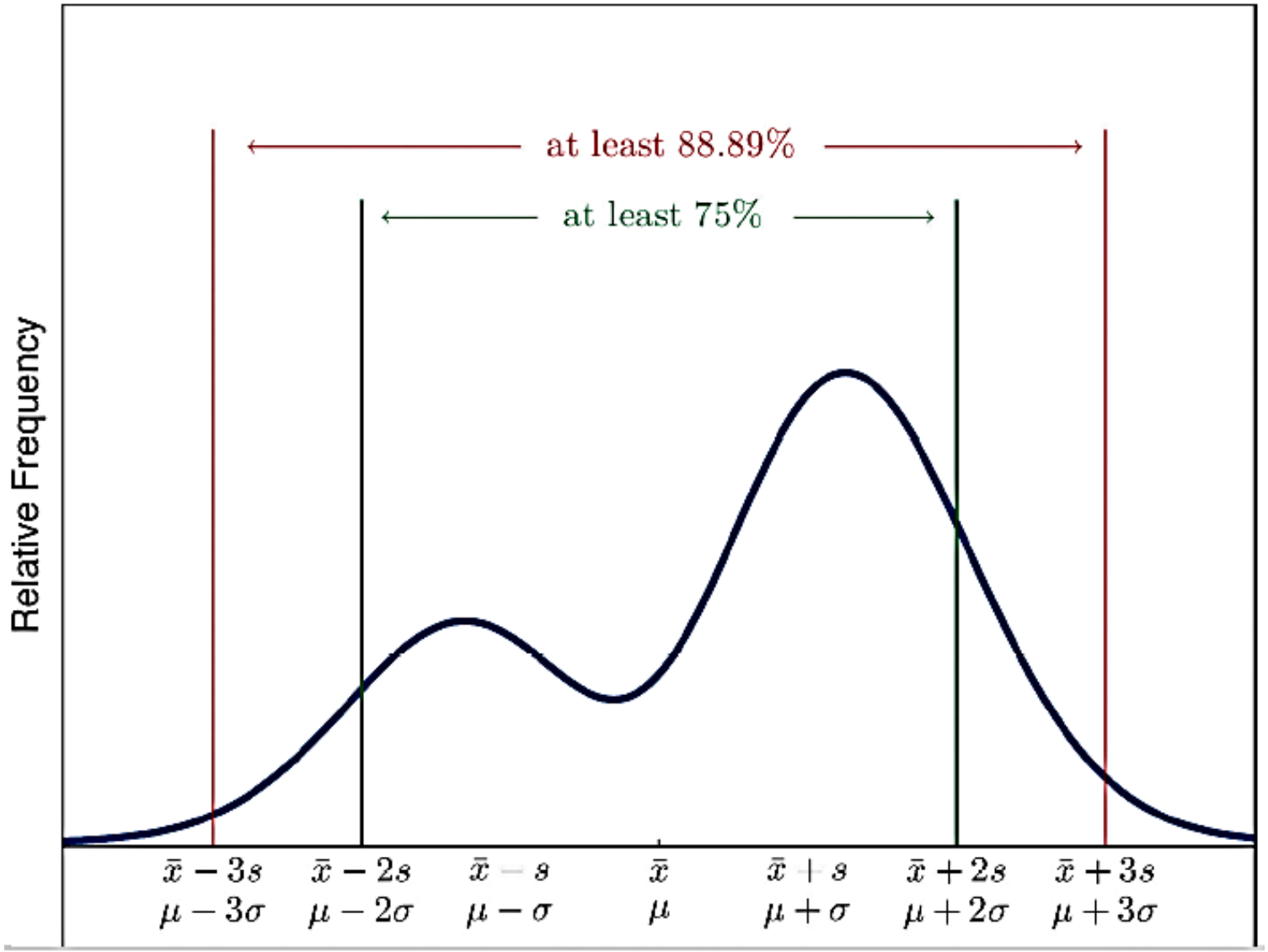
By a similar argument, only 15/100 of 1% of all recorded observations, or about one or two in every thousand, would have an precipitation average above 130mm. This fact makes the measurement 130mm extremely high.

Chebyshev's Theorem:

- As the Empirical Rule does not apply to all data sets, it is only applied when data sets have bell shaped, and even then is stated in terms of approximations. Chebyshev's Theorem can be applied on every data set.
- Chebyshev's Theorem is a fact that applies to all possible data sets. It describes the minimum proportion of the measurements that lie must within one, two, or more standard deviations of the mean.

For any numerical data set and based on the visual illustration below can be noted the following:

- At least 3/4 of the data lie within two standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 2s$ for samples and with endpoints $\mu \pm 2\sigma$ for populations.
- At least 8/9 of the data lie within three standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm 3s$ for samples and with endpoints $\mu \pm 3\sigma$ for populations.
- At least $1 - 1 / k^2$ of the data lie within k standard deviations of the mean, that is, in the interval with endpoints $\bar{x} \pm ks$ for samples and with endpoints $\mu \pm k\sigma$ for populations, where k is any positive whole number that is greater than 1.



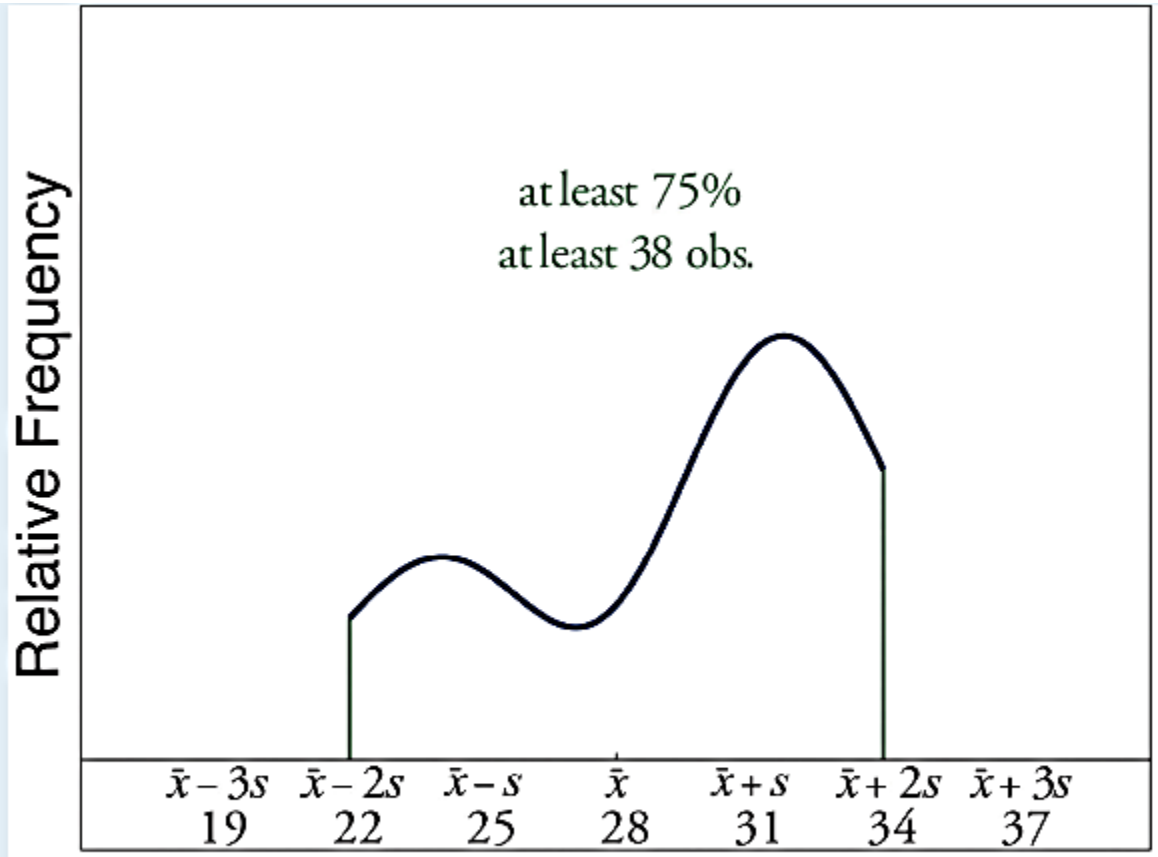
Visual illustration of Chebyshev's Theorem.

Case (3) A sample of size n = 50 has mean $\bar{x}$ = 28 and standard deviation s = 3.

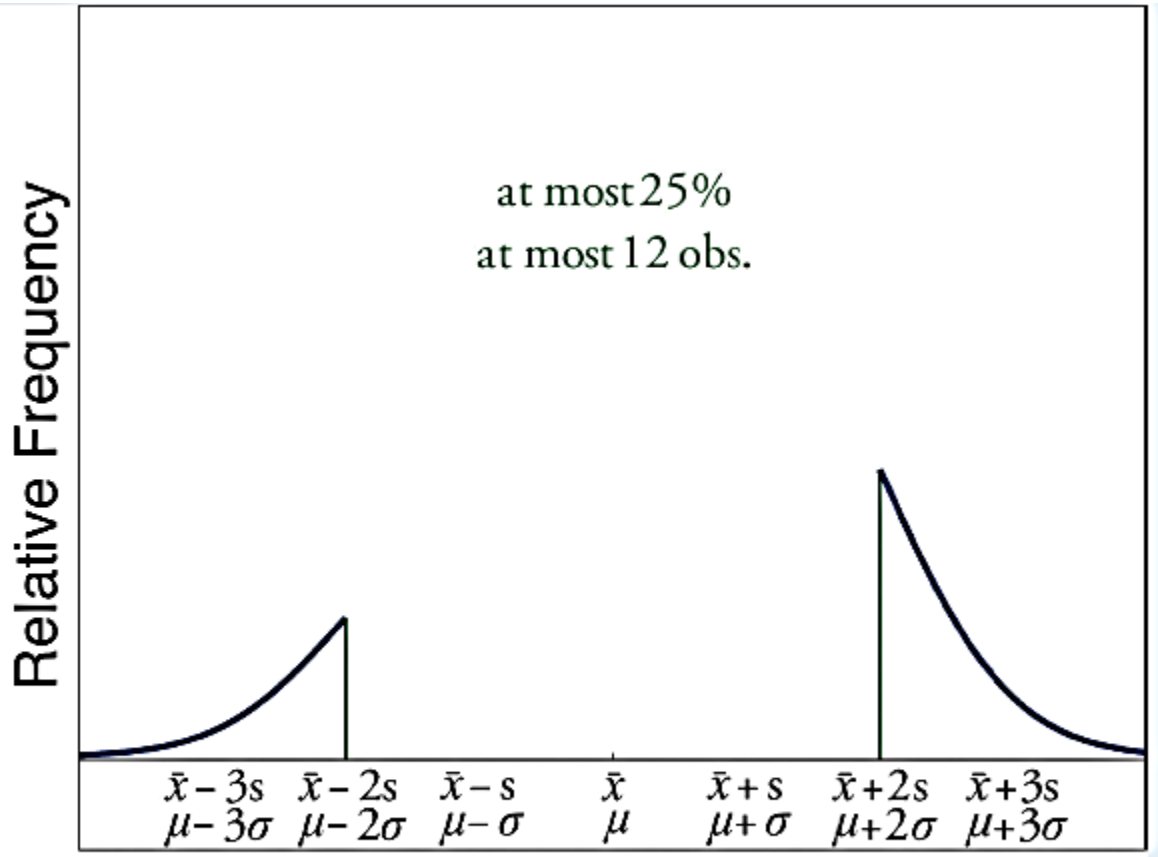
Without knowing anything else about the sample, what can be said about the number of observations that lie in the interval (22,34)? What can be said about the number of observations that lie outside that interval?

Solution:

- The interval (22,34) is the one that is formed by adding and subtracting two standard deviations from the mean. By Chebyshev's Theorem, at least 3/4 of the data are within this interval. Since 3/4 of 50 is 37.5, this means that at least 37.5 observations are in the interval. But one cannot take a fractional observation, so we conclude that at least 38 observations must lie inside the interval (22,34).
- If at least 3/4 of the observations are in the interval, then at most 1/4 of them are outside it. Since 1/4 of 50 is 12.5, at most 12.5 observations are outside the interval. Since again a fraction of an observation is impossible, x (22,34).
- See the figure below which used in solution.



(a) Within $\bar{x} \pm 2s$



(b) Outside $\bar{x} \pm 2$

Case (4) The number of vehicles passing through a busy intersection between 8:00AM and 10:00AM was observed and recorded on every weekday morning of the last year. The data set contains n = 251 numbers. The sample mean is $\bar{x}$ = 725 and the sample standard deviation is s = 25. Identify which of the following statements must be true.

1- On approximately 95% of the weekday mornings last year the number of vehicles passing through the intersection from 8:00AM to 10:00AM was between 675 and 775.

2. On at least 75% of the weekday mornings last year the number of vehicles passing through the intersection from 8:00AM to 10:00AM was between 675 and 775.

3. On at least 189 weekday mornings last year the number of vehicles passing through the intersection from 8:00AM to 10:00AM was between 675 and 775.

4. On at most 25% of the weekday mornings last year the number of vehicles passing through the intersection from 8:00AM to 10:00AM was either less than 675 or greater than 775.

5. On at most 12.5% of the weekday mornings last year the number of vehicles passing through the intersection from 8:00AM to 10:00AM was less than 675.

Solution:

1. Since it is not stated that the relative frequency histogram of the data is bell-shaped, the Empirical Rule does not apply. Statement (1) is based on the Empirical Rule and therefore it might not be correct.

2. Statement (2) is a direct application of part (1) of Chebyshev's Theorem because $(\bar{x} - 2s, \bar{x} + 2s) = (675,775)$. It must be correct.

3. Statement (3) says the same thing as statement (2) because 75% of 251 is 188.25, so the minimum whole number of observations in this interval is 189. Thus statement (3) is definitely correct. (the actual percentage of observations in any of the intervals specified by the rule could be either greater or less than those given in the rule).

4. Statement (4) says the same thing as statement (2) but in different words, and therefore is definitely correct.

5. Statement (4), which is definitely correct, states that at most 25% of the time either fewer than 675 or more than 775 vehicles passed through the intersection. Statement (5) says that half of that 25% corresponds to days of light traffic. This would be correct if the relative frequency histogram of the data were known to be symmetric. But this is not stated; perhaps all of the observations outside the interval (675,775) are less than 75. Thus statement (5) might not be correct.

EXERCISES

BASIC

1. State the Empirical Rule.
2. Describe the conditions under which the Empirical Rule may be applied.
3. State Chebyshev's Theorem.
4. Describe the conditions under which Chebyshev's Theorem may be applied.
5. A sample data set with a bell-shaped distribution has mean $\overline{x} = 6$ and standard deviation $s = 2$. Find the approximate proportion of observations in the data set that lie:
   a. between 4 and 8;
   b. between 2 and 10;
   c. between 0 and 12.
6. A population data set with a bell-shaped distribution has mean $\mu = 6$ and standard deviation $\sigma = 2$. Find the approximate proportion of observations in the data set that lie:
   a. between 4 and 8;
   b. between 2 and 10;
   c. between 0 and 12.
7. A population data set with a bell-shaped distribution has mean $\mu = 2$ and standard deviation $\sigma = 1.1$. Find the approximate proportion of observations in the data set that lie:
   a. above 2;
   b. above 3.1;
   c. between 2 and 3.1.
8. A sample data set with a bell-shaped distribution has mean $\overline{x} = 2$ and standard deviation $s = 1.1$. Find the approximate proportion of observations in the data set that lie:
   a. below –0.2;
   b. below 3.1;
   c. between –1.3 and 0.9.
9. A population data set with a bell-shaped distribution and size $N = 500$ has mean $\mu = 2$ and standard deviation $\sigma = 1.1$. Find the approximate number of observations in the data set that lie:
   a. above 2;
   b. above 3.1;
   c. between 2 and 3.1.
10. A sample data set with a bell-shaped distribution and size $n = 128$ has mean $\overline{x} = 2$ and standard deviation $s = 1.1$. Find the approximate number of observations in the data set that lie:
    a. below –0.2;
    b. below 3.1;
    c. between –1.3 and 0.9.
11. A sample data set has mean $\overline{x} = 6$ and standard deviation $s = 2$. Find the minimum proportion of observations in the data set that must lie:
    a. between 2 and 10;
    b. between 0 and 12;
    c. between 4 and 8.
12. A population data set has mean $\mu = 2$ and standard deviation $\sigma = 1.1$. Find the minimum proportion of observations in the data set that must lie:
    a. between –0.2 and 4.2;
    b. between –1.3 and 5.3.
13. A population data set of size $N = 500$ has mean $\mu = 5.2$ and standard deviation $\sigma = 1.1$. Find the minimum number of observations in the data set that must lie:
    a. between 3 and 7.4;
    b. between 1.9 and 8.5.
14. A sample data set of size $n = 128$ has mean $\overline{x} = 2$ and standard deviation $s = 2$. Find the minimum number of observations in the data set that must lie:
    a. between –2 and 6 (including –2 and 6);
    b. between –4 and 8 (including –4 and 8).
15. A sample data set of size $n = 30$ has mean $\overline{x} = 6$ and standard deviation $s = 2$.
    a. What is the maximum proportion of observations in the data set that can lie outside the interval (2,10)?
    b. What can be said about the proportion of observations in the data set that are below 2?

c. What can be said about the proportion of observations in the data set that are above 10?
d. What can be said about the number of observations in the data set that are above 10?

16. A population data set has mean $\mu = 2$ and standard deviation $\sigma = 1.1$.

a. What is the maximum proportion of observations in the data set that can lie outside the interval $(-1.3, 5.3)$?
b. What can be said about the proportion of observations in the data set that are below –1.3?
c. What can be said about the proportion of observations in the data set that are above 5.3?

APPLICATIONS

17. Scores on a final exam taken by 1,200 students have a bell-shaped distribution with mean 72 and standard deviation 9.

a. What is the median score on the exam?
b. About how many students scored between 63 and 81?
c. About how many students scored between 72 and 90?
d. About how many students scored below 54?

18. Lengths of fish caught by a commercial fishing boat have a bell-shaped distribution with mean 23 inches and standard deviation 1.5 inches.

a. About what proportion of all fish caught are between 20 inches and 26 inches long?
b. About what proportion of all fish caught are between 20 inches and 23 inches long?
c. About how long is the longest fish caught (only a small fraction of a percent are longer)?

19. Hockey pucks used in professional hockey games must weigh between 5.5 and 6 ounces. If the weight of pucks manufactured by a particular process is bell-shaped, has mean 5.75 ounces and standard deviation 0.125 ounce, what proportion of the pucks will be usable in professional games?

20. Hockey pucks used in professional hockey games must weigh between 5.5 and 6 ounces. If the weight of pucks manufactured by a particular process is bell-shaped and has mean 5.75 ounces, how large can the standard deviation be if 99.7% of the pucks are to be usable in professional games?

21. Speeds of vehicles on a section of highway have a bell-shaped distribution with mean 60 mph and standard deviation 2.5 mph.

a. If the speed limit is 55 mph, about what proportion of vehicles are speeding?
b. What is the median speed for vehicles on this highway?
c. What is the percentile rank of the speed 65 mph?
d. What speed corresponds to the 16th percentile?

22. Suppose that, as in the previous exercise, speeds of vehicles on a section of highway have mean 60 mph and standard deviation 2.5 mph, but now the distribution of speeds is unknown.

a. If the speed limit is 55 mph, at least what proportion of vehicles must speeding?
b. What can be said about the proportion of vehicles going 65 mph or faster?

23. An instructor announces to the class that the scores on a recent exam had a bell-shaped distribution with mean 75 and standard deviation 5.

a. What is the median score?
b. Approximately what proportion of students in the class scored between 70 and 80?
c. Approximately what proportion of students in the class scored above 85?
d. What is the percentile rank of the score 85?

24. The GPAs of all currently registered students at a large university have a bell-shaped distribution with mean 2.7 and standard deviation 0.6. Students with a GPA below 1.5 are placed on academic probation. Approximately what percentage of currently registered students at the university are on academic probation?

25. Thirty-six students took an exam on which the average was 80 and the standard deviation was 6. A rumor says that five students had scores 61 or below. Can the rumor be true? Why or why not?

ANSWERS

1. See the displayed statement in the text.

3. See the displayed statement in the text.

5.
   a. 0.68.
   b. 0.95.
   c. 0.997.

7.
   a. 0.5.
   b. 0.16.
   c. 0.34.

9.
   a. 250.
   b. 80.
   c. 170.

11.
   a. 3/4.
   b. 8/9.
   c. 0.

13.
   a. 375.
   b. 445.

15.
   a. At most 0.25.
   b. At most 0.25.
   c. At most 0.25.
   d. At most 7.

17.
   a. 72.
   b. 816.
   c. 570.
   d. 30.

19. 0.95.

21.
   a. 0.975.
   b. 60.
   c. 97.5.
   d. 57.5.

23.
   a. 75.
   b. 0.68.
   c. 0.025.
   d. 0.975.

25. By Chebyshev's Theorem at most 1/9 of the scores can be below 62, so the rumor is impossible.

Basic Concepts of Probability

Probability is how likely something is to happen.

Many events can't be predicted with total certainty. The best we can say is how likely they are to happen, using the probability.

Will adopt the experiments of die throwing and coin flipping/tossing which consider most examples using to introduce the basic of probability.

Tossing a Coin

When a coin is tossed, there are two possible outcomes:

- heads (H) or
- tails (T)

We say that the probability of the coin landing **H** is ½

And the probability of the coin landing **T** is ½

Throwing Dice

When a single die is thrown, there are six possible outcomes: **1, 2, 3, 4, 5, 6**.

The probability of any one of them is $\frac{1}{6}$

In general:

Probability of an event happening = $\frac{\text{Number of ways it can happen}}{\text{Total number of outcomes}}$

Case (5) The chances of rolling a "4" with a die?

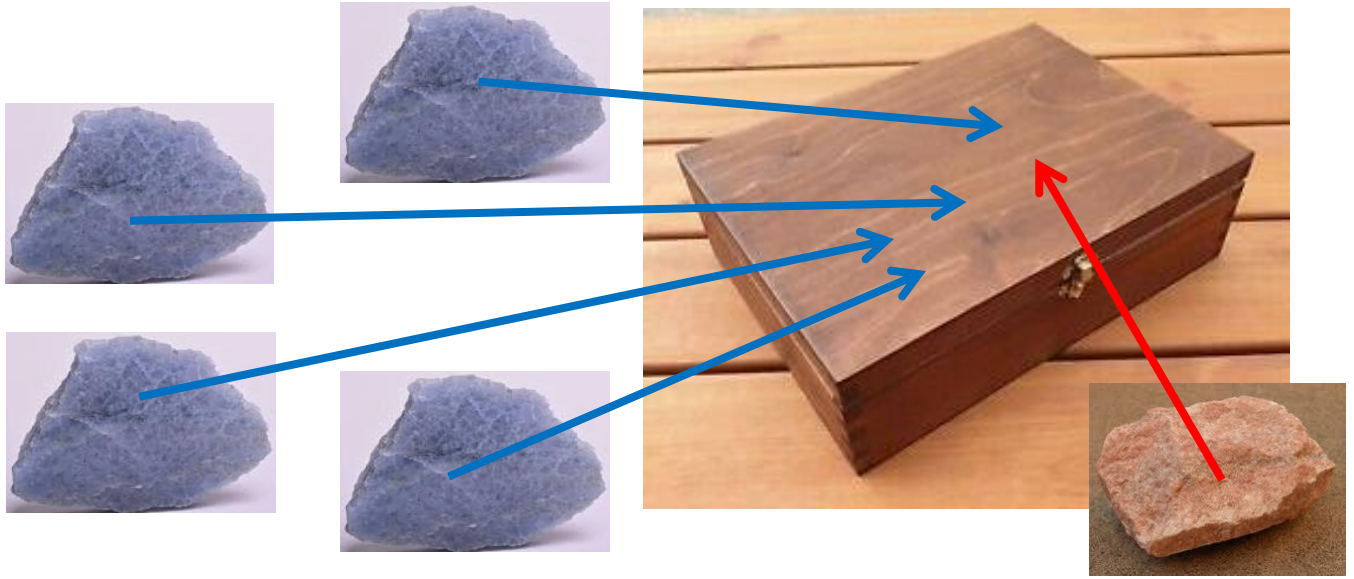
Solution:

Number of ways it can happen: 1 (there is only 1 face with a "4" on it)

Total number of outcomes: 6 (there are 6 faces altogether)

So the probability = $\frac{1}{6}$

Case (6) There are 5 samples of marble rocks in the lab casket: four are blue, and one is red. What is the probability that a blue marble gets picked?



Solution:

Number of ways it can happen: 4 (there are 4 blues)

Total number of outcomes: 5 (there are 5 marbles in total)

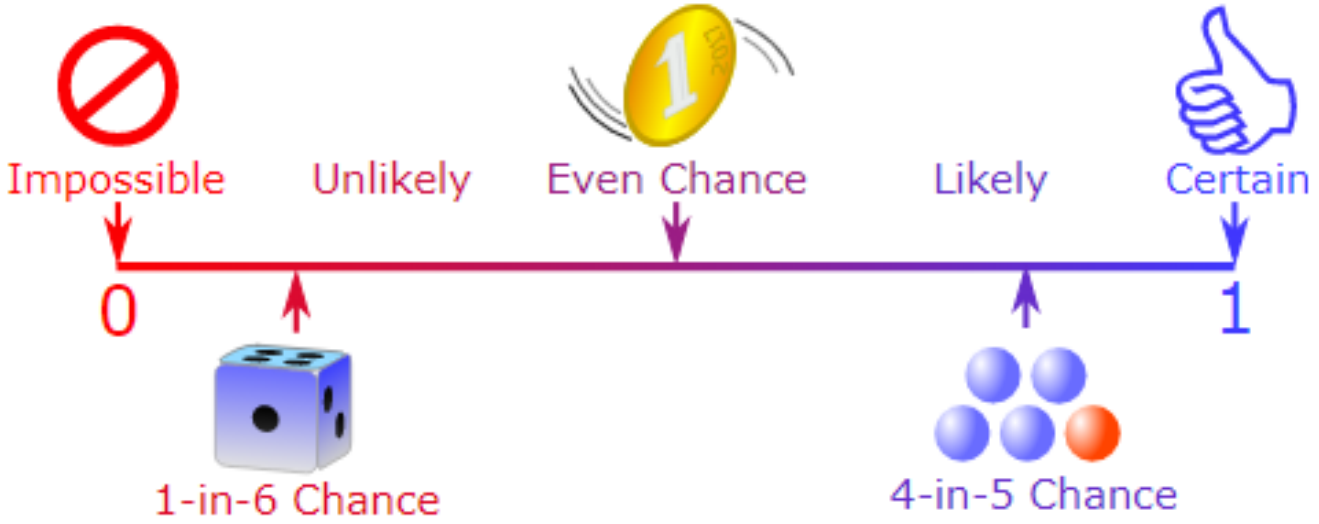
So the probability = $\frac{4}{5}$

Probability Line: Can be show the probability on a Probability Line:

- The probability of an event occurring is somewhere between impossible and certain.
- As well as words, we can use numbers to show the probability of something happening:

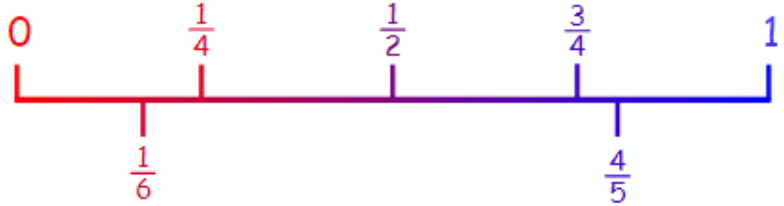
➢ Impossible is zero

➢ Certain is one

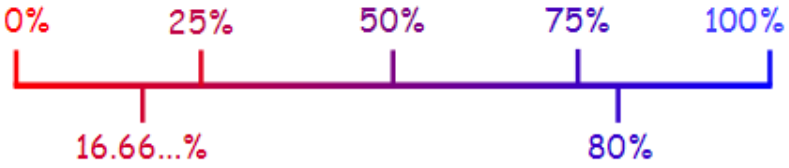


Probability Line

- We can use fraction



- We can use percent



- We can use decimals

Sample Spaces, Events, and Their Probabilities:

Sample Spaces and Events Meaning?

Rolling an ordinary six-sided die is a familiar example of a *random experiment*, an action for which all possible outcomes can be listed, but for which the actual outcome on any given trial of the experiment cannot be predicted with certainty. In such a situation we wish to assign to each outcome, such as rolling a two, a number, called the *probability* of the outcome, that indicates how likely it is that the outcome will occur. Similarly, we would like to assign a probability to any *event*, or collection of outcomes, such as rolling an even number, which indicates how likely it is that the event will occur if the experiment is performed.

Experiment: A repeatable procedure with a set of possible results.

Case (7) When throw the dice again and again, so what is the possible results from the repeatable throwing?.

Solution:

The set of possible results from any single throw is {1, 2, 3, 4, 5, 6}

Sample Space: All the possible outcomes of an experiment. Example: sample space is {1, 2, 3, 4, 5, 6}

Outcome: A possible result of an experiment. Example: getting “6”.

So, ***A random experiment*** is a mechanism that produces a definite outcome that cannot be predicted with certainty. **The sample space** associated with a random experiment is the set of all possible outcomes. **An event** is a subset of the sample space.

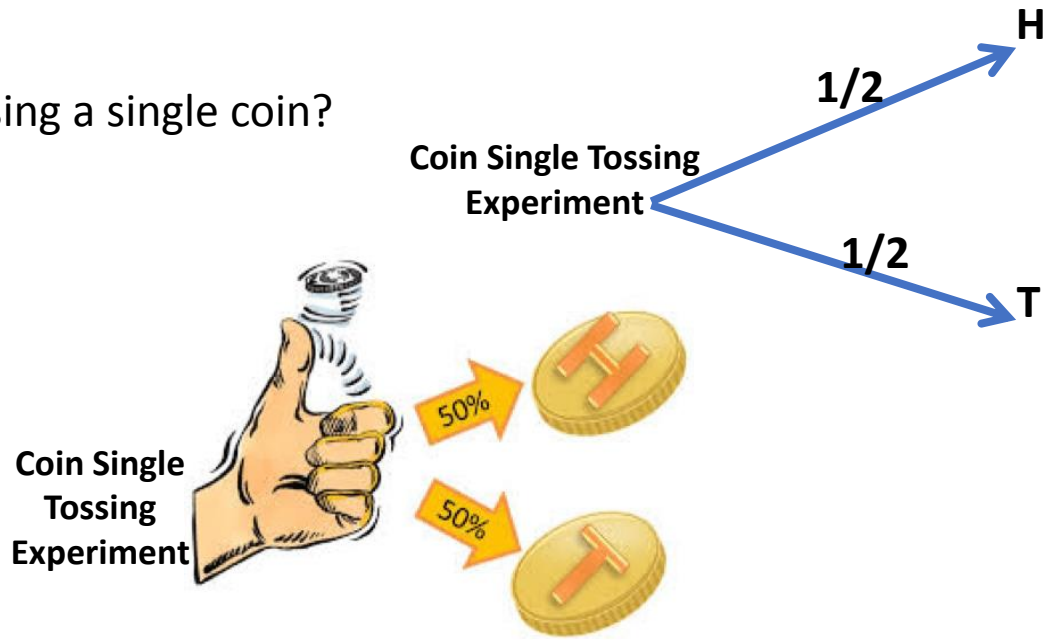
Event E is said to occur on a particular trial of the experiment if the outcome observed is an element of the set E. It is any set of outcomes.

Case (8) Construct a sample space for the experiment that consists of tossing a single coin?

Solution: The outcomes could be labeled *h* for heads and *t* for tails.

Then the sample space is the set $S = \{h, t\}$.

- If the coin single tossing is called balanced or fair, it means has equally likely to land up and the outcomes have the same probabilities, which must add up to 1, each outcome is assigned probability 1/2.



Case (9) Construct a sample space for the experiment that consists of rolling a single die. Find the events that correspond to the phrases “an even number is rolled” and “a number greater than two is rolled.”

Solution: The sample space is S = {1, 2, 3, 4, 5, 6}. The outcomes that are even are 2, 4, and 6, so the **even**t that corresponded to the even number after rolling is set {2,4,6}, which re-write $E = \{2,4,6\}$. Similarly the **event** that corresponded to the number greater than two after rolling is set $T = \{3,4,5,6\}$, which can be denoted by T to recognize it from the set of even numbers event.

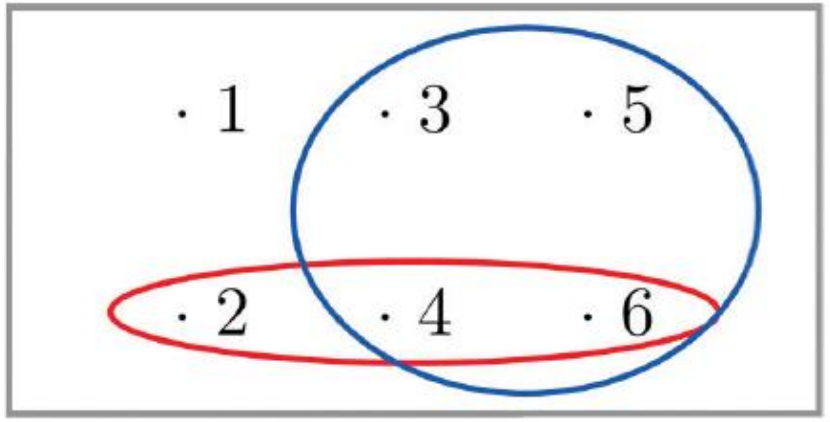
A graphical representation of a sample space and events is called **Venn diagram**, which applied on the cases 8 and 9 aforementioned and shown below:

In general the sample space *S* is represented by a rectangle, outcomes by points within the rectangle, and events by ovals that enclose the outcomes that compose them.

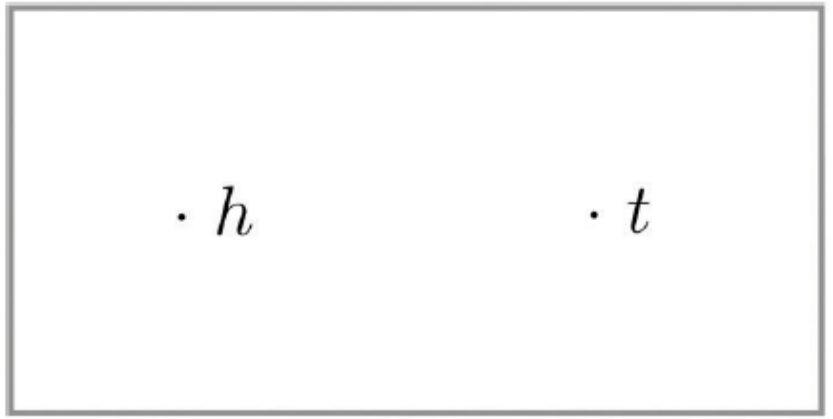
Tree Diagram a way that can be helpful in identifying all possible outcomes of a random experiment, particularly one that can be viewed as proceeding in stages. See the figure below and will be explained in details later in next case.

The line segments are called branches of the tree. The right ending point of each branch is called a node. The nodes on the extreme right are the final nodes; to each one there corresponds an outcome.

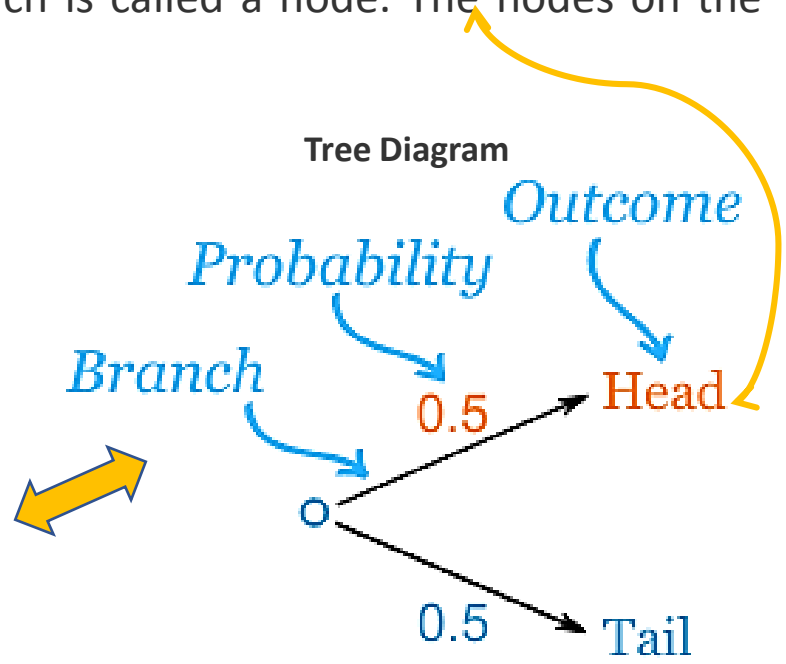
Venn Diagram for Sample Space of One Die Roll



Venn Diagram for Sample Space of One Coin Toss



Tree Diagram



Case (10) A random experiment consists of tossing two coins.

a. Construct a sample space for the situation that the coins are indistinguishable, such as two brand new pennies.

b. Construct a sample space for the situation that the coins are distinguishable, such as one a penny and the other a nickel.

Solution:

a. After the coins are tossed one sees either two heads, which could be labeled $2h$, two tails, which could be labeled $2t$, or coins that differ, which could be labeled d. Thus a sample space is $S = \{2h, 2t, d\}$.

b. Since we can tell the coins apart, there are now two ways for the coins to differ: the penny heads and the nickel tails, or the penny tails and the nickel heads. We can label each outcome as a pair of letters, the first of which indicates how the penny landed and the second of which indicates how the nickel landed. A sample space is then $S' = \{hh, ht, th, tt\}$.

Penny Coin

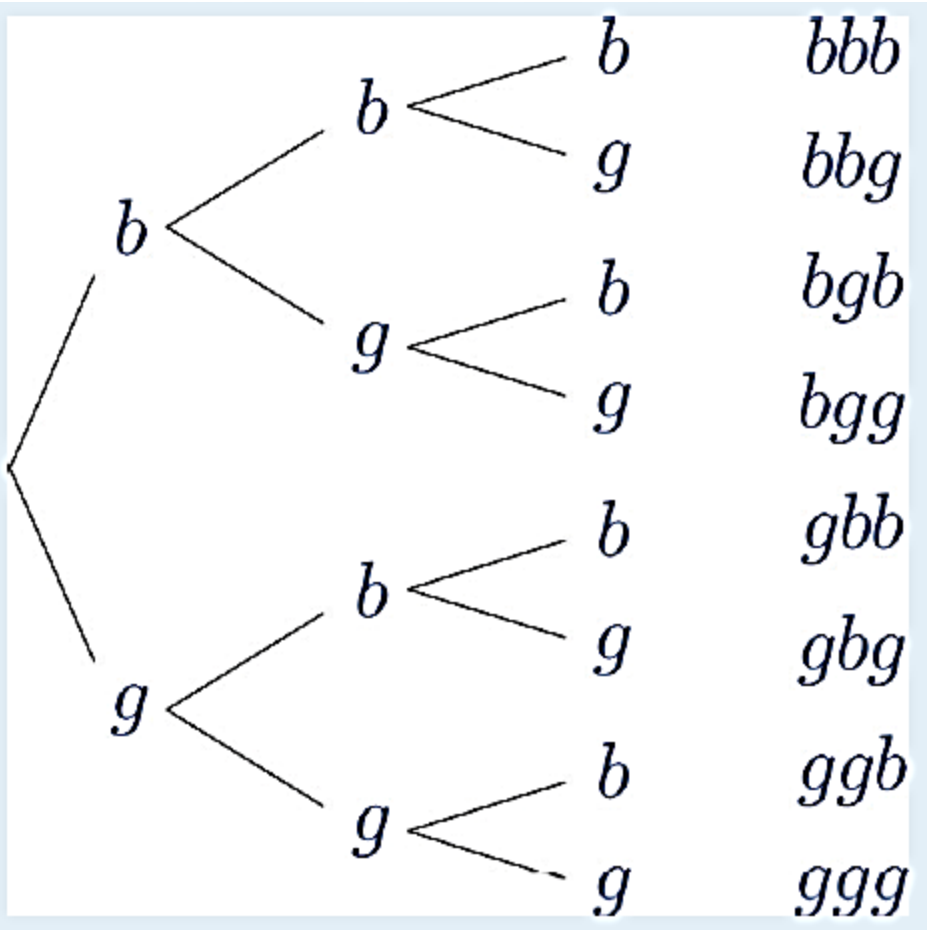
Observe or called Head **Reverse or called Tail**

Nickel Coin

Observe or called Head **Reverse or called Tail**

Case (11) Construct a sample space that describes all three-child families according to the genders of the children with respect to birth order.

Solution: Two of the outcomes are "two boys then a girl," which we might denote *bbg* , and "a girl then two boys," which we would denote *gbb*. Clearly there are many outcomes, and when we try to list all of them it could be difficult to be sure that we have found them all unless we proceed systematically. The tree diagram shown in the Figure below gives a systematic approach.



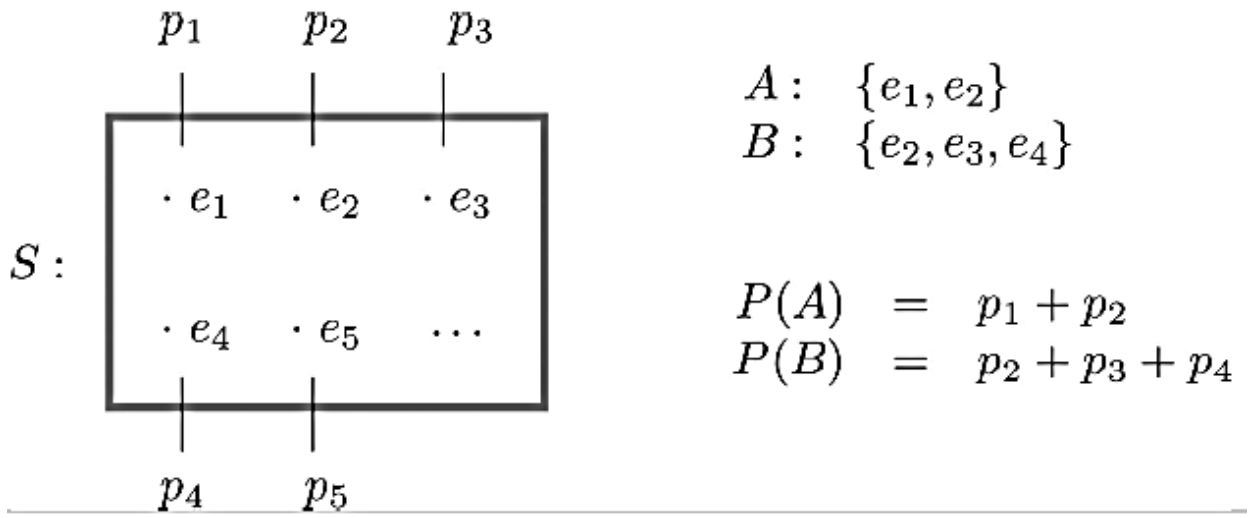
The diagram can be constructed based on there are two possibilities for the first child, boy or girl. So we draw two line segments coming out of a starting point, one ending in a *b* for "boy" and the other ending in a *g* for "girl." For each of these two possibilities for the first child there are two possibilities for the second child, "boy" or "girl," so from each of the *b* and *g* we draw two line segments, one segment ending in a *b* and one in a *g*. For each of the four ending points now in the diagram there are two possibilities for the third child, so we repeat the process more. From the tree it is easy to read off the eight outcomes of the experiment, so the sample space is, reading from the top to the bottom of the final nodes in the tree: $S = \{bbb, bbg, bgb, bgg, gbb, gbg, ggb, ggg\}$

So, based on previous discussion, we can summarize the following:

- ***The probability of an outcome*** e in a sample space S is a number p between 0 and 1 that measures the likelihood that e will occur on a single trial of the corresponding random experiment. The value p = 0 corresponds to the outcome e being impossible and the value p = 1 corresponds to the outcome e being certain. It is number that measures the likelihood of the outcome.
- ***The probability of an event*** *A is the sum of the probabilities of the individual outcomes of which it is composed. It is denoted* $P(A)$. It is number that measures the likelihood of the event. It has been defined by the following formula:

If an event E is $E = \{e_1, e_2, \dots, e_k\}$, then …….. $P(E) = P(e_1) + P(e_2) + \cdots + P(e_k)$

Sample Spaces and Probability can be illustrated graphically by the following figure:



$$A: \{e_1, e_2\}$$
$$B: \{e_2, e_3, e_4\}$$

$$P(A) = p_1 + p_2$$
$$P(B) = p_2 + p_3 + p_4$$

Since the whole sample space S is an event that is certain to occur, the sum of the probabilities of all the outcomes must be the number 1.

Probabilities are frequently expressed as percentages. For example, we say that there is a 70% chance of rain tomorrow, meaning that the probability of rain is 0.70. All computational formulas we should use 0.70 but expressed as 70%.

Case (12)

A die is called “balanced” or “fair” if each side is equally likely to land on top. Assign a probability to each outcome in the sample space for the experiment that consists of tossing a single fair die. Find the probabilities of the events *E*: “an even number is rolled” and *T*: “a number greater than two is rolled.”

Solution:

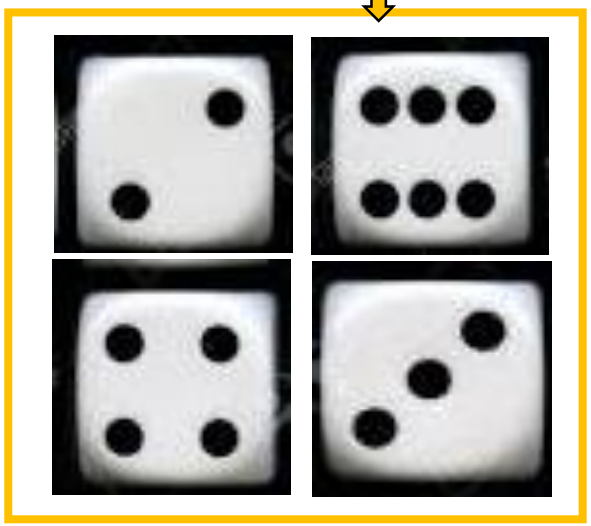
With outcomes labeled according to the number of dots on the top face of the die, the sample space is the set S = {1,2,3,4,5,6}.

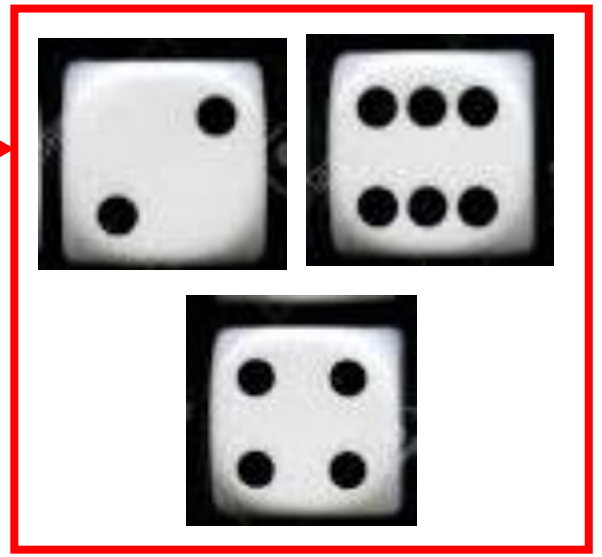
Since there are six equally likely outcomes, which must add up to 1, each is assigned probability 1/6.

Since *E* = {2,4,6},

P (*E*) = 1 / 6 + 1 / 6 + 1 / 6 = 3 / 6 = 1 / 2.

Since *T* = {3,4,5,6} , *P* (*T*) = 4 / 6 = 2 / 3.





All possible Results (All faces)

Case (13) Two fair coins are tossed. Find the probability that the coins match, i.e., either both land heads or both land tails.

Solution:

- Based on the solution of previous Case # (10), it was constructed the sample space S = {2h, 2t, d} for the situation in which the coins are identical and the sample space S′ = {hh, ht, th, tt} for the situation in which the two coins can be told apart.
- The theory of probability does not tell us *how* to assign probabilities to the outcomes, only what to do with them once they are assigned. Specifically, using sample space *S*, matching coins is the event *M* = {2*h*, 2*t*}, which has probability *P* (2*h*) + *P* (2*t*) . Using sample space *S′* , matching coins is the event *M′* = {*hh*, *tt*}, which has probability *P* (*hh*) + *P* (*tt*).
- In the physical world it should make no difference whether the coins are identical or not, and so we would like to assign probabilities to the outcomes so that the numbers *P* (*M*) and *P* (*M′*) are the same and best match what we observe when actual physical experiments are performed with coins that seem to be fair. Actual experience suggests that the outcomes in *S′* are equally likely, so we assign to each probability 1⁄4, and then:

$$P\,(M') = P\,(hh) + P\,(tt) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

- Similarly, from experience appropriate choices for the outcomes in *S* are:

$$P\,(2h) = \frac{1}{4} \qquad P\,(2t) = \frac{1}{4} \qquad P\,(d) = \frac{1}{2}$$ ⟹ which give the same final answer $P\,(M) = P\,(2h) + P\,(2\mathrm{t}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

Regarding to the last two cases it was computed the probabilities simply by counting when the sample space consists of a finite number of equally likely outcomes.

In some situations the individual outcomes of any sample space that represents the experiment are unavoidably unequally likely, in which case probabilities cannot be computed merely by counting, but the computational formula given in the definition of the probability of an event must be used.

Case (14) The breakdown of the student body in a local high school according to race and ethnicity is 51% white, 27% black, 11% Hispanic, 6% Asian, and 5% for all others. A student is randomly selected from this high school. (To select "randomly" means that every student has the same chance of being selected.) Find the probabilities of the following events:

a. *B*: the student is black, b. *M*: the student is minority (that is, not white), c. *N*: the student is not black.

Solution: The experiment is the action of randomly selecting a student from the student population of the high school. An obvious sample space is $S = \{w, b, h, a, o\}$. Since 51% of the students are white and all students have the same chance of being selected, $P(w) = 0.51$, and similarly for the other outcomes. This information is summarized in the following table:

Outcome	w	b	h	a	o
Probability	0.51	0.27	0.11	0.06	0.05

a. Since $B = \{b\}$, $P(B) = P(b) = 0.27$.

b. Since $M = \{b, h, a, o\}$,

$P(M) = P(b) + P(h) + P(a) + P(o) = 0.27 + 0.11 + 0.06 + 0.05 = 0.49$

c. Since $N = \{w, h, a, o\}$,

$P(N) = P(w) + P(h) + P(a) + P(o) = 0.51 + 0.11 + 0.06 + 0.05 = 0.73$

Case (15) The student body in the high school considered in the previous case # 14 may be broken down into ten categories as follows: 25% white male, 26% white female, 12% black male, 15% black female, 6% Hispanic male, 5% Hispanic female, 3% Asian male, 3% Asian female, 1% male of other minorities combined, and 4% female of other minorities combined. A student is randomly selected from this high school. Find the probabilities of the following events:

a. *B*: the student is black, b. *MF*: the student is minority female, c. *FN* : the student is female and is not black.

Solution:

The sample space is *S* = {*wm*, *bm*, *hm*, *am*, *om*, *wf* , *bf* , *hf* , *af* , *of* }

Based on the given information in this case can be summarized using following which called a *two-way contingency table*:

Gender	Race / Ethnicity				
	White	Black	Hispanic	Asian	Other
Male	0.25	0.12	0.06	0.03	0.01
Female	0.26	0.15	0.05	0.03	0.04

a. Since B = {bm, bf } ⇒ *P (B) = P {bm) + P (bf) = 0.12 + 0.15 = 0.27.*

b. Since MF = {wf, bf , hf , af , of } ⇒ *P (M) = P {wf) + P {bf) + P (hf) + P (af) + P (of) = 0.26 + 0.15 + 0.05 + 0.03 + 0.04 = 0.53.*

c. Since FN = {wf, hf , af , of } ⇒ *P (M) = P {wf) + P (hf) + P (af) + P (of) = 0.26 + 0.05 + 0.03 + 0.04 = 0.38.*

Keys:

- The sample space of a random experiment is the collection of all possible outcomes.
- An event associated with a random experiment is a subset of the sample space.
- The probability of any outcome is a number between 0 and 1. The probabilities of all the outcomes add up to 1.
- The probability of any event *A* is the sum of the probabilities of the outcomes in *A*.

EXERCISES

BASIC

1. A box contains 10 white and 10 black marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, two marbles in succession and noting the color each time. (To draw "with replacement" means that the first marble is put back before the second marble is drawn.)

2. A box contains 16 white and 16 black marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, three marbles in succession and noting the color each time. (To draw "with replacement" means that each marble is put back before the next marble is drawn.)

3. A box contains 8 red, 8 yellow, and 8 green marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, two marbles in succession and noting the color each time.

4. A box contains 6 red, 6 yellow, and 6 green marbles. Construct a sample space for the experiment of randomly drawing out, with replacement, three marbles in succession and noting the color each time.

5. In the situation of Exercise 1, list the outcomes that comprise each of the following events.

   a. At least one marble of each color is drawn.
   b. No white marble is drawn.

6. In the situation of Exercise 2, list the outcomes that comprise each of the following events.

   a. At least one marble of each color is drawn.
   b. No white marble is drawn.
   c. More black than white marbles are drawn.

7. In the situation of Exercise 3, list the outcomes that comprise each of the following events.

   a. No yellow marble is drawn.
   b. The two marbles drawn have the same color.
   c. At least one marble of each color is drawn.

8. In the situation of Exercise 4, list the outcomes that comprise each of the following events.

   a. No yellow marble is drawn.
   b. The three marbles drawn have the same color.
   c. At least one marble of each color is drawn.

9. Assuming that each outcome is equally likely, find the probability of each event in Exercise 5.

10. Assuming that each outcome is equally likely, find the probability of each event in Exercise 6.

11. Assuming that each outcome is equally likely, find the probability of each event in Exercise 7.

12. Assuming that each outcome is equally likely, find the probability of each event in Exercise 8.

13. A sample space is $S = \{a, b, c, d, e\}$. Identify two events as $U = \{a, b, d\}$ and $V = \{b, c, d\}$. Suppose $P(a)$ and $P(b)$ are each 0.2 and $P(c)$ and $P(d)$ are each 0.1.

    a. Determine what $P(e)$ must be.
    b. Find $P(U)$.
    c. Find $P(V)$.

14. A sample space is $S = \{u, v, w, x\}$. Identify two events as $A = \{v, w\}$ and $B = \{u, w, x\}$. Suppose $P(u) = 0.22, P(w) = 0.36$, and $P(x) = 0.27$.

    a. Determine what $P(v)$ must be.
    b. Find $P(A)$.
    c. Find $P(B)$.

15. A sample space is $S = \{m, n, q, r, s\}$. Identify two events as $U = \{m, q, s\}$ and $V = \{n, q, r\}$. The probabilities of some of the outcomes are given by the following table:

Outcome	m	n	q	r	s
Probablity	0.18	0.16		0.24	0.21

    a. Determine what $P(q)$ must be.
    b. Find $P(U)$.
    c. Find $P(V)$.

16. A sample space is $S = \{d, e, f, g, h\}$. Identify two events as $M = \{e, f, g, h\}$ and $N = \{d, g\}$. The probabilities of some of the outcomes are given by the following table:

Outcome	d	e	f	g	h
Probablity	0.22	0.13	0.27		0.19

a. Determine what $P(g)$ must be.
b. Find $P(M)$.
c. Find $P(N)$.

APPLICATIONS

17. The sample space that describes all three-child families according to the genders of the children with respect to birth order was constructed in Note 3.9 "Example 4". Identify the outcomes that comprise each of the following events in the experiment of selecting a three-child family at random.

    a. At least one child is a girl.
    b. At most one child is a girl.
    c. All of the children are girls.
    d. Exactly two of the children are girls.
    e. The first born is a girl.

18. The sample space that describes three tosses of a coin is the same as the one constructed in Note 3.9 "Example 4" with "boy" replaced by "heads" and "girl" replaced by "tails." Identify the outcomes that comprise each of the following events in the experiment of tossing a coin three times.

    a. The coin lands heads more often than tails.
    b. The coin lands heads the same number of times as it lands tails.
    c. The coin lands heads at least twice.
    d. The coin lands heads on the last toss.

19. Assuming that the outcomes are equally likely, find the probability of each event in Exercise 17.

20. Assuming that the outcomes are equally likely, find the probability of each event in Exercise 18.

ADDITIONAL EXERCISES

21. The following two-way contingency table gives the breakdown of the population in a particular locale according to age and tobacco usage:

Age	Tobacco Use	
	Smoker	Non-smoker
Under 30	0.05	0.20
Over 30	0.20	0.55

A person is selected at random. Find the probability of each of the following events.

a. The person is a smoker.
b. The person is under 30.
c. The person is a smoker who is under 30.

22. The following two-way contingency table gives the breakdown of the population in a particular locale according to party affiliation (*A*, *B*, *C*, or *None*) and opinion on a bond issue:

Affiliation	Opinion		
	Favors	Opposes	Undecided
A	0.12	0.09	0.07
B	0.16	0.12	0.14
C	0.04	0.03	0.06
None	0.08	0.06	0.03

A person is selected at random. Find the probability of each of the following events.

a. The person is affiliated with party *B*.
b. The person is affiliated with some party.
c. The person is in favor of the bond issue.
d. The person has no party affiliation and is undecided about the bond issue.

23. The following two-way contingency table gives the breakdown of the population of married or previously married women beyond child-bearing age in a particular locale according to age at first marriage and number of children:

Age	Number of Children		
	0	1 or 2	3 or More
Under 20	0.02	0.14	0.08
20–29	0.07	0.37	0.11

Age	Number of Children		
	0	1 or 2	3 or More
30 and above	0.10	0.10	0.01

A woman is selected at random. Find the probability of each of the following events.

a. The woman was in her twenties at her first marriage.
b. The woman was 20 or older at her first marriage.
c. The woman had no children.
d. The woman was in her twenties at her first marriage and had at least three children.

24. The following two-way contingency table gives the breakdown of the population of adults in a particular locale according to highest level of education and whether or not the individual regularly takes dietary supplements:

Education	Use of Supplements	
	Takes	Does Not Take
No High School Diploma	0.04	0.06
High School Diploma	0.06	0.44
Undergraduate Degree	0.09	0.28
Graduate Degree	0.01	0.02

An adult is selected at random. Find the probability of each of the following events.

a. The person has a high school diploma and takes dietary supplements regularly.
b. The person has an undergraduate degree and takes dietary supplements regularly.
c. The person takes dietary supplements regularly.
d. The person does not take dietary supplements regularly.

ANSWERS

1. $S = \{bb, bw, wb, ww\}$

3. $S = \{rr, ry, rg, yr, yy, yg, gr, gy, gg\}$

5. a. $\{bw, wb\}$
   b. $\{bb\}$

7. a. $\{rr, rg, gr, gg\}$
   b. $\{rr, yy, gg\}$
   c. $\emptyset$

9. a. 2/4
   b. 1/4

11. a. 4/9
    b. 3/9
    c. 0

13. a. 0.4
    b. 0.5
    c. 0.4

15. a. 0.21
    b. 0.6
    c. 0.61

17. a. $\{bbg, bgb, bgg, gbb, gbg, ggb, ggg\}$
    b. $\{bbb, bbg, bgb, gbb\}$
    c. $\{ggg\}$
    d. $\{bgg, gbg, ggb\}$
    e. $\{gbb, gbg, ggb, ggg\}$

19. a. 7/8
    b. 4/8
    c. 1/8
    d. 3/8
    e. 4/8

21. a. 0.25
    b. 0.25
    c. 0.05

23. a. 0.55
    b. 0.76
    c. 0.19
    d. 0.11

25. The relative frequencies for 1 through 6 are 0.16, 0.194, 0.162, 0.164, 0.154 and 0.166. It would appear that the die is not balanced.