Section 6.1: Counting Techniques

Although the idea of counting the number of objects seems straightforward, there are a few tricky bits to it when you are looking at large quantities. Let's start by looking at an example.

Example

How many three-digit natural numbers are there?

Answer: Let's look at the list: 100, 101, 102, … 999. There are three methods we can use to determine the total number here.

Method #1: We could, for example, write the list as a sequence. We see that it's an arithmetic sequence with $d = 1$. Then

$$\begin{aligned} a_n &= a_1 + (n-1)d \\ 999 &= 100 + (n-1)1 \\ 899 &= n-1 \\ n &= 900 \end{aligned}$$

so there are 900 three-digit numbers in total.

Method #2: Since the numbers go up one-by-one, can use the nice summation notation trick of (last – first + 1) = (999 – 100 + 1) = 900 as well.

Method #3: Consider the digits ___ ___ ___ . The first digit can be 1-9 for 9 choices, the second and third can be 0-9 for 10 choices. Then you get 9×10×10 = 900 numbers.

Example

How many three-digit natural numbers are divisible by 5?

Answer: The numbers are 100, 105, 110, … 995

Method #1 works very nicely:

$$\begin{aligned} a_n &= a_1 + (n-1)d \\ 995 &= 100 + (n-1)5 \\ 895 &= (n-1)5 \\ 179 &= n-1 \\ n &= 180 \end{aligned}$$

Method #2: we can rewrite the sequence as 20×5, 21×5, 22×5, …199×5. Then we use (last – first + 1) = (199 – 20 + 1) = 180. (We can only use this formula when we are counting by steps of one. So we have to force our sequence into a counting-by-one step to use it.)

Method #3: We note that if the number is divisible by 5, then the last digit is either 0 or 5, for two choices. We then get 9×10×2 = 180. However, be careful with this method! It will work well for numbers divisible by 1, 2, 5, and 10, because this eliminates digits in the last column. You can't use this method at all with most other divisors like 3, 4, 6, etc.

The reason Method #3 works is called the Multiplication Principle of Counting. If a question consists of a series of choices in which there are p possibilities for the first choice, q possibilities for the second choice, r for the third, etc., then the number of ways in which the question can be done is just $p \times q \times r \times \ldots$ In other words, you just multiply together the number of ways each step can be done.

Example

How many BC licence plates for cars are there (barring reserved words, etc.)?

Answer: The patterns allowable are

letter-letter-letter number-number-number

number-number-number letter-letter-letter

(we'll ignore personalized plates or reserved words, etc.) Looking at the first pattern, there are 26 choices for each letter and 10 for each number. So we've got 26×26×26×10×10×10 = 17,576,000 plates.

The second pattern will have the same number, for a total of 35,152,000 plates using the two patterns.

Example

How many (250) area code phone numbers are there?

Answer: Phone numbers are of the form (250) ### – ####. To look at all possibilities, there are 10 digits for each #, so 10^7 or ten million possibilities.

Example

How many (250) area code phone numbers are there that don't start with zero?

Answer: Now we have only 9 choices for the first #, so 9×10^6 or nine million.

Example

How many (250) area code phone numbers are there that don't start with 911?

Answer: This one's more tricky. What we'll do is take the total number of phone numbers and subtract the number that **do** begin with 911. Numbers beginning with 911 will look like: (250) 911–####, so we'll have 10^4 choices.

Since there are 10,000,000 numbers in total, the number that don't start with 911 is (10,000,000 – 1000) = 9,999,000.

There's another rule that we should know regarding counting if we're using the word "or".

Example

How many numbers from 1 to 30 are a) divisible by 3? b) divisible by 5? c) divisible by 3 or 5?

Answer: Well, let's try the brute force method.

Divisible by 3: 3, 6, 9, 12, 15, 18, 21, 24, 27, 30 total: 10

Divisible by 5: 5, 10, 15, 20, 25, 30 total: 6

Divisible by 3 or 5: number in first row (10) + number in second row (6) but we have to subtract 2, because otherwise we are counting 15 and 30 twice! So we get 10 + 6 – 2 = 14.

This brings up the addition rule:

$$n(A \text{ or } B) = n(A) + n(B) - n(AB)$$

Now, if the two situations don't have overlap, then n(AB) can be zero. We say then that the two situations are **mutually exclusive**.

Example

How many case-sensitive alpha-numeric passwords are there that have 6 or 7 characters?

Answer: First, look at what "case-sensitive, alpha-numeric" means. It means that capital letters are considered different from lowercase, so 52 letters instead of 26. Also, numbers are allowed, so 62 choices in total.

Number of 6-char passwords: $62\times62\times62\times62\times62\times62 = 62^6 = 5.68\times10^{10}$

Number of 7-char passwords: $62^7 = 3.52\times10^{12}$

Now, a password can either have 6 characters or 7 but not both, so to get the total, we just add $5.68\times10^{10} + 3.52\times10^{12} = 3.58\times10^{12}$.

Example

How many case-sensitive alpha-numeric passwords are there that have 6 characters and at least one number and letter?

Answer: This question is quite difficult as written. However, if we calculate instead the total number of passwords and subtract the number that don't have any numbers and the number that don't have any letters, we'll get the same result.

Number of 6-char passwords without any numbers: 52^6

Number of 6-char passwords without any letters: 10^6

So we get total = $62^6 - 52^6 - 10^6 = 3.70\times10^{10}$

Section 6.1: Counting Techniques

Exercises

1. How many 2-digit numbers are

   a) even?
   b) divisible by 7?
   c) not divisible by 7?

2. How many 4-digit numbers are

   a) divisible by 3?
   b) divisible by 5?
   c) divisible by 3 and 5?
   d) divisible by 3 or 5?
   e) divisible by neither 3 nor 5?

3. A computer system requires a case-sensitive, alpha-numeric password containing 4 or 5 characters. How many possible passwords are there?

4. A computer system requires a case-sensitive, alpha-numeric password containing 5 digits. How many possible passwords are there if

   a) you can repeat characters?
   b) you cannot repeat characters?
   c) you can repeat characters but the first character must be a letter and not a digit?

5. A computer system requires an eight-character, case-sensitive, alpha-numeric passwords.

   a) How many possible passwords are there?
   b) How many passwords are there that contain at least one digit?
   c) How many passwords are there that contain at least one letter?
   d) How many passwords are there that contain at least one digit and one letter?

6. A computer system requires a case-sensitive, alpha-numeric password containing six characters.

   a) How many passwords are there that contain no “A”s?
   b) How many passwords are there that contain no “a”s?
   c) How many passwords are there that contain no “A”s or “a”s?

7. For homework, Peter has assigned reading pages 25-37 inclusive. How many pages has he asked his class to read?

8. Gilles has assigned for homework all the odd questions from 7 to 89. How many homework questions has he assigned?

9. Canadian postal codes are of the form "letter-number-letter number-letter-number". The first letter shows which province or territory is from, for a total of 13 letters allowable. The remaining letters can be any letter of the alphabet except for O and I. All numbers are allowed. How many possible Canadian postal codes are there?

10. How many days of the week

    a) contain the letter "t"?
    b) contain the letter "s"?
    c) contain the letters "t" and "s"?
    d) contain the letters "t" or "s"?

11. Pat is writing up systems of equations containing two variables. She will be using lower-case letters for her variables, but doesn't want to use the letters "e", "i", and "o" (for obvious reasons!). How many possible letter combinations does she have to choose from?

12. The mythical Canadian province of Gondor has licence plates of the form "letter-letter number-number-number". Because of an odd superstition, you cannot repeat a letter on the licence plate, but you can repeat a number. How many possible Gondorian licence plates are there?

Section 6.1: Counting Techniques

Solutions

1. First, note that 2-digit numbers run from 10, 11, 12, … 99.

   a) The even ones are 10, 12, 14, … 98. You can do the really short method to count them: ___ ___ – the first slot can have the digits 1-9 for 9 choices, and the second can only have 2, 4, 6, 8, or 0 for 5 choices. Then the total number is 9×5 = 45 numbers.

   b) Unfortunately, you cannot use the above technique for dividing by 7, since 7 doesn't restrict the last digit. Instead, you have to note that the first 2-digit number that's divisible by 7 is 14, the next is 21, then 28, and so on. To find the last digit, you have to count backwards from 99 to find one that's divisible by 7. 99 does not divide evenly by 7, but with your calculator (sigh) you can quickly find that 98÷7 = 14.

      So our sequence is 14, 21, 28, … 98. This is just 2×7, 3×7, 4×7, … 14×7. So there are (last – first + 1) = 14 – 2 + 1 = 13 numbers divisible by 7.

   c) Total number of 2-digit numbers: last – first + 1 = 90 – 10 + 1 = 90. So the total number of 2-digit numbers **not** divisible by 7 is the total number minus the number that **are** divisible by 7. So, we get 90 – 13 = 77 for our answer.

2. First, note that 4-digit numbers run from 1000, 11, 12, … 9999.

   a) The first number that's divisible by 3 is 1002, the next is 1005, then 1008, and so on up to 9999, which also divides evenly by 3.

      So our sequence is 1002, 1005, 1008, … 9999. This is just 334×3, 335×3, 336×3, … 3333×3. So there are (last – first + 1) = 3333 – 334 + 1 = 3000 numbers divisible by 3.

   b) Numbers that divide evenly by 5 end in either 0 or 5. You can do the really short method to count them: ___ ___ ___ ___ – the first slot can have the digits 1-9 for 9 choices, the second and third slots can have 0-9 for 10 choices and the second can only have 0 or 5 for 2 choices. Then the total number is 9×10×10×2 = 1800 numbers.

   c) Numbers divisible by 3 **and** 5 must be divisible by 15. Looking at our sequence in a), we can see that 1005 must be the first number, then add 15 to get 1020, etc. Starting from 9999 and working downwards, we'll see that the first possibility is 9995, which doesn't divide, but 9990 does.

So our sequence is 1005, 1020, 1035, … 9990. This is just 67×15, 68×15, 69×15, … 666. So the total number is 666 – 67 + 1 = 600.

d) n(3 or 5) = n(3) + n(5) – n(3 and 5)
= 3000 + 1800 – 600
= 4200

e) total number with 4-digits: 9999-1000+1 = 9000 numbers
Then the total divisible by neither 3 nor 5 is 9000 – 4200 = 4800.

3. Case-sensitive, alpha-numeric passwords have 2×26 + 10 choices for characters, or 62 different possibilities. The number of passwords containing 4 characters is 62×62×62×62 = 14,776,336. The number of passwords containing 5 characters is 62×62×62×62×62 = 916,132,832. The total number of passwords is then the sum of these two (since you can't have four **and** five at the same time), = 930,909,168.

4. a) This is the same as in question #3: 916,132,832.

   b) If you can't repeat, then you get 62 choices for the first one, 61 for the second, etc., to give 62×61×60×59×58 = 776,520,240.

   c) If the first number must be a letter, then you only have 52 possibilities for the first slot: 52×62×62×62×62 = 768,369,472.

5. a) You have 62 choices for each slot, so result is $62^8 = 2.18\times10^{14}$.

   b) The number containing **no** digits is $52^8 = 5.35\times10^{13}$. So the number containing at least one digit is $2.18\times10^{14} – 5.35\times10^{13} = 1.65\times10^{14}$.

   c) The number containing **no** letters is 10^8. So the number containing at least one letter is $2.18\times10^{14} – 10^8 = 2.18\times10^{14}$ (essentially the same number, since 10^8 is so much smaller).

   d) The number containing at least one digit and one letter must be the total minus (the number containing no digits plus the number containing no letters). So we get $2.18\times10^{14} – 5.35\times10^{13} – 10^8 = 1.65\times10^{14}$ (very close to the answer to b – you'd have to write out a few more decimals to see the difference).

6. a) If no "A"s are allowed, then we are constrained to 61 choices from our original 62. Then we'll get $61^6 = 5.15\times10^{10}$ passwords.

   b) This will again give us 61 choices for each character, or $61^6 = 5.15\times10^{10}$ passwords.

   c) Now, we're down to 60 choices, since we can't have "A" or "a". We then get $60^6 = 4.67\times10^{10}$ passwords.

7. Peter has assigned 25-37 pages, so 25, 26, 27, … 37. # pages = last – first + 1 = 37 – 25 + 1 = 13 pages.

8. Gilles has assigned odd questions, so 7, 9, 11, … 89. It's a bit tricky to do the odd numbers, so I'm going to take all numbers from 7 to 89 and subtract the even numbers.

   total number from 7 to 89: $89 - 7 + 1 = 83$

   even numbers: 8, 10, 12, … 88 is the same as 4×2, 5×2, 6×2, … 44×2. So we get 44–4+1 = 41 even numbers

   odd numbers = 83 – 41 = 42 odd numbered questions

9. first letter: 13 choices, second and third letters: 24 choices, all numbers: 10 choices

   So we get 13 10 24 10 24 10 = 13×10×24×10×24×10 = 7,488,000 possible postal codes. (Note that since postal codes reference a geographical area and not a group of people, we're not likely to run out any time soon!)

10. a) Counting on my fingers, I get that Tuesday, Thursday, and Saturday contain the letter "t" for a total of 3.

    b) Counting on my fingers, I get that all days except for Monday and Friday have "s" in them for a total of 5.

    c) I see that all of the days containing "t" also contain "s" for a total of 3.

    d) Using my counting rules, I get that n(t or s) = n(t) + n(s) – n(t & s) = 3 + 5 – 3 = 5.

11. I will not be using 3 letters, leaving 23 lower-case letters to choose from. But they have to be different, so I'll get 23×22 choices = 506 choices.

12. 26 25 10 10 10 = $26 \times 25 \times 10^3 = 650{,}000$.

Section 6.2: Combinations and Permutations

Factorials

If you've ever seen 10!, a number with an exclamation point behind it, it means that we take the following product:

$$10! = 10\times 9\times 8\times 7\times \ \ldots\ 1\ .$$

More generally, $n!$ is called "n-factorial" and is the product of the number n and all of the positive integers less than or equal to n:

$$n! = n\times(n-1)\times(n-2)\times\ldots 1\ .$$

Example

Calculate $\dfrac{12!}{9!}$.

Answer: You probably have a factorial button on your calculator. But if you don't, you can shorten the calculation considerably by noting that

$$\begin{aligned}\frac{12!}{9!} &= \frac{12\times 11\times 10\times 9\times 8\times 7\times\ldots 1}{9\times 8\times 7\times\ldots 1}\\ &= \frac{12\times 11\times 10}{1}\\ &= 1320\end{aligned}$$

Permutations

A permutation is an **ordered** group of objects chosen **without repetition** from a set of possibilities. If the number of objects we are choosing is r and the number of possibilities that we can choose from is n (with $n \geq r$), then this permutation is has the symbol ${}_nP_r$, where it can be calculated by

$${}_nP_r = \frac{n!}{(n-r)!}$$

(Other commonly used symbols are P_r^n and $P(n,r)$.)

To see how this works (and to justify this strange equation), we should look at an example.

Example

How many four-digit PIN number for a bank card could you have if you are not allowed to repeat digits?

Answer: There are two ways to calculate this: you could say that there are 10 possibilities for the first digit, 9 for the second, 8 for the third, and 7 for the fourth (since you can't repeat numbers). Then your answer would be that the total number allowed would be 10×9×8×7, which equals 5040.

Alternatively, you could note that

$$10\times 9\times 8\times 7 = \frac{10\times 9\times 8\times 7\times 6\times 5\times ...1}{6\times 5\times 4\times ...1}$$

so that you could say that $10\times 9\times 8\times 7 = \frac{10!}{6!} = \frac{10!}{(10-4)!}$, where 10 is the number of digits you are selecting from (n)and 4 is the number you are selecting (r).

So, ${}_{10}P_4 = \frac{10!}{(10-4)!}$, and more generally ${}_nP_r = \frac{n!}{(n-r)!}$.

Combinations

A combination is an **unordered** group of objects chosen **without repetition** from a set of possibilities. If the number of objects we are choosing is r and the number of possibilities that we can choose from is n (with $n \geq r$), then this combination is has the symbol ${}_nC_r$, where it can be calculated by

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

(Other commonly used symbols are $C_r^{\ n}$, $C(n,r)$, and $\binom{n}{r}$.)

In other words, ${}_nC_r$ can be calculated by taking the number of ordered arrangements ${}_nP_r$ and dividing by the number of ways you can arrange those r objects (which turns out to be $r!$).

Example

If you went to the library and found that there were seven books on the subject of bathtub races, and you decided to check out three of them, how many different selections could you possibly make?

Answer: You are choosing 3 from 7, so $r = 3$ and $n = 7$. Since the order that you check them out doesn't matter, we want ${}_nC_r$, so:

$$ {}_nC_r = \frac{n!}{r!(n-r)!} $$

$$ {}_7C_3 = \frac{7!}{3!4!} = 35 $$

An important difference between permutations and combinations is whether order matters. For example, order matters when listing the characters in computer passwords, student numbers, and words in general. Order does not matter when you are looking at a committee of people, the cards in a poker hand, or the winning numbers for the Lotto 6/49 game.

Example

Consider the set of letters A, B, C, D, and E. If you were to list all possible selections of two letters drawn from this set without repetition, how many choices would you have if a) order matters and b) order doesn't matter?

Answer:

Method #1: Well, there is a very small list here, so we could do this by brute force and list all of the possible choices:

	AB	AC	AD	AE
BA		BC	BD	BE
CA	CB		CD	CE
DA	DB	DC		DE
EA	EB	EC	ED	

Notice that I've left the diagonal blank because you can't repeat letters and so AA would not be allowed. Then if you count the number of choices, there are 20 of them. So the answer to the first part is that there are 20 possibilities.

To answer part b), you have to notice that AB and BA are considered to be the same when order doesn't matter. So all of the entries below the diagonal are repetitions of the ones above. We can then cross out all the ones below, leaving 10 possibilities when order doesn't matter.

Method #2: To do this using ${}_nP_r$ and ${}_nC_r$, we notice that part a) will be a permutation and b) will be a combination. For both of these, $n = 5$ (letters to choose from) and $r = 2$ (number of letters we are choosing). Then we get

a) ${}_nP_r = \dfrac{n!}{(n-r)!}$ so ${}_5P_2 = \dfrac{5!}{3!} = 20$

b) ${}_nC_r = \dfrac{n!}{r!(n-r)!}$ so ${}_5C_2 = \dfrac{5!}{2!3!} = 10$

Example

Pat has a test bank of 20 multiple-choice questions and 15 word problems. She wishes to create a quiz of ten multiple-choice questions and two word problems. How many different quizzes could she potentially make?

Answer: In general, a quiz is considered the same if it contains the same questions as another, so order does not matter. We should treat the two parts, multiple-choice and word-problem questions, separately:

combinations for multiple-choice $= {}_{20}C_{10} = \dfrac{n!}{r!(n-r)!} = \dfrac{20!}{10!10!} = 184,756$

combinations for word problems $= {}_{15}C_2 = \dfrac{n!}{r!(n-r)!} = \dfrac{15!}{2!13!} = 105$

so total # quizzes = (multiple-choice combos)×(word problem combos)

$= 184,756 \times 105 = 19,399,380$

Example

Students are given a list of twelve computer games and asked to pick their three favourites. How many different lists could there be if

a) students just list their favourites in any order

b) students rank their favourite as #1, their second favourite as #2, etc.?

Answer:

In both of these questions, $r = 3$ and $n = 12$.

a) ${}_nC_r = \dfrac{n!}{r!(n-r)!}$ so ${}_{12}C_3 = \dfrac{12!}{3!9!} = 220$

b) ${}_nP_r = \dfrac{n!}{(n-r)!}$ so ${}_{12}P_3 = \dfrac{12!}{9!} = 1320$

Section 6.2: Combinations and Permutations

Exercises

Calculate the following quantities. (This is just to ensure that you know how to use the functions on your calculator correctly.)

1. $7!$

2. $8!$

3. $\dfrac{11!}{4!}$

4. $\dfrac{10!}{8!}$

5. $\dfrac{12!}{8!4!}$

6. $\dfrac{20!}{5!15!}$

7. ${}_{25}P_4$

8. ${}_4P_3$ and ${}_4P_1$

9. ${}_{25}C_4$

10. ${}_5C_3$ and ${}_5C_2$

Calculate the requested permutations.

11. In how many different ways can five boxes be stacked?

12. How many ways are there to seat eight people in a row?

13. A "combination" lock is unlocked by first rotating the dial clockwise to the first number, then rotating the dial counterclockwise to the second number, then clockwise to the third. If there are 50 numbers on the dial and you're not allowed to repeat a number, how many possible sequences can be formed to open the lock? (Bonus question: is it really a "combination" lock or is it a "permutation" lock?)

14. How many 4-digit PIN numbers are possible if

    a) digits cannot be repeated?

b) digits can be repeated?

Calculate the requested combinations.

15. For a poker game, each player is dealt five cards from the standard deck of 52 cards. How many different poker hands are possible?

16. For the Lotto 6/49, a player chooses 6 numbers from 1 through 49 (without repetition) and circles them on the card. The winning ticket will then have the same six numbers as the numbers drawn randomly from a barrel at Lotto 6/49 headquarters. How many different possible number selections could the player make?

17. There are twenty-five members of the Math Department at Camosun, and nine members of the Physics Department. If a committee is to be formed from three math faculty and two physics faculty, how many possible committees are there?

Calculate the following quantities.

18. How many different ways are there to rearrange the four letters in the word "math" (the arrangements don't have to form an actual word)?

19. In how many different ways can 3 people each have a different birthday, assuming that a year has exactly 365 days?

20. There are 24 seats on a particular bus. When 28 grumpy and tired people board the bus, four will be left standing. How many different groups of people left standing could there be?

21. Six horses run a race. Prizes are then awarded for first, second, and third place. How many winning arrangements are possible?

22. In a psychology experiment, a person claiming to have ESP draws 3 cards from a set of 6 cards. If the person is asked what are the three symbols on the cards he's drawn, how many possible answers could he give if a) the order of the symbols matters or b) the order doesn't matter.

23. A Math 161 class consists of twenty men and five women. How many ways can a committee be formed if

    a) the committee has four members?
    b) the committee has four members – three men and one woman?
    c) the committee has four members – president, vice-president, secretary, and treasurer?

24. A student committee is formed with two second-year students and two first-year students. If there are 25 second-year students and 52 first year students, how many different committees could there be?

Section 6.2: Combinations and Permutations

Solutions

1. 5040
2. 40320
3. 1 663 200
4. 90
5. 495
6. 15 504
7. 303 600
8. 24 and 4
9. 12 650
10. 10 and 10
11. ${}_5P_5 = 120$ (or you could just say this is 5×4×3×2×1=5!=120)
12. ${}_8P_8 = 40320$
13. ${}_{50}P_3 = 117600$ (Yes, for nitpickers from hell, it should be a "permutation" lock.)
14. a) ${}_{10}P_4 = 5040$ b) 10^4 = 10 000
15. ${}_{52}C_5 = 2598960$
16. ${}_{49}C_6 = 13,983,816$
17. total = (total choices math)×(total choices physics) = ${}_{25}C_3 \; {}_9C_2 = 82,800$
18. ${}_4P_4 = 24$
19. ${}_{365}P_3 = 48\ 228\ 180$
20. ${}_{28}C_4 = 20,475$
21. ${}_6P_3 = 120$

22. a) ${}_6P_3 = 120$ b) ${}_6C_3 = 20$

23.

a) ${}_{25}C_4 = 12650$
b) ${}_{20}C_3 \ {}_5C_1 = 5700$
c) ${}_{25}P_4 = 303,600$

24. ${}_{25}C_2 \ {}_{52}C_2 = 397,800$

Section 6.3: Probability

Suppose you were to roll a six-sided die (one die, two dice – die is the singular of dice). What's the probability of rolling a 1? Well, if the die is fair and all six numbers are equally likely, then the probability of getting any one number is just 1/6.

This is the idea of classical probability – if all outcomes are equally likely, then the probability of event *E* happening is just the number of outcomes in which *E* happens, *n(E),* divided by the total number of outcomes *n*:

$$P(E) = \frac{n(E)}{n}$$

Example

If you roll two four-sided dice, what's the probability of rolling a total of 5?

Answer: The brute force method involves writing out all possible outcomes.

11	12	13	14
21	22	23	24
31	32	33	34
41	42	43	44

Then, assuming that the dice are fair, there are sixteen equally likely outcomes, and four of them – 41, 32, 23, and 14 – lead to a total of 5. Then P(total of 5) = 4/16 = ¼ or 25%.

Example

What's the probability of winning the Lotto 6/49 if you buy only one ticket?

Answer: You choose six numbers from 49, so there are ${}_{49}C_6 = 13,983,816$ possible combinations. There's only one way to win, which is if all six numbers drawn match the ones you've chosen. So your probability is

P(winning) = 1/13,983,816

Example

The Saanich city council has four members: Alex, Barbara, Charlie, and Dorothy. Two of these members are to be selected to form a subcommittee to study the city's traffic problems.

a) How many different subcommittees are possible? What probability would you assign to each one if there is an equal chance of selecting each council member?

b) What is the probability that Dorothy is a member of the committee?

c) What is the probability that Charlie and Dorothy are both selected?

d) What is the probability either Charlie or Dorothy or both are selected?

Answer:

a) possible outcomes = {AB, AC, AD, BC, BD, CD} for six outcomes (note that AB=BA since order doesn't matter). Or you could just say ${}_4C_2 = 6$. If they are all equally likely, then each has a probability of 1/6.

b) P(D) = (3 subcommittees with Dorothy)/(6 in total) = ½

c) P(CD) = (1 subcommittee with C & D)/6 = 1/6

d) P(C or D) = (5 subcommittees with C or D)/6 = 5/6

or P(C or D) = P(C) + P(D) – P(CD) = ½ + ½ – 1/6 = 5/6

or P(C or D) = 1 – P(AB) = 1 – 1/6 = 5/6

(note that AB is the only committee with neither C nor D)

Which brings us to the addition rule for probability:

$$\begin{aligned} P(A \text{ or } B) &= \frac{n(A \text{ or } B)}{n} \\ &= \frac{n(A) + n(B) - n(AB)}{n} \\ &= P(A) + P(B) - P(AB) \end{aligned}$$

However, in real life, you frequently get situations where not all outcomes are equally likely. One tool that we can use in that situation is called a **contingency table**.

Contingency tables

To study contingency tables, it's easiest to look at an example.

To simplify matters, let's assume that students at Interurban are enrolled in either Technology or Business (but not both). Let's also assume that men and women are equally represented in Business, but that only 10% of Technology students are women (which, frankly, is being generous!). Let's also assume that there are the same number of Technology and Business students. Then our entire student population of 100 (to make the numbers easier) would look like this:

	Technology	Business	Total
Male	45	25	70
Female	5	25	30
Total	50	50	100

Let's calculate the probability that if a student were randomly selected from this group, that the student was enrolled in Technology. Then what we wish to calculate is

$$P(T) = \frac{n(T)}{n}$$

where *T* is technology, *P(T)* is the probability of being in technology, *n(T)* is the number of "technology events" or in this case the number of students in technology, and *n* is the total number of students. Then P(T) = 50/100 = ½ or 50%. (The number 50 came from the total at the bottom of the technology column.)

Let's calculate the probability that the student was a female business student. This would be P(FB) = 25/100 = ¼ or 25%. The way we find n(FB) is we look at the cell in the intersection of the "female" row and the "business" column.

Let's calculate the probability that the student was male or in business. P(M or B) can be calculated either by adding up all the students who are male or in business or in both: P(M or B) = (45 + 25 + 25)/100 = 95%. Or you could say that it's going to be the total minus the FB students: (100-5)/100 = 95%. Or you could say that it's P(M)+P(B)–P(MB) = 70% + 50% – 25% = 95%.

Let's calculate the probability that if the student were female, that she was enrolled in Technology. The way we write this in symbols is P(T | F), which we read as P(T "if" F). What we are really asking is that if we only look at the female students (we limit our population), what's the probability of getting a technology student from among those female students? Then $P(T \mid F) = \frac{n(FT)}{n(F)} = \frac{5}{30} = \frac{1}{6}$.

Let's calculate the probability that that if the student were in Technology, that she were female. This seems like it's the same question in the last paragraph, but it's not. We're now limiting our population to the technology students, and calculating:

$$P(F \mid T) = \frac{n(FT)}{n(T)} = \frac{5}{50} = \frac{1}{10}$$

We can also ask the question: Are the events "student is female" and "student is enrolled in Technology" independent? What this is asking is "are the probabilities of being female the same for the entire population and for the technology population?". The way we tell is to calculate P(F) and P(F | B). If these two probabilities are the same, then the probability of being female **does not depend** on whether the student is in technology and we say the events are **independent**. Otherwise, we say that one depends on the other and the events are **dependent**.

So, P(F) = 30/100 = 30%. We already found that P(F | T) = 10%. So these probability are not the same, and these events are **dependent**. Notice that we could instead calculate P(T) and P(T | F) and compare those probabilities. P(T) = 50% = ½ and P(T | F) = 1/6, so we will reach the same conclusion.

Section 6.3: Probability

Exercises

1. A fair twelve-sided die is rolled. What is the probability that the roll is

   a) a 7?
   b) even?
   c) greater than 5?
   d) not a 7?
   e) a 1 or a 2?

2. Two four-sided dice are rolled. What is the probability that the roll

   a) results in the same number on both dice?
   b) results in different numbers on both dice?
   c) has a sum of 6?
   d) has at least one die rolling a 3?

3. (You may wish to consult your work for section 4.2 for this question) A two-digit number is generated at random. What's the probability that this number is divisible by 7?

4. (Again, consult your work for 4.2) In a psychology experiment, a person claiming to have ESP draws 3 cards from a set of six. If the person is asked what are the three symbols (in any order) on the cards he's drawn, what's the probability that he'll get it right if he guesses randomly?

5. An individual is presented with three different glasses of soft drink, labeled A, B, and C. He is asked to taste all three and then list them in order of preference. Suppose that the same soft drink has actually been put into all three glasses.

   a) How many outcomes are there in this experiment? What probability would you assign to each one?
   b) What is the probability that A is ranked first?
   c) What is the probability that either B or C is ranked first?
   d) What is the probability that A is ranked first and B is ranked last?

6. Your ATM/debit card has a four-digit PIN number associated with it. If there are no restrictions on what digits or what order you can pick them, then

   a) how many PIN numbers are possible?
   b) what is the probability that someone could guess your PIN randomly?
   c) if that person saw you input the first two digits when you were at the grocery checkout counter, what are their chances of guessing your PIN correctly now?

Complete the following exercises involving contingency tables

7. One hundred students each from the Computing Systems Technology program and from the English department were asked who is the greatest fictional wizard ever, with the following results.

	Gandalf	Dumbledore	total
CST	90	10	
English	40	60	
total			

a) Calculate P(G).
b) Calculate P(C|G).
c) Calculate P(G|C).
d) Calculate P(E or D).

8. A sampling of CST faculty and students were asked what operating system they used on their home computer, with the following results.

	Windows	Linux
faculty	6	2
students	24	8

a) What's the probability that a random CST user (faculty or student) will have Linux on their home machine?
b) What's the probability that a random CST **student** will have Linux on their home machine?
c) Are the events "student" and "Linux user" independent?

9. One thousand television watchers from BC and Alberta were asked if they watched the Rick Mercer Report on CBC with the following results.

	Yes	No
BC	500	500
AB	250	750

a) What's the probability that one of these people, when selected randomly, is from BC or watches the RMR?
b) What's the probability that one of these people, when selected randomly, is from BC and watches the RMR?

c) What's the probability that one of these people, when selected randomly, is from Alberta and does not watch the RMR?
d) What's the probability that a Rick Mercer watcher is from BC?
e) What's the probability that a British Columbian watches Rick Mercer?

10. A roving reporter surveyed all of the patrons inside the Starbucks and the Moka House coffee houses in Cook Street Village (it was a slow news day). The beverage each patron was drinking was noted and summarized in the following table.

	coffee	tea	other
Starbucks	45	9	6
Moka House	30	8	2

a) Are the events "drinking coffee" and "Starbucks" independent?
b) Are the events "tea" and "Moka House" independent?

11. StatsCan surveyed one hundred Canadians and found that 60 of them exercise regularly, 75 of them eat healthy diets, and 45 of them do both.

a) Complete the following contingency table using the above information

	exercise regularly	don't exercise regularly	total
healthy diet			
unhealthy diet			
total			

b) If one of these Canadians is selected randomly, what is the probability that this person exercises regularly but does not eat a healthy diet?
c) If one of these Canadians is selected randomly, what is the probability that this person exercises regularly or eats a healthy diet?
d) Is eating a healthy diet independent of exercising regularly for this sample of Canadians?

Section 6.3: Probability

Solutions

1. A fair twelve-sided die is rolled.

   a) P(7) = 1/12 (only one way to get a 7, and there are 12 outcomes)
   b) Even numbers from 1 to 12: 2, 4, 6, 8, 10, 12, so six possibilities out of 12 outcomes. P(even) = 6/12 = ½. (Or you could note that exactly half of the outcomes gave an even number to get an even shorter solution.)
   c) P(>5)=P(6 or 7 or 8 or 9 or 10 or 11 or 12) = 7/12
   d) P(not 7) = 1 – P(7) = 11/12
   e) P(1 or 2) = 2/12 = 1/6

2. I'm going to use the brute force method here and list all possible rolls:

11	12	13	14
21	22	23	24
31	32	33	34
41	42	43	44

   a) We can see that there are sixteen possibilities in total, and four of them will result in the same number on both dice, so P(both same) = 4/16 = ¼.
   b) P(different) = 1 – P(same) = ¾.
   c) P(sum of 6) = 3/16 (need 42, 33, or 24).
   d) You can count them up to find P(at least one 3) = 7/16.
   [Or you could say that's P(at least one 3) = 1 – P(no 3s). And the number of rolls with no 3s is _3_ _3__ = 9 possibilities, so then you'd get 1 – 9/16 = 7/16.]

3. P(2 digits divisible by 7) = n(2 digits divisible by 7) / n(2 digits) = 13/90.

4. We found from 4.2 that there are ${}_6C_3 = 20$ different possible sets of 3 cards from six. So if the person claiming to have ESP guesses randomly, he has a 1/20 chance.

5. This is a permutation question, since order matters.

   a) You can either just calculate ${}_3P_3 = 6$ or list the possible outcomes: {ABC,ACB,BAC,BCA,CAB,CBA}. Since it's the same soft drink in each glass, the lists should all be equally probable at 1/6 each.
   b) Only 2 of the 6 outcomes have A first, so P(A first) = 2/6 = 1/3. Or you could say that if A is first, there are two choices for second place and one for third (or ${}_2P_2 = 2$, if you insist).
   c) If either B or C is ranked first, then A is not. So P(B or C) = 1 – P(A) = 2/3.
   d) If A is first and B is last, then C is in the middle. P(ACB) = 1/6.

6. Your ATM/debit card has a four-digit PIN number associated with it. If there are no restrictions on what digits or what order you can pick them, then

   a) There are 10^4 PIN numbers possible (or 10,000).
   b) If that person is only given one guess (I should have said that!), then their chance is 1/10,000.
   c) If they know the first two numbers, then there are only 10×10 possible PIN numbers left. So their chances are now 1/100.

7.

	Gandalf	Dumbledore	total
CST	90	10	100
English	40	60	100
total	130	70	200

   a) P(G) = 130/200 = 13/20 (or 65%)
   b) P(C|G) = n(CG)/n(G) = 90/130 = 9/13 (which is roughly 69%)
   c) P(G|C) = n(CG)/n(C) = 90/100 = 9/10 (or 90%)
   d) P(E or D) = (40 + 60 + 10)/200 = 11/20 (or 55%)

8.

	Windows	Linux	total
faculty	6	2	8
students	24	8	32
total	30	10	40

   a) P(L)=10/40 = ¼ or 25%
   b) P(L|S) = 8/32 = ¼ or 25%
   c) Yes, "student" and "Linux user" are independent because P(L) = P(L|S).

9. One thousand television watchers from BC and Alberta were asked if they watched the Rick Mercer Report on CBC with the following results.

	Yes	No	total
BC	500	500	1000
AB	250	750	1000

total	750	1250	2000

a) P(BC or Y) = (500 + 500 + 250)/2000 = 5/8 (or 62.5%).
b) P(BC and Y) = 500/2000 = ¼ (or 25%).
c) P(AB and N) = 750/2000 = 3/8 (or 37.5%)
d) P(BC|Y) = n(BC and Y)/n(Y) = 500/750 = 2/3
e) P(Y|BC) = n (Y and BC)/n(BC) = 500/1000 = ½ (or 50%)

10.

	coffee	tea	other	total
Starbucks	45	9	6	60
Moka House	30	8	2	40
total	75	17	8	100

a) P(C) = 75/100 = ¾ = 75%. P(C|S) = 45/60 = ¾ = 75%. Yes, these events are independent. (You could alternatively calculate P(S)=60% and P(S|C)=60% to reach the same conclusion.)
b) P(T) = 17/100 = 17%. P(T|M) = 8/40 = 1/5 = 20%. As these are not the same, the events are dependent.

11.

a)

	exercise regularly	don't exercise regularly	total
healthy diet	45	30	75
unhealthy diet	15	10	25
total	60	40	100

b) $P(E\overline{H}) = 15/100 = 15\%$.
c) P(E or H) = (15 + 45 + 30)/100 = 90/100 = 90%.
(or $P(E \text{ or } H) = 1 - P(\overline{E}\,\overline{H}) = 1 - 10/100 = 90\%$)
d) P(H) = 75/100 = 75%. P(H|E) = 45/60 = 75 %. Since these probabilities are the same, eating a healthy diet is independent of exercising regularly for this sample of Canadians.

Section 6.4: Statistical Quantities

There are two ways to summarize a set of data: by a graph or by some numerical quantities. In this section, we will examine two types of statistical quantities: measures of centre and measures of spread.

Before we get there, though, we should first talk about the concept of sampling.

Samples and Populations

We are often interested in measuring a characteristic of a population. For example, we might be interested in the salaries of working Canadians, or the diameter or a particular widget, or the number of times a particular operating system crashes per day. Sometimes, if we care enough, we can attempt to measure that characteristic for the entire population – such as when we hold a federal election in which each citizen of voting age could potentially vote, or when Statistics Canada holds a national census. However, most of the time we are satisfied with measuring a sample which is much smaller than the entire population, and then generalizing the results of the sample to that population. This brings us to the concepts of

> **sample** – the measurements taken on a small group (in other words, a subset of the complete set of measurements that could be taken)
>
> **population** – the measurements taken on the entire group of interest

Although the sample and population, when taken strictly, refer to the measurements on a group, we often in speech use sample and population to refer to the groups themselves.

> ***Example***
>
> The Math Department is interested to find out how many students are using the Math Lab at Interurban. They therefore ask Gilles to survey his Math 222 class to find out how many of them have used the Math Lab. Identify the sample and population for this survey.
>
> Answer:
>
> sample: Gilles' Math 222 class (if you want to nitpick, you could even say the students in Gilles' Math 222 class that happened to show up that day)
>
> population: Interurban students (again, you could nitpick by staying Interurban students taking math courses, but you get the idea)

Now that we've identified the difference between sample and population, we can move on to discussing the quantities that summarize a sample data set.

Measures of centre

Measures of centre are statistical quantities that describe the location of the centre of the data set of interest. The two we will be looking at are the mean and median.

Mean

The mean, or average, of a set of data points is equal to the sum of the data points divided by the number of data points. So, writing this in summation notation, the mean of a sample is written as $\bar{x}$ and given by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} .$$

If we are looking at the mean of a population, it has the same formula but is given the symbol μ.

Example

Calculate the mean of the following data set: 2, 6, 7, 3, 3.

Answer: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+6+7+3+3}{5} = 4.2$

Median

The median is the middle value in the set of measurements when written as an ordered list. Remember, data sets can have repetition! If there is an even number of measurements, then the median is the average of the two middle values.

Example

Calculate the median of the following data set: 2, 6, 7, 3, 3.

Answer: Writing the data set as an ordered list, we get 2, 3, 3, 6, 7. The middle value is clearly the third one, so the median is 3.

Example

Calculate the median of the following data set: 12, 26, 39, 14, 16, 15.

Answer: Writing the data set as an ordered list, we get 12, 14, 15, 16, 26, 39. The two middle measurements are 15 and 16, so the median is the average of the two, or 15.5.

For larger data sets, it can become harder to determine which is the middle value (is it the 54^{th} value? the 55^{th}?). The rule is that you take the number of measurements, divide by two, and add ½. For example, when there are five measurements, 5/2 + 1/2 = 3, so the third measurement is the median. When there are six measurements, 6/2 + 1/2 = 3.5, so we need to take the average between the 3^{rd} and 4^{th} measurements.

Measures of variability

Measures of variability (or spread) are statistical quantities that describe the width of the distribution. The two we will be looking at are the range and the standard deviation.

Range

The range of a set of data points is the difference between the highest and lowest value.

Example

Calculate the range of the following data set: 2, 6, 7, 3, 3.

Answer: range = highest – lowest = 7 – 2 = 5.

The range is very easy to calculate and visualize, but can be problematic: if you have an outlying point, it can distort the range quantity and make it not as meaningful as you might want. That's why we use another quantity which has a more slightly complicated calculation, the standard deviation.

Standard Deviation

Let's consider our data set 2, 6, 7, 3, 3, which has a mean of 4.2. One way we could determine how "wide" the distribution is would be to calculate how far each data point is from the centre-line (the mean). So we could calculate $\left(x_i - \bar{x}\right)$ as a measure of how far each data point is from the centre. For example, when x_i=2, then $\left(x_i - \bar{x}\right) = 2 - 4.2 = -2.2$. If we were to add up all of these quantities, we should get zero, because all of the points on the left side of the distribution should cancel out the points on the right side of the distribution. But if we were to square this quantity to get all positive values, then the sum would be a direct measure of the distance of each point from the centre line (since distances are always positive). The sum of these squares divided by the number of data points gives a value called the **variance**. However, since the units of the variance are the squares of the units of the original measurements, we prefer to take the square root of this quantity, which is called the standard deviation.

For a sample of measurements, the standard deviation *s* is equal to

$$s = \sqrt{\frac{\sum\left(x_i - \bar{x}\right)^2}{n-1}}$$

where $\bar{x}$ is the mean of the sample and n is the number of measurements.

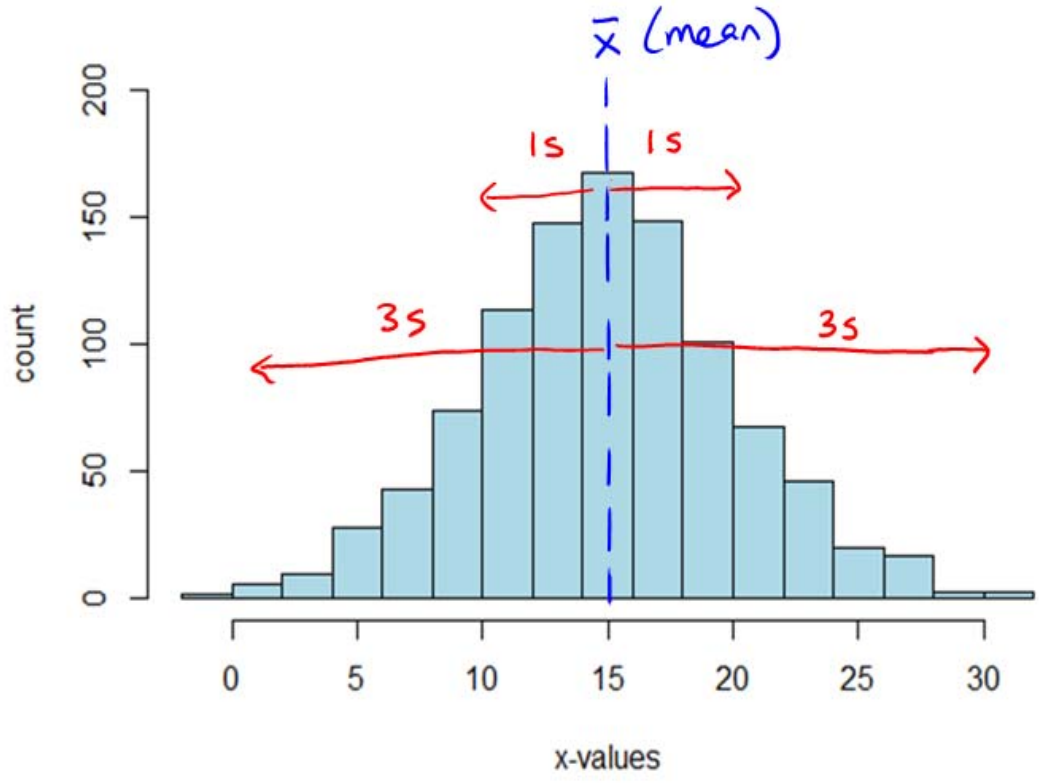
For measurements on an entire population, the standard deviation is given the symbol σ and is given by

$$\sigma = \sqrt{\frac{\sum\left(x_i - \mu\right)^2}{n}}$$

where μ is the mean of the population and n is the number of measurements.

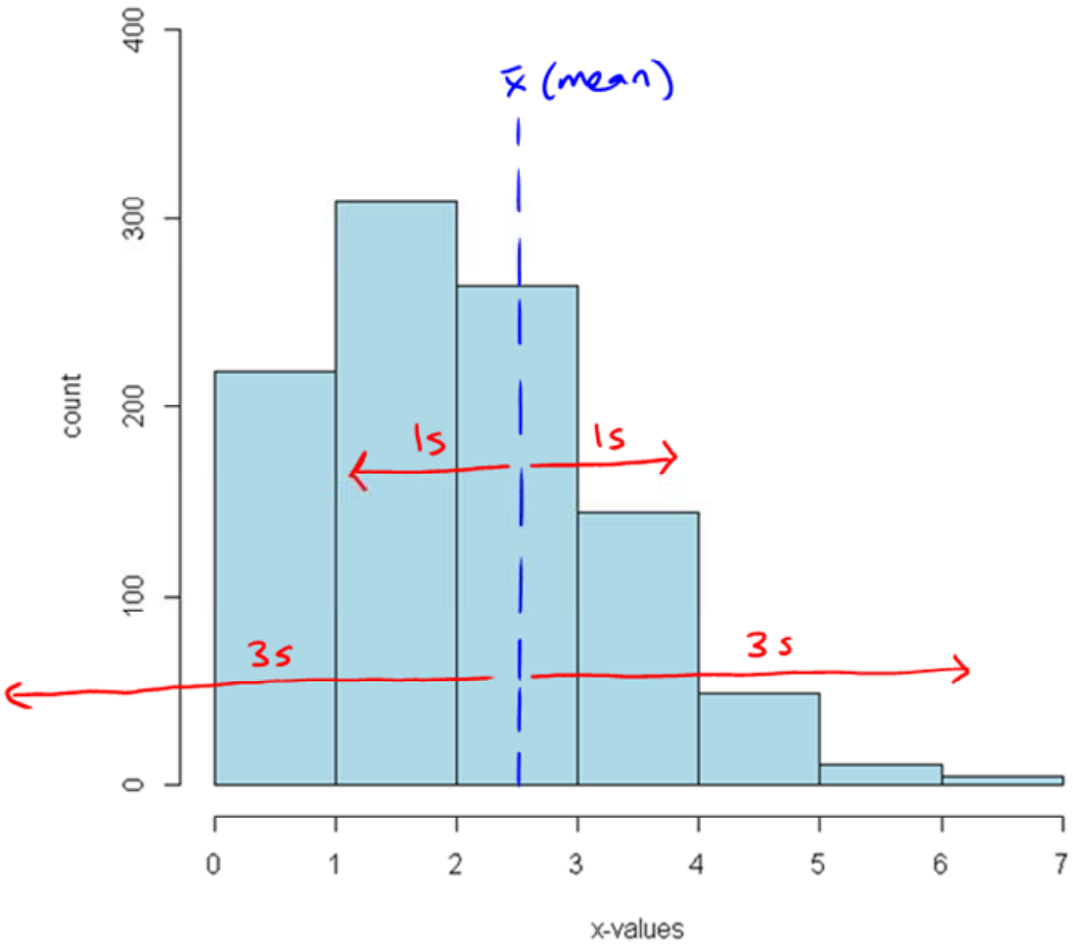
Let's see what this looks like with some data. Suppose we have a population whose data is symmetrically distributed with only one peak. (Distributions with only one peak are said to be unimodal.) This data set has a mean $\bar{x}$ of 15, a median of 15, and a standard deviation s of 5. You can see that the bulk of the measurements lie within one standard deviation of the mean, and almost all of them lie within three standard deviations from the mean.

Figure 1: A Symmetrical, Unimodal Distribution



Compare the previous histogram with the next one, which has an asymmetrical distribution with mean $\bar{x}$ of 2.5 and standard deviation s of 1.4. Even though this one has considerable asymmetry (the median is 2), the bulk of the data still lies within one standard deviation of the mean, and virtually all of the data lies within three standard deviations of the mean.

Figure 2: An Asymmetrical Distribution



Section 6.4: Measures of Centre and of Variability

Exercises

1. The top ten movies (Skyfall, The Hobbit, etc.) and their profits (in millions of dollars) from last weekend are reported in Monday Magazine. Calculate the mean and median for this data.

   profits: 1.4, 4.1, 1.2, 1.3, 5.8, 5.0, 2.6, 1.8, 2.9, 5.9, 2.5, 5.3

2. Calculate the mean and median for the data set: 35, 47, 29, 42, 38, 39, 42.

3. Pat finds the mean height of all twelve students in her physics class to be 68.0 inches. Just as she's finished that calculation, one more student walks in late. If that student is 63.0 inches tall, what is the mean height of all thirteen students?

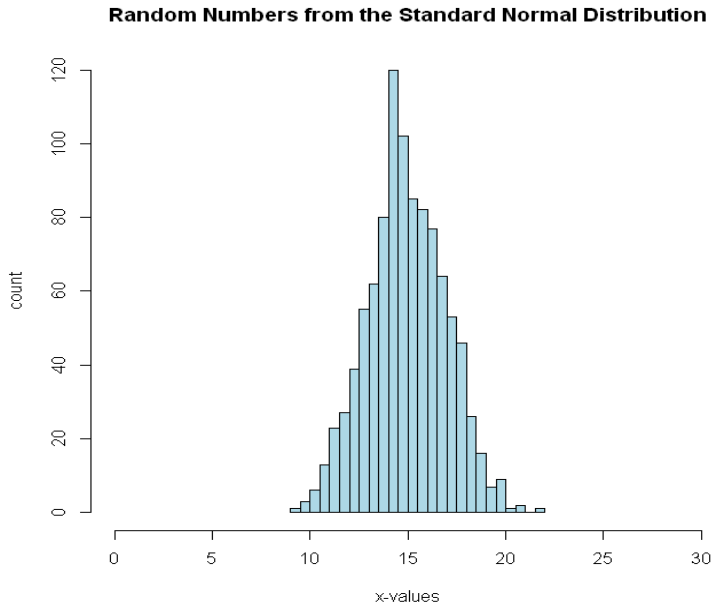
4. The Victoria Real Estate Board claims that in October of 2012, the average cost of a single-family home in Greater Victoria was $592,000, while the median was $527,000. Why is the mean greater than the median for housing prices? Explain.

5. Tom is running a small business with five employees, including himself. The salaries of the five people (in thousands of dollars) are 30, 45, 50, 55, and 75, with Tom making the highest salary.

   a) calculate the mean and median of these salaries
   b) if Tom gives everyone a $2000 bonus, what happens to the mean and median?
   c) if Tom gives everyone a 5% raise, what happens to the mean and median?
   d) if Tom decides to keep everyone else's salary the same, but raise his own salary by $10,000, what happens to the mean and the median?

6. Consider the following histogram. Is the standard deviation equal to

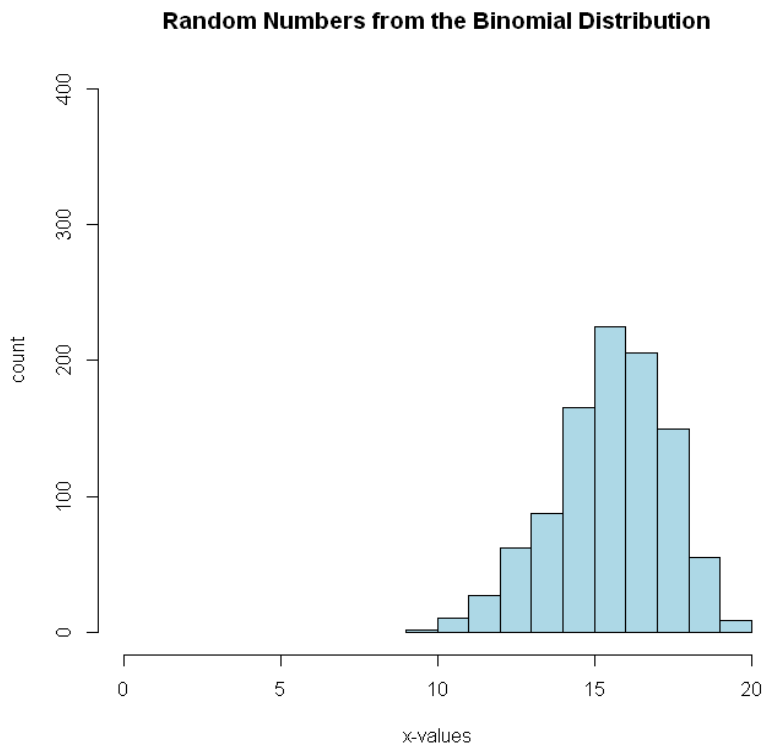
   a) 0.5
   b) 2
   c) 15
   d) 20



7. Consider the following data set: 7, 7, 7, 7, 7, and 7. What is the mean and the median? What is the range? Without calculating it, what would be the standard deviation?

8. Consider the following histogram. Is the standard deviation equal to

   a) 1
   b) 2
   c) 5
   d) 10
   e) 15
   f) 20



9. Pat, when entering quiz scores into her spreadsheet, accidentally put an extra zero on the end of one student's score (making it 380/40 instead of 38/40), and then calculated the mean, median, range, and standard deviation for the section. She then noticed her mistake and recalculated all of the quantities. For the following quantities, state whether the corrected value will be higher, lower, or the same as the value calculated with the incorrect quiz score:

   a) mean
   b) median
   c) range
   d) standard deviation

10. Consider the following sets of data. Without calculating any values, state which set will have the higher standard deviation (or will they be the same?).

    a) Set 1: 2, 3, 9, 16, 17 Set 2: 2, 8, 9, 10, 17
    b) Set 1: 2, 3, 9, 16, 17 Set 2: 3, 4, 10, 17, 18

Section 6.4: Statistical Quantities

Solutions

1. The mean is 3.31667, or just 3.3. There are twelve points, so the median is the 12/2+1/2=6.5th point, which means the average of the 6th and 7th points. Therefore, the median is (2.6+2.9)/2 = 2.75.

2. The mean is 38.8571. (You can round to 38.9 if you like.)

3. To find the mean, we want the sum of all of the heights divided by the total number of students. Since the average of the twelve students is 68.0 inches, the total of all of those heights is just 68.0 times 12, which is 816.0 inches. Adding the height of the thirteenth student brings the total to 879.0 inches, then dividing by 13 gives a mean of 67.6 inches.

4. The histogram of Victoria housing prices will not be symmetrical: there is a lower limit for the price of single-family homes, while there can be house prices in the millions of dollars. Just a few very expensive homes will bring up the mean but not affect the median in any way, which is why the mean is greater than the median.

5. The means and medians are:

   a) mean = \$51,000 and median = \$50,000
   b) the mean and median will each increase by \$2000: mean is now \$53,000 and the median \$52,000
   c) the mean and median will both increase by a factor of 1.05 (they are multiplied by 1.05): mean is now \$53,550 and median is \$52,500
   d) the mean will become \$53,000 but the median will stay the same

6. Looking at the histogram, you can estimate the standard deviation by picking a "width" about the mean/average that most of the data points fall within. From this histogram, the standard deviation is about half of 5, since most of the data falls between approximately 12.5 and 17.5 (ish). And the closest value given that matches that is (b) 2.

7. The mean and median are both 7. The range is 0. The standard deviation is also 0, since all points lie exactly on the mean and $(x - \overline{x})$ is zero for each point.

8. Using the same reasoning as for question 6, most of the data seems to fall between 12.5 and 17.5, so the standard deviation is around 2.5 (ish). So the closest option given is (b) again.

9. New values:

   a) The corrected mean will be lower, since one value was lowered.

b) The median will remain unchanged (assuming that the 38/40 was in the upper half of the scores to begin with, so changing it to 380 and back won't affect that)
c) The corrected range will be lower, since the highest point has changed.
d) The standard deviation will be lower, since the corrected point's distance from the mean is lower than the uncorrected value.

10. a) Set 1's values are farther from the mean on average than Set 2's data points. So Set 1 will have a higher standard deviation.

b) Set 2's data points are just Set 1's points moved up by 1 unit. So each point's distance from the mean will be the same as Set 1, and the standard deviations will be the same also.

Section 6.5: The Empirical Rule

Now that we know how to calculate the mean and standard deviation, how can we use these quantities to say something more specific about a set of data? The following rules will allow us to do that.

Tchebysheff's Theorem

Tchebysheff's theorem can be applied to *any data set*, including both samples and populations. What it says is that *at least* 3/4 (75%) of the measurements lie within two standard deviations of the mean and that *at least* 8/9 (approximately 89%) of the measurements lie within three standard deviations of the mean. Interestingly, the theorem has nothing to say about how many of the measurements fall within one standard deviation from the mean. In fact, it's entirely possible that none do.

(The full theorem is considerably more complicated, but this is a useful summary for now.)

Example:

A sample of Douglas fir trees was measured, and the average diameter $\overline{x}$ was found to be 32 cm with a standard deviation *s* of 3 cm. What can you say about the fraction of measurements between 26 and 38 cm?

Answer:

38 – 32 cm = 6 cm = $2 \times$ (3 cm), so the measurement 38 cm is two standard deviations above the mean. Similarly, 26 cm is two standard deviations below the mean. Tchebysheff says that *at least* 75% of the measurements are within this range. (It could be more than 75%, but cannot be less than that.)

Example:

Consider the same sample of Douglas fir trees as in the previous problem. What can you say about the fraction of measurements between 23 and 41 cm? Is it possible that all measurements lie in this range?

Answer:

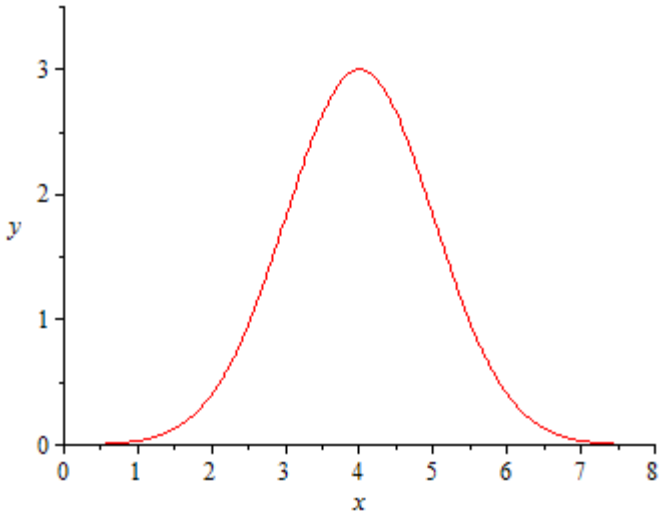
41 – 32 = 9 cm = $3 \times$ (3 cm), so the measurement 41 cm is three standard deviations above the mean. Similarly, 23 cm is three standard deviations below the mean. Tchebysheff says that *at least* 8/9 (or 89%) of the measurements lie within this range. However, this is a *lower* limit, so it is indeed possible for *all* measurements to lie in this range. (Whether or not it is probable is a different question!)

Note that the estimates from Tchebysheff's theorem tend to be very conservative. This is because the theorem works for *all* distributions, and only gives lower limits.

The Empirical Rule

There is a particular shape for a data distribution that occurs frequently in nature, as is shown in the figure below. As you can see, the curve is symmetrical and unimodal (has only one peak). This distribution is said to be "mound-shaped", and there is a rule that describes data in that form called the Empirical Rule.

Figure 1: A Mound-shaped Curve



The Empirical Rule says that if a set of measurements can be said to be mound-shaped, then

- approximately 68% of the data will lie within 1 standard deviation of the mean
- approximately 95% of the data will lie within 2 standard deviations of the mean
- approximately 99.7% of the data will lie within 3 standard deviations of the mean

Example

Twelve software engineers in the Greater Victoria area were picked randomly from an industry list and asked what their yearly salary was (in thousands of dollars), with results displayed in the list below.

79, 83, 94, 88, 98, 106, 76, 71, 82, 86, 63, 90

The mean and standard deviation of this sample data are 84.7 and 11.8 thousand dollars, respectively, and the data are plotted in the histogram below.

Figure 2: Salaries of Victoria Engineers

a) Describe the shape and symmetry of the resulting distribution. Is it mound-shaped?

b) Complete the table below by finding the percentage of measurements in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$ and $\bar{x} \pm 3s$. Also, state the percentages you'd expect to find in each interval according to the Empirical Rule and Tchebysheff's Theorem.

	interval	# of points	% of points	Empirical	Tchebysheff
$\bar{x} \pm s$					
$\bar{x} \pm 2s$					
$\bar{x} \pm 3s$					

c) Do your percentages in the table above agree with those given by the Empirical Rule and Tchebysheff's Theorem? Explain briefly.

Answer

a) The distribution is symmetrical and unimodal, so it is roughly mound-shaped.

b)

	interval	# of points	% of points	Empirical	Tchebysheff
$\bar{x} \pm s$	72.9 – 96.5	8	66.7%	~68%	---
$\bar{x} \pm 2s$	61.1 – 108.3	12	100%	~95%	$\geq 75\%$
$\bar{x} \pm 3s$	49.3 – 120.1	12	100%	~99.7%	$\geq 89\%$

c) Yes, they do. Since the distribution is mound-shaped, the Empirical Rule applies, the Empirical predictions fit pretty well with the actual numbers (the ~95% is a little off, but there aren't a lot of data points so the fit isn't too bad). Tchebysheff's theorem works for all distributions, so we expect the predictions to be accurate in this case, as they are.

Example

In the 2000 Olympic Games in Sydney, the mean jump on the long jump event was 610 cm with a standard deviation of 30 cm. Suppose the Canadian competitor jumped a distance of 640 cm. Assuming that these distances have a mound-shaped distribution and that there were 25 competitors in total, on average how many of the competitors would you expect to finish ahead of the Canadian?

Answer: 640 cm is one standard deviation above the mean. The Empirical Rule states that about 68% of the measurements fall within one standard deviation, so that leaves 32% either above 640 cm or below 580 cm. The distribution is symmetrical, leaving 16% above. 16% of 25 is 4, so we'd expect around 4 jumpers to finish ahead of the Canadian.

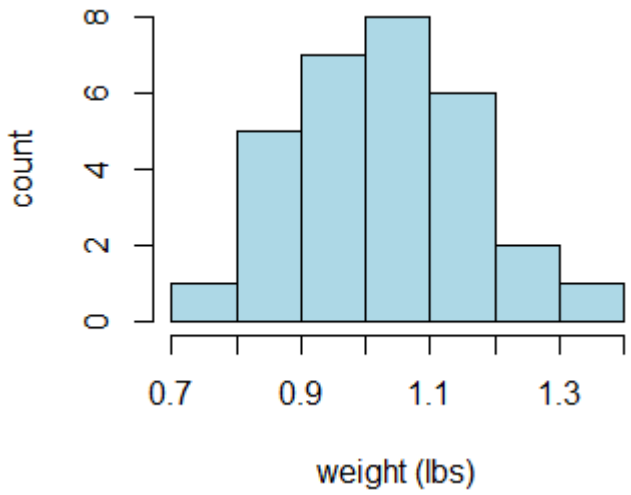
Section 6.5: The Empirical Rule

Exercises

1. A market researcher listed the weights (in pounds, sadly, considering that we are supposed to be a metric country) for 30 packages of ground beef in Thrifty's meat display.

   1.08, 0.99, 0.97, 1.18, 1.09, 1.28, 0.83, 1.06, 1.14, 1.38, 0.75, 0.96, 1.08, 0.87, 0.89, 0.89, 0.96, 1.01, 1.12, 1.06, 0.93, 1.24, 0.89, 0.98, 1.14, 0.92, 1.18, 1.17, 1.02, 1.03

   The mean and standard deviation of this sample are 1.04 and 0.14 lbs, respectively, and the data is graphed in the plot below.



   a) Describe the shape of the resulting distribution. Is the distribution mound-shaped?
   b) Find the percentage of measurements in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$. In your table, also state what percentages you expect to see in these intervals using either Tchebysheff or the Empirical Rule.
   c) Do the percentages obtained in part b) agree with those given by the Empirical Rule? By Tchebysheff? Should they?

2. A set of data has a mean of 75 and a standard deviation of 5. The histogram shows a more-or-less symmetric, mound-shaped distribution.

   a) What can you <u>estimate</u> about the proportion of measurements that fall between 70 and 80? Between 65 and 85?
   b) What can you say <u>with certainty</u> about the proportion of measurements that fall between 70 and 80? Between 65 and 85?

3. The top ten movies (Skyfall, Men in Black III, etc.) and their profits (in millions of dollars) from last weeked are reported in Monday Magazine. Consider this data to be a sample of movie profits for weekends in the year 2012.

   profits: 1.1, 1.1, 1.2, 1.3, 5.1, 6.0, 9.6, 9.8, 9.9, 9.9

   The mean and standard deviation of this data set are 5.5 and 4.1 million dollars, respectively.

   a) By looking at the data, do you think that this distribution relatively mound-shaped?
   b) Find the percentage of measurements in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$.
   c) How do the percentages obtained in part c compare with those given by the Empirical Rule and Tchebysheff's Theorem? Do they agree? Do you expect them to?

4. In the 2000 Olympic Games in Sydney, the mean time on the 800 m event was 137 seconds with a standard deviation of 4 seconds. Gertrud Bacher of Italy finished in 129 seconds. Assuming that these running times have a mound-shaped distribution and that there were 38 competitors in total, on average how many of the competitors would you expect to finish ahead of the Italian? (She won the event, by the way.)

Section 6.5: The Empirical Rule

Solutions

1. Thrifty Foods:

   a) The distribution is roughly symmetrical and only has one peak (it is unimodal). Yes, it is mound-shaped.
   b) Consider the following table.

	interval	# of points	% of points	Empirical	Tchebysheff
$\bar{x} \pm s$	0.90 – 1.18	21	70.0%	~68%	---
$\bar{x} \pm 2s$	0.76 – 1.32	28	93.3%	~95%	$\geq 75\%$
$\bar{x} \pm 3s$	0.62 – 1.46	30	100%	~99.7%	$\geq 89\%$

   c) See the table above.
   d) The agreement between the actual numbers and the Empirical Rule is pretty good (considering that there are only 30 points in the distribution, so each point represents 3% of the measurements), which is to be expected since the distribution is mound-shaped. As always, Tchebysheff is accurate since the predictions are true for all distributions.

2. Since the distribution is symmetrical, the Empirical Rule applies which allows you to estimate percentages. And Tchebysheff's theorem, which applies to all distributions, allows you to calculate with certainty lower limits on percentages.

   a) Between 70 and 80 is one standard deviation away from the mean, so by the Empirical Rule we can estimate that ~68% of measurements lie in this range. Between 65 and 85 is two standard deviations away from the mean, so ~95% of measurements will lie in this range.
   b) Tchebysheff's theorem has nothing to say about the percentage of measurements within one standard deviation of the mean, so we can't say anything with certainty about those that lie between 70 and 80. But *at least* 75% will fall between 65 and 85.

3. Movies:

   a) There's a group of measurements down around 1.1, then a couple in the middle, then a big group up at 9.9. So this distribution has more than one peak, and is not mound-shaped.

b) Here's a table:

	interval	# of points	% of points	Empirical	Tchebysheff
$\bar{x} \pm s$	1.4 – 9.6	3	30.0%	~68%	---
$\bar{x} \pm 2s$	(-2.7) – 13.7	10	100%	~95%	$\geq 75\%$
$\bar{x} \pm 3s$	(-6.8) – 17.8	10	100%	~99.7%	$\geq 89\%$

c) The actual percentages don't agree with the Empirical Rule predications at all, which is not surprising since the distribution is not mound-shaped and the Empirical Rule does not apply. As always, Tchebysheff applies and is accurate.

4. Bacher finished 8 seconds ahead of the mean, which is two standard deviations. Since the distribution is mound-shaped, the Empirical Rule applies, and ~95% of the competitors' times should fall within two standard deviations of the mean. So, ~5% of the times will lie outside of this range, and since the distribution is symmetrical, we can assume that ~2.5% will lie below Bacher's time. Multiplying this by the number of competitors gives 0.95 competitors who will finish ahead of Bacher, which rounds to 1 person on average finishing ahead of her.