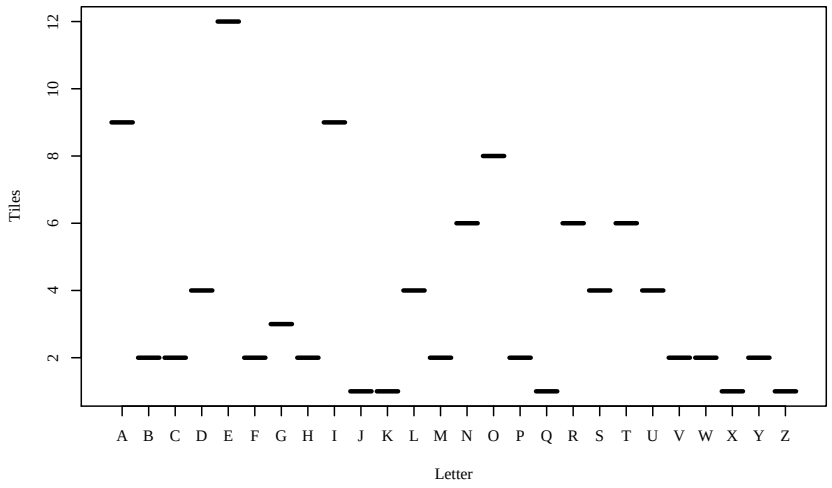
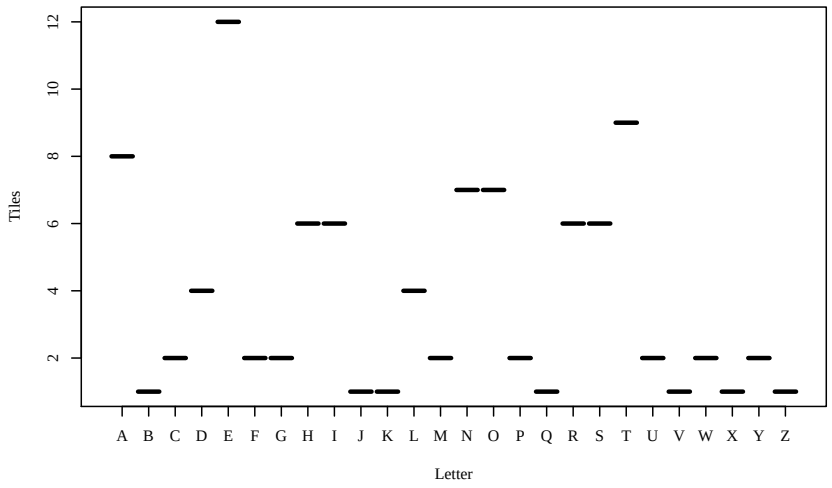
## Scrabble letters

When Alfred Mosher Butts developed Scrabble beginning in 1933, he chose the distribution of letters after long and careful consideration. He ultimately decided there should be 100 tiles, with two blanks, and the other 98 divided among the letters like so:




If Butts had simply counted the letters on the front page of *The New York Times* as is commonly believed, his letter distribution would have been more like this:



For the questions that follow, let's assume that the actual Scrabble distribution is perfect and that the alternate distribution is wrong.

(a) Which five letters' counts change the most between the two distributions?

(b) The Brown Corpus consists of over one million words of text taken from a variety of sources and genres. We will pretend it is a reasonable approximation to the front page of *The New York Times*.

The twenty words that occur most frequently in the Brown Corpus are *the*, *of*, *and*, *to*, *a*, *in*, *that*, *is*, *was*, *he*, *for*, *it*, *with*, *as*, *his*, *on*, *be*, *at*, *by*, and *I*, in that order. These twenty words comprise about 31% of the word tokens in the corpus. Here *token* refers to an instance of a word in the text.

For which of the letters in your answer to (a) does this list help explain why the two distributions assign it a different number of tiles?

(c) The words in (b) are not equally frequent. Rather, frequency decreases rapidly with rank:

| | Word | Frequency (%) |
|---|---|---|
| 1. | the | 6.8872 |
| 2. | of | 3.5839 |
| 3. | and | 2.8401 |
| 4. | to | 2.5744 |
| 5. | a | 2.2996 |
| 6. | in | 2.1010 |
| 7. | that | 1.0428 |
| 8. | is | 0.9943 |
| 9. | was | 0.9661 |
| 10. | he | 0.9392 |
| 11. | for | 0.9340 |
| 12. | it | 0.8623 |
| 13. | with | 0.7176 |
| 14. | as | 0.7137 |
| 15. | his | 0.6886 |
| 16. | on | 0.6636 |
| 17. | be | 0.6276 |
| 18. | at | 0.5293 |
| 19. | by | 0.5224 |
| 20. | I | 0.5099 |

How does this new information change your answer to part (b)?