

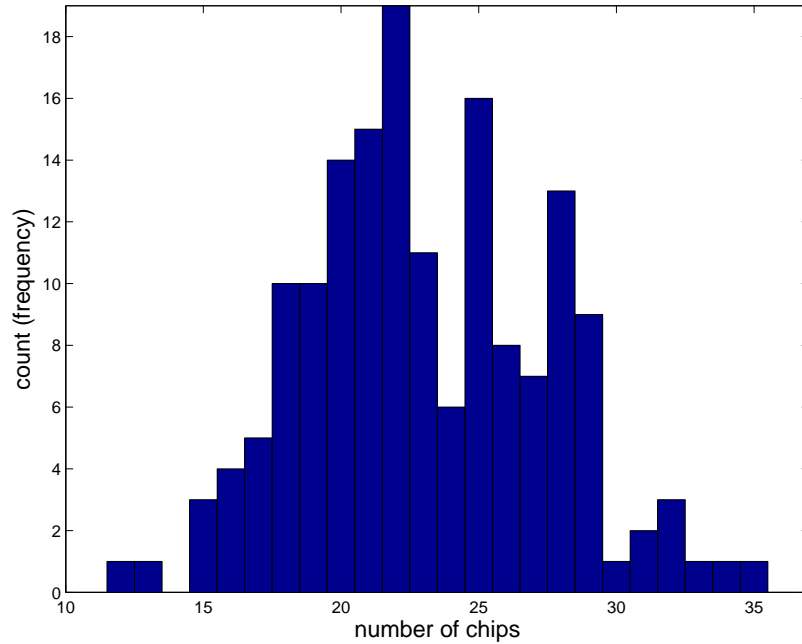
Chocolate Chip Histograms

Robin Morris

- Histograms are a visual display of variability.
- We recorded the number of chocolate chips in the cookies in class. The first thing to do is to look at the **range** of the data. **the smallest number of chips was 12**, and **the largest number was 35**. Most of the cookies had between 15 and 30 chips.
- To plot the histogram, we divide the range of the data (12-35) into **class intervals**. To begin with, we are going to use class intervals of size one, so that each chip-count gets its own interval.
- We draw up a table which shows the frequency of each number of chips.

# chips	frequency
12	1
13	1
14	0
15	3
16	4
17	5
18	10
19	10
20	14
21	15
22	19
23	11
24	6
25	16
26	8
27	7
28	13
29	9
30	1
31	2
32	3
33	1
34	1
35	1

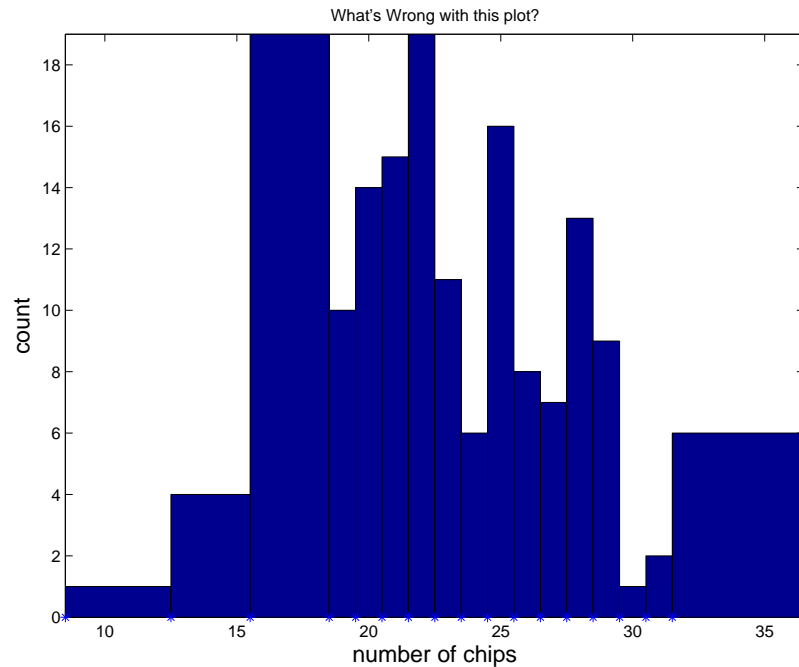
- Then we draw a bar plot, where the length of the bar is proportional to the number (frequency) in each class interval. This is shown below.



- Because the **class intervals** were all the same, the height of the block gives all the relevant information. What happens when this is not the case?
- In the table below, instead of having each number of chip being a separate class interval, I have defined class intervals that are larger for small and large numbers of chips, where the frequencies are lower.

#chips	count
9-13	1
13-16	4
16-19	19
19-20	10
20-21	14
21-22	15
22-23	19
23-24	11
24-25	6
25-26	16
26-27	8
27-28	7
28-29	13
29-30	9
30-31	1
31-32	2
32-37	6

- Note that we need to specify which **endpoint** is included in the interval, and which endpoint is excluded. In this table we have included the **lower** endpoint, and excluded the upper one.
- In the figure below, I have drawn a bar plot similar to the earlier one, where the height of the bar is proportional to the frequency in this table.



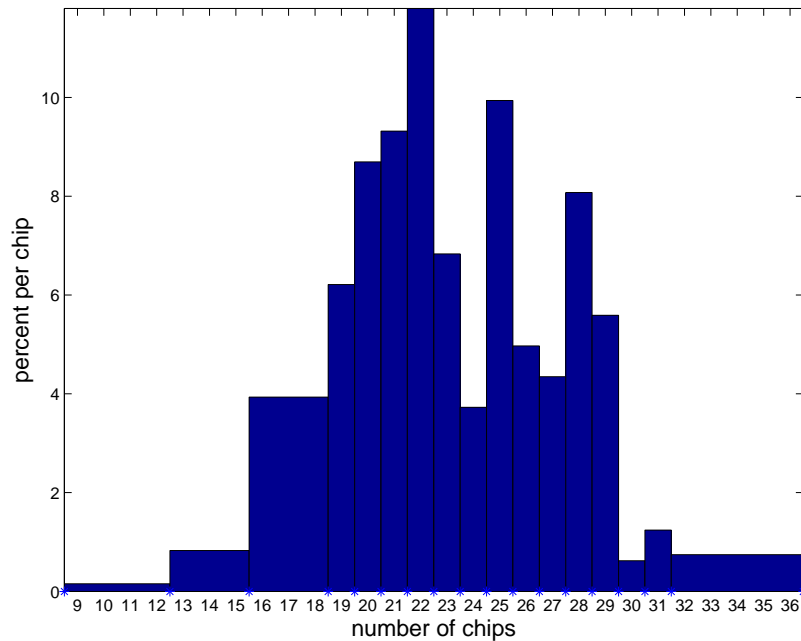
- What is wrong with this plot?
- Clearly it is **visually misleading** –
 - The 16-19 bar looks much bigger than the 22-23 bar, whereas the number of cookies in each of these two class intervals is the same.
 - The bar for the class interval 32-37 looks much larger than the bar for the interval 24-25, whereas the number of cookies in each interval is the same
- It is also hard to use this diagram to answer questions of the form “what the number of cookies with between 16 and 18 chips?”
- We need to scale the blocks by the size of the class intervals so that **The AREA of the block represents percentage**, not the height.
- To do this we **divide** the number of counts by (total number of counts) to convert to percentage, and then divide by **the length of the class interval**
- In the next table, I show the percentage in each class interval.

#chips	count	percentage
9-13	1	0.6
13-16	4	2.5
16-19	19	11.8
19-20	10	6.2
20-21	14	8.7
21-22	15	9.3
22-23	19	11.8
23-24	11	6.8
24-25	6	3.7
25-26	16	9.9
26-27	8	5.0
27-28	7	4.3
28-29	13	8.1
29-30	9	5.6
30-31	1	0.6
31-32	2	1.2
32-37	6	3.7

- And in the following table, I also show the result of dividing the percentage by the width of the class interval. We need to figure out **what units** this last column is in. The values are computed by dividing a percentage by a number of chips, so the units are **percent per chip**

#chips	count	pct	pct-per-chip
9-13	1	0.6	0.2
13-16	4	2.5	0.8
16-19	19	11.8	3.9
19-20	10	6.2	6.2
20-21	14	8.7	8.7
21-22	15	9.3	9.3
22-23	19	11.8	11.8
23-24	11	6.8	6.8
24-25	6	3.7	3.7
25-26	16	9.9	9.9
26-27	8	5.0	5.0
27-28	7	4.3	4.3
28-29	13	8.1	8.1
29-30	9	5.6	5.6
30-31	1	0.6	0.6
31-32	2	1.2	1.2
32-37	6	3.7	0.7

- Again, the class intervals include the lower edge, but not the upper edge.
- Plotting this in the form of a bar chart gives

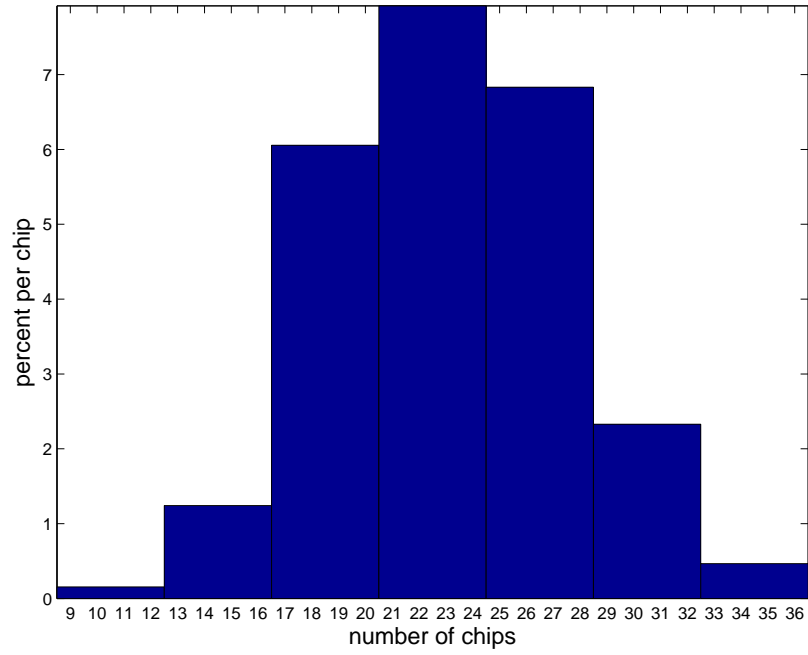


which gives a better visual indication of the distribution of chips within cookies.

- Now we can go back and answer the question “what the number of cookies with between 16 and 18 chips?”

We find that the bar that covers the interval 16-19 has height 3.9 percent-per-chip, so that the percentage of cookies with between 16 and 18 chips is 3.9×2 (percent-per-chip * chips) (Where we have again included the lower endpoint but not the upper endpoint in defining the interval of interest.)

- Finally, using class intervals of width 4, we can plot the following histogram



This shows that careful choice of class interval can have a smoothing effect, removing some of the variation in the data, and showing the main trends.