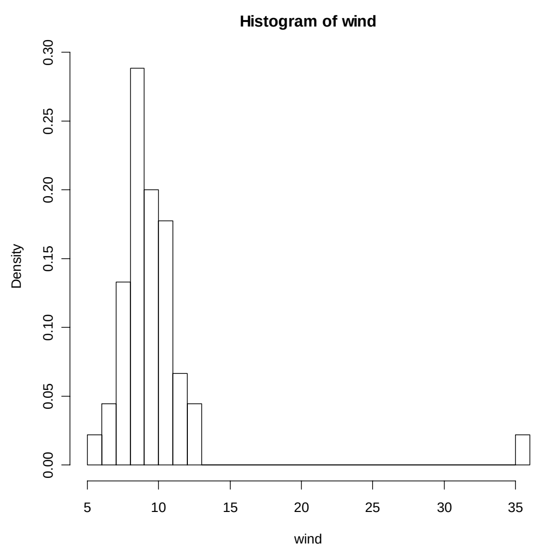
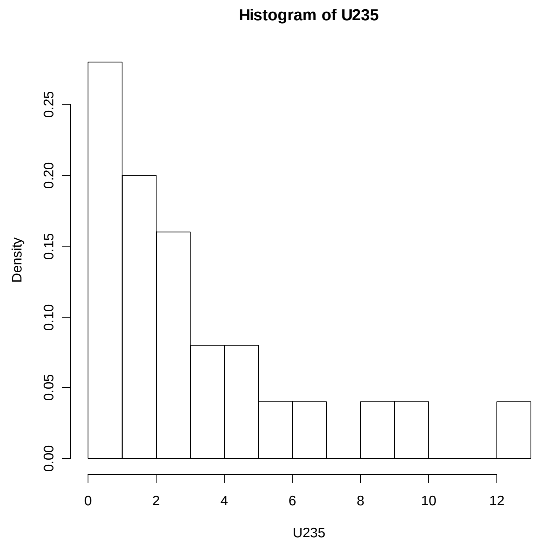
# Chapter 1: What is Statistics?

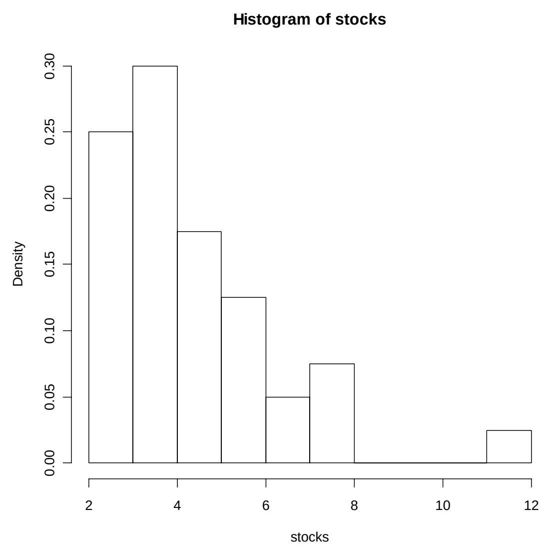
**1.1** **a.** Population: all generation X age US citizens (specifically, assign a ‘1’ to those who want to start their own business and a ‘0’ to those who do not, so that the population is the set of 1’s and 0’s). Objective: to estimate the proportion of generation X age US citizens who want to start their own business.
**b.** Population: all healthy adults in the US. Objective: to estimate the true mean body temperature
**c.** Population: single family dwelling units in the city. Objective: to estimate the true mean water consumption
**d.** Population: all tires manufactured by the company for the specific year. Objective: to estimate the proportion of tires with unsafe tread.
**e.** Population: all adult residents of the particular state. Objective: to estimate the proportion who favor a unicameral legislature.
**f.** Population: times until recurrence for all people who have had a particular disease. Objective: to estimate the true average time until recurrence.
**g.** Population: lifetime measurements for all resistors of this type. Objective: to estimate the true mean lifetime (in hours).




**1.2** **a.** This histogram is above.
**b.** Yes, it is quite windy there.
**c.** 11/45, or approx. 24.4%
**d.** it is not especially windy in the overall sample.




**1.3** The histogram is above.

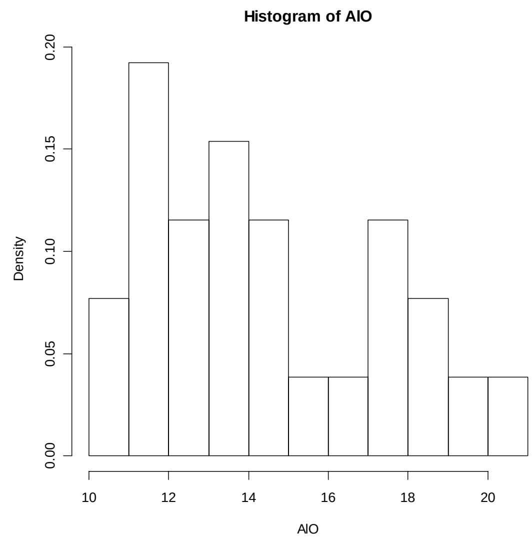



**1.4** **a.** The histogram is above.
**b.** 18/40 = 45%
**c.** 29/40 = 72.5%

**1.5** **a.** The categories with the largest grouping of students are 2.45 to 2.65 and 2.65 to 2.85. (both have 7 students).
**b.** 7/30
**c.** 7/30 + 3/30 + 3/30 + 3/30 = 16/30

**1.6** **a.** The modal category is 2 (quarts of milk). About 36% (9 people) of the 25 are in this category.
**b.** .2 + .12 + .04 = .36
**c.** Note that 8% purchased 0 while 4% purchased 5. Thus, 1 – .08 – .04 = .88 purchased between 1 and 4 quarts.

**1.7** **a.** There is a possibility of bimodality in the distribution.
**b.** There is a dip in heights at 68 inches.
**c.** If all of the students are roughly the same age, the bimodality could be a result of the men/women distributions.




**1.8** **a.** The histogram is above.
**b.** The data appears to be bimodal. Llanederyn and Caldicot have lower sample values than the other two.

**1.9** **a.** Note that 9.7 = 12 – 2.3 and 14.3 = 12 + 2.3. So, (9.7, 14.3) should contain approximately 68% of the values.
**b.** Note that 7.4 = 12 – 2(2.3) and 16.6 = 12 + 2(2.3). So, (7.4, 16.6) should contain approximately 95% of the values.
**c.** From parts (a) and (b) above, 95% - 68% = 27% lie in both (14.3. 16.6) and (7.4, 9.7). By symmetry, 13.5% should lie in (14.3, 16.6) so that 68% + 13.5% = 81.5% are in (9.7, 16.6)
**d.** Since 5.1 and 18.9 represent three standard deviations away from the mean, the proportion outside of these limits is approximately 0.

**1.10** **a.** 14 – 17 = -3.
**b.** Since 68% lie within one standard deviation of the mean, 32% should lie outside. By symmetry, 16% should lie below one standard deviation from the mean.
**c.** If normally distributed, approximately 16% of people would spend less than –3 hours on the internet. Since this doesn't make sense, the population is not normal.

**1.11** **a.** $\sum_{i=1}^{n} c = c + c + \ldots + c = nc$.

**b.** $\sum_{i=1}^{n} cy_i = c(y_1 + \ldots + y_n) = c\sum_{i=1}^{n} y_i$

**c.** $\sum_{i=1}^{n} (x_i + y_i) = x_1 + y_1 + x_2 + y_2 + \ldots + x_n + y_n = (x_1 + x_2 + \ldots + x_n) + (y_1 + y_2 + \ldots + y_n)$

Using the above, the numerator of $s^2$ is $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i^2 - 2y_i\bar{y} + \bar{y}^2) = \sum_{i=1}^{n} y_i^2 -$ $2\bar{y}\sum_{i=1}^{n} y_i + n\bar{y}^2$ Since $n\bar{y} = \sum_{i=1}^{n} y_i$, we have $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$. Let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ to get the result.

**1.12** Using the data, $\sum_{i=1}^{6} y_i = 14$ and $\sum_{i=1}^{6} y_i^2 = 40$. So, $s^2 = (40 - 14^2/6)/5 = 1.47$. So, $s = 1.21$.

**1.13** **a.** With $\sum_{i=1}^{45} y_i = 440.6$ and $\sum_{i=1}^{45} y_i^2 = 5067.38$, we have that $\bar{y} = 9.79$ and $s = 4.14$.

**b.**

| *k* | interval | frequency | Exp. frequency |
|---|---|---|---|
| 1 | 5.65, 13.93 | 44 | 30.6 |
| 2 | 1.51, 18.07 | 44 | 42.75 |
| 3 | -2.63, 22.21 | 44 | 45 |

**1.14** **a.** With $\sum_{i=1}^{25} y_i = 80.63$ and $\sum_{i=1}^{25} y_i^2 = 500.7459$, we have that $\bar{y} = 3.23$ and $s = 3.17$.

**b.**

| *k* | interval | frequency | Exp. frequency |
|---|---|---|---|
| 1 | 0.063, 6.397 | 21 | 17 |
| 2 | -3.104, 9.564 | 23 | 23.75 |
| 3 | -6.271, 12.731 | 25 | 25 |

**1.15** **a.** With $\sum_{i=1}^{40} y_i = 175.48$ and $\sum_{i=1}^{40} y_i^2 = 906.4118$, we have that $\bar{y} = 4.39$ and $s = 1.87$.

**b.**

| *k* | interval | frequency | Exp. frequency |
|---|---|---|---|
| 1 | 2.52, 6.26 | 35 | 27.2 |
| 2 | 0.65, 8.13 | 39 | 38 |
| 3 | -1.22, 10 | 39 | 40 |

**1.16** **a.** Without the extreme value, $\bar{y} = 4.19$ and $s = 1.44$.
**b.** These counts compare more favorably:

| *k* | interval | frequency | Exp. frequency |
|---|---|---|---|
| 1 | 2.75, 5.63 | 25 | 26.52 |
| 2 | 1.31, 7.07 | 36 | 37.05 |
| 3 | -0.13, 8.51 | 39 | 39 |

**1.17** For Ex. 1.2, range/4 = 7.35, while $s$ = 4.14. For Ex. 1.3, range/4 = 3.04, while = s = 3.17. For Ex. 1.4, range/4 = 2.32, while s = 1.87.

**1.18** The approximation is (800–200)/4 = 150.

**1.19** One standard deviation below the mean is 34 – 53 = –19. The empirical rule suggests that 16% of all measurements should lie one standard deviation below the mean. Since chloroform measurements cannot be negative, this population cannot be normally distributed.

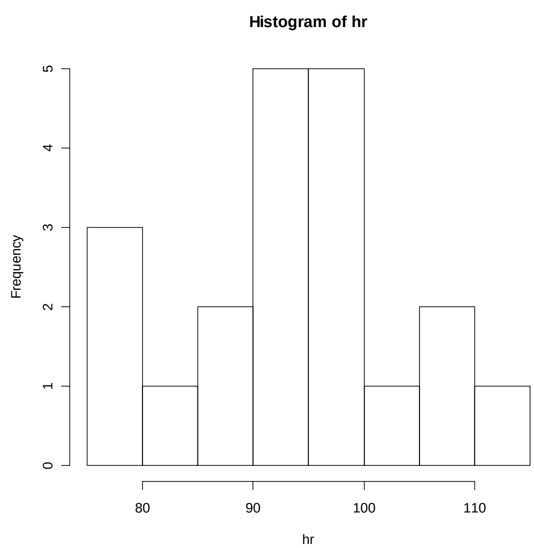
**1.20** Since approximately 68% will fall between \$390 (\$420 – \$30) to \$450 (\$420 + \$30), the proportion above \$450 is approximately 16%.

**1.21** (Similar to exercise 1.20) Having a gain of more than 20 pounds represents all measurements greater than one standard deviation below the mean. By the empirical rule, the proportion above this value is approximately 84%, so the manufacturer is probably correct.

**1.22** (See exercise 1.11) $\sum_{i=1}^{n}(y_i - \overline{y}) = \sum_{i=1}^{n} y_i - n\overline{y} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i = 0$.

**1.23** **a.** (Similar to exercise 1.20) 95 sec = 1 standard deviation above 75 sec, so this percentage is 16% by the empirical rule.
**b.** (35 sec., 115 sec) represents an interval of 2 standard deviations about the mean, so approximately 95%
**c.** 2 minutes = 120 sec = 2.5 standard deviations above the mean. This is unlikely.

**1.24** **a.** (112-78)/4 = 8.5




**b.** The histogram is above.

**c.** With $\sum_{i=1}^{20} y_i$ = 1874.0 and $\sum_{i=1}^{20} y_i^2$ = 117,328.0, we have that $\overline{y}$ = 93.7 and $s$ = 9.55.

**d.**

| $k$ | interval | frequency | Exp. frequency |
|---|---|---|---|
| 1 | 84.1, 103.2 | 13 | 13.6 |
| 2 | 74.6, 112.8 | 20 | 19 |
| 3 | 65.0, 122.4 | 20 | 20 |

**1.25** **a.** (716-8)/4 = 177
**b.** The figure is omitted.
**c.** With $\sum_{i=1}^{88} y_i$ = 18,550 and $\sum_{i=1}^{88} y_i^2$ = 6,198,356, we have that $\bar{y}$ = 210.8 and $s$ = 162.17.
**d.**

| $k$ | interval | frequency | Exp. frequency |
|---|---|---|---|
| 1 | 48.6, 373 | 63 | 59.84 |
| 2 | -113.5, 535.1 | 82 | 83.6 |
| 3 | -275.7, 697.3 | 87 | 88 |

**1.26** For Ex. 1.12, 3/1.21 = 2.48. For Ex. 1.24, 34/9.55 = 3.56. For Ex. 1.25, 708/162.17 = 4.37. The ratio increases as the sample size increases.

**1.27** (64, 80) is one standard deviation about the mean, so 68% of 340 or approx. 231 scores. (56, 88) is two standard deviations about the mean, so 95% of 340 or 323 scores.

**1.28** (Similar to 1.23) 13 mg/L is one standard deviation below the mean, so 16%.

**1.29** If the empirical rule is assumed, approximately 95% of all bearing should lie in (2.98, 3.02) – this interval represents two standard deviations about the mean. So, approximately 5% will lie outside of this interval.

**1.30** If $\mu = 0$ and $\sigma = 1.2$, we expect 34% to be between 0 and 0 + 1.2 = 1.2. Also, approximately 95%/2 = 47.5% will lie between 0 and 2.4. So, 47.5% – 34% = 13.5% should lie between 1.2 and 2.4.

**1.31** Assuming normality, approximately 95% will lie between 40 and 80 (the standard deviation is 10). The percent below 40 is approximately 2.5% which is relatively unlikely.

**1.32** For a sample of size $n$, let $n'$ denote the number of measurements that fall outside the interval $\bar{y} \pm ks$, so that $(n - n')/n$ is the fraction that falls inside the interval. To show this fraction is greater than or equal to $1 - 1/k^2$, note that

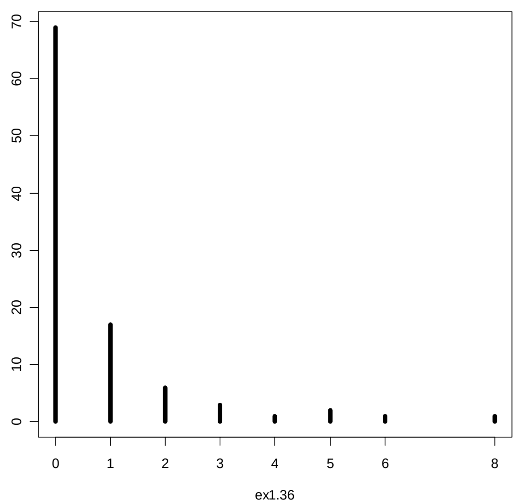
$$(n-1)s^2 = \sum_{i\in A}(y_i - \bar{y})^2 + \sum_{i\in b}(y_i - \bar{y})^2 \text{, (both sums must be positive)}$$

where $A = \{i\colon |y_i - \bar{y}| \geq ks\}$ and $B = \{i\colon |y_i - \bar{y}| < ks\}$. We have that $\sum_{i\in A}(y_i - \bar{y})^2 \geq \sum_{i\in A} k^2 s^2 = n'k^2s^2$, since if $i$ is in $A$, $|y_i - \bar{y}| \geq ks$ and there are $n'$ elements in $A$. Thus, we have that $s^2 \geq k^2s^2n'/(n-1)$, or $1 \geq k^2n'/(n-1) \geq k^2n'/n$. Thus, $1/k^2 \geq n'/n$ or $(n - n')/n \geq 1 - 1/k^2$.

**1.33** With $k$ =2, at least 1 – 1/4 = 75% should lie within 2 standard deviations of the mean. The interval is (0.5, 10.5).

**1.34** The point 13 is 13 – 5.5 = 7.5 units above the mean, or 7.5/2.5 = 3 standard deviations above the mean. By Tchebysheff's theorem, at least $1 - 1/3^2 = 8/9$ will lie within 3 standard deviations of the mean. Thus, at most 1/9 of the values will exceed 13.

**1.35** **a.** (172 – 108)/4 =16
**b.** With $\sum_{i=1}^{15} y_i = 2041$ and $\sum_{i=1}^{15} y_i^2 = 281{,}807$ we have that $\bar{y} = 136.1$ and $s = 17.1$
**c.** $a = 136.1 - 2(17.1) = 101.9$, $b = 136.1 + 2(17.1) = 170.3$.
**d.** There are 14 observations contained in this interval, and 14/15 = 93.3%. 75% is a lower bound.




**1.36** **a.** The histogram is above.
**b.** With $\sum_{i=1}^{100} y_i = 66$ and $\sum_{i=1}^{100} y_i^2 = 234$ we have that $\bar{y} = 0.66$ and $s = 1.39$.
**c.** Within two standard deviations: 95, within three standard deviations: 96. The calculations agree with Tchebysheff's theorem.

**1.37** Since the lead readings must be non negative, 0 (the smallest possible value) is only 0.33 standard deviations from the mean. This indicates that the distribution is skewed.

**1.38** By Tchebysheff's theorem, at least 3/4 = 75% lie between (0, 140), at least 8/9 lie between (0, 193), and at least 15/16 lie between (0, 246). The lower bounds are all truncated a 0 since the measurement cannot be negative.