Measures of Variability.

**Range.** = maximum data value - minimum data value.

cookies: including outlier: 69 - 14

\[
\text{range} = 55
\]

excluding outlier: 42 - 14

\[
\text{range} = 28
\]

Single outlier can have a large effect.

Standard Deviation + Variance.

How far from the mean is a typical datapoint?

How spread out are typical datapoints?

\[
\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}
\]

Population standard deviation.
Variance  \[ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \]

Standard deviation is in the same units as the data (inches, kg, # chips, etc) - more interpretable than the variance (which is mathematically convenient).

Aug 2005 tropical storms
max wind speed. 40, 50, 65, 105, 175 mph.
(Range 175 - 40 = 135 mph).

These are measurements of a population (all the storms).

\[ \mu = \frac{\sum x_i}{n} = 87. \]

\[ \sigma^2 = \frac{1}{5} \left[ (40-87)^2 + (50-87)^2 + (65-87)^2 + (105-87)^2 + (175-87)^2 \right] \]

\[ = \frac{1}{5} \left[ (-47)^2 + (-37)^2 + (-22)^2 + (18)^2 + (88)^2 \right] \]

\[ = \frac{1}{5} \left[ 2209 + 1369 + 484 + 324 + 7744 \right] \]

\[ = \frac{1}{5} \left[ 12130 \right] = 2426. \text{ mph}^2 \]

\[ \sigma = 49.25. \]
Histogram of #Chips per cookie

mean = 27.7
standard deviation = 6.3

Frequency

10 20 30 40 50 60 70

x$V3

outlier
Histogram of #Chips per cookie

mean = 27.5

SD = 5.64
Average

Typical storm has max wind speed which is 50 mph different from the mean.

Sample standard deviation.

\[ S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \]
The histogram shows the distribution of chips with the following details:

- **Mean**: \( \bar{x} = 27.5 \)
- **Standard Deviation**: \( \sigma = 5.64 \)

95% of the chips lie within the interval (16.2 to 38.8). This is calculated as follows:

- The interval from \( \bar{x} - 2\sigma \) to \( \bar{x} + 2\sigma \) contains 68% of the data.
- The interval from \( \bar{x} - \sigma \) to \( \bar{x} + \sigma \) contains 95% of the data.

12 out of 213 chips lie outside the interval, and 94.4% are within 2\(\sigma\) of the mean.
Histogram of chips

\[ \bar{x} = 21.7 \]
\[ \sigma = 6.3 \]
\[ x \pm 2\sigma \]
\[ \frac{2.06}{2.14} = 96.3\% \]
\[ \text{within } x \pm 2\sigma \]
For data with a "bell shaped" distribution

**Empirical rule.**

About 68% of the data will be within one s.d. of the mean.

About 95% of the data will be within two s.d. of the mean.

---

**Standardized Scores.**

- putting different data sets on the same scale.

\[ z_i = \frac{x_i - \mu}{\sigma} \]  
\[ z_i = \frac{x_i - \bar{x}}{s} \]

\( z \) in dimensionless.

Eg. comparing hurricanes over different time periods

- 2005 Katrina: \[ \frac{175 - 87.5}{49.25} = 1.79 \]
- 2004 Karl: \[ \frac{145 - 90.6}{38.28} = 1.42 \]
Even after adjusting for the increased variability in 2005, Kahina stands out as more extreme.

eq: SAT scores are rescaled to adjust for varying difficulty of the exam.

\[ Z < -2 \text{ or } Z > 2 \text{ is unusual} \leq 5\% \text{ of the observations} \]

---

**Percentiles + Quartiles.**

Median - separates the data into two halves.

- 50% are below
- 50% are above

Median is the 50th percentile.

Generalize this:

10% of the data is below 1st percentile value
20%
30%
40%

---

etc.
To find the $k^{th}$ percentile:

1) sort the data.

2) compute $L = \frac{k}{100} \times n$

3) if $L$ is a whole number, take the mean of the $L^{th}$ and $(L+1)^{th}$ value.

b) otherwise, round $L$ up to the next whole number, and take that value.

2005 storms: 40, 50, 65, 105, 125

25$^{th}$ percentile.

$L = \frac{25}{100} \times 5 = \frac{5}{4} = 1.25$

round up to 2.

25$^{th}$ percentile wind speed is 50 mph

(2nd entry on sorted list).

2004 storms 40, 45, 65, 70, 105, 120, 135, 145

25$^{th}$ percentile

$L = \frac{25}{100} \times 8 = \frac{8}{4} = 2.$

25$^{th}$ percentile value is $\frac{45+65}{2} = 55$.
25th percentile is called 1st Quartile $Q_1$.
75th

3rd Quartile $Q_3$. Inter-quartile range

$IO = Q_3 - Q_1$ - measure of the spread in the data.

<table>
<thead>
<tr>
<th>Cookie data.</th>
<th>$Q_1$</th>
<th>$Q_2$</th>
<th>$Q_3$ (median)</th>
</tr>
</thead>
<tbody>
<tr>
<td>with outliers</td>
<td>24</td>
<td>28</td>
<td>31</td>
</tr>
<tr>
<td>without</td>
<td>24</td>
<td>28</td>
<td>31</td>
</tr>
</tbody>
</table>

Box plot.

**Exploratory Data Analysis.**

- center
- variation
- distribution shape
- outliers
- change over time.
# chips in cookies.
# chips excluding the outlier (69).
Hurricanes.

Easy visual comparison of the distributions of many data sets.
Probability -

What is the probability of H when flipping a fair coin?

What is the probability of rolling 10 on two dice?

What is the probability of Sarah Palin being elected the next president?

What is the probability that the person sitting next to you will say yes if you ask them on a date?

What is the probability that it will rain tomorrow?
Event.

A simple event is a possible outcome. e.g. H, T

The sample space is the set of all possible simple events

e.g. \{H, T\}
\{1, 2, 3, \ldots, 11, 12\}
\{Palin elected, not elected\}
\{rain, no rain\}

Event - a set of simple events.

e.g. \{H\}
\{10\}
\{roll an even number\}

Classical probability.

If all simple events are equally likely, each has probability \(\frac{1}{n}\)

If an event \(A\) is a set of \(S\) simple events, then it has probability \(P(A) = \frac{S}{n}\)
\[
P(H) = \frac{1}{2},
\]
\[
p(\text{roll 5 on one dice}) = \frac{1}{6}
\]
\[
p(\text{roll a total of 4 on two dice}) = \frac{3}{36} = \frac{1}{12}.
\]
- enumerate all the simple events:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
P(2H \text{ on 3 coin flips}) = \frac{3}{8}
\]
Relative frequency definition
- prob. of an event in the long-run average fraction of times it occurs.

Subjective (Bayesian) definition.
- probability is derived from our personal knowledge + perspective