Estimating parameters of the normal distribution \((\mu, \sigma)\) or the binomial distribution \((p)\) from data

We will assume throughout that the SAT math test was designed to have a normal distribution.

Secretly, \(\mu = 500\) and \(\sigma = 100\), but we don’t know those are the values so we want to estimate them from data.

- **Chapter 5.3:** Pretend we know \(\sigma\) but not \(\mu\) and we want to estimate \(\mu\) from experimental data.
- **Chapter 5.4:** Estimate both \(\mu\) and \(\sigma\) from experimental data.
5.3 Estimating parameters from data

Basic experiment

1. Sample $n$ random students from the whole population of SAT takers. The scores of these students are $x_1, \ldots, x_n$.

2. Compute the **sample mean** of these scores:

   $$ m = \bar{x} = \frac{x_1 + \cdots + x_n}{n} $$

   The sample mean is a **point estimate** of $\mu$; it just gives one number, without an indication of how far away it might be from $\mu$.

3. Repeat the above with many independent samples, getting different sample means each time.

The long-term average of the sample means will be approximately

$$ E(\bar{X}) = E \left( \frac{X_1 + \cdots + X_n}{n} \right) = \frac{\mu + \cdots + \mu}{n} = \frac{n\mu}{n} = \mu $$

These estimates will be distributed with variance $\text{Var}(\bar{X}) = \sigma^2/n$. 
## Sample data

<table>
<thead>
<tr>
<th>Trial #</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$m = \bar{x}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>720</td>
<td>490</td>
<td>660</td>
<td>520</td>
<td>390</td>
<td>390</td>
<td>528.33</td>
</tr>
<tr>
<td>2</td>
<td>380</td>
<td>260</td>
<td>390</td>
<td>630</td>
<td>540</td>
<td>440</td>
<td>440.00</td>
</tr>
<tr>
<td>3</td>
<td>800</td>
<td>450</td>
<td>580</td>
<td>520</td>
<td>650</td>
<td>390</td>
<td>565.00</td>
</tr>
<tr>
<td>4</td>
<td>510</td>
<td>370</td>
<td>530</td>
<td>290</td>
<td>460</td>
<td>540</td>
<td>450.00</td>
</tr>
<tr>
<td>5</td>
<td>580</td>
<td>500</td>
<td>540</td>
<td>540</td>
<td>340</td>
<td>340</td>
<td>473.33</td>
</tr>
<tr>
<td>6</td>
<td>500</td>
<td>490</td>
<td>480</td>
<td>550</td>
<td>390</td>
<td>450</td>
<td>476.67</td>
</tr>
<tr>
<td>7</td>
<td>530</td>
<td>680</td>
<td>540</td>
<td>510</td>
<td>520</td>
<td>590</td>
<td>561.67</td>
</tr>
<tr>
<td>8</td>
<td>480</td>
<td>600</td>
<td>520</td>
<td>600</td>
<td>520</td>
<td>390</td>
<td>518.33</td>
</tr>
<tr>
<td>9</td>
<td>340</td>
<td>520</td>
<td>500</td>
<td>650</td>
<td>400</td>
<td>530</td>
<td>490.00</td>
</tr>
<tr>
<td>10</td>
<td>460</td>
<td>450</td>
<td>500</td>
<td>360</td>
<td>600</td>
<td>440</td>
<td>468.33</td>
</tr>
<tr>
<td>11</td>
<td>540</td>
<td>520</td>
<td>360</td>
<td>500</td>
<td>520</td>
<td>640</td>
<td>513.33</td>
</tr>
<tr>
<td>12</td>
<td>440</td>
<td>420</td>
<td>610</td>
<td>530</td>
<td>490</td>
<td>570</td>
<td>510.00</td>
</tr>
<tr>
<td>13</td>
<td>520</td>
<td>570</td>
<td>430</td>
<td>320</td>
<td>650</td>
<td>540</td>
<td>505.00</td>
</tr>
<tr>
<td>14</td>
<td>560</td>
<td>380</td>
<td>440</td>
<td>610</td>
<td>680</td>
<td>460</td>
<td>521.67</td>
</tr>
<tr>
<td>15</td>
<td>460</td>
<td>590</td>
<td>350</td>
<td>470</td>
<td>420</td>
<td>740</td>
<td>505.00</td>
</tr>
<tr>
<td>16</td>
<td>430</td>
<td>490</td>
<td>370</td>
<td>350</td>
<td>360</td>
<td>470</td>
<td>411.67</td>
</tr>
<tr>
<td>17</td>
<td>570</td>
<td>610</td>
<td>460</td>
<td>410</td>
<td>550</td>
<td>510</td>
<td>518.33</td>
</tr>
<tr>
<td>18</td>
<td>380</td>
<td>540</td>
<td>570</td>
<td>400</td>
<td>360</td>
<td>500</td>
<td>458.33</td>
</tr>
<tr>
<td>19</td>
<td>410</td>
<td>730</td>
<td>480</td>
<td>600</td>
<td>270</td>
<td>320</td>
<td>468.33</td>
</tr>
<tr>
<td>20</td>
<td>490</td>
<td>390</td>
<td>450</td>
<td>610</td>
<td>320</td>
<td>440</td>
<td>450.00</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>491.67</strong></td>
</tr>
</tbody>
</table>
Sample mean notation

## Variable names

<table>
<thead>
<tr>
<th>Actual distribution (Greek letters)</th>
<th>Point estimate from a sample (Latin letters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X ): random variable</td>
<td>( x_1, \ldots, x_n ): sample data</td>
</tr>
<tr>
<td>( \mu ): mean</td>
<td>( m ) or ( \bar{x} ): sample mean</td>
</tr>
<tr>
<td>( \sigma^2 ): variance</td>
<td>( s^2 ): sample variance</td>
</tr>
<tr>
<td>( \sigma ): standard deviation</td>
<td>( s ): sample standard deviation</td>
</tr>
</tbody>
</table>

### Lowercase/Uppercase

- **Lowercase**: Given specific numbers \( x_1, \ldots, x_n \), the sample mean evaluates to a number as well.

- **Uppercase**: We will study performing this computation repeatedly with different data, treating the data \( X_1, \ldots, X_n \) as random variables. This makes the sample mean a random variable.

\[
m = \bar{x} = \frac{x_1 + \cdots + x_n}{n} \quad M = \bar{X} = \frac{X_1 + \cdots + X_n}{n}
\]
Z-scores

- How often is the sample mean “close” to the secret value of $\mu$?

- The sample mean is a random variable $\overline{X}$ with mean $E(\overline{X}) = \mu$ and standard deviation $SD(\overline{X}) = \sigma/\sqrt{n}$. So

  $$z = \frac{m - \mu}{\sigma/\sqrt{n}}$$

  if we knew secret: $= \frac{m - 500}{100/\sqrt{n}}$

- Exclude the top 2.5% and bottom 2.5% of values of $Z$ and regard the middle 95% as “close.” So

  $$P(-z_{0.025} \leq Z \leq z_{0.025}) = P(-1.96 \leq Z \leq 1.96) = .95$$
Confidence intervals

- We will rearrange this equation to isolate \( \mu \):
  \[
P(-1.96 \leq Z \leq 1.96) = P(-1.96 \leq \frac{M - \mu}{\sigma/\sqrt{n}} \leq 1.96) = .95
  \]

- **Interpretation:** in \( \approx 95\% \) of the trials of this experiment, the value \( M = m \) satisfies
  \[
  -1.96 \leq \frac{m - \mu}{\sigma/\sqrt{n}} \leq 1.96
  \]

- Solve for bounds on \( \mu \) from the upper limit on \( Z \):
  \[
  \frac{m - \mu}{\sigma/\sqrt{n}} \leq 1.96 \iff m - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}} \iff m - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu
  \]
  Notice the 1.96 turned into \(-1.96\) and we get a lower limit on \( \mu \).

- Also solve for an upper bound on \( \mu \) from the lower limit on \( Z \):
  \[
  -1.96 \leq \frac{m - \mu}{\sigma/\sqrt{n}} \iff -1.96 \frac{\sigma}{\sqrt{n}} \leq m - \mu \iff \mu \leq m + 1.96 \frac{\sigma}{\sqrt{n}}
  \]

- Together,
  \[
  m - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq m + 1.96 \frac{\sigma}{\sqrt{n}}
  \]
Confidence intervals

In $\approx 95\%$ of the trials of this experiment, the value $M = m$ satisfies

$$m - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq m + 1.96 \frac{\sigma}{\sqrt{n}}$$

So, 95\% of the time we perform this experiment, the true mean $\mu$ is in the interval

$$\left(m - 1.96 \frac{\sigma}{\sqrt{n}}, m + 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

which is called a (two-sided) 95\% confidence interval.

For a $100(1 - \alpha)\%$ C.I., use $\pm z_{\alpha/2}$ instead of $\pm 1.96$.

Other commonly used percentages:
For a 99\% confidence interval, use $\pm 2.58$ instead of $\pm 1.96$.
For a 90\% confidence interval, use $\pm 1.64$ instead of $\pm 1.96$.

For demo purposes:
For a 75\% confidence interval, use $\pm 1.15$ instead of $\pm 1.96$. 
Example: Six scores 380, 260, 390, 630, 540, 440

Sample mean: \[ m = \frac{380 + 260 + 390 + 630 + 540 + 440}{6} = 440 \]

\[ \sigma: \text{ We assumed } \sigma = 100 \text{ at the beginning } \]

95% CI half-width: \[ 1.96 \frac{\sigma}{\sqrt{n}} = \frac{(1.96)(100)}{\sqrt{6}} \approx 80.02 \]

95% CI: 
\[ (440 - 80.02, 440 + 80.02) = (359.98, 520.02) \]

Has the true mean, \( \mu = 500 \).

75% CI half-width: \[ 1.15 \frac{\sigma}{\sqrt{n}} = \frac{(1.15)(100)}{\sqrt{6}} \approx 46.95 \]

75% CI: 
\[ (440 - 46.95, 440 + 46.95) = (393.05, 486.95) \]

Doesn’t have the true mean, \( \mu = 500 \).
Confidence intervals

$\sigma = 100$ known, $\mu = 500$ unknown, $n = 6$ points per trial, 20 trials

Confidence intervals not containing point $\mu = 500$ are marked [93.05,486.95]*.

<table>
<thead>
<tr>
<th>Trial #</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$m = \bar{x}$</th>
<th>75% conf. int.</th>
<th>95% conf. int.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>720</td>
<td>490</td>
<td>660</td>
<td>520</td>
<td>390</td>
<td>390</td>
<td>528.33</td>
<td>(481.38,575.28)</td>
<td>(448.32,608.35)</td>
</tr>
<tr>
<td>2</td>
<td>380</td>
<td>260</td>
<td>390</td>
<td>630</td>
<td>540</td>
<td>440</td>
<td>440.00</td>
<td><em>(393.05,486.95)</em></td>
<td><em>(359.98,520.02)</em></td>
</tr>
<tr>
<td>3</td>
<td>800</td>
<td>450</td>
<td>580</td>
<td>520</td>
<td>650</td>
<td>390</td>
<td>565.00</td>
<td><em>(518.05,611.95)</em></td>
<td><em>(484.98,645.02)</em></td>
</tr>
<tr>
<td>4</td>
<td>510</td>
<td>370</td>
<td>530</td>
<td>290</td>
<td>460</td>
<td>540</td>
<td>450.00</td>
<td><em>(403.05,496.95)</em></td>
<td><em>(369.98,530.02)</em></td>
</tr>
<tr>
<td>5</td>
<td>580</td>
<td>500</td>
<td>540</td>
<td>340</td>
<td>340</td>
<td>340</td>
<td>473.33</td>
<td>(426.38,520.28)</td>
<td>(393.32,553.35)</td>
</tr>
<tr>
<td>6</td>
<td>500</td>
<td>490</td>
<td>480</td>
<td>550</td>
<td>390</td>
<td>450</td>
<td>476.67</td>
<td>(429.72,523.62)</td>
<td>(396.65,556.68)</td>
</tr>
<tr>
<td>7</td>
<td>530</td>
<td>680</td>
<td>540</td>
<td>510</td>
<td>520</td>
<td>590</td>
<td>561.67</td>
<td><em>(514.72,608.62)</em></td>
<td><em>(481.65,641.68)</em></td>
</tr>
<tr>
<td>8</td>
<td>480</td>
<td>600</td>
<td>520</td>
<td>600</td>
<td>520</td>
<td>390</td>
<td>518.33</td>
<td>(471.38,565.28)</td>
<td>(438.32,598.35)</td>
</tr>
<tr>
<td>9</td>
<td>340</td>
<td>520</td>
<td>500</td>
<td>650</td>
<td>400</td>
<td>530</td>
<td>490.00</td>
<td>(443.05,536.95)</td>
<td>(409.98,570.02)</td>
</tr>
<tr>
<td>10</td>
<td>460</td>
<td>450</td>
<td>500</td>
<td>360</td>
<td>600</td>
<td>440</td>
<td>468.33</td>
<td>(421.38,515.28)</td>
<td>(388.32,548.35)</td>
</tr>
<tr>
<td>11</td>
<td>540</td>
<td>520</td>
<td>360</td>
<td>500</td>
<td>520</td>
<td>640</td>
<td>513.33</td>
<td>(466.38,560.28)</td>
<td>(433.32,593.35)</td>
</tr>
<tr>
<td>12</td>
<td>440</td>
<td>420</td>
<td>610</td>
<td>530</td>
<td>490</td>
<td>570</td>
<td>510.00</td>
<td>(463.05,556.95)</td>
<td>(429.98,590.02)</td>
</tr>
<tr>
<td>13</td>
<td>520</td>
<td>570</td>
<td>430</td>
<td>320</td>
<td>650</td>
<td>540</td>
<td>505.00</td>
<td>(458.05,551.95)</td>
<td>(424.98,585.02)</td>
</tr>
<tr>
<td>14</td>
<td>560</td>
<td>380</td>
<td>440</td>
<td>610</td>
<td>680</td>
<td>460</td>
<td>521.67</td>
<td>(474.72,568.62)</td>
<td>(441.65,601.68)</td>
</tr>
<tr>
<td>15</td>
<td>460</td>
<td>590</td>
<td>350</td>
<td>470</td>
<td>420</td>
<td>740</td>
<td>505.00</td>
<td>(458.05,551.95)</td>
<td>(424.98,585.02)</td>
</tr>
<tr>
<td>16</td>
<td>430</td>
<td>490</td>
<td>370</td>
<td>350</td>
<td>360</td>
<td>470</td>
<td>411.67</td>
<td><em>(364.72,458.62)</em></td>
<td><em>(331.65,491.68)</em></td>
</tr>
<tr>
<td>17</td>
<td>570</td>
<td>610</td>
<td>460</td>
<td>410</td>
<td>550</td>
<td>510</td>
<td>518.33</td>
<td>(471.38,565.28)</td>
<td>(438.32,598.35)</td>
</tr>
<tr>
<td>18</td>
<td>380</td>
<td>540</td>
<td>570</td>
<td>400</td>
<td>360</td>
<td>500</td>
<td>458.33</td>
<td>(411.38,505.28)</td>
<td>(378.32,538.35)</td>
</tr>
<tr>
<td>19</td>
<td>410</td>
<td>730</td>
<td>480</td>
<td>600</td>
<td>270</td>
<td>320</td>
<td>468.33</td>
<td>(421.38,515.28)</td>
<td>(388.32,548.35)</td>
</tr>
<tr>
<td>20</td>
<td>490</td>
<td>390</td>
<td>450</td>
<td>610</td>
<td>320</td>
<td>440</td>
<td>450.00</td>
<td><em>(403.05,496.95)</em></td>
<td><em>(369.98,530.02)</em></td>
</tr>
</tbody>
</table>
Confidence intervals

$\sigma = 100$ known, $\mu = 500$ unknown, $n = 6$ points per trial, 20 trials

- In the 75% confidence interval column, 14 out of 20 (70%) intervals contain the mean ($\mu = 500$).
  This is close to 75%.

- In the 95% confidence interval column, 19 out of 20 (95%) intervals contain the mean ($\mu = 500$).
  This is exactly 95% (though if you do it 20 more times, it wouldn’t necessarily be exactly 19 the next time).

- A $k\%$ confidence interval means if we repeat the experiment a lot of times, *approximately* $k\%$ of the intervals will contain $\mu$.
  It is *not* a guarantee that exactly $k\%$ will contain it.

- *Note*: If you really don’t know the true value of $\mu$, you can’t actually mark the intervals that do or don’t contain it.
Confidence intervals: choosing $n$

- For a smaller width 95% confidence interval, increase $n$.
- For example, to make the 95% confidence interval be $(m - 10, m + 10)$ or smaller, we need

$$1.96\sigma / \sqrt{n} \leq 10$$

so

$$\sqrt{n} \geq 1.96\sigma / 10 = 1.96(100)/10 = 19.6$$

$$n \geq 19.6^2 = 384.16$$

$$n \geq 385$$
One-sided confidence intervals

- In a two-sided 95% confidence interval, we excluded the highest and lowest 2.5% of values and keep the middle 95%. One-sided removes the whole 5% from one side.

- **One-sided to the right:** remove the highest (right) 5% values of $Z$

\[
P(Z \leq z_{0.05}) = P(Z \leq 1.64) = .95
\]

95% of experiments have
\[
\frac{m - \mu}{\sigma/\sqrt{n}} \leq 1.64 \quad \text{so} \quad \mu \geq m - 1.64 \frac{\sigma}{\sqrt{n}}
\]

So the one-sided (right) 95% CI for $\mu$ is $(m - 1.64 \frac{\sigma}{\sqrt{n}}, \infty)$

- **One-sided to the left:** remove lowest (left) 5% of values of $Z$

\[
P(-z_{0.05} \leq Z) = P(-1.64 \leq Z) = .95
\]

The one-sided (left) 95% CI for $\mu$ is $(-\infty, m + 1.64 \frac{\sigma}{\sqrt{n}})$
5.3 Confidence intervals for $p$ in the binomial distribution

- An election has two options, $A$ and $B$.
- There are no other options and no write-ins.
- In the election: $p$ is the fraction of votes cast for $A$, $1 - p$ is the fraction of votes cast for $B$.

In a poll beforehand: $\hat{p}$ is the fraction polled who say they’ll vote for $A$.

- A single point estimate of $p$ is denoted $\hat{p}$.
  We also want a 95% confidence interval for it.
- We model this by sampling from an urn
  - without replacement (hypergeometric distribution)
  - or with replacement (binomial distribution).

However, as previously discussed, this an imperfect model for a poll (sample may not be representative; sample may have non-voters; people may change their minds after the poll; etc.)
Estimating $p$ for a poll with binomial distribution

A poll should use the hypergeometric distribution (sampling without replacement), but we approximate it by the binomial distribution (sampling with replacement).

Let $p$ be the fraction of votes for $A$ out of all votes. The probability $k$ out of $n$ in the sample say they’ll vote for $A$ is

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$ 

The fraction of people polled who say they’ll vote for $A$ is $\hat{P} = \bar{X} = X/n$, with $E(\bar{X}) = p$ and $\text{Var}(\bar{X}) = p(1 - p)/n$.

The $\hat{}$ (caret) notation indicates it’s a point estimate. We already use $P$ for too many things, so we’ll use the $\bar{X}$ notation.
Estimating $p$

**Point estimate of $p$**

Poll 1000 people out of a much larger population.
Get 700 voting for $A$, 300 for $B$.
A point estimate of $p$ (the fraction voting for $A$) is
$$\hat{p} = \frac{700}{1000} = .7$$

**Interval estimate of $p$**

- We could get a 95% confidence interval for $p$ by using the formula
  $$(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}) = \left(\hat{p} - 1.96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}, \hat{p} + 1.96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$
  where we plugged in $\bar{x} = \hat{p}$ and $\sigma = SD(X_i) = \sqrt{p(1-p)}$.

- But that involves $p$, which is unknown! We’ll use two methods to deal with that. First, estimate $p$ by $\hat{p}$ in the SD to get
  $$(\hat{p} - 1.96 \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + 1.96 \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}})$$
  as an approximate 95% confidence interval for $p$.

- For $\hat{p} = .7$, we get $\sqrt{\hat{p}(1-\hat{p})}/n = \sqrt{.7(.3)}/1000 \approx .01449$. This gives 95% CI $(.7 - 1.96(.01449), .7 + 1.96(.01449)) = (.672, .728)$
Polls often report a margin of error instead of a confidence interval.

The half-width of the 95% confidence interval is \(1.96 \sqrt{p(1 - p)/n}\), and before we estimated \(p\) by the point estimate \(\hat{p}\).

The margin of error is the maximum that this half-width could be over all possible values of \(p\) \((0 \leq p \leq 1)\); this is at \(p = 1/2\), giving margin of error \(1.96 \sqrt{(1/2)(1/2)/n} = 1.96/(2 \sqrt{n})\).

Maximize \(p(1 - p)\) on \(0 \leq p \leq 1\):
\[
0 = \frac{d}{dp} (p - p^2) = 1 - 2p \text{ at } p = \frac{1}{2}
\]
\[
\frac{d^2}{dp^2} (p - p^2) = -2 < 0 \Rightarrow \text{ maximum}
\]
Interval estimate of $p$ using margin of error

- The **margin of error** is the maximum possible half-width, 
  \[ 1.96 \sqrt{\frac{1/2}{2/n}} = \frac{1.96}{2 \sqrt{n}}. \]

- With 1000 people, the margin of error is \( \frac{1.96}{2 \sqrt{1000}} \approx 0.03099 \), or about 3%. With 700 A’s, report \( \hat{p} = .70 \pm .03 \).

- A 3% margin of error means that if a large number of polls are conducted, each on 1000 people, then at least 95% of the polls will give values of \( \hat{p} \) such that the true \( p \) is between \( \hat{p} \pm 0.03 \).

- The reason it is “at least 95%” is that \( 1.96 \sqrt{\frac{p(1-p)}{n}} \leq 0.03 \) and only \( = 0.03 \) when \( p = 1/2 \) exactly.
  
  If the true \( p \) is not equal to 1/2, then \( \frac{0.03}{\sqrt{p(1-p)/n}} > 1.96 \) so it would be a higher percent confidence interval than 95%.
Choosing $n$ to get desired margin of error

- **Question:** How many people should be polled for a 2% margin of error?
- **Answer:** Solve $1.96/(2 \sqrt{n}) = .02$:

\[
n = \left(\frac{1.96}{2(0.02)}\right)^2 = 49^2 = 2401
\]

This means that if many polls are conducted, each with 2401 people, at least 95% of the polls will give values of $\hat{p}$ such that the true value of $p$ is between $\hat{p} \pm 0.02$. 
Consider data 1, 2, 12.

The sample mean is \( \bar{x} = \frac{1 + 2 + 12}{3} = 5 \).

The deviations of the data from the mean are \( x_i - \bar{x} \):
\[
1 - 5, \quad 2 - 5, \quad 12 - 5 = -4, -3, 7
\]

The deviations must sum to 0 since \( \sum_{i=1}^{n} x_i - n \bar{x} = 0 \).
Knowing any \( n - 1 \) of the deviations determines the missing one.

We say there are \( n - 1 \) degrees of freedom, or \( df = n - 1 \).

Here, there are 2 degrees of freedom, and the sum of squared deviations is
\[
ss = (-4)^2 + (-3)^2 + 7^2 = 16 + 9 + 49 = 74
\]

The **sample variance** is \( s^2 = ss/df = 74/2 = 37 \).
It is a point estimate of \( \sigma^2 \).

The **sample standard deviation** is \( s = \sqrt{s^2} = \sqrt{37} \approx 6.08 \), which is a point estimate of \( \sigma \).
Sample variance: estimating $\sigma^2$ from data

Definitions

**Sum of squared deviations:**

$$ss = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

**Sample variance:**

$$s^2 = \frac{ss}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

**Sample standard deviation:**

$$s = \sqrt{s^2}$$

- It turns out that $E(S^2) = \sigma^2$, so $s^2$ is an **unbiased estimator** of $\sigma^2$.
- For the sake of demonstration, let $u^2 = \frac{ss}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$.
  - It turns out that $E(U^2) = \frac{n-1}{n} \sigma^2$, so $u^2$ is a **biased estimator** of $\sigma^2$.
- This is because $\sum_{i=1}^{n} (x_i - \bar{x})^2$ underestimates $\sum_{i=1}^{n} (x_i - \mu)^2$. 
Estimating $\mu$ and $\sigma^2$ from sample data (secret: $\mu = 500$, $\sigma = 100$)

<table>
<thead>
<tr>
<th>Exp. #</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$\bar{x}$</th>
<th>$s^2$</th>
<th>$u^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>550</td>
<td>600</td>
<td>450</td>
<td>400</td>
<td>610</td>
<td>500</td>
<td>518.33</td>
<td>7016.67</td>
<td>5847.22</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td>520</td>
<td>370</td>
<td>520</td>
<td>480</td>
<td>440</td>
<td>471.67</td>
<td>3376.67</td>
<td>2813.89</td>
</tr>
<tr>
<td>3</td>
<td>470</td>
<td>530</td>
<td>610</td>
<td>370</td>
<td>350</td>
<td>710</td>
<td>506.67</td>
<td>19426.67</td>
<td>16188.89</td>
</tr>
<tr>
<td>4</td>
<td>630</td>
<td>620</td>
<td>430</td>
<td>470</td>
<td>500</td>
<td>470</td>
<td>520.00</td>
<td>7120.00</td>
<td>5933.33</td>
</tr>
<tr>
<td>5</td>
<td>690</td>
<td>470</td>
<td>500</td>
<td>410</td>
<td>510</td>
<td>360</td>
<td>490.00</td>
<td>12840.00</td>
<td>10700.00</td>
</tr>
<tr>
<td>6</td>
<td>450</td>
<td>490</td>
<td>500</td>
<td>380</td>
<td>530</td>
<td>680</td>
<td>505.00</td>
<td>10030.00</td>
<td>8358.33</td>
</tr>
<tr>
<td>7</td>
<td>510</td>
<td>370</td>
<td>480</td>
<td>400</td>
<td>550</td>
<td>530</td>
<td>473.33</td>
<td>5306.67</td>
<td>4422.22</td>
</tr>
<tr>
<td>8</td>
<td>420</td>
<td>330</td>
<td>540</td>
<td>460</td>
<td>630</td>
<td>390</td>
<td>461.67</td>
<td>11736.67</td>
<td>9780.56</td>
</tr>
<tr>
<td>9</td>
<td>570</td>
<td>430</td>
<td>470</td>
<td>520</td>
<td>450</td>
<td>560</td>
<td>500.00</td>
<td>3440.00</td>
<td>2866.67</td>
</tr>
<tr>
<td>10</td>
<td>260</td>
<td>530</td>
<td>330</td>
<td>490</td>
<td>530</td>
<td>630</td>
<td>461.67</td>
<td>19296.67</td>
<td>16080.56</td>
</tr>
</tbody>
</table>

Average $490.83$ $9959.00$ $8299.17$

- We used $n = 6$, repeated for 10 trials, to fit the slide. Larger values of $n$ would be better in practice.
- Average of sample means: $490.83 \approx \mu = 500$.
- Average of sample variances: $9959.00 \approx \sigma^2 = 10000$.
- $u^2$, using the wrong denominator $n = 6$ instead of $n - 1 = 5$, gave an average $8299.17 \approx \frac{n-1}{n}\sigma^2 = 8333.33$. 
Proof that denominator $n - 1$ makes $s^2$ unbiased

- Expand the $i = 1$ term of $SS = \sum_{i=1}^{n} (X_i - \bar{X})^2$:
  \[
  E((X_1 - \bar{X})^2) = E(X_1^2) + E(\bar{X}^2) - 2E(X_1\bar{X})
  \]

- $\text{Var}(X) = E(X^2) - E(X)^2 \Rightarrow E(X^2) = \text{Var}(X) + E(X)^2$. So
  \[
  E(X_1^2) = \sigma^2 + \mu^2 \quad E(\bar{X}^2) = \text{Var}(\bar{X}) + E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2
  \]

- Cross-term:
  \[
  E(X_1\bar{X}) = \frac{E(X_1^2) + E(X_1)E(X_2) + \cdots + E(X_1)E(X_n)}{n}
  \]
  \[
  = \frac{(\sigma^2 + \mu^2) + (n - 1)\mu^2}{n} = \frac{\sigma^2}{n} + \mu^2
  \]

- Total for $i = 1$ term:
  \[
  E((X_1 - \bar{X})^2) = (\sigma^2 + \mu^2) + \left(\frac{\sigma^2}{n} + \mu^2\right) - 2 \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n - 1}{n} \sigma^2
  \]
Proof that denominator \( n - 1 \) makes \( s^2 \) unbiased

Similarly, term \( i \) of \( SS = \sum_{i=1}^{n} (X_i - \bar{X})^2 \) expands to

\[
E((X_i - \bar{X})^2) = \frac{n - 1}{n} \sigma^2
\]

The total is

\[
E(SS) = (n - 1) \sigma^2
\]

Thus we must divide \( SS \) by \( n - 1 \) instead of \( n \) to get an estimate of \( \sigma^2 \) (called an \textit{unbiased estimator} of \( \sigma^2 \)).

\[
E \left( \frac{SS}{n - 1} \right) = \sigma^2
\]

If we divided by \( n \) instead, it would come out to

\[
E \left( \frac{SS}{n} \right) = \frac{n - 1}{n} \sigma^2
\]

which is called a \textit{biased estimator}.
Let $x_1, \ldots, x_n$ be $n$ data points. We already saw these formulas:

**Sample mean:**

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

**Sample variance:**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2$$

**Sample standard deviation:**

$$s = \sqrt{s^2}$$

By plugging the formula for $m$ into the formula for $s^2$ and manipulating it, it can be shown that

$$s^2 = \frac{n \left( \sum_{i=1}^{n} x_i^2 \right) - \left( \sum_{i=1}^{n} x_i \right)^2}{n(n-1)}$$

This is a useful shortcut in calculators and statistical software.
Efficient formula for sample variance

- Some calculators have a feature to let you type in a list of numbers and compute their sample mean and sample standard deviation.
- For the numbers 10, 20, 30, 40:

<table>
<thead>
<tr>
<th>n</th>
<th>$x_n$</th>
<th>$\sum_{i=1}^{n} x_i$</th>
<th>$\sum_{i=1}^{n} x_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>30</td>
<td>500</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>60</td>
<td>1400</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>100</td>
<td>3000</td>
</tr>
</tbody>
</table>

The calculator only keeps track of $n$ and running totals $\sum x_i$, $\sum x_i^2$.
- The sample mean is $m = (\sum_{i=1}^{n} x_i)/n = 100/4 = 25$.
- The sample variance and sample standard deviation are

$$s^2 = \frac{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}{n(n-1)} = \frac{4(3000)-(100)^2}{4(3)} \approx 166.67$$

$$s = \sqrt{\frac{500}{3}} \approx 12.91$$

- With the formula $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2$, the calculator has to store all the numbers, then compute $m$, then compute $s$. 
The CAPE questionnaire asks how many hours a week you spend on a class. Suppose the number of answers in each category is

<table>
<thead>
<tr>
<th># hours/week</th>
<th>Frequency ($f_i$)</th>
<th>Midpoint of interval ($m_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–1</td>
<td>2</td>
<td>.5</td>
</tr>
<tr>
<td>2–3</td>
<td>20</td>
<td>2.5</td>
</tr>
<tr>
<td>4–5</td>
<td>31</td>
<td>4.5</td>
</tr>
<tr>
<td>6–7</td>
<td>11</td>
<td>6.5</td>
</tr>
<tr>
<td>8–9</td>
<td>3</td>
<td>8.5</td>
</tr>
<tr>
<td>10–11</td>
<td>1</td>
<td>10.5</td>
</tr>
<tr>
<td>12–13</td>
<td>5</td>
<td>12.5</td>
</tr>
</tbody>
</table>

Total: $n = 73$

This question on the survey has $k = 7$ groups into which the $n = 73$ students are placed.

Assume all students in the 0–1 hrs/wk category are .5 hrs/wk; all students in the 2–3 hrs/wk category are 2.5 hrs/wk; etc.

Treat it as a list of two .5’s, twenty 2.5’s, thirty one 4.5’s, etc.
### Grouped data (also called binned data)

<table>
<thead>
<tr>
<th># hours/week</th>
<th>Frequency ((f_i))</th>
<th>Midpoint of interval ((m_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–1</td>
<td>2</td>
<td>.5</td>
</tr>
<tr>
<td>2–3</td>
<td>20</td>
<td>2.5</td>
</tr>
<tr>
<td>4–5</td>
<td>31</td>
<td>4.5</td>
</tr>
<tr>
<td>6–7</td>
<td>11</td>
<td>6.5</td>
</tr>
<tr>
<td>8–9</td>
<td>3</td>
<td>8.5</td>
</tr>
<tr>
<td>10–11</td>
<td>1</td>
<td>10.5</td>
</tr>
<tr>
<td>12–13</td>
<td>5</td>
<td>12.5</td>
</tr>
</tbody>
</table>

Total: \(n = 73\)

- **Sample mean:**
  \[
  \frac{1}{73} \left( 2(.5) + 20(2.5) + 31(4.5) + 11(6.5) + 3(8.5) + 1(10.5) + 5(12.5) \right)
  = 4.9384 \text{ hours/week}
  
- **Sample variance and SD:**
  \[
  s^2 = \frac{1}{72} \left( 2(.5 - 4.94)^2 + 20(2.5 - 4.94)^2 + \cdots + 5(12.5 - 4.94)^2 \right)
  = 7.5830 \text{ hours}^2/\text{week}^2
  
  \[ s = \sqrt{7.5830} = 2.7537 \text{ hours/week} \]
The bins on the CAPE survey should be widened to cover all possibilities (for example, where does 7.25 go?)
Fix it by expanding the bins: e.g., 2–3 becomes 1.5–3.5.

Treating all students in the 2–3 hours/week category (which should be 1.5–3.5) as 2.5 hours/week is only an approximation; for each student in this category, this is off by up to ±1.

- In computing the grouped sample mean, it is assumed that such errors balance out.
- In computing the grouped sample variance, these errors are not taken into consideration. A different formula could be used to take that into account.