Estimating parameters
5.3 Confidence Intervals
5.4 Sample Variance

Prof. Tesler

Math 186
Winter 2019

Estimating parameters of the normal distribution (μ, σ) 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 σ but not μ and we want to estimate μ from experimental data.
- **Chapter 5.4:** Estimate both μ and σ 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 μ; it just gives one number, without an indication of how far away it might be from μ.
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
$$E(\overline{X}) = E\left(\tfrac{X_1+\cdots+X_n}{n}\right) = \tfrac{\mu+\cdots+\mu}{n} = \frac{n\mu}{n} = \mu$$
These estimates will be distributed with variance $\mathrm{Var}(\overline{X}) = \sigma^2/n$.

Sample data

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

Sample mean notation

Variable names

Actual distribution (Greek letters)	**Point estimate from a sample (Latin letters)**
X: random variable	$x_1, \ldots, x_n$: sample data
μ: mean	m or $\bar{x}$: sample mean (or Y; $y_1, \ldots, y_n$; $\bar{y}$)
σ^2: variance	s^2: sample variance
σ: standard deviation	s: sample standard deviation

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.

$$m = \bar{x} = \frac{x_1 + \cdots + x_n}{n} \qquad M = \overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

Sample data

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

- $\mu = 500$, $\sigma = 100$
- Are the sample means close or far to the true mean $\mu = 500$?
- What does "close" mean?
  Within ± 0.01?
  ± 1?
  ± 10?
- The scale for measuring "closeness" is based on standard deviations.

Z-scores

- How often is the sample mean "close" to the secret value of μ?
- The sample mean is a random variable $\overline{X}$ with mean $E(\overline{X}) = \mu$ and standard deviation $\mathrm{SD}(\overline{X}) = \sigma/\sqrt{n}$. So
$$z = \frac{m-\mu}{\sigma/\sqrt{n}} \qquad \text{if we knew secret: } = \frac{m-500}{100/\sqrt{n}}$$
- Exclude the top 2.5% and bottom 2.5% of values of Z and regard the middle 95% as "close." So
$$P(-z_{.025} < Z < z_{.025}) = P(-1.96 < Z < 1.96) = .95$$
- For $m = 411.67$ (one of the trials in our demo):
$$z = \frac{411.67-500}{100/\sqrt{6}} = \frac{-88.33}{40.82} = -2.16$$
which is outside of ± 1.96, so $m = 411.67$ is "far" from $\mu = 500$.

Confidence intervals

- We will rearrange this equation to isolate μ:

$$P(-1.96 < Z < 1.96) = P(-1.96 < \frac{M-\mu}{\sigma/\sqrt{n}} < 1.96) = .95$$

- **Interpretation:** in $\approx 95\%$ of the trials of this experiment, the value $M = m$ satisfies

$$-1.96 < \tfrac{m-\mu}{\sigma/\sqrt{n}} < 1.96$$

- Solve for bounds on μ from the upper limit on Z:

$$\tfrac{m-\mu}{\sigma/\sqrt{n}} < 1.96 \quad \Leftrightarrow \quad m - \mu < 1.96\tfrac{\sigma}{\sqrt{n}} \quad \Leftrightarrow \quad m - 1.96\tfrac{\sigma}{\sqrt{n}} < \mu$$

  Notice the 1.96 turned into -1.96 and we get a lower limit on μ.

- Also solve for an upper bound on μ from the lower limit on Z:

$$-1.96 < \tfrac{m-\mu}{\sigma/\sqrt{n}} \quad \Leftrightarrow \quad -1.96\tfrac{\sigma}{\sqrt{n}} < m - \mu \quad \Leftrightarrow \quad \mu < m + 1.96\tfrac{\sigma}{\sqrt{n}}$$

- Together, $\qquad m - 1.96\tfrac{\sigma}{\sqrt{n}} < \mu < m + 1.96\tfrac{\sigma}{\sqrt{n}}$

Confidence intervals

- In $\approx 95\%$ of the trials of this experiment, the value $M = m$ satisfies
$$m - 1.96\frac{\sigma}{\sqrt{n}} < \mu < m + 1.96\frac{\sigma}{\sqrt{n}}$$
So, $\approx 95\%$ of the time we perform this experiment, the true mean μ 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 ± 1.96.

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

  For demo purposes:
  For a 75% confidence interval, use ± 1.15 instead of ± 1.96.

Confidence intervals

Example:	Six scores 380, 260, 390, 630, 540, 440
Sample mean:	$m = \frac{380+260+390+630+540+440}{6} = 440$
σ:	We assumed $\sigma = 100$ 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 *(393.05,486.95)*.

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

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 μ.
  It is *not* a guarantee that exactly k% will contain it.

- *Note:* If you really don't know the true value of μ, 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

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

so

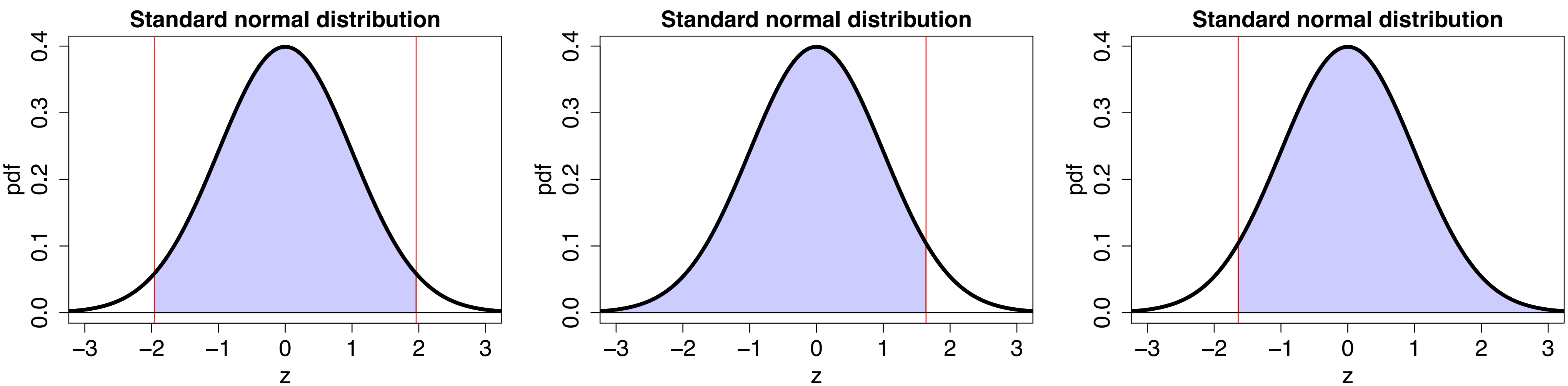
$$\sqrt{n} \geqslant 1.96\sigma/10 = 1.96(100)/10 = 19.6$$

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

$$n \geqslant 385$$

One-sided confidence intervals

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



	Two-sided	One-sided (right)	One-sided (left)
Remove:			
	2.5% on each side	5% on right	5% on left
Keep:			
	$-z_{.025} < Z < z_{.025}$	$Z < z_{.05}$	$Z > -z_{.05}$
	$-1.96 < Z < 1.96$	$Z < 1.64$	$Z > -1.64$

One-sided confidence intervals

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

$$P(Z < z_{.05}) = P(Z < 1.64) = .95$$

$$\approx 95\% \text{ of experiments have } \frac{m-\mu}{\sigma/\sqrt{n}} < 1.64, \text{ so } \mu > m - 1.64\frac{\sigma}{\sqrt{n}}$$

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

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

$$P(-z_{.05} < Z) = P(-1.64 < Z) = .95$$

The one-sided (left) 95% CI for μ 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, 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
  $\widehat{P}=\overline{X}=X/n$, with $E(\overline{X})=p$ and $\mathrm{Var}(\overline{X})=p(1-p)/n$.
- The $\widehat{\ }$ (caret) notation indicates it's a point estimate.
  We already use P for too many things, so we'll use the $\overline{X}$ 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
$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = \left(\hat{p} - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}, \hat{p} + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$
where we plugged in $\bar{x} = \hat{p}$ and $\sigma = \mathrm{SD}(X_i) = \sqrt{p(1-p)}$.
- 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
$$\left(\hat{p} - 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right)$$
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 $(.7 - 1.96(.01449), .7 + 1.96(.01449)) = (.672, .728)$

Interval estimate of p using margin of error

- Polls often report a *margin of error* instead of a confidence interval.

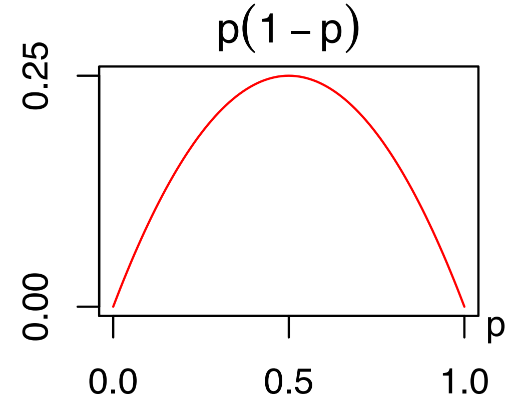
- The half-width of the 95% confidence interval is $1.96\sqrt{p(1-p)/n}$, and before we estimated p by the point estimate $\hat{p}$.

- The *margin of error* is the maximum that this half-width could be over all possible values of p $(0 \leqslant p \leqslant 1)$; 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 $0 \leqslant p \leqslant 1$:

  $0 = \frac{d}{dp}(p - p^2) = 1 - 2p$ at $p = \frac{1}{2}$

  $\frac{d^2}{dp^2}(p - p^2) = -2 < 0 \Rightarrow$ maximum



Interval estimate of p using margin of error

- The *margin of error* is the maximum possible half-width, $1.96\sqrt{(1/2)(1/2)/n} = 1.96/(2\sqrt{n})$.
- With 1000 people, the margin of error is $1.96/(2\sqrt{1000}) \approx .03099$, 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 $1.96\sqrt{p(1-p)/n} \leqslant 0.03$ 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 $1.96/(2\sqrt{n}) = .02$:
  $$n = (1.96/(2(0.02)))^2 = 49^2 = 2401$$
- 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 $\hat{p} \pm 0.02$.

5.4 Sample variance s^2: estimating σ^2 from data

- 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 $x_i - \bar{x}$:

$$1-5,\quad 2-5,\quad 12-5 \quad = \quad -4, -3, 7$$

- The deviations must sum to 0 since $\left(\sum_{i=1}^{n} x_i\right) - n\bar{x} = 0$.
  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 σ^2.
- The *sample standard deviation* is $s = \sqrt{s^2} = \sqrt{37} \approx 6.08$, which is a point estimate of σ.

Sample variance: estimating σ^2 from data

Definitions

Sum of squared deviations: $ss = \sum_{i=1}^{n} (x_i - \bar{x})^2$

Sample variance: $s^2 = \dfrac{ss}{n-1} = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Sample standard deviation: $s = \sqrt{s^2}$

- It turns out that $E(S^2) = \sigma^2$, so s^2 is an *unbiased estimator* of σ^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 σ^2.
- This is because $\sum_{i=1}^{n} (x_i - \bar{x})^2$ underestimates $\sum_{i=1}^{n} (x_i - \mu)^2$.

Estimating μ and σ^2 from sample data (secret: $\mu = 500$, $\sigma = 100$)

Exp. #	x_1	x_2	x_3	x_4	x_5	x_6	$\bar{x}$	s^2	u^2
1	550	600	450	400	610	500	518.33	7016.67	5847.22
2	500	520	370	520	480	440	471.67	3376.67	2813.89
3	470	530	610	370	350	710	506.67	19426.67	16188.89
4	630	620	430	470	500	470	520.00	7120.00	5933.33
5	690	470	500	410	510	360	490.00	12840.00	10700.00
6	450	490	500	380	530	680	505.00	10030.00	8358.33
7	510	370	480	400	550	530	473.33	5306.67	4422.22
8	420	330	540	460	630	390	461.67	11736.67	9780.56
9	570	430	470	520	450	560	500.00	3440.00	2866.67
10	260	530	330	490	530	630	461.67	19296.67	16080.56
						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 $SS=\sum_{i=1}^{n}(X_i-\overline{X})^2$:
$$E((X_1-\overline{X})^2)=E({X_1}^2)+E(\overline{X}^2)-2E(X_1\overline{X})$$

- $\mathrm{Var}(X)=E(X^2)-E(X)^2 \quad\Rightarrow\quad E(X^2)=\mathrm{Var}(X)+E(X)^2$. So
$$E({X_1}^2)=\sigma^2+\mu^2 \qquad E(\overline{X}^2)=\mathrm{Var}(\overline{X})+E(\overline{X})^2=\frac{\sigma^2}{n}+\mu^2$$

- Cross-term:
$$\begin{aligned} E(X_1\overline{X}) &= \frac{E({X_1}^2)+E(X_1)E(X_2)+\cdots+E(X_1)E(X_n)}{n} \\ &= \frac{(\sigma^2+\mu^2)+(n-1)\mu^2}{n} = \frac{\sigma^2}{n}+\mu^2 \end{aligned}$$

- Total for $i=1$ term:
$$E((X_1-\overline{X})^2)=(\sigma^2+\mu^2)+\left(\frac{\sigma^2}{n}+\mu^2\right)-2\left(\frac{\sigma^2}{n}+\mu^2\right)=\frac{n-1}{n}\sigma^2$$

Proof that denominator $n-1$ makes s^2 unbiased

- Similarly, term i of $SS = \sum_{i=1}^{n}(X_i - \overline{X})^2$ expands to
$$E((X_i - \overline{X})^2) = \frac{n-1}{n}\sigma^2$$
- 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 σ^2 (called an *unbiased estimator* of σ^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 *biased estimator*.

More formulas for sample mean and variance

- 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:

n	x_n	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} {x_i}^2$
1	10	10	100
2	20	30	500
3	30	60	1400
4	40	100	3000

  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

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

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

- 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.

Grouped data (also called binned data)

- The CAPE questionnaire asks how many hours a week you spend on a class. Suppose the number of answers in each category is

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

# hours/week	Frequency (f_i)	Midpoint of interval (m_i)
0–1	2	.5
2–3	20	2.5
4–5	31	4.5
6–7	11	6.5
8–9	3	8.5
10–11	1	10.5
12–13	5	12.5
Total:	$n = 73$	

- **Sample mean:**
$$\tfrac{1}{73}\left(2(.5)+20(2.5)+31(4.5)+11(6.5)+3(8.5)+1(10.5)+5(12.5)\right)$$
$$= 4.9384 \text{ hours/week}$$

- **Sample variance and SD:**
$$s^2 = \tfrac{1}{72}\left(2(.5-4.94)^2 + 20(2.5-4.94)^2 + \cdots + 5(12.5-4.94)^2\right)$$
$$= 7.5830 \text{ hours}^2\text{/week}^2$$
$$s = \sqrt{7.5830} = 2.7537 \text{ hours/week}$$

Grouped data — errors in this method

- 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.