I. Confidence interval for $\pi$: Motivation

• Want to estimate a population proportion $\pi$.

• The sample proportion $p$ is a good estimator.

• But we need to quantify the “uncertainty” in $p$.

• We’ve done this in an ad-hoc way, by computing the probability that $p$ is within 0.03 or 0.05 or … of $\pi$.

• It would be more satisfying if we could construct an interval of values that we’re sure contains the population proportion $\pi$.

• For example, we’d like to be able to say “We’re sure that the population proportion is between 0.04 and 0.07.”
• Unfortunately, we can’t do this (unless we choose “useless” intervals like (0, 1)).

• Why not? Because there’s always a small chance that we’ll get a sample that is very unrepresentative of the population.

• For example:
  – Population has 1000000 members.
  – 10000 are female (so $\pi = 0.01$)
  – Choose $n = 500$ at random.
  – There’s a (very small) chance that all 500 will be female, leading to $p = 1$!

• So we’ll have to be satisfied with something like “We’re pretty sure” in place of “We’re sure.”
II. Confidence Interval for \( \pi \): Specifics

We’ll concentrate first on a “98% confidence interval”.

- Our goal: Come up with a procedure that
  (a) Yields an interval of values for \( \pi \)
  (b) Works with probability 0.98, where “works” means produces an interval that contains \( \pi \).

- The interval will look like

\[(p - \text{something}, \ p + \text{something})\]

- The hard part: What “margin of error” should replace “something?”

- We’ll make all the usual (Chapter 5) assumptions about independent observations, etc.
III. Confidence interval for \( \pi \): Derivation

- We know (as long as \( n \) is large enough) that \( p \) is approximately normal with
  - Mean \( \pi \)
  - Standard deviation \( \sqrt{\pi(1 - \pi)/n} \).

- So

\[
\frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}}
\]

is approximately standard normal.

- Chapter 4 problem: Find \(-c\) and \(c\) such that

\[
P \left( -c \leq \frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}} \leq c \right) \approx 0.98.
\]

- Answer: \( c = 2.33 \). See Figure 1.
Figure 1: For a 98% confidence interval, we need to find $c$.
• So

\[
P \left( -2.33 \leq \frac{p - \pi}{\sqrt{\pi(1 - \pi)/n}} \leq 2.33 \right) \approx 0.98.
\]

• Algebra:

\[
P \left[ -2.33 \sqrt{\pi(1 - \pi)/n} \right.
\]

\[
\leq p - \pi
\]

\[
\leq 2.33 \sqrt{\pi(1 - \pi)/n} \right] \approx 0.98.
\]

• More algebra:

\[
P \left[ p - 2.33 \sqrt{\pi(1 - \pi)/n} \right.
\]

\[
\leq \pi
\]

\[
\leq p + 2.33 \sqrt{\pi(1 - \pi)/n} \right] \approx 0.98.
\]
• Success? The interval
\[ p \pm 2.33\sqrt{\pi(1 - \pi)/n} \]
will work.

• Failure? We can’t use this interval, because we don’t know \( \pi \).

• Redemption? We’ll plug in \( p \) for \( \pi \) to get the interval
\[ p \pm 2.33\sqrt{p(1 - p)/n} \]

• Note: Now we have two potential inaccuracies:
  – Approximating the distribution of \( p \) by a normal.
  – Replacing \( \pi \) by \( p \).

• Rule of thumb: If \( np > 10 \) and \( n(1 - p) > 10 \), then this method is typically valid, in the sense that the probability is close to 0.98.
IV. Confidence Interval for $\pi$: Computation

Computing confidence intervals is easy!

Example: Death penalty opinion poll from last week.

- Poll of 1003 U.S. adults.
- Found $p = 0.46$ supported the death penalty instead of life in prison for convicted murderers.
- A 98% confidence interval for the proportion $\pi$ of all U.S. adults who favor the death penalty is

$$p \pm 2.33 \sqrt{p(1-p)/n}$$

$$= 0.46 \pm 2.33 \sqrt{0.46(1-0.46)/1003}$$

$$\approx 0.46 \pm (0.0367)$$

$$= (0.4233, 0.4967).$$

- So we are “98% confident that the proportion of U.S. adults who favor the death penalty is between 0.4233 and 0.4967.”
V. Other confidence levels

- What about 95% confidence intervals, or 90% confidence intervals, or 99% confidence intervals?
- We need not repeat the derivation, just the computations, replacing the probability 0.98 by 0.95 or 0.90 or 0.99, etc.
• For example, to get a 95% confidence interval:
  – Find \( c \) and \( -c \) such that

  \[
P(-c \leq Z \leq c) = 0.95,
  \]

  where \( Z \) is standard normal.
  – The answer is \( c = 1.96 \). See Figure 2.
  – Then replace 2.33 by 1.96 in the formula:

  \[
p \pm 1.96 \sqrt{p(1-p)/n}
  \]
Figure 2: For a 95% confidence interval, we need to find $c$. The shaded area is 0.95.
• For the death penalty data this becomes

\[ 0.46 \pm 1.96 \sqrt{0.46(1 - 0.46)/1003} \]

\[ \approx 0.46 \pm (0.0308) \]

\[ = (0.4292, 0.4908). \]

• So we are “95% confident that the proportion of U.S. adults who favor the death penalty is between 0.4292 and 0.4908.”
• Changing confidence levels just changes the multiplier $c$. Some common values for $c$:

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>Multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>$c = 1.64$</td>
</tr>
<tr>
<td>95%</td>
<td>$c = 1.96$</td>
</tr>
<tr>
<td>98%</td>
<td>$c = 2.33$</td>
</tr>
<tr>
<td>99%</td>
<td>$c = 2.58$</td>
</tr>
</tbody>
</table>

• Increasing the confidence level increases the width of the interval. This makes intuitive sense!
VI. Sample size determination

- The “margin of error” of an interval is the “something” we add and subtract.
- In the present case, the margin of error is

$$c\sqrt{p(1-p)/n},$$

where $c$ is the multiplier we get from Table B.2, which depends on the confidence level.

- A common question: How much data do I need to get a confidence interval with margin of error $B$ or less?

- It’s easy to answer this.
- Want
  \[ B = c \sqrt{\frac{p(1 - p)}{n}}. \]
- Solve for \( n \) to get
  \[ n = \frac{c^2 p(1 - p)}{B^2}. \]

- **A problem:** We want to figure this out before we collect the data, but \( p \) can’t be computed until we collect the data.

- **A solution:** This is maximized when \( p = 1/2 \). So we’ll always be safe (although maybe a bit inefficient) using
  \[ n = \frac{c^2 (1/2)(1/2)}{B^2} \]

- The book gives some different approaches, but we’ll always be “conservative” and plug in \( p = 1/2 \).
Example revisited In the death penalty example we computed a 98% confidence interval to be $0.46 \pm (0.0367)$, so the margin of error is 0.0367.

- **Question:** What sample size $n$ guarantees a margin of error of 0.02 or less?

- **Answer:** Plug $c = 2.33$ and $B = 0.02$ into the formula:

$$n = \frac{(2.33)^2(1/2)(1/2)}{(0.02)^2} = 3393.0625.$$

- To be safe we always round up. In this case, we get $n = 3394$. 
Example: A Washington Post/ABC News poll asked 1513 randomly selected adults “Would you support or oppose the federal government giving parents money to send their children to private or religious schools instead?” Of those surveyed, 48% said YES.

- Find a 96% confidence interval for the proportion of adults who support this.
- First find $c = 2.05$ from Table B.2 as in Figure 3.
- Then plug in to the formula:

$$0.48 \pm (2.05)\sqrt{0.48(0.52)/1513}$$

$$= 0.48 \pm 0.026$$

$$= (0.454, 0.506).$$
Figure 3: For a 96% confidence interval, we need to find $c$
• How large would $n$ have to be to assure that a 97% confidence interval would have margin of error 0.01 or less?

• First find $c = 2.17$ from Table B.2 as in Figure 4.

• Then plug $c = 2.17$ and $B = 0.01$ into the formula:

$$n = \frac{(2.17)^2(1/2)(1/2)}{(0.01)^2} = 11772.25.$$  

• Round up to 11773.
Figure 4: For a 97% confidence interval, we need to find $c$.
VII. Confidence intervals for \( \pi \):

**Interpretation**

- The (for example) 95\% confidence interval procedure has a 0.95 probability of working (giving an interval containing \( \pi \)) before the data are collected. (Before the data are collected, \( p \) is a random variable.)

- Once the data are collected and the interval is computed, the procedure has either worked or not! (Once the data are collected, \( p \) is a number.)

- So we cannot make meaningful probability statements about the actual interval, only about the method that we used to construct the interval.
VIII. Confidence intervals for $\pi$: Properties

- As we expect, as the confidence level goes up, the width of the interval goes up.

- Mathematically, this is due to the fact that the interval $(-c, c)$ needs to contain a larger area as the confidence level goes up.

- As we expect, for a fixed confidence level and $p$, as the sample size $n$ goes up, the width of the interval goes down.

- This is due to the $\sqrt{n}$ in the denominator.
IX. Confidence interval for $\mu$ when the population standard deviation $\sigma$ is known (Section 7.2)

- Same setting as Section 5.2:
  - Normally distributed population or large sample size $n$
  - Population mean $\mu$ (unknown)
  - Population standard deviation $\sigma$ known.
  - The sample mean is denoted $\overline{X}$.

- Now we want to find confidence intervals for $\mu$. 
- We know that

\[
\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}
\]

is (at least approximately) standard normal.

- Repeat the derivation from last time to get the confidence interval formula

\[
\bar{X} \pm c \frac{\sigma}{\sqrt{n}}
\]

- Here \(c\) is computed from the standard normal density in exactly the same way as for proportions.
Example: Recall Newcomb’s experiment to estimate the speed of light.

- He took \( n = 66 \) measurements.
- The mean of his 66 measurements is \( \bar{X} = 298072100 \) meters per second.
- (We’ll assume that) the standard deviation \( \sigma \) of Newton’s measurements is 150000 meters per second.
• To find a 95% confidence interval:
  – Find the number $c$ such that the area between $-c$ and $c$ under the standard normal density is 0.95. Answer:
    \[ c = 1.96. \]
  – Plug $c, \sigma, n, \overline{X}$ into the formula:
    \[
    298072100 \pm (1.96) \frac{150000}{\sqrt{66}}
    = 298072100 \pm 36189
    \]
    – This yields the interval
    \[
    (298035911, 298108289).
    \]
**Intuition:** Here’s the formula for a confidence interval:

\[ \bar{X} \pm c \frac{\sigma}{\sqrt{n}}. \]

- If we increase the confidence level, we increase \( c \), so the interval becomes wider.
- Larger \( \sigma \) (less precise data) leads to wider intervals.
- Larger sample sizes \( n \) lead to narrower intervals.
Required sample size

- **Question**: What sample size do we need to get a margin of error of $B$ or less?

- **Answer**: Solve

$$B = c \frac{\sigma}{\sqrt{n}}$$

for $n$ to get

$$n = \left( \frac{c \sigma}{B} \right)^2.$$
Example: In Newcomb’s experiment, how many measurements do we need to get a 99% confidence interval with margin of error 20000 or less?

- Here we have $B = 20000$ and $c = 2.57$ (from Table B.2).
- Plug into formula:
  
  $$n = \left( \frac{(2.57)(150000)}{20000} \right)^2 \approx 371.5.$$

- Round up to $n = 372.$
X. Confidence interval for $\mu$ when the population standard deviation $\sigma$ is unknown (Section 7.3)

- Section 7.2 was easy, but unrealistic.
- It’s rare that we actually know $\sigma$.
- How can we proceed when $\sigma$ is unknown?
• One idea:
  – Estimate $\sigma$ (the population standard deviation) by $S$ (the sample standard deviation).
  – Replace $\sigma$ by $S$ in the confidence interval formula, and don’t change anything else.

• This might work if $n$ is large, because then we’d expect that $S$ will be close to $\sigma$.

• But if $n$ is small, this does not work well: the confidence interval does not have the correct confidence level!

• The problem: Inserting an estimate ($S$) for $\sigma$ introduces extra variability (or extra uncertainty), and we need to change the multiplier $c$ in the confidence interval formula to take this into account.
• What should we expect?
  – The multiplier $c$ should be larger than the multiplier from the standard normal distribution.
  – As $n$ gets larger, there is less added variability, so the multiplier $c$ should get closer to the multiplier from the standard normal distribution.
Formalities: The $t$ distribution

For now we’ll assume:

- The population distribution is normal, with unknown mean $\mu$ and unknown standard deviation $\sigma$.
- Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a distribution called a “$t$ distribution with $n - 1$ degrees of freedom.”

- So to compute confidence intervals we need to find $c$ from the $t$ table rather than from the standard normal table.
- Next week we’ll learn how to do this.