Today's plan:

- Section 4.4.2: Capture-Recapture method revisited and Section 4.4.3: Public Opinion Polls

Section 4.4.2: Capture-Recapture method revisited

Let's use statistical inference to get a better estimate of a population size.

Example

- Estimate the population of fish in a lake.

Example

- Estimate the population of fish in a lake.
- Catch a sample of 150 fish. Tag and release them.

Example

- Estimate the population of fish in a lake.
- Catch a sample of 150 fish. Tag and release them.
- A week later, catch a new sample of 100 fish. The number of tagged fish is 12.

Example

- Estimate the population of fish in a lake.
- Catch a sample of 150 fish. Tag and release them.
- A week later, catch a new sample of 100 fish. The number of tagged fish is 12.
- Get a 95% confidence level estimate of the fish population.

The second sample is a repeated two-outcome experiment, done 100 times:

The second sample is a repeated two-outcome experiment, done 100 times:

- Take a fish and check for a tag

The second sample is a repeated two-outcome experiment, done 100 times:

- Take a fish and check for a tag
- The two outcomes are: *tagged* and *not tagged*

The number k of successes is the number of tagged fish in the sample.

The number k of successes is the number of tagged fish in the sample. The statistic $\hat{p}$ is

$$\hat{p} = \frac{k}{n} = \frac{12}{100} = 0.12$$

With $\hat{p} = 0.12$ and $n = 100$ in hand, we compute:

$$\textbf{st.err.} \approx \sqrt{\frac{0.12 \times (1 - 0.12)}{100}}$$
$$\approx 0.0325$$

With $\hat{p} = 0.12$ and $n = 100$ in hand, we compute:

$$\textbf{st.err.} \approx \sqrt{\frac{0.12 \times (1 - 0.12)}{100}}$$
$$\approx 0.0325$$

So what's p, with 95% confidence?

$$\hat{p} - (2 \times \frac{\sigma}{n}) \leq \quad p \quad \leq \hat{p} + (2 \times \frac{\sigma}{n})$$

$$
\begin{aligned}
\hat{p} - (2 \times \frac{\sigma}{n}) \leq \quad & p \quad \leq \hat{p} + (2 \times \frac{\sigma}{n}) \\
0.12 - (2 \times 0.0325) \leq \quad & p \quad \leq 0.12 + (2 \times 0.03
\end{aligned}
$$

$$
\begin{aligned}
\hat{p} - (2 \times \frac{\sigma}{n}) \leq & \quad p \quad \leq \hat{p} + (2 \times \frac{\sigma}{n}) \\
0.12 - (2 \times 0.0325) \leq & \quad p \quad \leq 0.12 + (2 \times 0.03 \\
0.055 \leq & \quad \frac{150}{N} \quad \leq 0.185
\end{aligned}
$$

$$
\begin{aligned}
\hat{p} - (2 \times \frac{\sigma}{n}) \leq &\quad p \quad \leq \hat{p} + (2 \times \frac{\sigma}{n}) \\
0.12 - (2 \times 0.0325) \leq &\quad p \quad \leq 0.12 + (2 \times 0.03 \\
0.055 \leq &\quad \frac{150}{N} \quad \leq 0.185 \\
\frac{0.055}{150} \leq &\quad \frac{1}{N} \quad \leq \frac{0.185}{150}
\end{aligned}
$$

$$
\begin{aligned}
\hat{p} - (2 \times \frac{\sigma}{n}) \leq & \quad p \quad \leq \hat{p} + (2 \times \frac{\sigma}{n}) \\
0.12 - (2 \times 0.0325) \leq & \quad p \quad \leq 0.12 + (2 \times 0.03 \\
0.055 \leq & \quad \frac{150}{N} \quad \leq 0.185 \\
\frac{0.055}{150} \leq & \quad \frac{1}{N} \quad \leq \frac{0.185}{150} \\
\frac{150}{0.055} \geq & \quad N \quad \geq \frac{150}{0.185}
\end{aligned}
$$

$$
\begin{aligned}
\hat{p} - (2 \times \frac{\sigma}{n}) \leq & \quad p \quad \leq \hat{p} + (2 \times \frac{\sigma}{n}) \\
0.12 - (2 \times 0.0325) \leq & \quad p \quad \leq 0.12 + (2 \times 0.03 \\
0.055 \leq & \quad \frac{150}{N} \quad \leq 0.185 \\
\frac{0.055}{150} \leq & \quad \frac{1}{N} \quad \leq \frac{0.185}{150} \\
\frac{150}{0.055} \geq & \quad N \quad \geq \frac{150}{0.185} \\
2727.27 \geq & \quad N \quad \geq 810.81
\end{aligned}
$$

We can say with 95% confidence that the population is somewhere between **811** and **2,727**.

- This interval is very wide

- This interval is very wide
- We can narrow the interval at the cost of reducing the confidence level.

- This interval is very wide
- We can narrow the interval at the cost of reducing the confidence level.
- or increasing the sample size

- With 68% confidence, we conclude the population is between 984 and 1,714.

- With 68% confidence, we conclude the population is between 984 and 1,714.
- The original estimate 1250 (when $\mathbf{st.err.} = 0$) is not the middle of the interval [811,2,727]

- With 68% confidence, we conclude the population is between 984 and 1,714.
- The original estimate 1250 (when $\mathbf{st.err.} = 0$) is not the middle of the interval [811,2,727]
- This is an artifact of estimating $1/N$ to get N.

Section 4.4.3: Public opinion polls

Example

The results of a poll (of 1350 people) for a mayoral election are

- 648 in favor of Candidate A

Example

The results of a poll (of 1350 people) for a mayoral election are

- 648 in favor of Candidate A
- 702 in favor of Candidate B

Example

The results of a poll (of 1350 people) for a mayoral election are

- 648 in favor of Candidate A
- 702 in favor of Candidate B

What predictions can we make about the election?

Let's begin with Candidate A.

- Sample size $n = 1350$

Let's begin with Candidate A.

- Sample size $n = 1350$
- Favorable voters $k = 648$

Let's begin with Candidate A.

- Sample size $n = 1350$
- Favorable voters $k = 648$
- Therefore $\hat{p} = \dfrac{648}{1350} = 0.48$ or 48%

Let's begin with Candidate A.

- Sample size $n = 1350$
- Favorable voters $k = 648$
- Therefore $\hat{p} = \dfrac{648}{1350} = 0.48$ or 48%
- $\sigma \approx \sqrt{1350 \times 0.48 \times (1 - 0.48)} \approx 18.3565$

- so the standard error is

$$\mathbf{st.err.} \approx \frac{18.3565}{1350} \approx 0.0136$$

$$\text{or} \quad 1.36\%$$

- so the standard error is

$$\textbf{st.err.} \approx \frac{18.3565}{1350} \approx 0.0136$$

$$\text{or} \quad 1.36\%$$

- Thus, the 95% confidence interval is

$$[48 - 2 \times 1.36, \quad 48 + 2 \times 1.36]$$

or

$$[45.28\%, 50.72\%]$$

Similarly, for Candidate B:

- Sample size $n = 1350$

Similarly, for Candidate B:

- Sample size $n = 1350$
- favorable voters $k = 702$

Similarly, for Candidate B:

- Sample size $n = 1350$
- favorable voters $k = 702$
- Therefore $\hat{p} = \dfrac{702}{1350} = 0.52$ or 52%

Similarly, for Candidate B:

- Sample size $n = 1350$
- favorable voters $k = 702$
- Therefore $\hat{p} = \dfrac{702}{1350} = 0.52$ or 52%
- $\sigma \approx \sqrt{1350 \times 0.52 \times (1 - 0.52)} \approx 18.3565$

- so the standard error is

$$\mathbf{st.err.} \approx \frac{18.3565}{1350} \approx 0.0136$$

$$\text{or} \quad 1.36\%$$

- so the standard error is

$$\textbf{st.err.} \approx \frac{18.3565}{1350} \approx 0.0136$$

or $\quad 1.36\%$

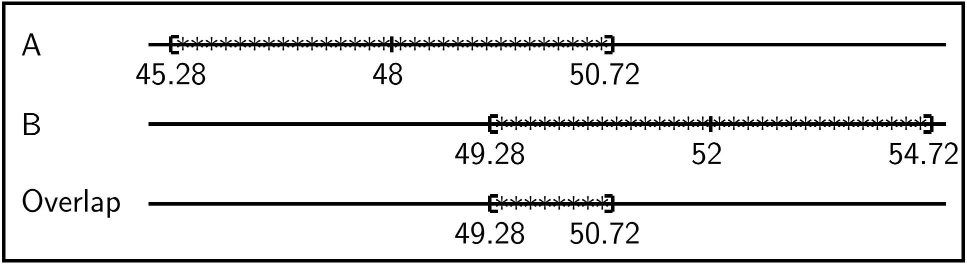
- Thus, the 95% confidence interval is

$$[52 - 2 \times 1.36, \quad 52 + 2 \times 1.36]$$

or

$$[49.28\%, 54.72\%]$$

When we draw these two intervals we clearly see they overlap.



- So with 95% confidence, we can't say who will win.

- So with 95% confidence, we can't say who will win.
- We call this a **statistical tie**, or we say the difference is not **statistically significant**.

Remarks:

- For both candidates the standard error was exactly the same.

Remarks:

- For both candidates the standard error was exactly the same.
- That is always the case when there are only two options.

Remarks:

- For both candidates the standard error was exactly the same.
- That is always the case when there are only two options.

$$\sigma \approx \sqrt{1350 \times 0.48 \times (1 - 0.48)}$$

$$= \sqrt{1350 \times 0.52 \times (1 - 0.52)}$$

- Even with three options, say, A, B and No preference, if not many people pick the third option then the standard error for both candidates will be almost the same.

- Even with three options, say, A, B and No preference, if not many people pick the third option then the standard error for both candidates will be almost the same.
- In such cases we can get away with only computing one standard error.

Example

Now a new poll is taken, and the numbers are:

- 581 in favor of Candidate A
- 769 in favor of Candidate B

Is the difference statistically significant now?

The sample size is $n = 1350$, and the poll has only two options, so there is a **common standard error**.

For Candidate A, we have

- $k = 581$

For Candidate A, we have

- $k = 581$
- so $\hat{p} = \dfrac{581}{1350} \approx 0.4303$ or 43.03%.

For Candidate B, we have

- $k = 769$

For Candidate B, we have

- $k = 769$
- so $\hat{p} = \dfrac{769}{1350} \approx 0.5696$ or 56.96%.

The standard error is

$$\textbf{st.err.} \approx \sqrt{\frac{0.4304 \times (1 - 0.4304)}{1350}}$$

$$\approx 0.0135 \quad \text{or} \quad 1.35\%$$

The 95% confidence interval for Candidate A is

$$[43.03 - 2 \times 1.35, \quad 43.03 + 2 \times 1.35]$$

The 95% confidence interval for Candidate A is

$$[43.03 - 2 \times 1.35, \quad 43.03 + 2 \times 1.35]$$

or

$$[40.33\%, \quad 45.73\%]$$

The 95% confidence interval for Candidate B

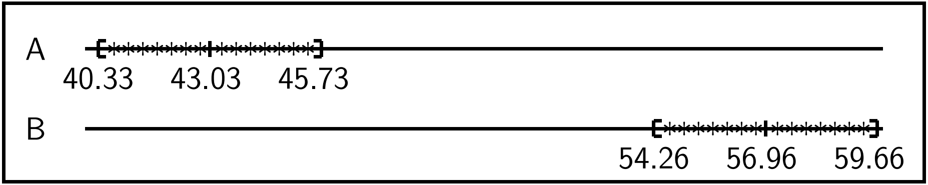
$$[56.96 - 2 \times 1.35, \quad 56.96 + 2 \times 1.35]$$

The 95% confidence interval for Candidate B

$$[56.96 - 2 \times 1.35, \quad 56.96 + 2 \times 1.35]$$

or

$$[54.26\%, \quad 59.66\%]$$


A
40.33
43.03
45.73
B
54.26
56.96
59.66

Remarks:

- Now they don't overlap at all.

Remarks:

- Now they don't overlap at all.
- Candidate B now has a statistically significant advantage over Candidate A.

Another way to see whether the difference between the candidates is statistically significant is whether their levels of support in the poll differ by **more than 4 standard errors**.

Another way to see whether the difference between the candidates is statistically significant is whether their levels of support in the poll differ by **more than 4 standard errors**.

$$\hat{p}_B - \hat{p}_A \approx 57\% - 43\% = 14\%$$

Another way to see whether the difference between the candidates is statistically significant is whether their levels of support in the poll differ by **more than 4 standard errors**.

$$\hat{p}_B - \hat{p}_A \approx 57\% - 43\% = 14\%$$

whereas

$$4 \times \textbf{st.err.} = 4 \times 1.35\% = 5.4\%$$

Next time: Section 4.4.4: Clinical Studies