INTRODUCTION TO STATISTICS

LECTURE 9

CONFIDENCE INTERVALS: RECAP

- A confidence interval is a *random* interval defined in terms of upper and lower confidence limits.

CONFIDENCE INTERVALS: RECAP

- A confidence interval is a *random* interval defined in terms of upper and lower confidence limits.

- Covers the parameter with probability $1 - \alpha$.

CONFIDENCE INTERVALS: RECAP

- A confidence interval is a *random* interval defined in terms of upper and lower confidence limits.

- Covers the parameter with probability $1 - \alpha$.

- Often constructed as follows:

$$\textit{point estimate} \pm \textit{quantile} \cdot \textit{variance of point estimate}$$

CONFIDENCE INTERVALS: RECAP

- NORMAL DISTRIBUTION:

$$X_1, X_2, \dots, X_n - \text{i.i.d. samples}$$

CONFIDENCE INTERVALS: RECAP

- NORMAL DISTRIBUTION:

$$X_1, X_2, \dots, X_n - \text{i.i.d. samples}$$

  - CI for μ, σ is known: **z-interval**

$$\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \quad z_{1-\alpha/2} - \text{quantile from } N(0,1)$$

CONFIDENCE INTERVALS: RECAP

- NORMAL DISTRIBUTION:

$$X_1, X_2, \ldots, X_n - \text{i.i.d. samples}$$

  - CI for μ, σ is known: **z-interval**

$$\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \quad z_{1-\alpha/2} - \text{quantile from } N(0,1)$$

  - CI for μ, σ is known: **t-interval**

$$\bar{X} \pm \frac{\sigma}{\sqrt{n}} t_{1-\alpha/2}, \quad t_{1-\alpha/2} - \text{quantile from Student distribution } (n-1) \text{ d.f.}$$

CONFIDENCE INTERVALS: RECAP

- NORMAL DISTRIBUTION:

$$X_1, X_2, \dots, X_n - \text{i.i.d. samples}$$

  - CI for μ, σ is known: **z-interval**

$$\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \quad z_{1-\alpha/2} - \text{quantile from } N(0,1)$$

  - CI for μ, σ is known: **t-interval**

$$\bar{X} \pm \frac{\sigma}{\sqrt{n}} t_{1-\alpha/2}, \quad t_{1-\alpha/2} - \text{quantile from Student distribution } (n-1) \text{ d.f.}$$

  - CI for σ, μ is unknown: $\boldsymbol{\chi^2}$**-interval**

$$\left(\sqrt{\frac{(n-1)s^2}{q_{1-\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{q_{\alpha/2}}}\right), q - \text{ quantiles from } \chi^2 \text{ distribution } (n-1) \text{ d.f.}$$

CONFIDENCE INTERVALS: RECAP

- OTHER / UNKNOWN DISTRIBUTION:

CONFIDENCE INTERVALS: RECAP

- OTHER / UNKNOWN DISTRIBUTION:

$X_1, X_2, \ldots, X_n -$ i.i.d. samples from a distribution with finite mean and variance

$$\Rightarrow$$

CONFIDENCE INTERVALS: RECAP

- OTHER / UNKNOWN DISTRIBUTION:

$X_1, X_2, \dots, X_n -$ i.i.d. samples from a distribution with finite mean and variance

$$\Rightarrow$$

Central Limit Theorem: for large n,

CONFIDENCE INTERVALS: RECAP

- OTHER / UNKNOWN DISTRIBUTION:

$X_1, X_2, \dots, X_n -$ i.i.d. samples from a distribution with finite mean and variance

$$\Rightarrow$$

Central Limit Theorem: for large n,

$$\frac{(\bar{X}-\mu)s}{\sqrt{n}} \sim N(0,1), \quad s - \text{sample std.}$$

CONFIDENCE INTERVALS: RECAP

- OTHER / UNKNOWN DISTRIBUTION:

$X_1, X_2, \dots, X_n -$ i.i.d. samples from a distribution with finite mean and variance

$$\Rightarrow$$

Central Limit Theorem: for large n,

$$\frac{(\bar{X}-\mu)s}{\sqrt{n}} \sim N(0,1), \quad s - \text{sample std.}$$

$$\mu = \bar{X} \pm \frac{s}{\sqrt{n}} z_{1-\alpha/2}$$

CI FOR BERNOULLI DISTRIBUTION

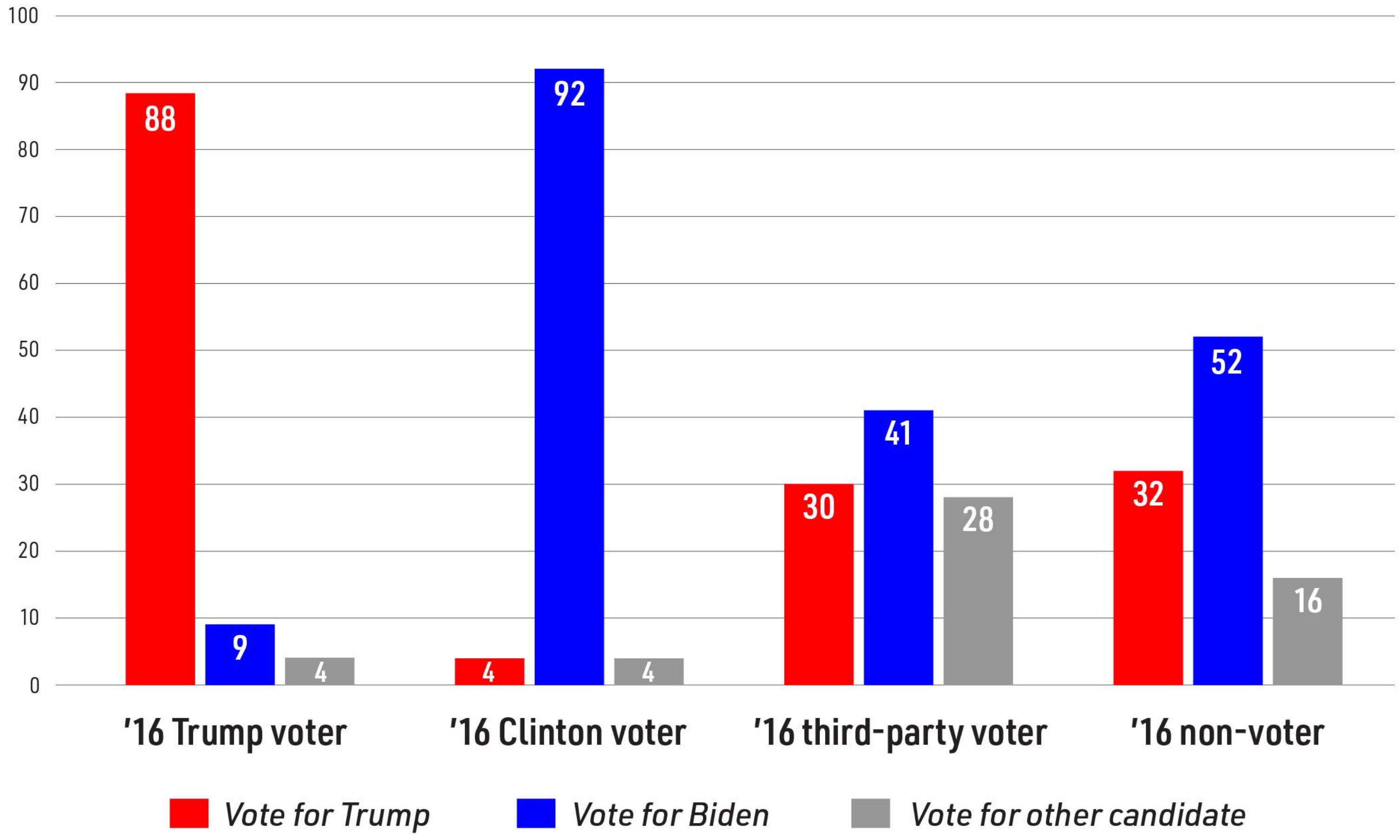
POLITICAL POLLS

- Political polls are often reported as a value with a margin-of-error.

52% favor candidate A with a margin-of-error of ±5%.

Presidential Election Preview:

How Clinton, Trump Voters from 2016 Plan to Vote in 2020



A total of 1,510 eligible voters, who are adult members of the USC Dornsife Center for Economic and Social Research's Understanding America Study internet panel, participated from August 11 – 16, 2020. Margin of sampling error for this preliminary sample is +/-3 percentage points. Tracking graphs will be updated every day. For full question text, methodology, and other information, visit https://uasdata.usc.edu/index.php

POLITICAL POLLS

- Political polls are often reported as a value with a margin-of-error.

52% favor candidate A with a margin-of-error of ±5%.

- The actual precise meaning:

POLITICAL POLLS

- Political polls are often reported as a value with a margin-of-error.

52% favor candidate A with a margin-of-error of ±5%.

- The actual precise meaning:

If p is the proportion of the population that supports A, then

POLITICAL POLLS

- Political polls are often reported as a value with a margin-of-error.

52% favor candidate A with a margin-of-error of ±5%.

- The actual precise meaning:

If p is the proportion of the population that supports A, then
the point estimate for p is $\hat{p} = 0.52$

POLITICAL POLLS

- Political polls are often reported as a value with a margin-of-error.

52% favor candidate A with a margin-of-error of ±5%.

- The actual precise meaning:

If p is the proportion of the population that supports A, then
the point estimate for p is $\hat{p} = 0.52$
and the 95% - CI is $52\% \pm 5\%$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \dots, X_n$.

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \ldots, X_n$.

$$\hat{p} = \bar{X}$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \dots, X_n$.

$$\text{CLT:} \quad \hat{p} = \bar{X} \;\approx\; N\left(\;\;,\qquad\quad \right)$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \dots, X_n$.

$$\text{CLT:} \quad \hat{p} = \bar{X} \approx N\left(p, \qquad \right)$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \ldots, X_n$.

$$\text{CLT:} \quad \hat{p} = \bar{X} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \dots, X_n$.

CLT: $\hat{p} = \bar{X} \approx N\left(p, \frac{p(1-p)}{n}\right)$

Approximation: $\frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n}$ (not valid for small samples)

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \ldots, X_n$.

CLT: $\hat{p} = \bar{X} \approx N\left(p, \frac{p(1-p)}{n}\right)$

Approximation: $\frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n}$ (not valid for small samples)

$\approx N(0,1)$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \dots, X_n$.

$$\text{CLT:} \quad \hat{p} = \bar{X} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

Approximation: $\frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n}$ (not valid for small samples)

$$\frac{(\hat{p} - p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} \approx N(0,1)$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \ldots, X_n$.

$$\frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} \approx N(0,1)$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \ldots, X_n$.

$$\frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} \approx N(0,1)$$

$$P\left(-z_{1-\alpha/2} < \frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} < z_{1-\alpha/2}\right) = 1-\alpha$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \dots, X_n$.

$$\frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} \approx N(0,1)$$

$$P\left(-z_{1-\alpha/2} < \frac{(\hat{p}-p)\sqrt{n}}{\sqrt{\hat{p}(1-\hat{p})}} < z_{1-\alpha/2}\right) = 1-\alpha$$

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \dots, X_n$.

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \dots, X_n$.

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

- Example: toss a coin 100 times, 60 heads. CI for p ($\alpha = 0.05$):

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \ldots, X_n$.

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

- Example: toss a coin 100 times, 60 heads. CI for p ($\alpha = 0.05$):

$$p \in (0.6 - 1.96 \cdot 0.049\, ; 0.6 + 1.96 \cdot 0.049)$$

CI FOR p (BERNOULLI)

- Construct a confidence interval for the parameter p of the Bernoulli distribution based on samples $X_1, \dots, X_n$.

$$P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

- Example: toss a coin 100 times, 60 heads. CI for p ($\alpha = 0.05$):

$$p \in (0.6 - 1.96 \cdot 0.049\,; 0.6 + 1.96 \cdot 0.049)$$

$$p \in (0.504; 0.696)$$

THE 95%-CI: RULE-OF-THUMB

$$p \approx \hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

THE 95%-CI: RULE-OF-THUMB

$$p \approx \hat{p} \pm\ z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

What is the maximum possible value of $\hat{p}(1-\hat{p})$?

THE 95%-CI: RULE-OF-THUMB

$$p \approx \hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

What is the maximum possible value of $\hat{p}(1-\hat{p})$? 1/4!

THE 95%-CI: RULE-OF-THUMB

$$p \approx \hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

What is the maximum possible value of $\hat{p}(1-\hat{p})$? 1/4!

$$z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq$$

THE 95%-CI: RULE-OF-THUMB

$$p \approx \hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

What is the maximum possible value of $\hat{p}(1-\hat{p})$? 1/4!

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq z_{1-\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}} =$$

THE 95%-CI: RULE-OF-THUMB

$$p \approx \hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

What is the maximum possible value of $\hat{p}(1-\hat{p})$? 1/4!

$$z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq z_{1-\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}} = 1.96 \cdot \frac{1}{2\sqrt{n}} \approx$$

THE 95%-CI: RULE-OF-THUMB

$$p \approx \hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

What is the maximum possible value of $\hat{p}(1-\hat{p})$? 1/4!

$$z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq z_{1-\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}} = 1.96 \cdot \frac{1}{2\sqrt{n}} \approx \frac{1}{\sqrt{n}}$$

THE 95%-CI: RULE-OF-THUMB

$$p \approx \hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

What is the maximum possible value of $\hat{p}(1-\hat{p})$? 1/4!

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq z_{1-\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}} = 1.96 \cdot \frac{1}{2\sqrt{n}} \approx \frac{1}{\sqrt{n}}$$

$$p \approx$$

THE 95%-CI: RULE-OF-THUMB

$$p \approx \hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

What is the maximum possible value of $\hat{p}(1-\hat{p})$? 1/4!

$$z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq z_{1-\frac{\alpha}{2}} \cdot \frac{1}{2\sqrt{n}} = 1.96 \cdot \frac{1}{2\sqrt{n}} \approx \frac{1}{\sqrt{n}}$$

$$p \approx \hat{p} \pm \frac{1}{\sqrt{n}}$$

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

- Let’s find the point estimates and 95% rule-of-thumb confidence intervals for each poll:

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

- Let's find the point estimates and 95% rule-of-thumb confidence intervals for each poll:

$$p = \hat{p} \pm \frac{1}{\sqrt{n}}$$

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

- Poll 1:

$$\hat{p} =$$

- Poll 2:

$$\hat{p} =$$

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

- Poll 1:

$$\hat{p} = 22/40$$

- Poll 2:

$$\hat{p} =$$

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

- Poll 1:

$$\hat{p} = 22/40$$

- Poll 2:

$$\hat{p} = 190/400$$

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

- Poll 1:

$$\hat{p} = 22/40 \qquad p = \hat{p} \pm \frac{1}{\sqrt{n}} =$$

- Poll 2:

$$\hat{p} = 190/400 \qquad p = \hat{p} \pm \frac{1}{\sqrt{n}} =$$

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

- Poll 1:

$$\hat{p} = 22/40 \qquad p = \hat{p} \pm \frac{1}{\sqrt{n}} = 0.55 \pm 0.16$$

- Poll 2:

$$\hat{p} = 190/400 \qquad p = \hat{p} \pm \frac{1}{\sqrt{n}} =$$

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

- Poll 1:

$$\hat{p} = 22/40 \qquad p = \hat{p} \pm \frac{1}{\sqrt{n}} = 0.55 \pm 0.16$$

- Poll 2:

$$\hat{p} = 190/400 \qquad p = \hat{p} \pm \frac{1}{\sqrt{n}} = 0.475 \pm 0.05$$

EXAMPLE: POLITICAL POLLS

- Two polls:
  - Fast and First:
    polls 40 random voters and finds 22 support A.
  - Quick but Cautious:
    polls 400 random voters and finds 190 support A.

- Poll 1:

$$\hat{p} = 22/40 \qquad p = \hat{p} \pm \frac{1}{\sqrt{n}} = 0.55 \pm 0.16$$

- Poll 2:

$$\hat{p} = 190/400 \qquad p = \hat{p} \pm \frac{1}{\sqrt{n}} = 0.475 \pm 0.05$$

SAMPLE SIZE

How large should my sample be?

SAMPLE SIZE DETERMINATION

- Data collection is difficult.
- How much is ‘just enough’?

- Example: estimating CI for the mean, σ is know.

SAMPLE SIZE DETERMINATION

- Data collection is difficult.
- How much is ‘just enough’?

- Example: estimating CI for the mean, σ is know.

$$\mu \in \bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

SAMPLE SIZE DETERMINATION

- Data collection is difficult.
- How much is ‘just enough’?

- Example: estimating CI for the mean, σ is know.

$$\mu \in \bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

SAMPLE SIZE DETERMINATION

- Data collection is difficult.
- How much is ‘just enough’?

- Example: estimating CI for the mean, σ is know.

$$\mu \in \bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

Limit the width of the interval: $\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \epsilon$

SAMPLE SIZE DETERMINATION

- Data collection is difficult.
- How much is ‘just enough’?

- Example: estimating CI for the mean, σ is know.

$$\mu \in \bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

Limit the width of the interval: $\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \epsilon$

$$n \geq$$

SAMPLE SIZE DETERMINATION

- Data collection is difficult.
- How much is 'just enough'?

- Example: estimating CI for the mean, σ is know.

$$\mu \in \bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

Limit the width of the interval: $\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \epsilon$

$$n \geq \frac{\sigma^2 z_{1-\alpha/2}^2}{\epsilon^2}$$

SAMPLE SIZE DETERMINATION

- At a juice factory machine is set up as follows:
  - the average content of juice per bottle equals μ;
  - σ = 5cl.

SAMPLE SIZE DETERMINATION

- At a juice factory machine is set up as follows:
  - the average content of juice per bottle equals μ;
  - σ = 5cl.

- What sample size is required to estimate the average contents to within 0.5cl at the 95% confidence level?

SAMPLE SIZE DETERMINATION

- At a juice factory machine is set up as follows:
  - the average content of juice per bottle equals μ;
  - σ = 5cl.

- What sample size is required to estimate the average contents to within 0.5cl at the 95% confidence level?

$$n \geq \frac{\sigma^2 z^2_{1-\frac{\alpha}{2}}}{\epsilon^2} =$$

SAMPLE SIZE DETERMINATION

- At a juice factory machine is set up as follows:
  - the average content of juice per bottle equals μ;
  - σ = 5cl.

- What sample size is required to estimate the average contents to within 0.5cl at the 95% confidence level?

$$n \geq \frac{\sigma^2 z^2_{1-\frac{\alpha}{2}}}{\epsilon^2} = \frac{5^2 \cdot 1.96^2}{0.5^2} \approx$$

SAMPLE SIZE DETERMINATION

- At a juice factory machine is set up as follows:
  - the average content of juice per bottle equals μ;
  - σ = 5cl.

- What sample size is required to estimate the average contents to within 0.5cl at the 95% confidence level?

$$n \geq \frac{\sigma^2 z^2_{1-\frac{\alpha}{2}}}{\epsilon^2} = \frac{5^2 \cdot 1.96^2}{0.5^2} \approx 384.16$$

SAMPLE SIZE DETERMINATION

- At a juice factory machine is set up as follows:
  - the average content of juice per bottle equals μ;
  - σ = 5cl.

- What sample size is required to estimate the average contents to within 0.5cl at the 95% confidence level?

$$n \geq \frac{\sigma^2 z^2_{1-\frac{\alpha}{2}}}{\epsilon^2} = \frac{5^2 \cdot 1.96^2}{0.5^2} \approx 384.16$$

$$n^* = 385$$

SAMPLE SIZE DETERMINATION

- Similar reasoning applies to the CI for Bernoulli.

- *How many customers should be surveyed in order to estimate the share of satisfied customers to within 3% at the 95% confidence level?*

SAMPLE SIZE DETERMINATION

- Similar reasoning applies to the CI for Bernoulli.
- *How many customers should be surveyed in order to estimate the share of satisfied customers to within 3% at the 95% confidence level?*

$$\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

SAMPLE SIZE DETERMINATION

- Similar reasoning applies to the CI for Bernoulli.

- *How many customers should be surveyed in order to estimate the share of satisfied customers to within 3% at the 95% confidence level?*

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}}$$

SAMPLE SIZE DETERMINATION

- Similar reasoning applies to the CI for Bernoulli.

- *How many customers should be surveyed in order to estimate the share of satisfied customers to within 3% at the 95% confidence level?*

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}}$$

Let $p_0 = 0.5$ – the worst possible variance

SAMPLE SIZE DETERMINATION

- Similar reasoning applies to the CI for Bernoulli.
- *How many customers should be surveyed in order to estimate the share of satisfied customers to within 3% at the 95% confidence level?*

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}}$$

Let $p_0 = 0.5$ – the worst possible variance

$$1.96 \sqrt{\frac{0.25}{n}} \leq 0.03 \iff$$

SAMPLE SIZE DETERMINATION

- Similar reasoning applies to the CI for Bernoulli.

- *How many customers should be surveyed in order to estimate the share of satisfied customers to within 3% at the 95% confidence level?*

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}}$$

Let $p_0 = 0.5$ – the worst possible variance

$$1.96 \sqrt{\frac{0.25}{n}} \leq 0.03 \iff n \geq 1067.1$$

HYPOTHESIS TESTING

HYPOTHESIS TESTING

- After tossing a coin 100 times, we observed H only 30 times. Is it a fair coin?

- Did extra tuition help students?

- Is drug A more efficient than drug B?

HOW IT ALL STARTED

https://youtu.be/lgs7d5saFFc

LADY TASTING TEA

Ronald Fisher, 1913

LADY TASTING TEA

- 8 cups of tea
  - 4 cups: milk first
  - 4 cups: tea first

- The lady must select 4 cups prepared by one method.

- *How to check her ability to distinguish the teas?*

Ronald Fisher, 1913

LADY TASTING TEA

- The default assumption:

H_0: the lady can't distinguish the teas

LADY TASTING TEA

- The default assumption:

H_0: the lady can't distinguish the teas

- How unlikely would it be to randomly guess all 4?

LADY TASTING TEA

- The default assumption:

H_0: the lady can't distinguish the teas

- How unlikely would it be to randomly guess all 4?

Tea-Tasting Distribution Assuming H_0		
Success count	Combinations of selection	Number of Combinations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
Total		70

LADY TASTING TEA

- The default assumption:

H_0: the lady can't distinguish the teas

- How unlikely would it be to randomly guess all 4?

$$1/70 \approx 0.014$$

Tea-Tasting Distribution Assuming H_0		
Success count	Combinations of selection	Number of Combinations
0	oooo	$1 \times 1 = 1$
1	ooox, ooxo, oxoo, xooo	$4 \times 4 = 16$
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	$6 \times 6 = 36$
3	oxxx, xoxx, xxox, xxxo	$4 \times 4 = 16$
4	xxxx	$1 \times 1 = 1$
Total		70

LADY TASTING TEA

- The default assumption:

H_0: the lady can’t distinguish the teas

- How unlikely would it be to randomly guess all 4?

$$1/70 \approx 0.014$$

That’s *surprising enough* to reject H_0

Tea-Tasting Distribution Assuming H_0		
Success count	Combinations of selection	Number of Combinations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
Total		70