Introduction to Data Science

Foundations for Inference

A Survey

JULY 6, 1999

New Poll Gauges Americans' General Knowledge Levels

Four-fifths know earth revolves around sun

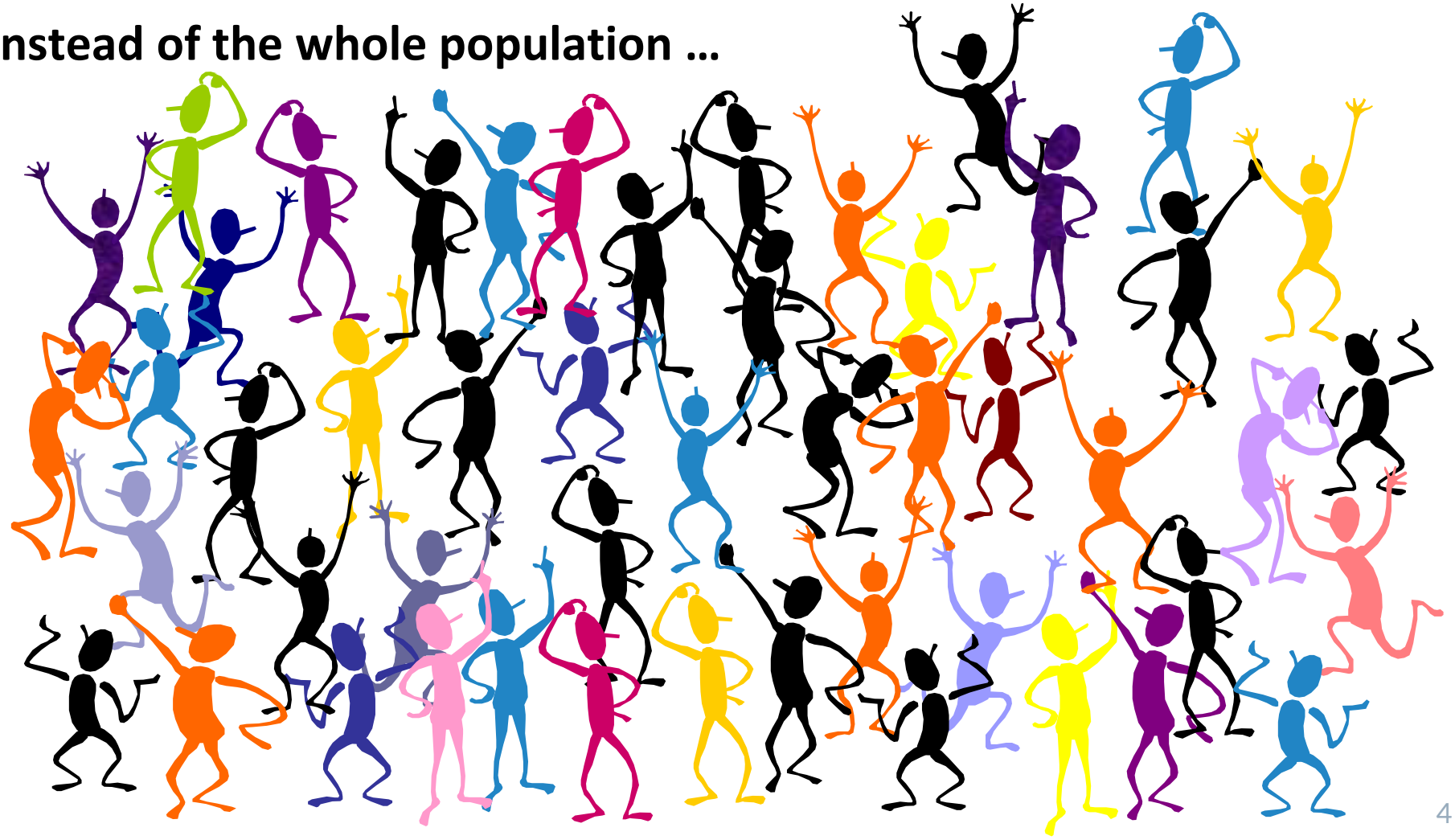
BY STEVE CRABTREE

- Probing a universal measure of knowledge, Gallup asked the basic science question: "As far as you know, does the earth revolve around the sun or does the sun revolve around the earth?"
  - 79% of Americans correctly respond that the earth revolves around the sun, while 18% say it is the other way around.

Results of the Survey

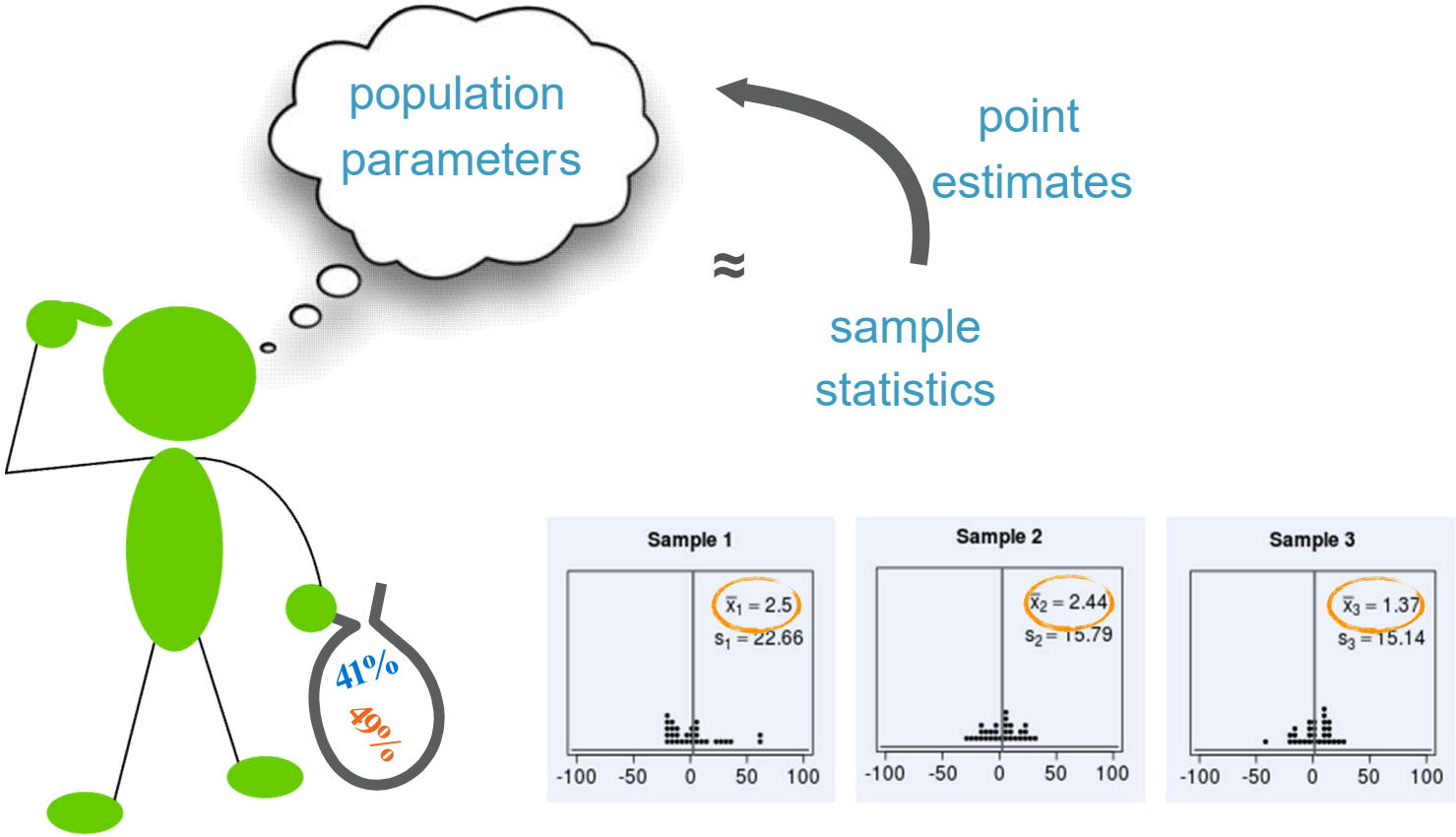
- The general public survey is based on telephone interviews conducted June 25-27, 1999, with a nationally representative sample of 1,016 adults ages 18 and older living in the continental United States.
- Margin of sampling error is plus or minus 3 percentage points for results at the 95% confidence level.
- 18% ± 3%: We are 95% confident that 15% to 21% of the adult Americans believe that the sun revolve around the earth.

Instead of the whole population …



We can use a much smaller sample to infer parameters.

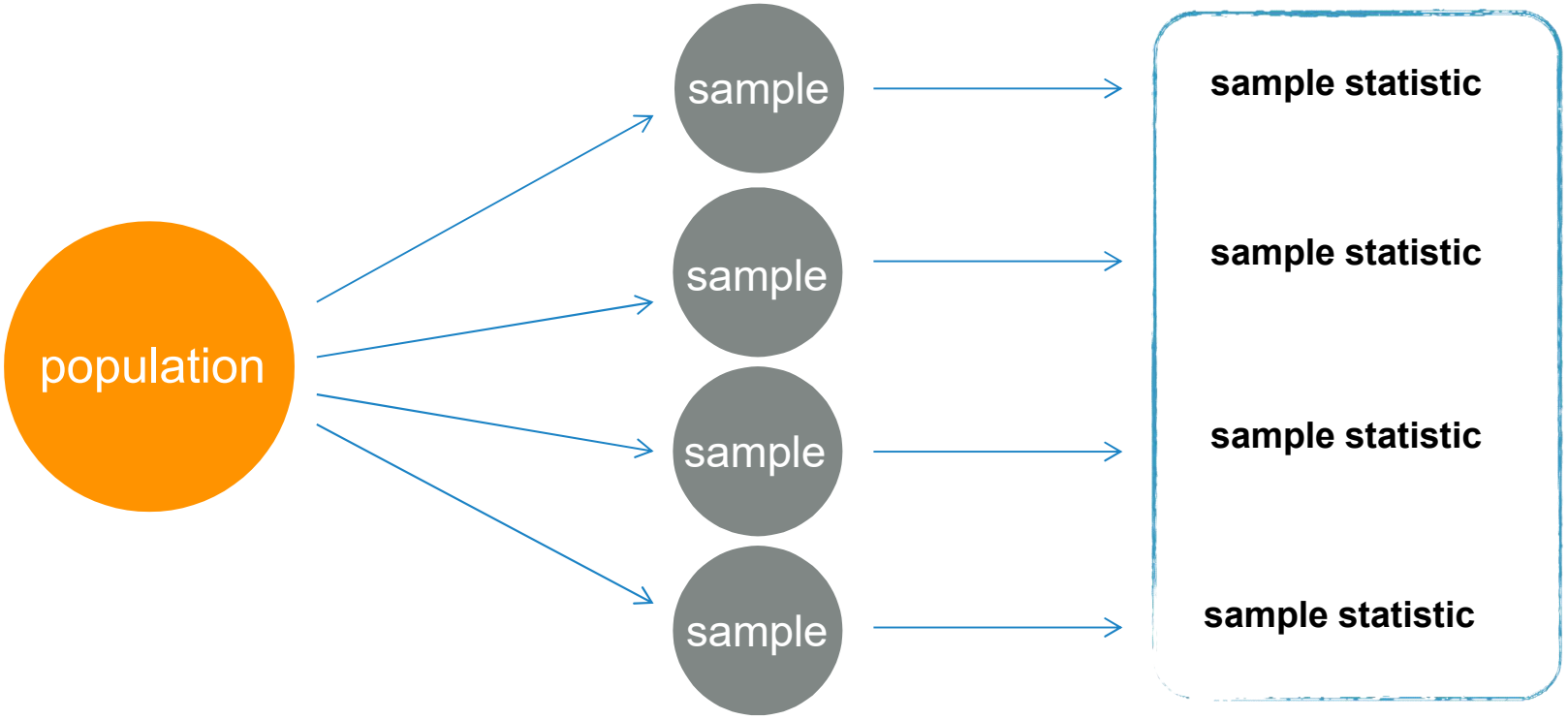
Parameter Estimation



Parameter Estimation

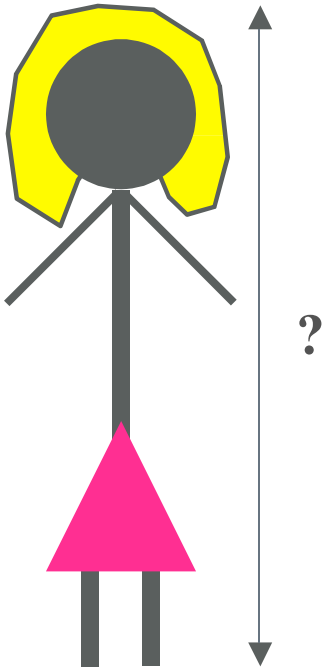
- We are often interested in population parameters.
- Since complete populations are difficult (or impossible) to collect data on, we use sample statistics as point estimates for the unknown population parameters of interest.
- Sample statistics vary from sample to sample.
- Quantifying how sample statistics vary provides a way to estimate the margin of error associated with our point estimate.

Sample vs. Sampling Distribution



Sample Distributions $\neq$ **Sampling Distribution**

Example



?

US
women
N = pop size
μ

S1 $x_{1,1}, x_{1,2}, \dots, x_{1,1000}$

…

S20 $x_{20,1}, x_{20,2}, \dots, x_{20,1000}$

…

S50 $x_{50,1}, x_{50,2}, \dots, x_{50,1000}$

$\bar{x}_1$ … $\bar{x}_{20}$ … $\bar{x}_{50}$

$\bar{x}$: sampling distribution

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

$mean(\bar{x}) = \mu$

$SD(\bar{x}) < \sigma$ → standard error

Central Limit Theorem

- Central Limit Theorem (CLT): The distribution of the sample mean is well approximated by a normal model:

$$\bar{x} \sim N(mean = \mu\ , SE = \frac{\sigma}{\sqrt{n}})$$

  where SE represents standard error, which is defined as the standard deviation of the sampling distribution.

- Note that as n increases SE decreases.
- If σ is unknown, use s (the sample standard deviation).
  - s : the standard deviation of one sample that we have at hand.

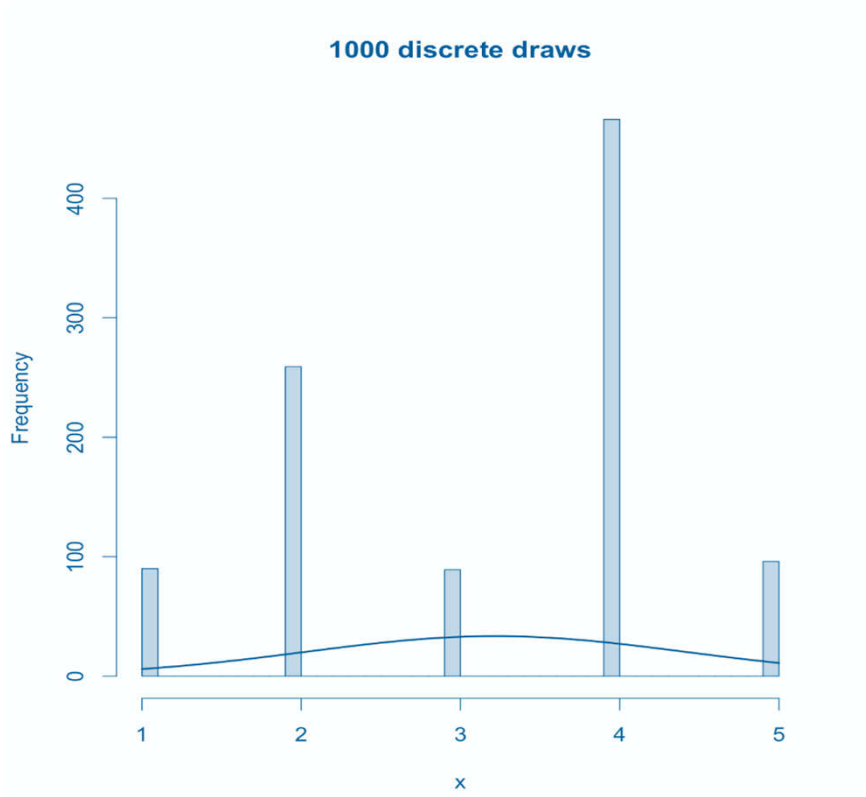
Example

X	1	2	3	4	5
$P(X)$	0.1	0.25	0.1	0.45	0.1

$$E[X] = 3.2 \ , \qquad X_i \sim X$$

Sample Size: $n = 1$

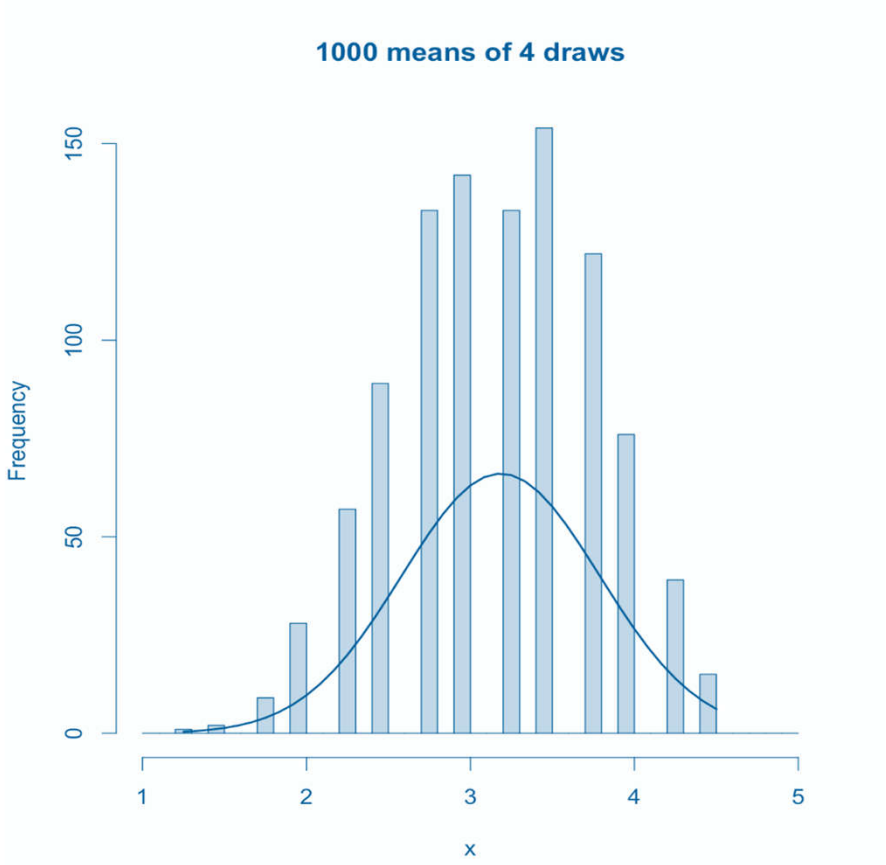
$$\bar{X} = \frac{X_1}{1}$$



Example

Sample Size: $n = 4$

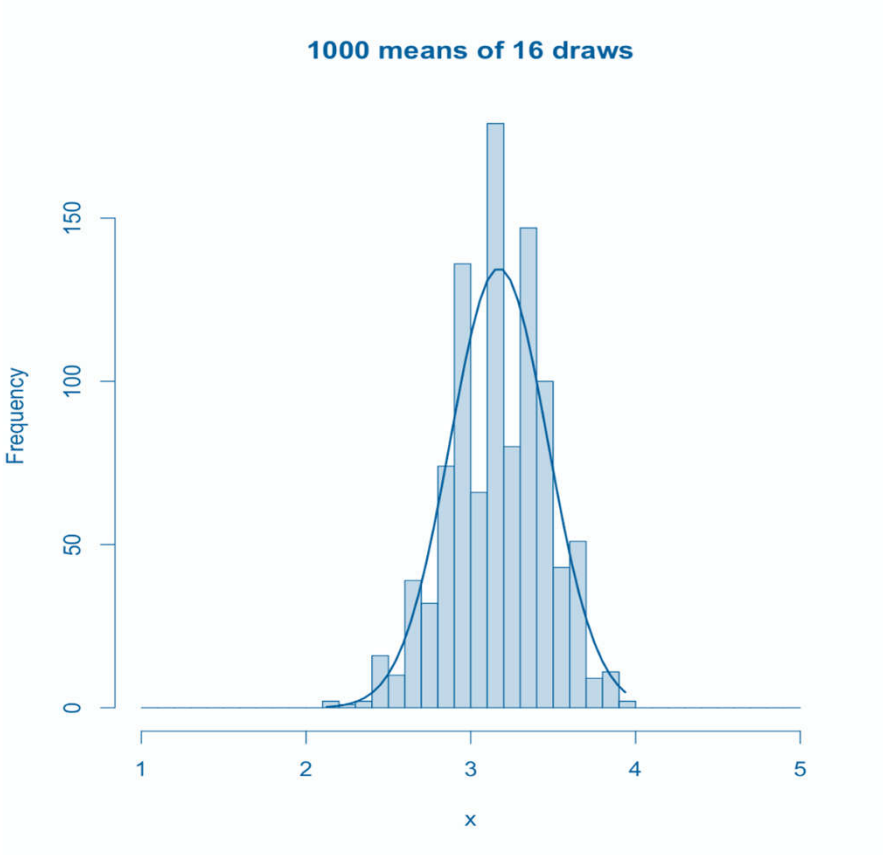
$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4}{4}$$



Example

Sample Size: $n = 16$

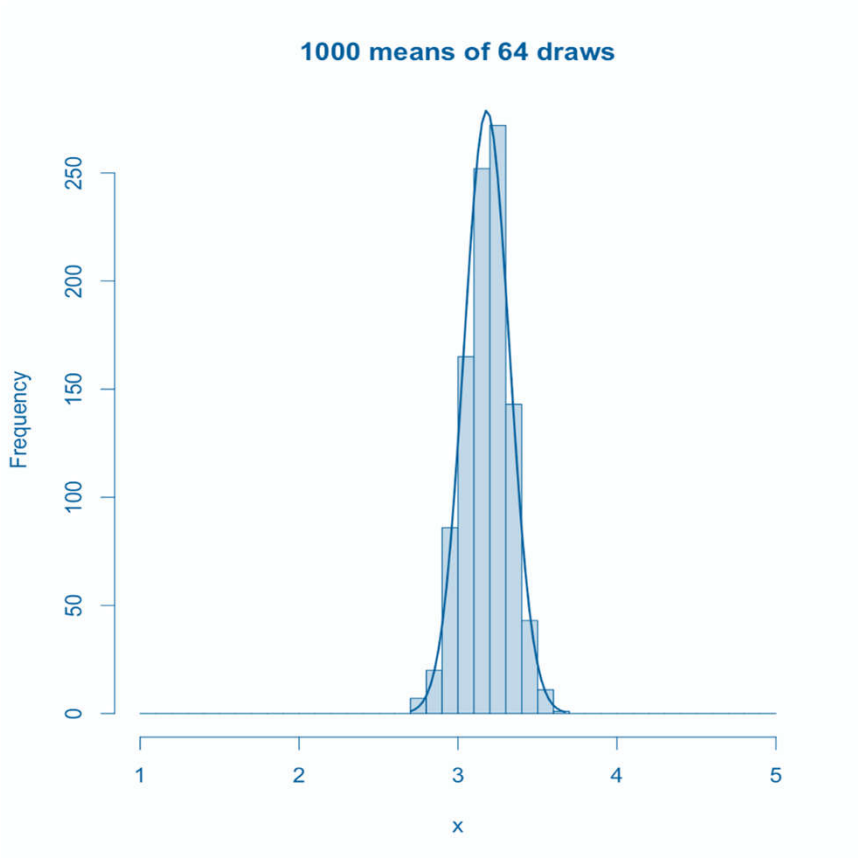
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{16}}{16}$$



Example

Sample Size: $n = 64$

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{64}}{64}$$



CLT Conditions

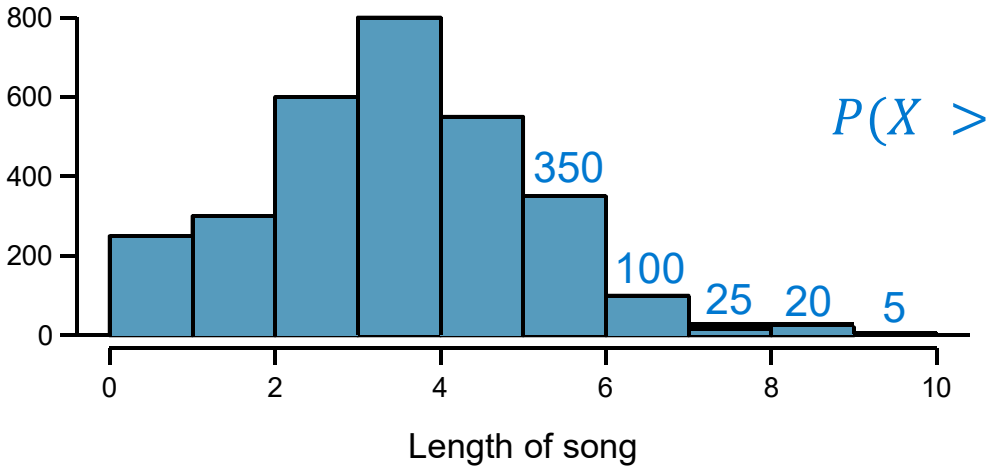
- Certain conditions must be met for the CLT to apply:

  1. Independence: Sampled observations must be independent. This is difficult to verify, but is more likely if:
     - random sampling/assignment is used, and
     - if sampling without replacement, $n < 10\%$ of the population.

  2. Sample size/skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.
     - the more skewed the population distribution, the larger sample size we need for the CLT to apply
     - for moderately skewed distributions $n > 30$ is a widely used rule of thumb

Example 1

- Suppose my phone has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this phone, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. Calculate the probability that a randomly selected song lasts more than 5 minutes.



X = length of one song

$$P(X > 5) = \frac{350 + 100 + 25 + 20 + 5}{3000}$$

$$= 500 / 3000$$

$$\approx 0.17$$

Example 2

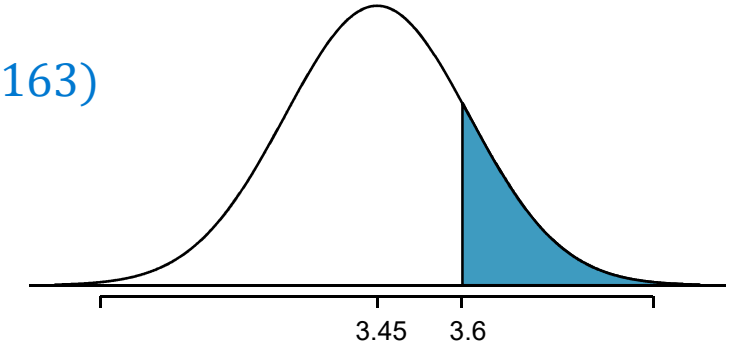
- I'm about to take a trip to visit my parents and the drive is 6 hours. I make a random playlist of 100 songs. What is the probability that my playlist lasts the entire drive?

$$P\{X_1 + X_2 + \ldots + X_{100} \geq 360 \text{ min}\} = ? \qquad \Rightarrow P\{\bar{X} \geq 3.6\} = ?$$

$$\bar{X} \sim N(mean = \mu = 3.45\ , SE = \frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{100}} = 0.163)$$

$$Z = \frac{3.6 - 3.45}{0.163} = 0.92$$

$$P(Z > 0.92) = 0.179$$



Confidence Interval

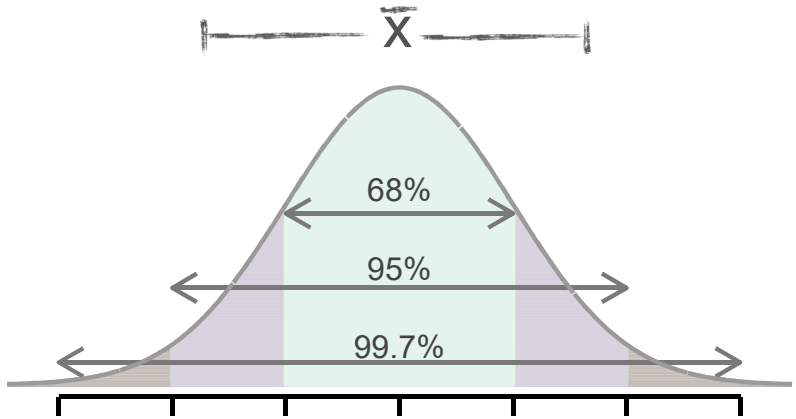
- A plausible range of values for the population parameter is called a confidence interval.

- If we report a point estimate, we probably won't hit the exact population parameter.
- If we report a range of plausible values we have a good shot at capturing the parameter.

Confidence Interval

Central Limit Theorem (CLT)

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$



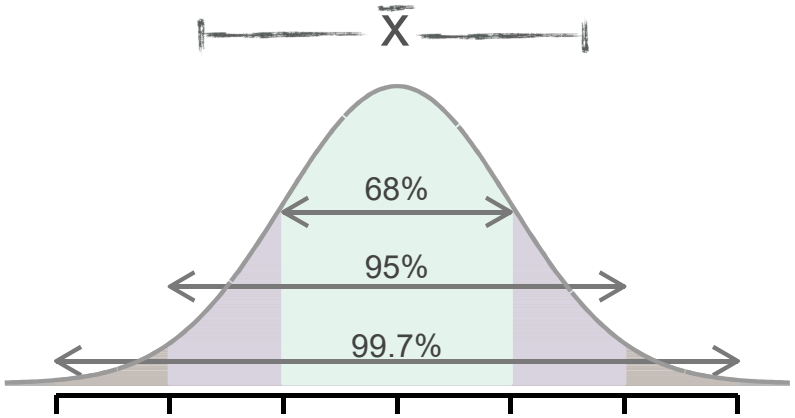
approximate 95% CI: $\bar{x} \pm 2SE$

margin of error (ME)

Confidence Interval

Central Limit Theorem (CLT)

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$



$$\Rightarrow P\{\mu - 2SE < \bar{x} < \mu + 2SE\} \approx 0.95$$

$$\Rightarrow P\{-2SE < \bar{x} - \mu < 2SE\} \approx 0.95$$

$$\Rightarrow P\{-2SE < \mu - \bar{x} < 2SE\} \approx 0.95$$

$$\Rightarrow P\{\bar{x} - 2SE < \mu < \bar{x} + 2SE\} \approx 0.95$$

approximate 95% CI for μ: $\bar{x} \pm 2SE$

Confidence Interval for a Population Mean

- Confidence interval for a population mean: Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution):

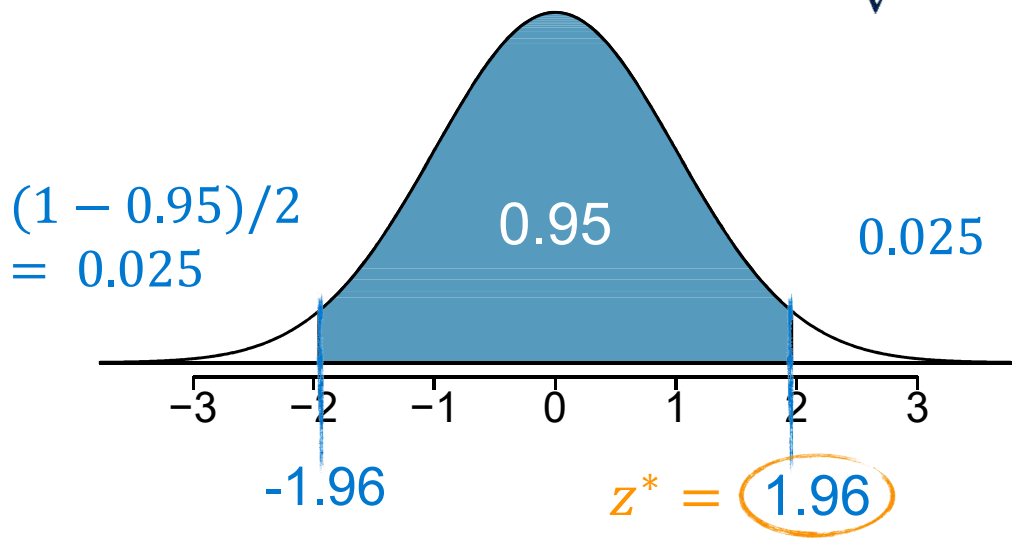
$$\bar{x} \pm z^\star \frac{s}{\sqrt{n}}$$

- **Conditions for this confidence interval:**
  1. **Independence:** Sampled observations must be independent.
     - random sample/assignment
     - if sampling without replacement, $n < 10\%$ of population
  2. **Sample size/skew:** $n \geq 30$, larger if the population distribution is very skewed.

Finding Critical Value

- Finding the critical value 95% confidence

$$\bar{x} \pm z^{\star} \frac{s}{\sqrt{n}}$$



	Second decimal place				
0.07	0.06	0.05	0.04	0.00	z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0004	0.0005	−3.3
0.0005	0.0006	0.0006	0.0006	0.0007	−3.2
0.0008	0.0008	0.0008	0.0008	0.0010	−3.1
0.0011	0.0011	0.0011	0.0012	0.0013	−3.0
0.0015	0.0015	0.0016	0.0016	0.0019	−2.9
0.0021	0.0021	0.0022	0.0023	0.0026	−2.8
0.0028	0.0029	0.0030	0.0031	0.0035	−2.7
0.0038	0.0039	0.0040	0.0041	0.0047	−2.6
0.0051	0.0052	0.0054	0.0055	0.0062	−2.5
0.0068	0.0069	0.0071	0.0073	0.0082	−2.4
0.0089	0.0091	0.0094	0.0096	0.0107	−2.3
0.0116	0.0119	0.0122	0.0125	0.0139	−2.2
0.0150	0.0154	0.0158	0.0162	0.0179	−2.1
0.0192	0.0197	0.0202	0.0207	0.0228	−2.0
0.0244	0.0250	0.0256	0.0262	0.0287	−1.9
0.0307	0.0314	0.0322	0.0329	0.0359	−1.8

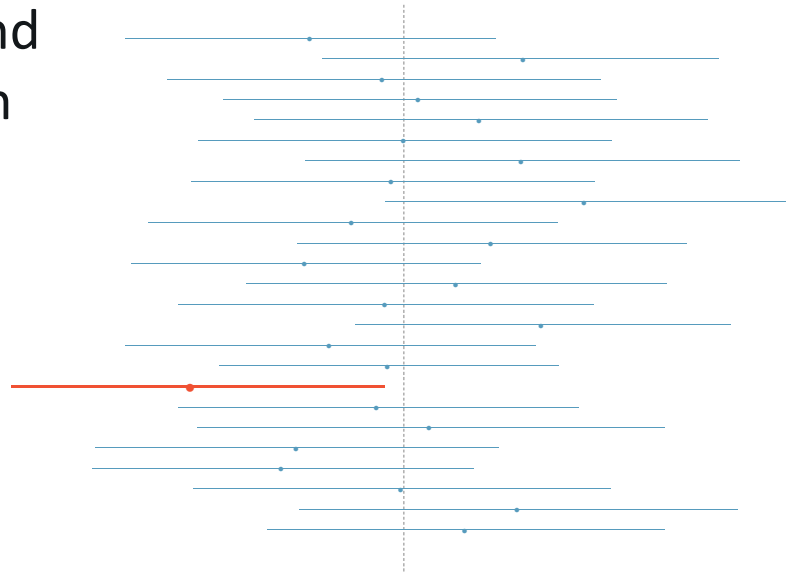
```
> from scipy.stats import norm
> norm.ppf(0.975)
```

Confidence Level

- Suppose we took many samples and built a confidence interval from each sample using the equation:

$$point\ estimate \pm 1.96 \times SE$$

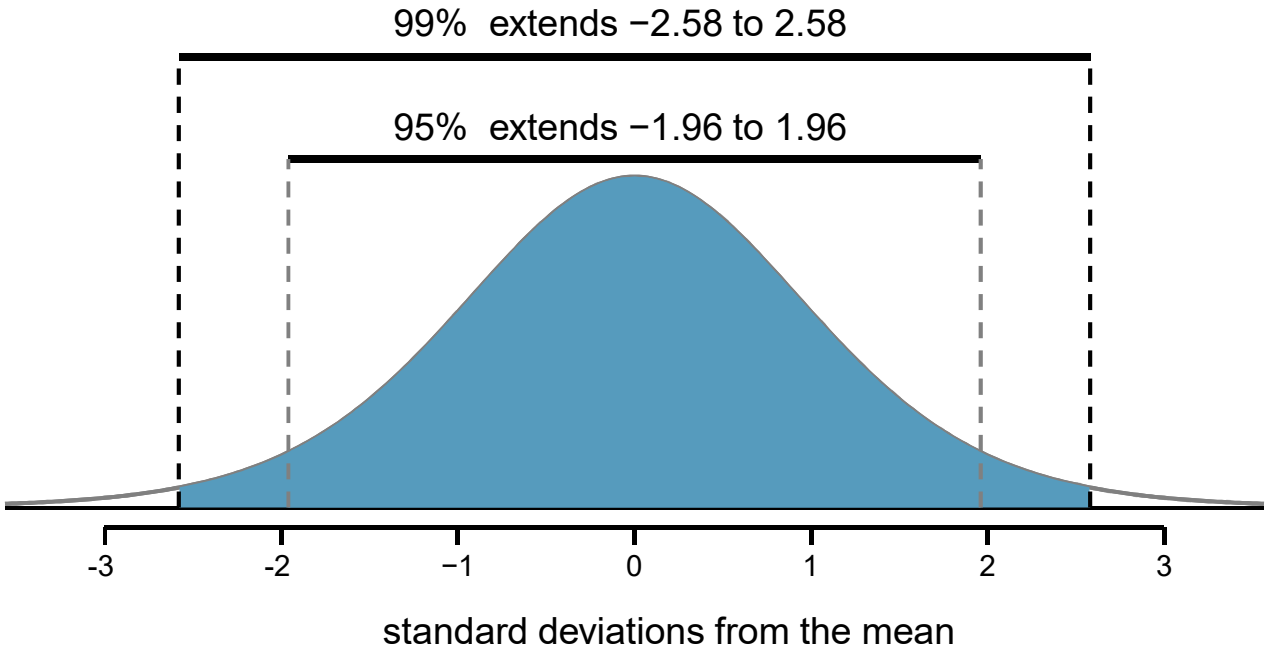
- Then about 95% of those intervals would contain the true population mean (μ).
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.



$24/25 = 0.96$

Width of Confidence Interval

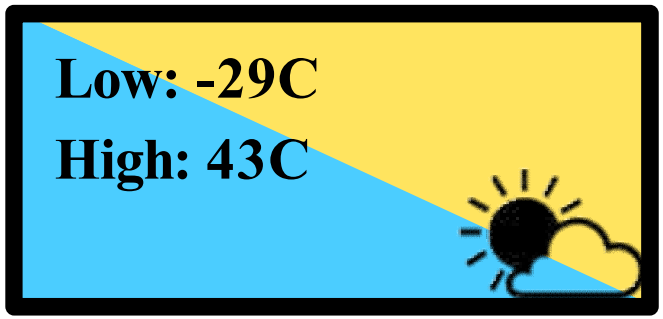
- If we want to be very certain that we capture the population parameter, should we use a wider interval or a narrower interval?



CL ↑ width ↑

Why not a wider interval?

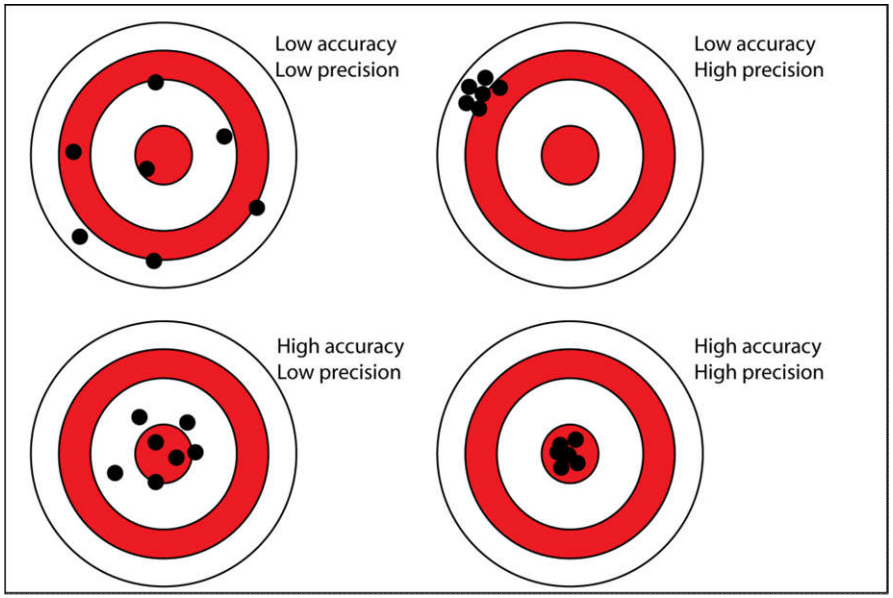
- What drawbacks are associated with using a wider interval?

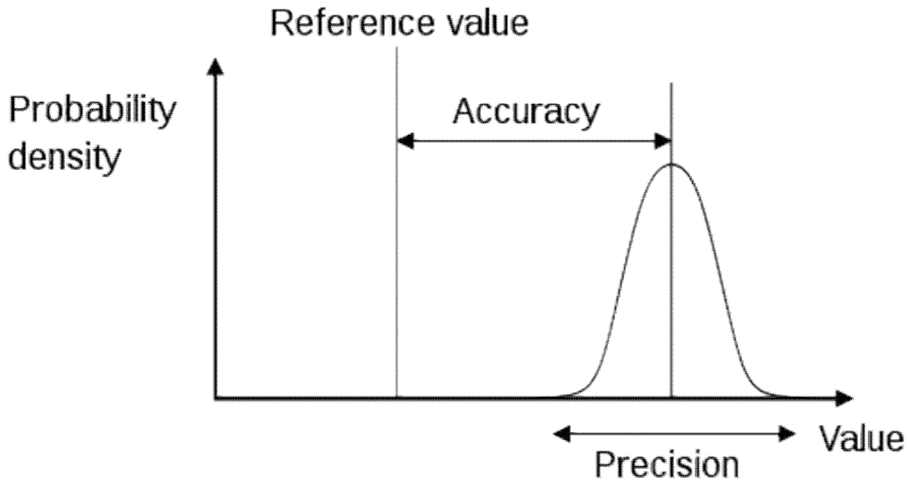


- Is this accurate? Most likely, yes.

- Is it informative, or, in other words, is it precise? Not really.

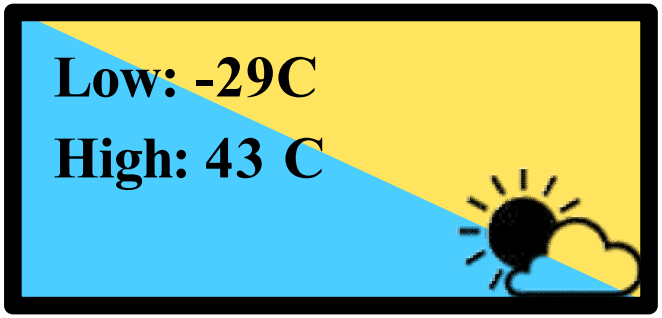
Accuracy vs. Precision





Precision vs. Accuracy

- We define accuracy in terms of whether or not the confidence interval contains the true population parameter.
- Precision refers to the width of a confidence interval.



CL ↑ width ↑ accuracy ↑

precision ↓

- How can we get the best of both worlds: higher precision and higher accuracy? increase sample size

Example

- A sample of 50 college students were asked how much money they spend on textbooks each semester. The students in the sample paid an average of $320 for textbooks, with a standard deviation of $174. In addition, the sample distribution was only slightly skewed to the right. Estimate the true average textbooks cost for college students based on this sample using a 95% confidence interval.

Checking conditions:

1. random sample & 50 < 10% of all college students

We can assume that the textbooks one student in the sample has bought is independent of another.

Example

2. $n > 30$ & not so skewed sample

We can assume that the sampling distribution of average textbooks cost from samples of size 50 will be nearly normal.

$$n = 50\ ,\qquad \bar{x} = 320\ ,\qquad s = 174$$

$$SE = \frac{s}{\sqrt{n}} = \frac{174}{\sqrt{50}} \approx 24.6$$

$$\bar{x} \pm z^* SE = 320 \pm 1.96 \times 24.6 = (272\ , 368)$$

We are 95% confident that college students on average have paid \$272 to \$368 for textbooks.

Hypothesis Testing Framework

- We start with a null hypothesis (H_0) that represents the status quo.
- We also have an alternative hypothesis (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods — methods that rely on the CLT.
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Hypotheses

null - H_0	Often either a skeptical perspective or a claim to be tested	=
alternative - H_A	Represents an alternative claim under consideration and is often represented by a range of possible parameter values.	<, >, ≠

- The skeptic will not abandon the H_0 unless the evidence in favor of the H_A is so strong that she rejects H_0 in favor of H_A.

Hypothesis Testing

$H_0: \mu = 300$ College students have paid \$300 for textbooks, on average.

$H_A: \mu > 300$ College students have paid more than \$300 for textbooks, on average.

always about pop. parameters,
never about sample statistics

p-value

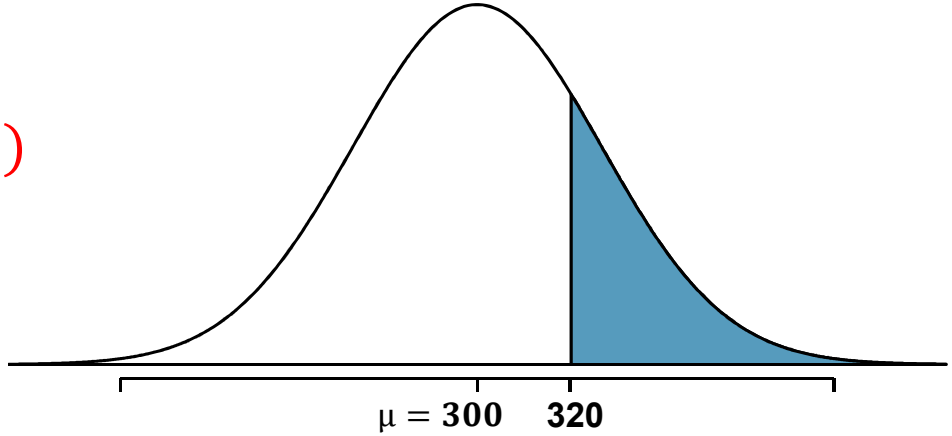
p-value = P(observed or more extreme outcome | H_0 true)

For previous example:

p-value = $P(\bar{x} > 320 | H_0: \mu = 300)$

$s = 174\,, n = 50 \;\Rightarrow\; SE = 24.6$

$\bar{x} \sim N(\mu = 300\,, SE = 24.6)$



test statistics: $Z = \frac{320-300}{24.6} = 0.81$

p-value = $\mathrm{P}(Z > 0.81) = 0.209$

Decision based on the p-value

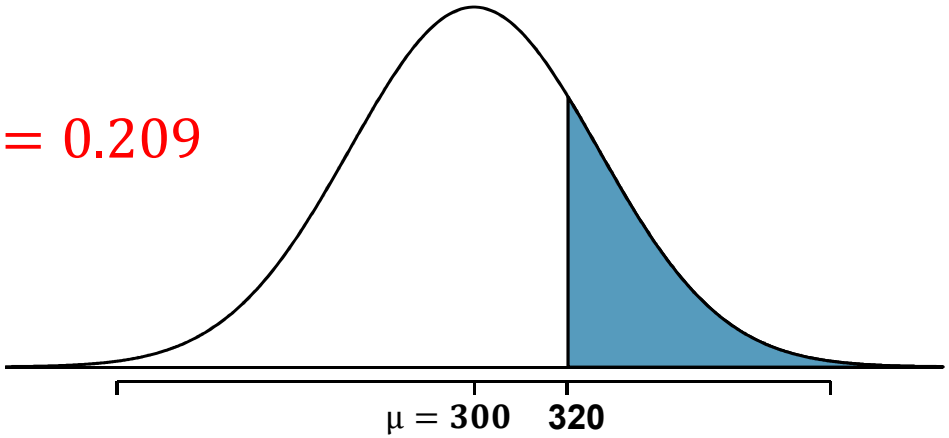
- We use the test statistic to calculate the p-value: the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.
- If the p-value is low (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence reject H_0.
- If the p-value is high (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence do not reject H_0.

Example

p-value = $P(\bar{x} > 320 | \mu = 300) = 0.209$

significance level: $\alpha = 0.05$

p-value = $0.209 > 0.05$

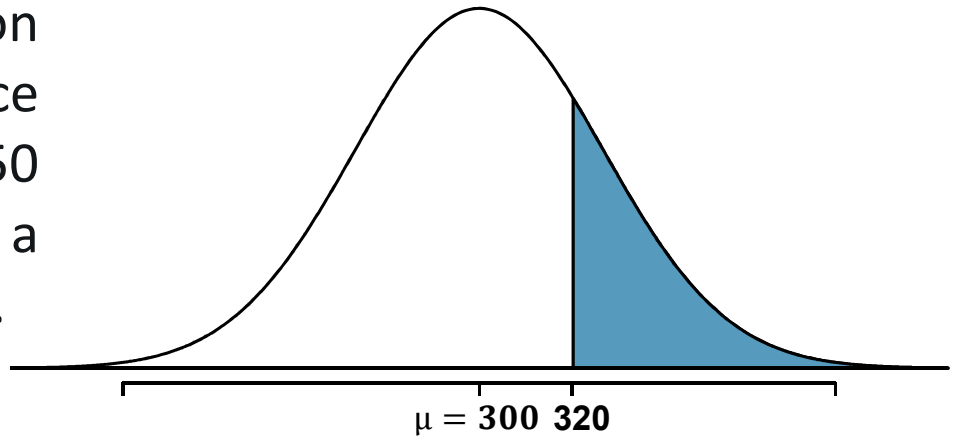


Since p-value is high, we do not reject H_0.

Interpreting the p-value

- If in fact college students have paid $300 for textbooks on average, there is a 21% chance that a random sample of 50 college students would yield a sample mean of $320 or higher.

p-value = $0.209 \approx 0.21$



- This is a pretty high probability, so we think that a sample mean of $320 or more is likely to happen simply by chance.

Making a Decision

- Since p-value is high (higher than 5%) we fail to reject H_0.
- The data do not provide convincing evidence that college students have paid more than $300 for textbooks on average.
- The difference between the null value of $300 textbooks cost and the observed sample mean of $320 is due to chance or sampling variability.

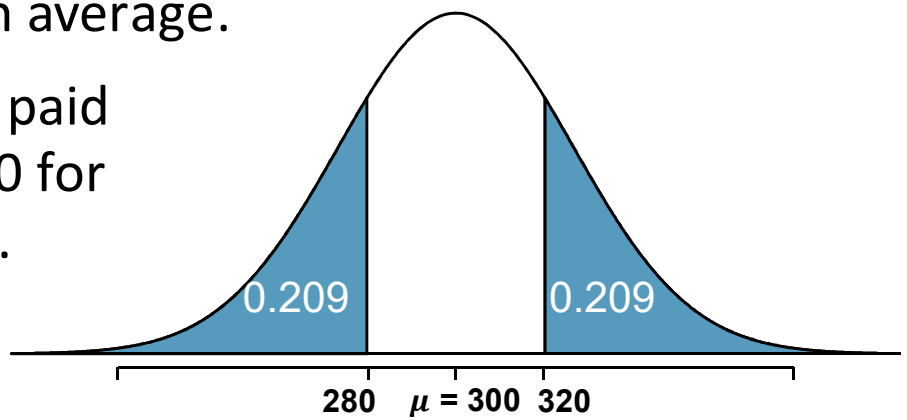
Two-sided Test

- Often instead of looking for a divergence from the null in a specific direction, we might be interested in divergence in any direction.
- We call such hypothesis tests two-sided (or two-tailed).
- The definition of a p-value is the same regardless of doing a one or two-sided test, however the calculation is slightly different since we need to consider “at least as extreme as the observed outcome” in both directions.

Example

$H_0: \mu = 300$ College students have paid $300 for textbooks, on average.

$H_A: \mu \neq 300$ College students have paid more or less than $300 for textbooks, on average.



p-value = $P(\bar{x} > 320 \text{ or } \bar{x} < 280 | H_0: \mu = 300)$

p-value = $P(Z > 0.81) + P(Z < -0.81) = 0.209 + 0.209 = 0.418$

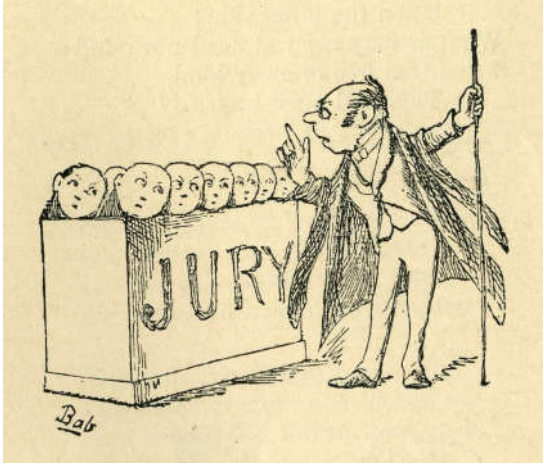
Decision Error

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✔	Type 1 error
	H_A true	Type 2 error	✔

- Type 1 error is rejecting H_0 when H_0 is true.
- Type 2 error is failing to reject H_0 when H_A is true.
- We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Hypothesis test as a trial

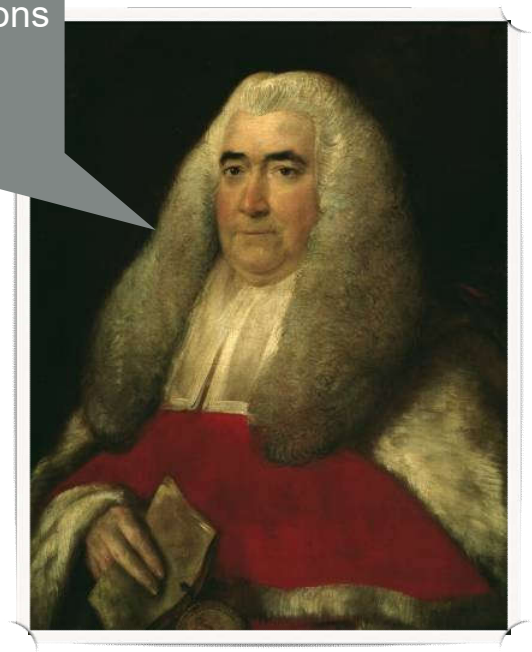
- If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:
  - H_0 : Defendant is innocent
  - H_A : Defendant is guilty



- Which type of error is being committed in the following circumstances?
  - Declaring the defendant innocent when they are actually guilty: *Type 2 error*
  - Declaring the defendant guilty when they are actually innocent: *Type 1 error*

Which error is the worse error to make?

- Which error is the worst error to make?
  - Type 2 : Declaring the defendant innocent when they are actually guilty
  - Type 1 : Declaring the defendant guilty when they are actually innocent



William Blackstone

Truth vs. Decision Table

goal: keep α and β low

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	Type 1 error, α
	H_A true	Type 2 error, β	$1-\beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level).
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β.
- Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$.

Recap: Hypothesis testing for a single mean

1. Set the hypotheses: H_0: $\mu = null\ value$
   H_A: $\mu < or > or \neq null\ value$
2. Calculate the point estimate: $\bar{x}$
3. Check conditions:
   - ***Independence:*** Sampled observations must be independent (random sample/ assignment & if sampling without replacement, $n < 10\%$ of population)
   - ***Sample size/skew:*** $n \geq 30$, larger if the population distribution is very skewed.
4. Draw sampling distribution, shade p-value, calculate test statistic:

$$Z = \frac{\bar{x} - \mu}{SE} \quad , \qquad SE = \frac{s}{\sqrt{n}}$$

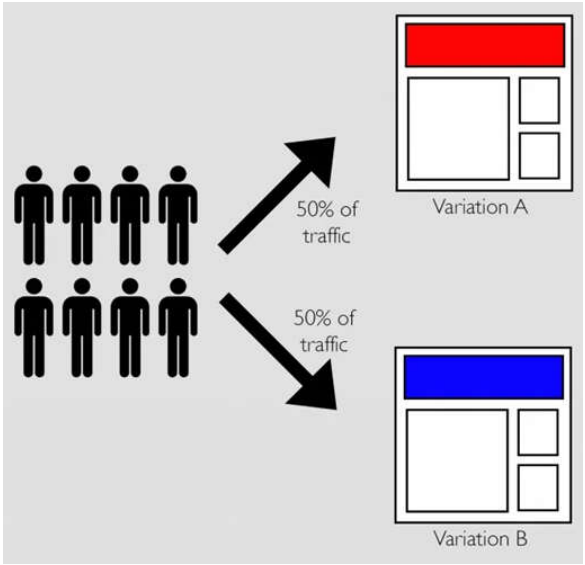
5. Make a decision, and interpret it in context of the research question:
   - If p-value $< \alpha$, reject H_0; the data provide convincing evidence for H_A.
   - If p-value $> \alpha$, fail to reject H_0; the data *do not* provide convincing evidence for H_A.

Statistical vs. Practical Significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not practically significant.
- The sample size is something the researcher has control over, because we can decide how many observations we want to sample.
  - when you see highly statistically significant results make sure that you also inquire whether the effect size is reported and what the sample size is as well.

Example

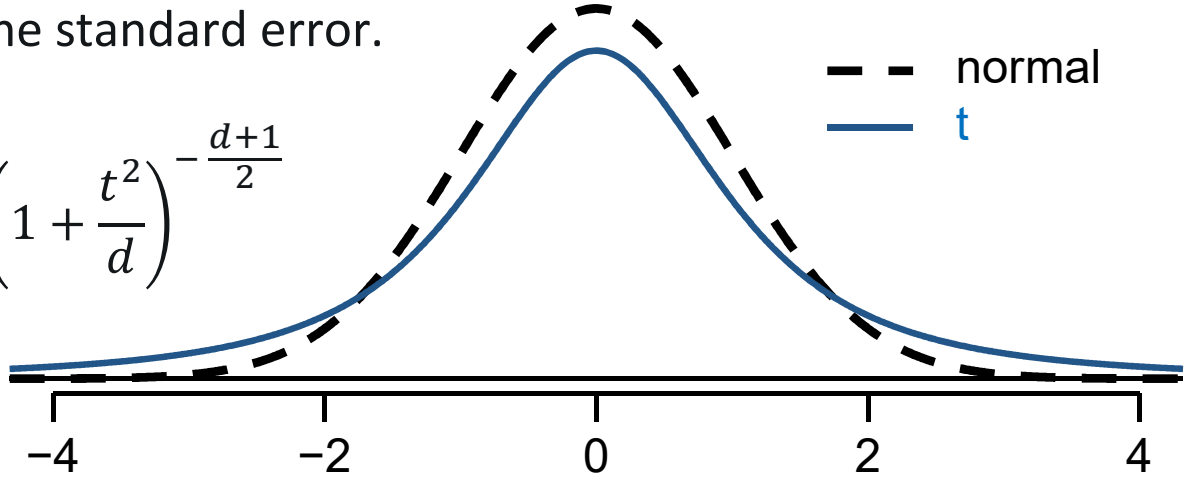
- **Scenario: A/B Testing for Click-Through Rate (CTR)**
  - Version A (Control): CTR = 4.00%
  - Version B (Treatment): CTR = 4.05%
  - p-value: 0.03
- **Statistical Significance:**
  - p-value < 0.05 → Difference unlikely due to chance
- **Practical Significance:**
  - Absolute CTR increase = 0.05%
  - Extra revenue = $1,000/month
  - Costly redesign required?



Student's t-distribution

- When σ is unknown, use the t-distribution to address uncertainty of the standard error estimate
- t-distribution is bell shaped but thicker tails than the normal
  - Observations more likely to fall beyond 2 SDs from the mean
  - Extra thick tails helpful for mitigating the effect of a less reliable estimate for the standard error.

$$f(t) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{\pi d}\,\Gamma\left(\frac{d}{2}\right)}\left(1+\frac{t^2}{d}\right)^{-\frac{d+1}{2}}$$



Example

- Suppose you have a two sided hypothesis test, and your test statistic is 2. Under which of these scenarios would you be able to reject the null hypothesis at the 5% significance level?

a. $P(|Z| > 2)$ 0.0455 → reject

b. $P(|t_{df=50}| > 2)$ 0.0509 → fail to reject?

c. $P(|t_{df=10}| > 2)$ 0.0734 → fail to reject

- Degrees of freedom for t statistic for inference on one sample mean:

$$df = n - 1$$