

A/B Testing & Statistical Thinking with 2022 BRFSS Data

Welcome to this comprehensive exploration of A/B testing in public health using the 2022 Behavioral Risk Factor Surveillance System (BRFSS) data. This presentation will ground you in the core statistical concepts needed to design, implement, and analyze effective A/B tests for public health interventions.



Why A/B Testing Matters in Public Health

1

Effective Intervention Design

Public health agencies create countless campaigns to encourage behavior change: smoking cessation, physical activity, cancer screening. Each intervention combines specific messaging, delivery methods, timing, and targeting—but how do we know which approach works best?

2

Evidence-Based Decision Making

A/B testing moves us from opinion-based to evidence-based decision making by systematically comparing approaches. Should we use fear-based messages or positive ones? Daily reminders or weekly? Testing provides definitive answers.

3

Real Health Outcomes

Public health isn't just about good intentions—it's about outcomes. When interventions fail, real people continue risky behaviors. When they succeed, lives improve. A/B testing ensures we focus resources on what actually works.

A/B testing helps us design more effective interventions by providing concrete evidence of what works. It can also address health equity by analyzing effects across demographic subgroups to identify which interventions work best for specific populations.

What Is A/B Testing?

A/B testing is a systematic method to compare two versions of an intervention to determine which performs better. The process involves:

1. Identifying your target audience
2. Randomly assigning participants to groups
3. Exposing one group to the control (current approach)
4. Exposing the other group to the treatment (new approach)
5. Measuring and comparing outcomes

Random assignment is crucial because it ensures groups are similar in all ways except for the intervention they receive, allowing for fair comparison.

Example: Testing cancer screening interventions by sending a basic postcard (control) versus personalized text messages with direct booking links (treatment), then measuring which approach yields higher screening rates.



The beauty of A/B testing is its versatility across public health interventions—you can test different message content, delivery channels, timing, or design elements. The key requirement is random assignment and a clear, measurable outcome.

Population vs. Sample in BRFSS



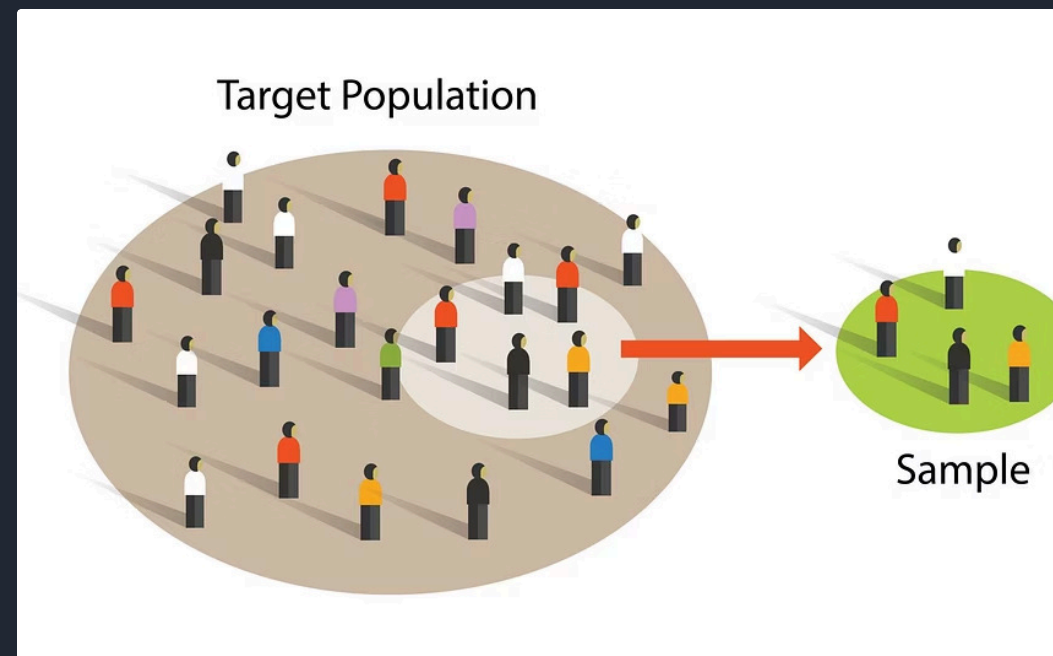
Population

The entire group we care about: could be all 260 million U.S. adults aged 18+. Impossible to survey everyone due to time and cost constraints.

Simple Random Sampling

Like stirring soup and taking a random ladle—you might get mostly broth or mostly vegetables just by chance.

The key insight: We don't need to survey everyone—we just need the right mix of people to make valid inferences about the whole population.



Sample

A carefully selected subset: ~400,000 people (less than 0.2% of adults) surveyed by BRFSS annually, chosen to represent the full population.

Stratified Sampling

BRFSS uses this approach—deliberately selecting the right proportions of each "ingredient" (age groups, geographic areas, racial/ethnic categories) to ensure the sample accurately represents the whole.

Mean, Variance & Standard Deviation

Mean

The average value—like finding that BRFSS respondents exercise an average of 3.2 days per week.

Standard Deviation

Measures how spread out values are around the mean. Low standard deviation means most people are close to the average; high standard deviation means values are widely dispersed.

Variance

The square of standard deviation—often used in calculations but harder to interpret directly since it's not in the original units.

Real-World Example

Community A: Low Variability

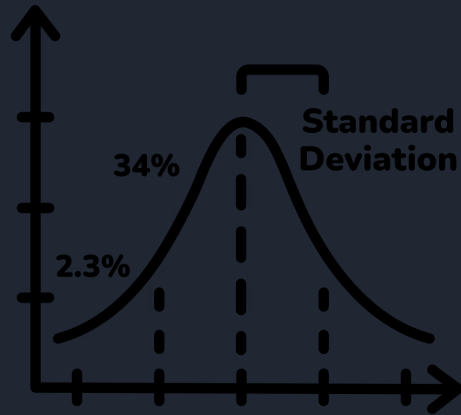
Most people exercise 2-4 days per week (close to 3.2 average). A single message encouraging "one more day of activity" might work for most.

Community B: High Variability

Half never exercise, half exercise daily (same 3.2 average). Requires completely different intervention approaches for each subgroup.

Understanding how consistent or variable behaviors are across a population is crucial for designing effective, targeted interventions.

Normal Distribution & Z-scores



Many health measures—like BMI, blood pressure, and cholesterol—often follow a normal distribution, creating a bell-shaped curve when graphed.

Comparing Different Measures

Z-scores allow comparison across completely different health measures. Someone 1.5 standard deviations above average for BMI can be meaningfully compared to someone 1.5 standard deviations above average for blood pressure.

Z-scores: Standardizing Comparisons

A z-score tells you how many standard deviations a value is from the mean, allowing comparison across different measurements.

Example: If average BMI is 27 with standard deviation 3.5, someone with BMI 33 has a z-score of:

$$(33 - 27) \div 3.5 = 1.7$$

This person's BMI is 1.7 standard deviations above average.

Identifying Risk

Z-scores help identify outliers who may need intervention. A z-score above 2 represents the most extreme 5% of the population—often a signal for public health professionals to intervene.

Central Limit Theorem (CLT)

The Central Limit Theorem is one of statistics' most powerful concepts: even when individual data is messy or skewed, the distribution of sample means becomes normal (bell-shaped) as sample size increases.

Individual Data

Individual alcohol consumption reports might be skewed—most people drink little or nothing, while a few drink heavily.

Take Multiple Samples

Randomly select 100 people and calculate their average alcohol consumption. Repeat this process 50 times with different groups.

Magic Happens

When you plot these 50 sample averages, they form a bell curve—even though the original data was skewed.

Why This Matters for Public Health

The CLT allows us to apply powerful statistical tools even when our original data isn't perfectly normal. It's why BRFSS can survey 400,000 people and make confident statements about all 260 million U.S. adults. The key insight: individual behaviors can be unpredictable, but averages of groups become remarkably predictable.

Confidence Intervals

Expressing Uncertainty

Rather than claiming "The average is exactly 3.2 days of exercise per week," a confidence interval gives a range: "The average is between 2.9 and 3.5 days per week."

What a 95% Confidence Interval Means

When we say "95% confidence interval," we mean we're 95% confident the true population average falls within this range. In other words, if we conducted this study 100 times, the true average would fall within our interval approximately 95 times.

Narrow Interval

A narrow interval (like 3.1 to 3.3 days) indicates high precision—we're fairly certain about our estimate.

Wide Interval

A wide interval (like 2.5 to 3.9 days) shows greater uncertainty in our estimate.

What Controls Interval Width?

1. Variability in responses—more consistent answers create narrower intervals
2. Sample size—larger samples produce narrower, more precise intervals

Comparing Two Groups – BRFSS Example

The Key Question

How do we know if an observed difference between groups represents a real effect or just random variation?

Example: Testing two flu vaccination strategies

- Group A (500 people): Standard postcard reminders
- Group B (500 people): Personalized text messages

The Results

Group	Vaccinated	Total People	Vaccination Rate
Group A (Postcards)	225	500	45%
Group B (Texts)	260	500	52%
Difference	35	-	7 percentage points

The Critical Question

Is this 7 percentage point difference real, or could it have happened by chance even if both methods were equally effective? This is where hypothesis testing comes in—it gives us a structured, mathematical approach to determine whether observed differences represent genuine effects or just random variation.

Answering the Question with Confidence Intervals

Interpreting Confidence Intervals for Differences

If both methods were truly identical, the true difference between them would be zero. But our sample showed a 7 percentage point difference.

1

Confidence Interval: 2 to 12 percentage points

This interval does NOT include zero, suggesting the difference is probably not just random chance. The text messages likely provide a real benefit between 2 and 12 percentage points.

2

Confidence Interval: -1 to 8 percentage points

This interval DOES include zero, meaning we can't be sure there's any real effect. The true effect could be zero (no difference) or even negative.

3

Confidence Interval: -8 to -2 percentage points

This interval does NOT include zero, but all values are negative, suggesting the intervention probably makes things worse.

What Confidence Intervals Tell Us

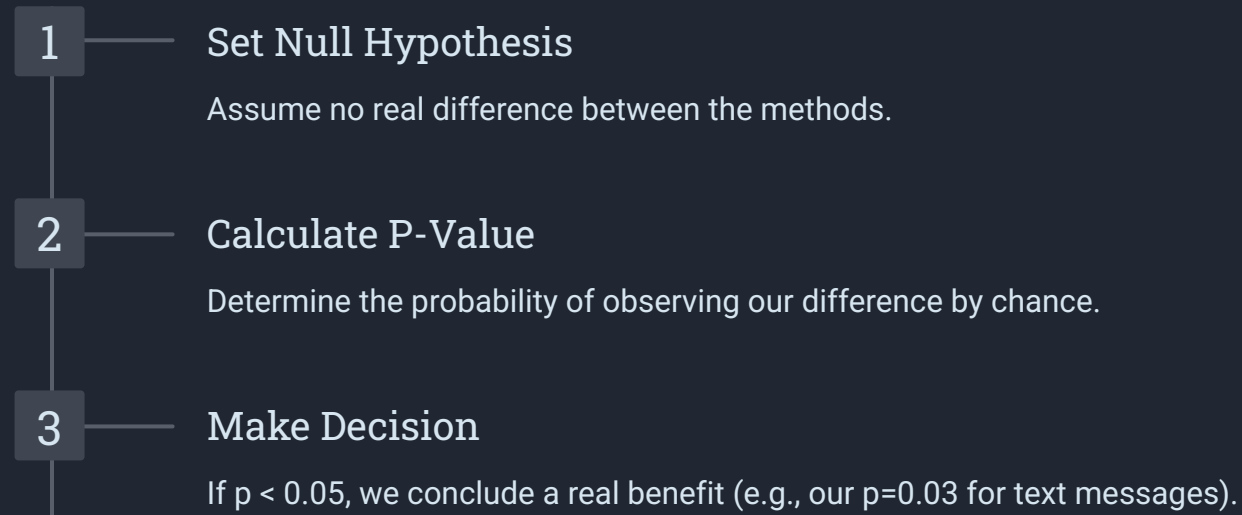
1. Is there an effect? If the interval excludes zero, then yes, there's probably a real effect.
2. How big might that effect be? The range shows the plausible size of the impact.
3. How precise is our estimate? Narrow intervals indicate more precision.

Answering the Question with P-Values

What P-Values Tell Us

A p-value tells us the probability of observing a difference as large as ours (or larger) if there were no real difference between the groups. For our flu vaccination example, a p-value of 0.03 means there's only a 3% chance of seeing a 7-percentage-point difference (or more) if texts and postcards were equally effective.

The P-Value Process



Interpreting P-Values

A lower p-value indicates stronger evidence against the null hypothesis (i.e., a real effect). For instance:

- **p = 0.01:** Very strong evidence.
- **p = 0.03:** Good evidence (our result).
- **p > 0.05:** Not significant; the difference is likely due to chance.

Putting P-values, Confidence Intervals, and Effect Size Together

Complete Statistical Analysis

P-Value: $p = 0.03$

"If text messages and postcards were equally effective, we'd only see a 7-point difference or larger about 3% of the time by random assignment."

Since 3% is below our 5% threshold, this difference is probably NOT just random chance.

Confidence Interval: 2 to 12 percentage points

"We're 95% confident the true difference is between 2 and 12 percentage points."

Since zero falls outside our range, this confirms the difference is probably real.

Effect Size: 7 percentage points

The magnitude of the difference we observed—our best estimate of the real-world impact.

This tells us switching from postcards to text messages would yield a meaningful improvement.

Practical Decision Making

Combining all three pieces of information: "We have strong statistical evidence that text message reminders work better than postcards ($p = 0.03$). We observed a 7 percentage point improvement, and the true improvement is likely between 2 and 12 percentage points. This suggests text messages provide a meaningful benefit worth implementing."

Choosing the Right Statistical Test

Z-Tests vs. T-Tests

Both tests can give us p-values and confidence intervals for comparing groups, but they're designed for different situations based on two key factors:

1. How many people are in our study?
2. How confident are we about the variability in our data?

1	2
Z-Test Use when you have BOTH: • Large samples (30+ people per group) • Good estimate of population variability Example: Comparing flu vaccination strategies across Texas using BRFSS data with thousands of respondents.	T-Test Use when you have EITHER: • Small samples (fewer than 30 per group) • Uncertain about population variability Example: Studying diabetes screening in rural Vermont with only 25 people per group.

Practical Decision Rules

When in doubt, use a t-test—it's the more conservative choice and will protect you from overconfident conclusions. With very large samples, z-tests and t-tests give almost identical results.

Common Mistakes and How to Avoid Them

1

Confusing Statistical Significance with Practical Importance

The mistake: "We got $p = 0.02$, so this intervention is definitely worth implementing!"

The solution: Always look at BOTH statistical significance AND effect size. A statistically significant but tiny effect may not be worth implementing.

2

Assuming Non-significant Results Mean "No Effect"

The mistake: "We got $p = 0.08$, so the intervention doesn't work."

The solution: Consider whether your study had enough STATISTICAL POWER to detect meaningful effects. Non-significance doesn't prove no effect—it just means insufficient evidence.

3

P-hacking or Cherry-picking Results

The mistake: Running multiple analyses until finding something significant, then only reporting that result.

The solution: Decide on your analysis plan BEFORE looking at the data. If exploring subgroups, adjust significance thresholds accordingly.

4

Ignoring Confidence Intervals

The mistake: Only looking at p-values and ignoring what confidence intervals tell us about precision.

The solution: Always examine confidence intervals. Wide intervals mean more uncertainty. Narrow intervals mean more precision.

5

Forgetting about Type I and Type II Errors

The mistake: Not considering the consequences of being wrong in either direction.

The solution: Consider the context and consequences. For expensive interventions, be more conservative. For low-risk, high-potential interventions, you might accept weaker evidence.

Statistical Power and Sample Size Planning

What is Statistical Power?

Statistical power is the probability that our test will detect a real effect when there actually is one.

High power (80-90%): Very likely to catch real effects

Low power (30-40%): Often miss real effects, even when they exist

Power = 1 - (probability of Type II error)

What Determines Power?



Sample Size

More participants = more power. For a 7 percentage point effect:

100 people per group: ~35% power

500 people per group: ~85% power



Effect Size

Bigger effects are easier to detect. With 500 people per group:

2 percentage point difference: ~25% power

7 percentage point difference: ~85% power



Variability

Less variable data gives more power. Highly consistent responses make differences easier to detect.

Power Analysis Before the Study

Calculate power BEFORE collecting data to ensure your study can reliably detect meaningful effects. Underpowered studies waste resources and can miss effective interventions.

Effect Size and Practical Significance

Beyond Statistical Significance

Effect size measures HOW BIG a difference is, independent of sample size and statistical significance.

In our flu vaccination example:

- Raw difference: 7 percentage points (52% - 45%)
- Relative effect: 15.6% increase $((52-45)/45)$
- Cohen's d: 0.35 (between small and medium effect)

Population-Level Impact

Even "small" effects can be hugely important in public health when applied to large populations:

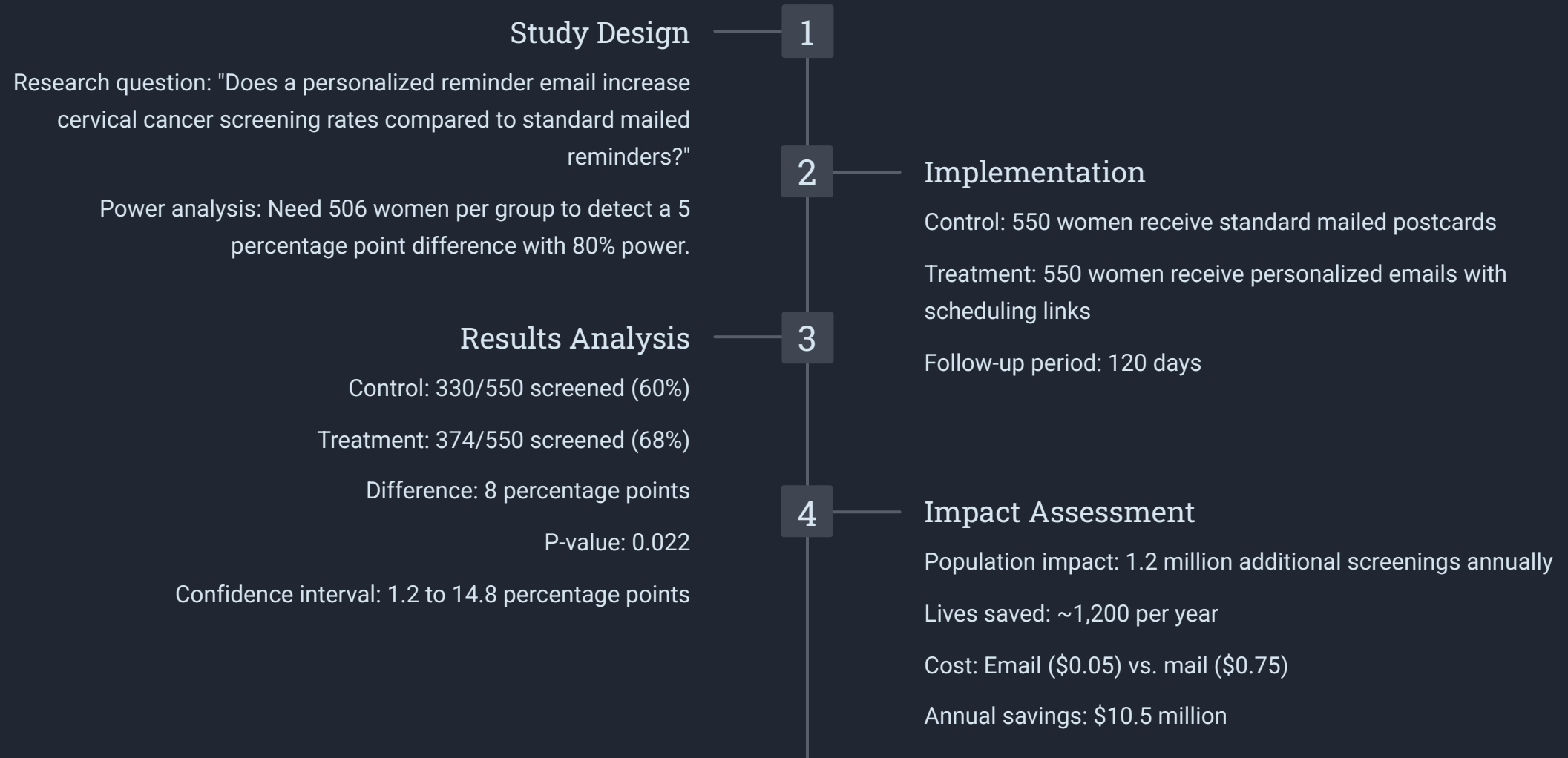
260M	18M	18K	\$144M
Eligible Adults	Additional Vaccinations	Hospitalizations Prevented	Healthcare Savings
U.S. adults eligible for flu vaccination	With 7 percentage point improvement (45% → 52%)	At 1 prevention per 1,000 vaccinations	At \$8,000 per hospitalization

Context Matters

When evaluating practical significance, consider health impact, population size, intervention cost, alternatives, and implementation difficulty. A 7 percentage point improvement might be life-changing for cancer screening but trivial for hand sanitizer use.

Complete A/B Test Analysis - Putting It All Together

Case Study: Increasing Cervical Cancer Screening



Final Recommendation

Implement email reminders as the new standard for cervical cancer screening outreach, with consideration for backup mailed reminders for older women or those without reliable email access.

Summary and Next Steps

What You've Accomplished

Throughout this presentation, you've built a comprehensive toolkit for evidence-based public health decision making:

- Understanding foundational statistical concepts
- Learning hypothesis testing frameworks
- Mastering study design principles
- Applying statistical thinking to real-world problems
- Translating results into policy recommendations

Key Takeaways



Evidence Matters

A/B testing transforms public health from opinion-based to evidence-based decision making



Statistical Rigor

Proper statistical analysis separates real effects from random variation



Context Counts

Effect size and practical significance depend on health impact, population, cost, and alternatives



Apply with Purpose

Use these tools to design better interventions that improve health outcomes

Conclusion

We've journeyed from the foundational concepts of statistical thinking to the practical application of A/B testing in public health. You now understand how to design robust studies, analyze data with statistical rigor, and interpret results like p-values and confidence intervals. More importantly, you've seen how to translate these statistical findings into meaningful insights and actionable recommendations that can genuinely improve population health outcomes.