J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Sampling and Sample Size

Rohit Naimpally
Research & Training Manager
J-PAL Global

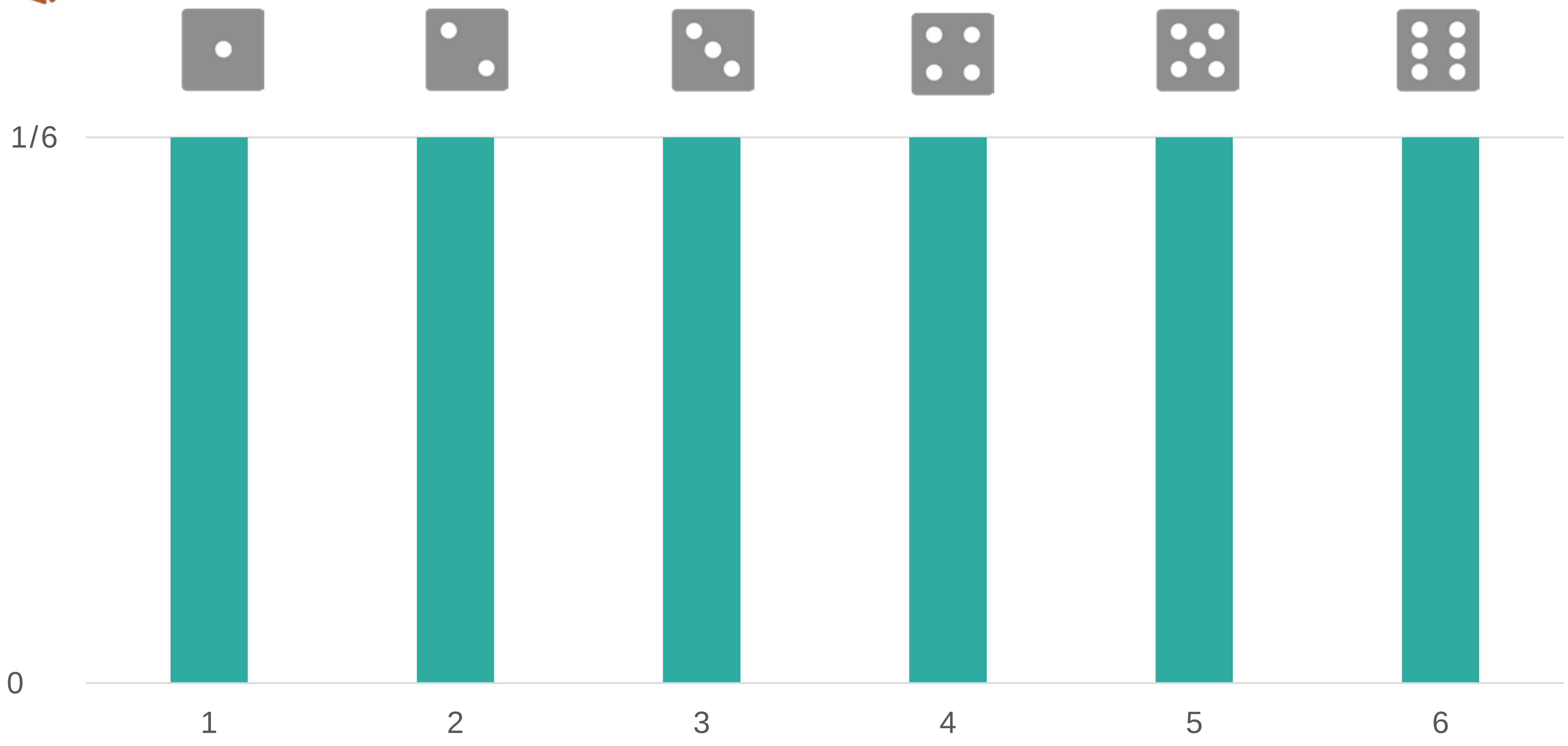
Course Overview

1. What is Evaluation?
2. Outcomes, Impact, and Indicators
3. Why Randomize?
4. How to Randomize
5. Sampling and Sample Size
6. Threats and Analysis
7. Evaluation from Start to Finish
8. Evidence from Community-Driven Development, Health, and Education Programs
9. Using Evidence from Randomized Evaluations

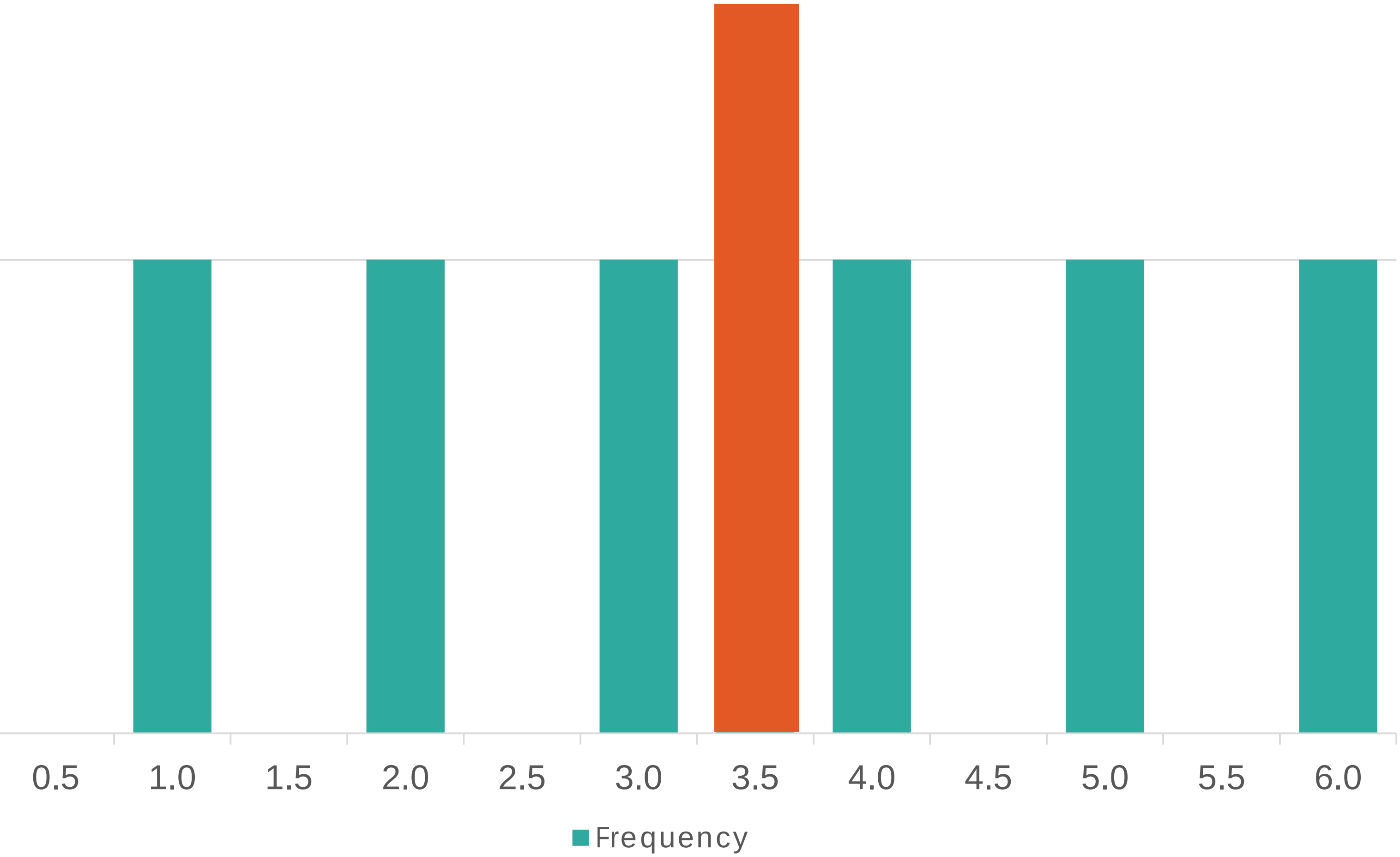
What's the average result?

- If you were to roll a die once, what is the "expected result"? (i.e. the average)

Possible results & probability: 1 die



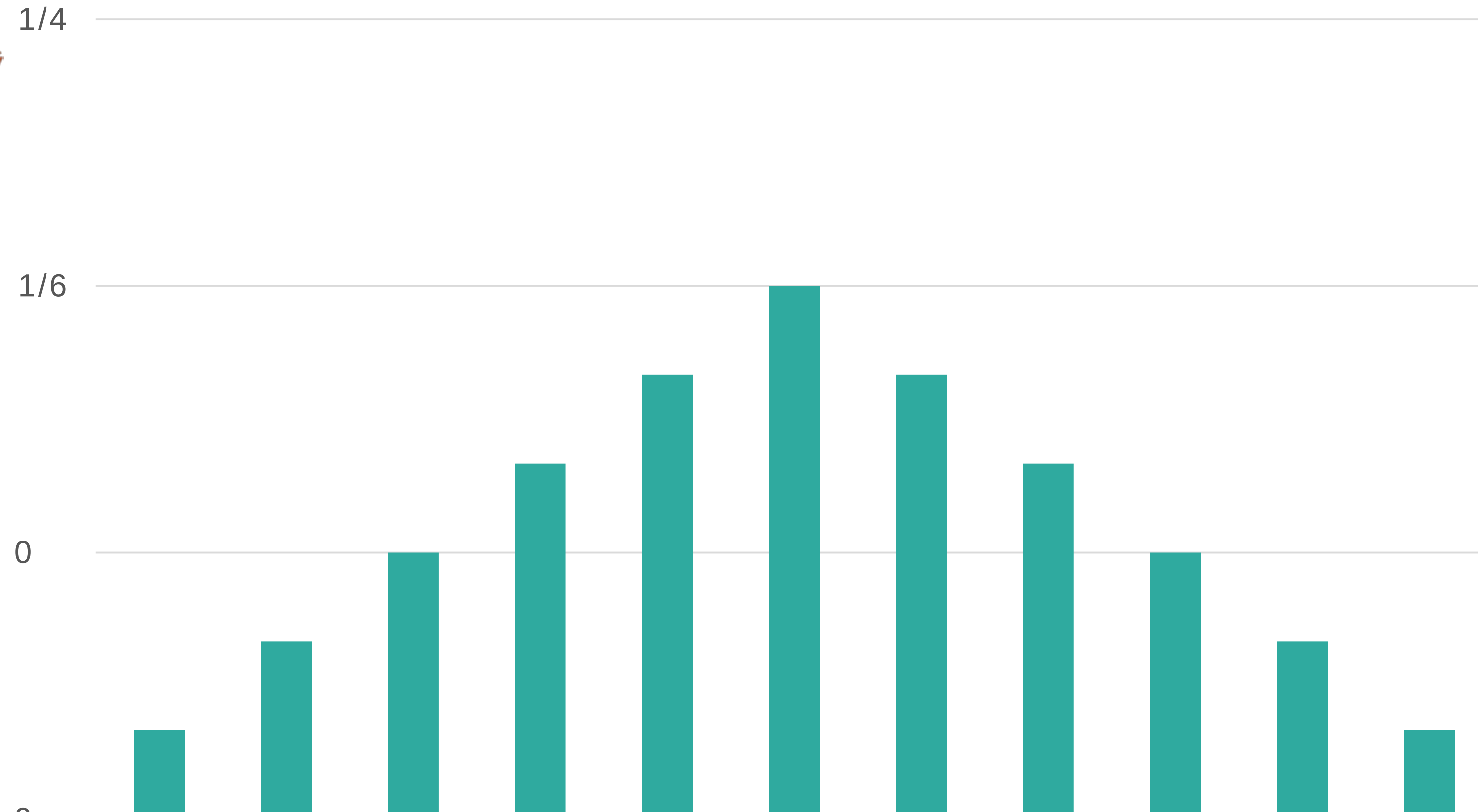
Rolling 1 die: possible results & average



What's the average result?

- If you were to roll two dice once, what is the expected average of the two dice?

Rolling 2 dice:
Possible totals & likelihood

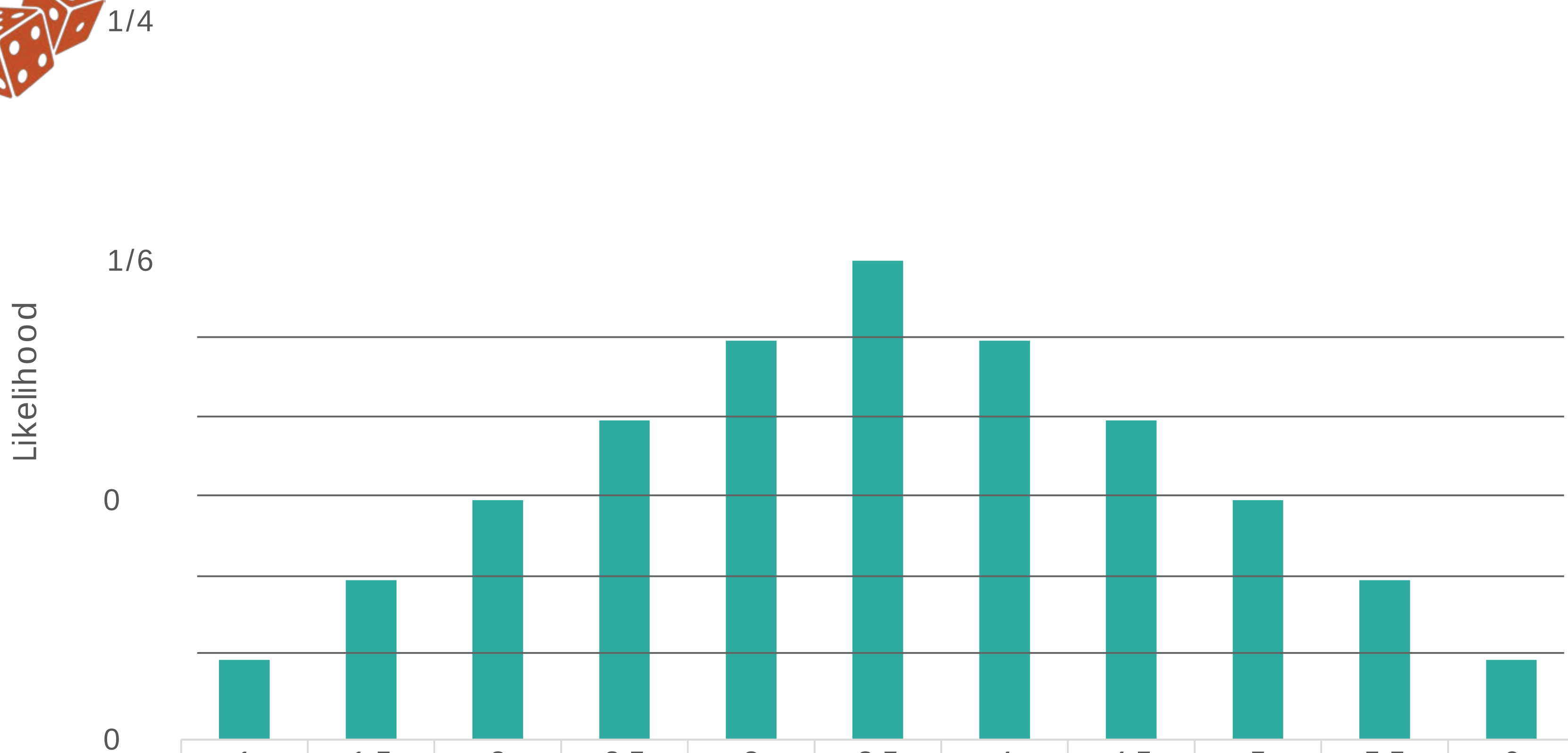


	2	3	4	5	6	7	8	9	10	11	12
Frequency	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

Rolling 2 dice: possible totals
12 possible totals, 36 permutations

		Die 1					
		⚀	⚁	⚂	⚃	⚄	⚅
Die 2	⚀	2	3	4	5	6	7
	⚁	3	4	5	6	7	8
	⚂	4	5	6	7	8	9
	⚃	5	6	7	8	9	10
	⚄	6	7	8	9	10	11
	⚅	7	8	9	10	11	12

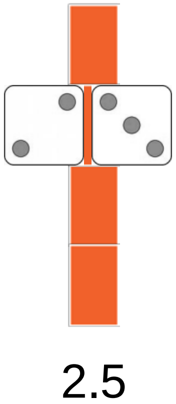
Rolling 2 dice:
Average score of dice & likelihood



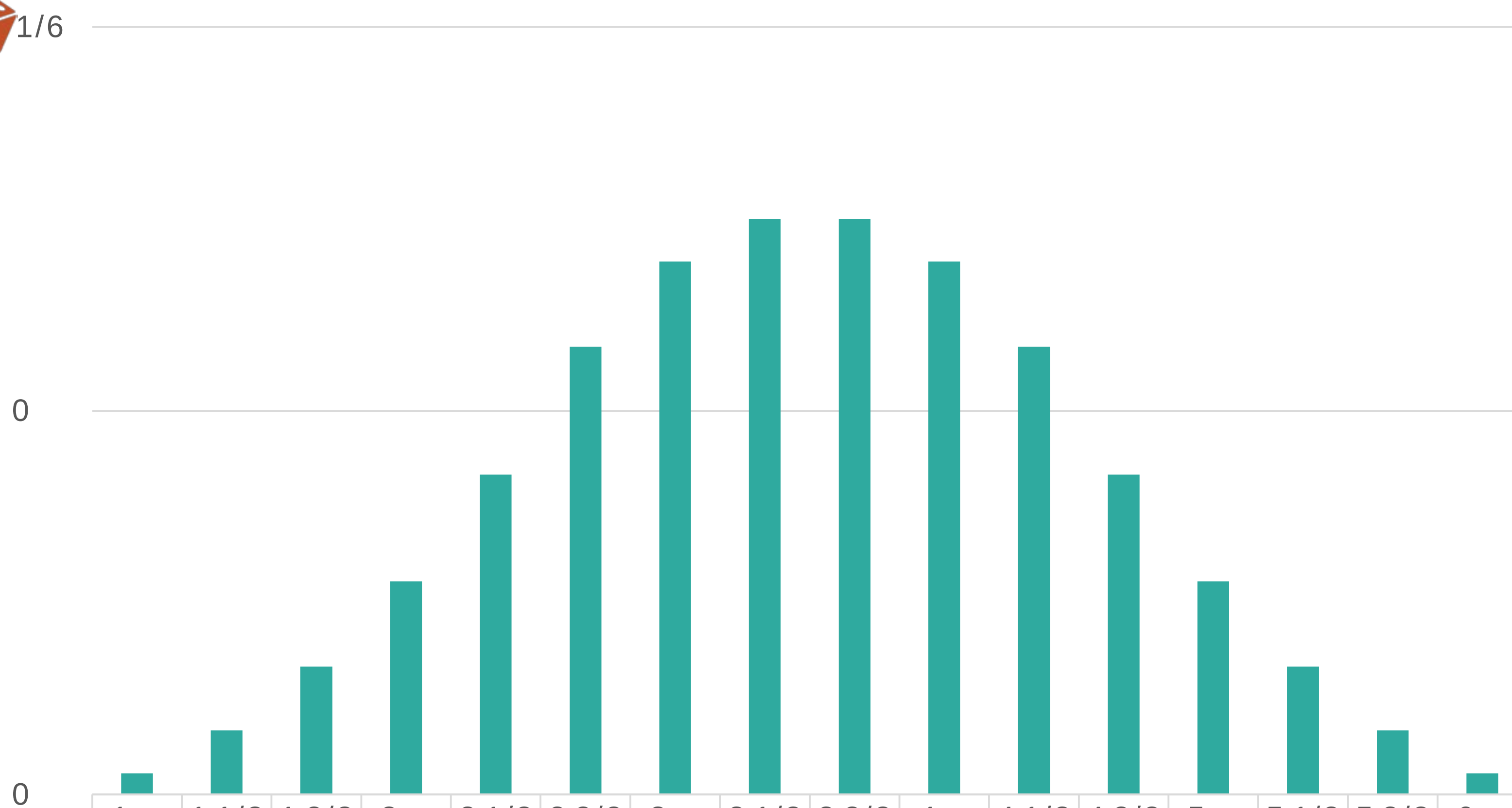
	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6
Frequency	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

Outcomes and Permutations

- Putting together permutations, you get:
  1. All possible outcomes
  2. The likelihood of each of those outcomes
     - Each column represents one possible outcome (average result)
     - Each block within a column represents one possible permutation (to obtain that average)

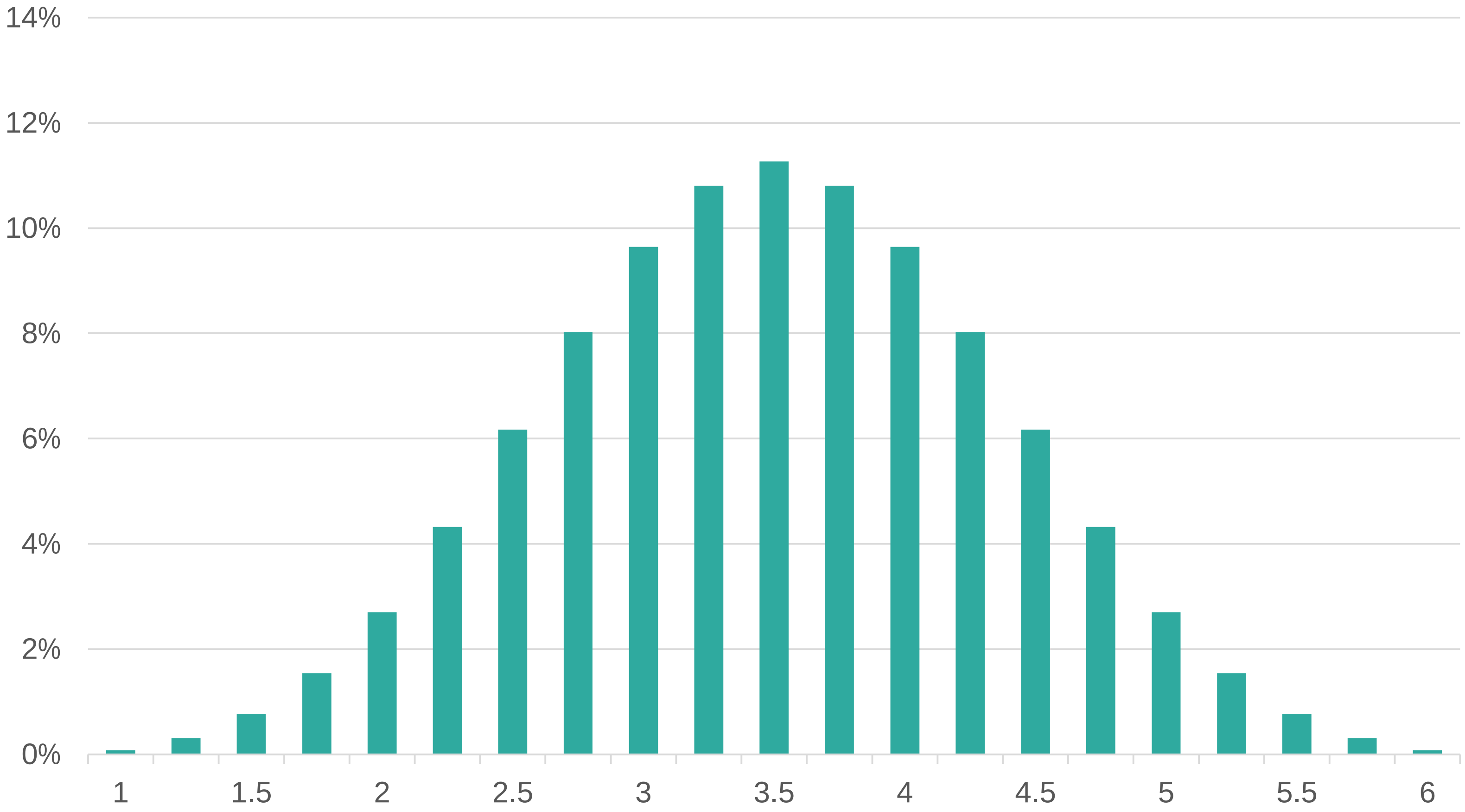


Rolling 3 dice:
16 results 3→18, 216 permutations

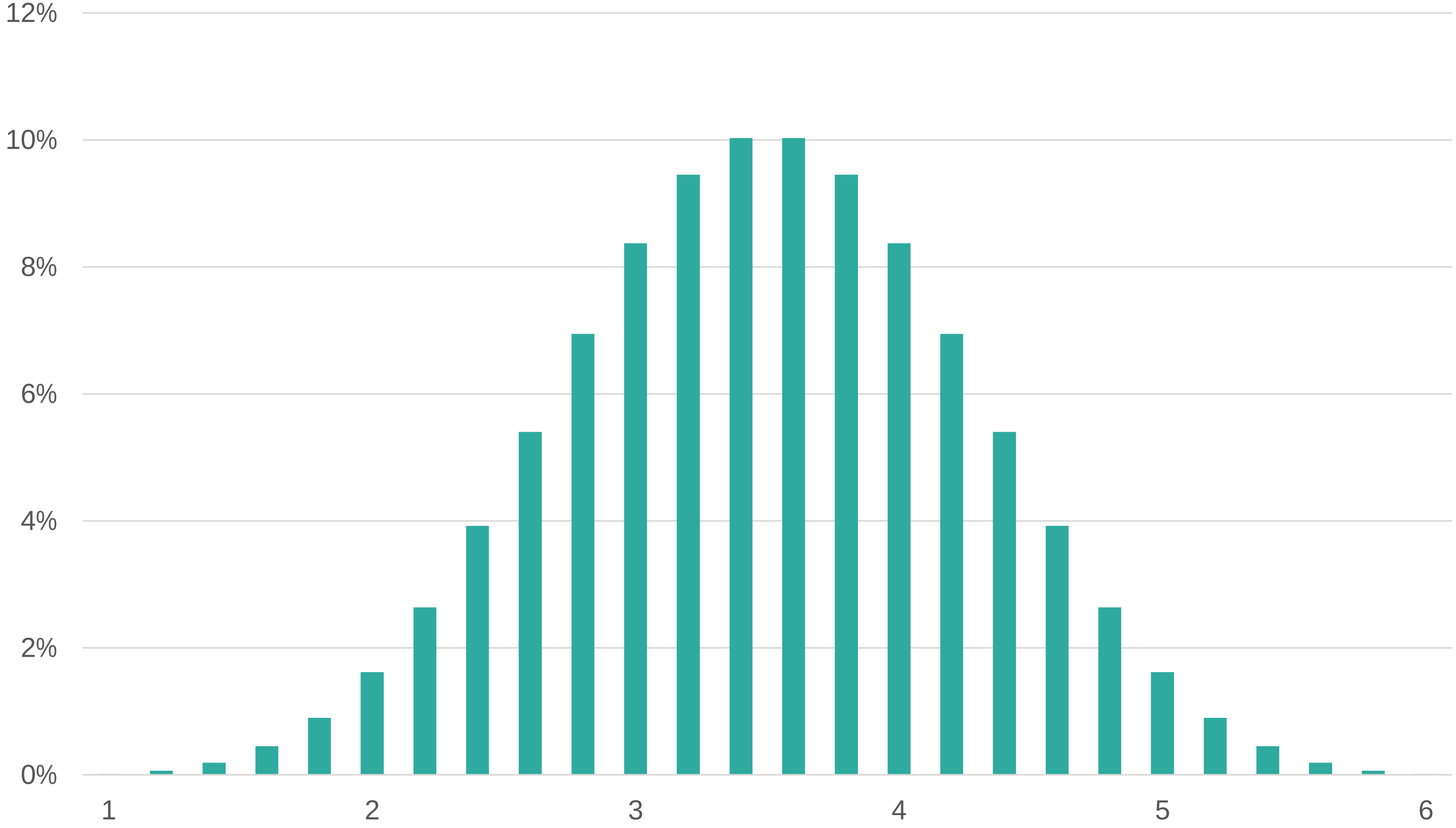


	1	1 1/3	1 2/3	2	2 1/3	2 2/3	3	3 1/3	3 2/3	4	4 1/3	4 2/3	5	5 1/3	5 2/3	6
Frequency	0%	1%	3%	5%	7%	10%	12%	13%	13%	12%	10%	7%	5%	3%	1%	0%

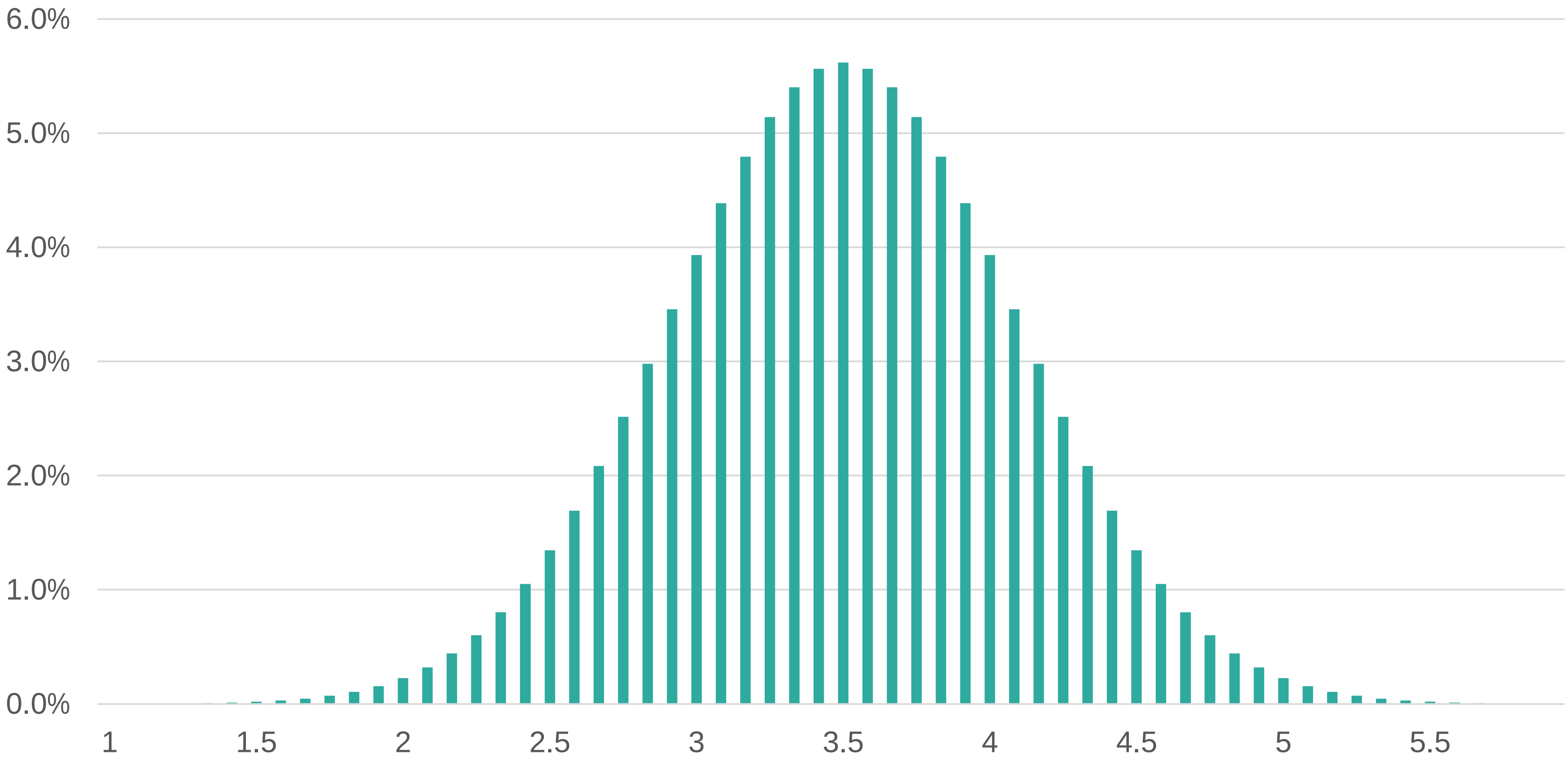
Rolling 4 dice:
21 results, 1296 permutations



Rolling 5 dice:
26 results, 7776 permutations

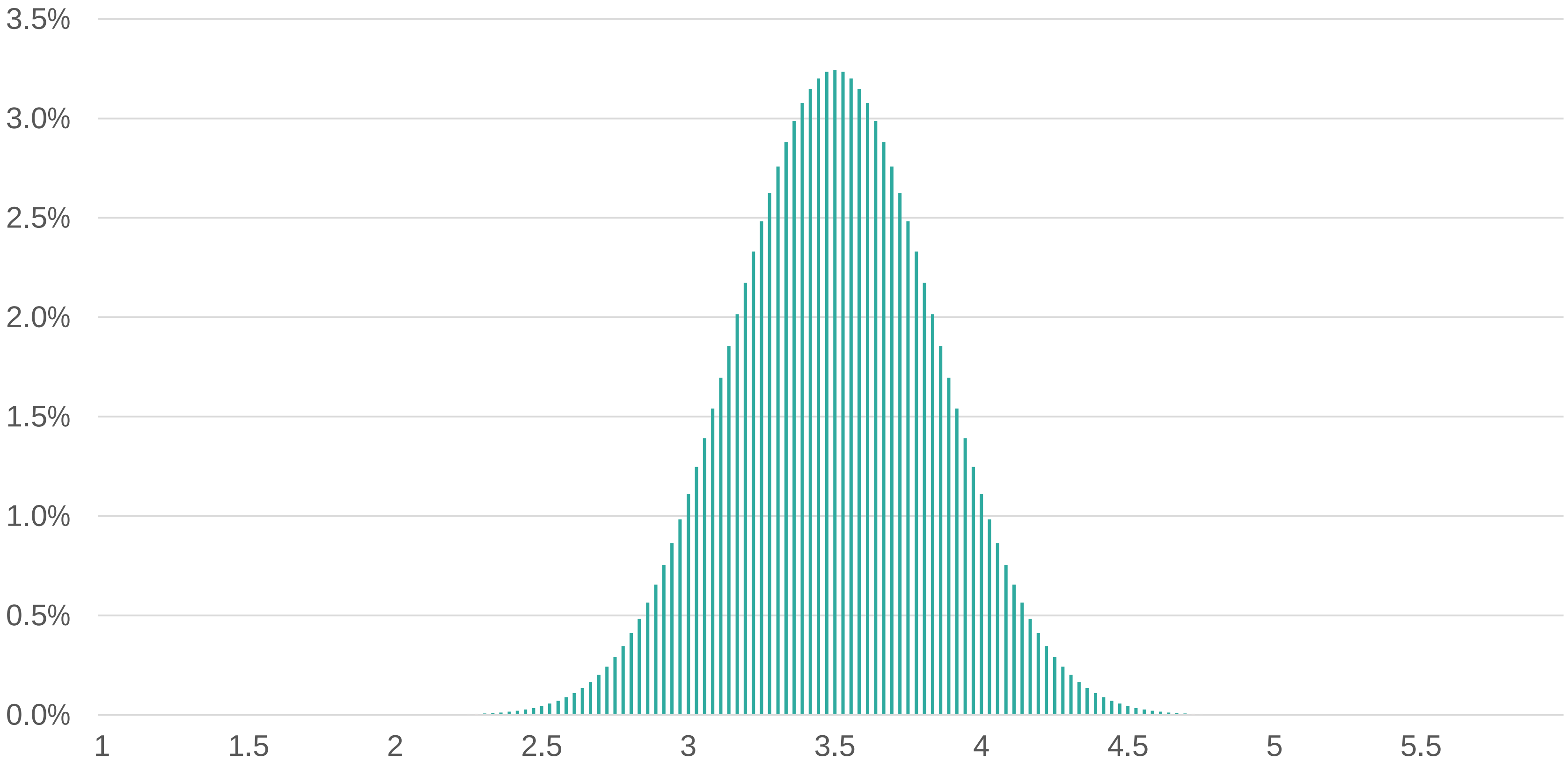


Rolling 10 dice: 50 results, >60 million permutations



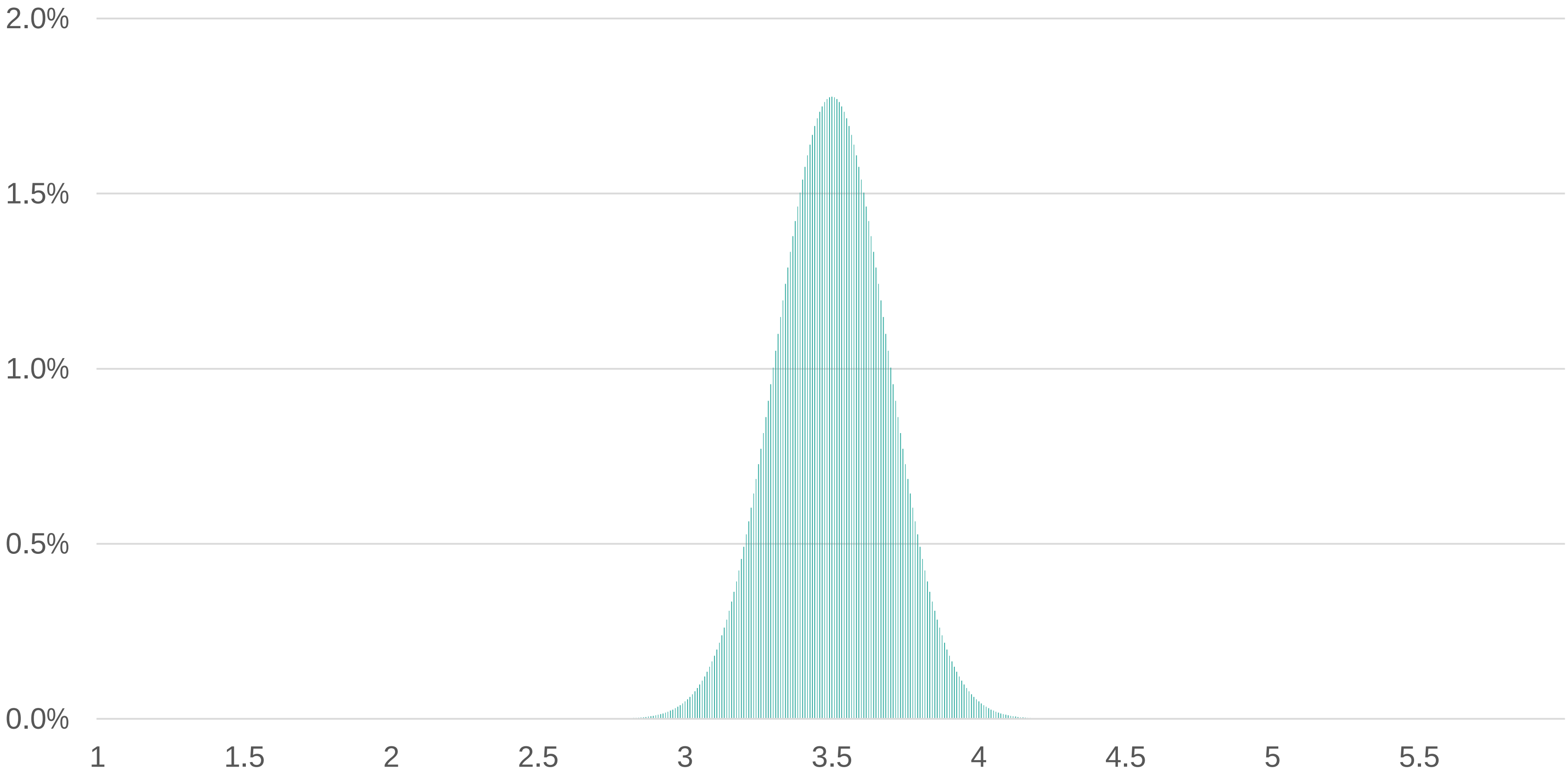
Looks like a bell curve, or a *normal* distribution

Rolling 30 dice:
150 results, 2 x 10 ^{23} permutations*



>95% of all rolls will yield an average between 3 and 4

Rolling 100 dice:
500 results, 6 x 10 77 permutations



>99% of all rolls will yield an average between 3 and 4

Rolling dice: 2 lessons

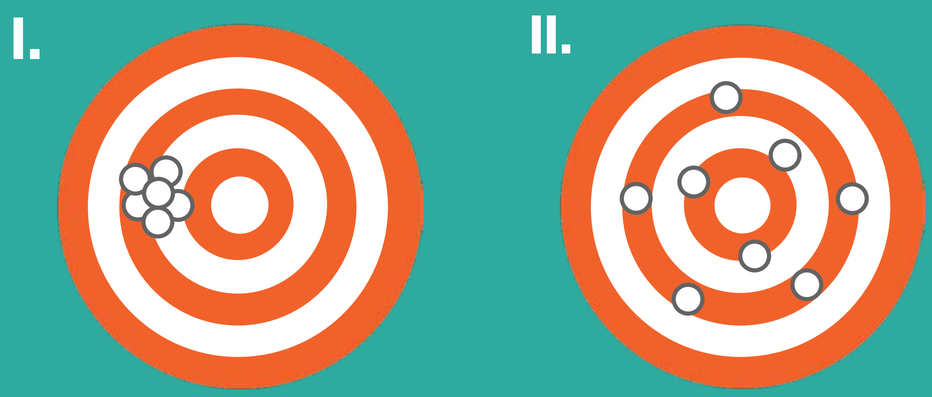
1. The more dice you roll, the closer most averages are to the true average (the distribution gets "tighter")

   -THE LAW OF LARGE NUMBERS-

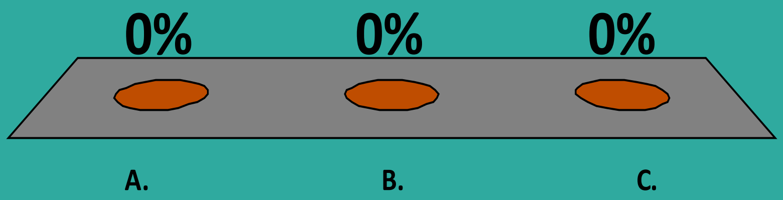
2. The more dice you roll, the more the distribution of possible averages (the sampling distribution) looks like a bell curve (a normal distribution)

   -THE CENTRAL LIMIT THEOREM-

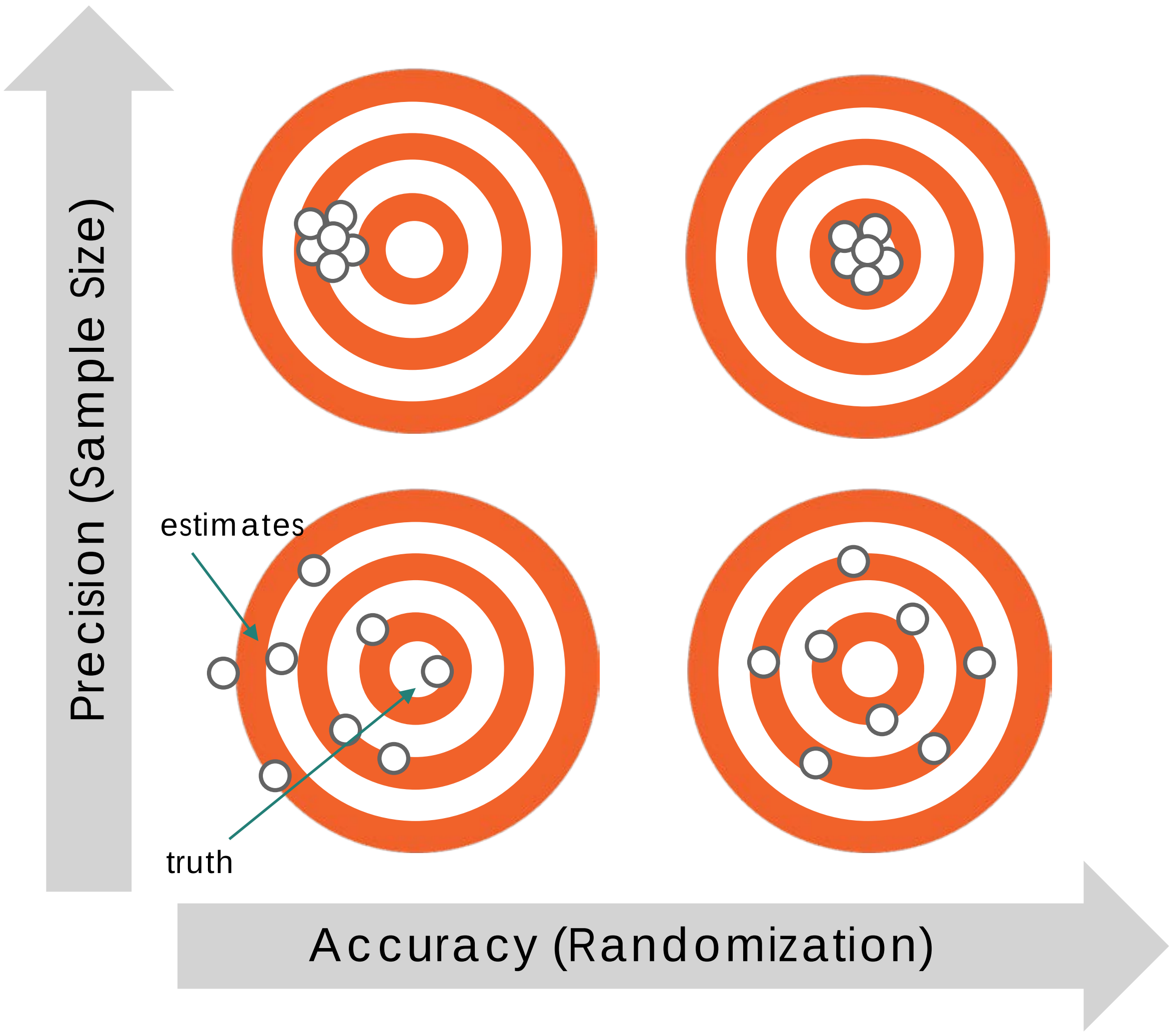
Which of these is more accurate?



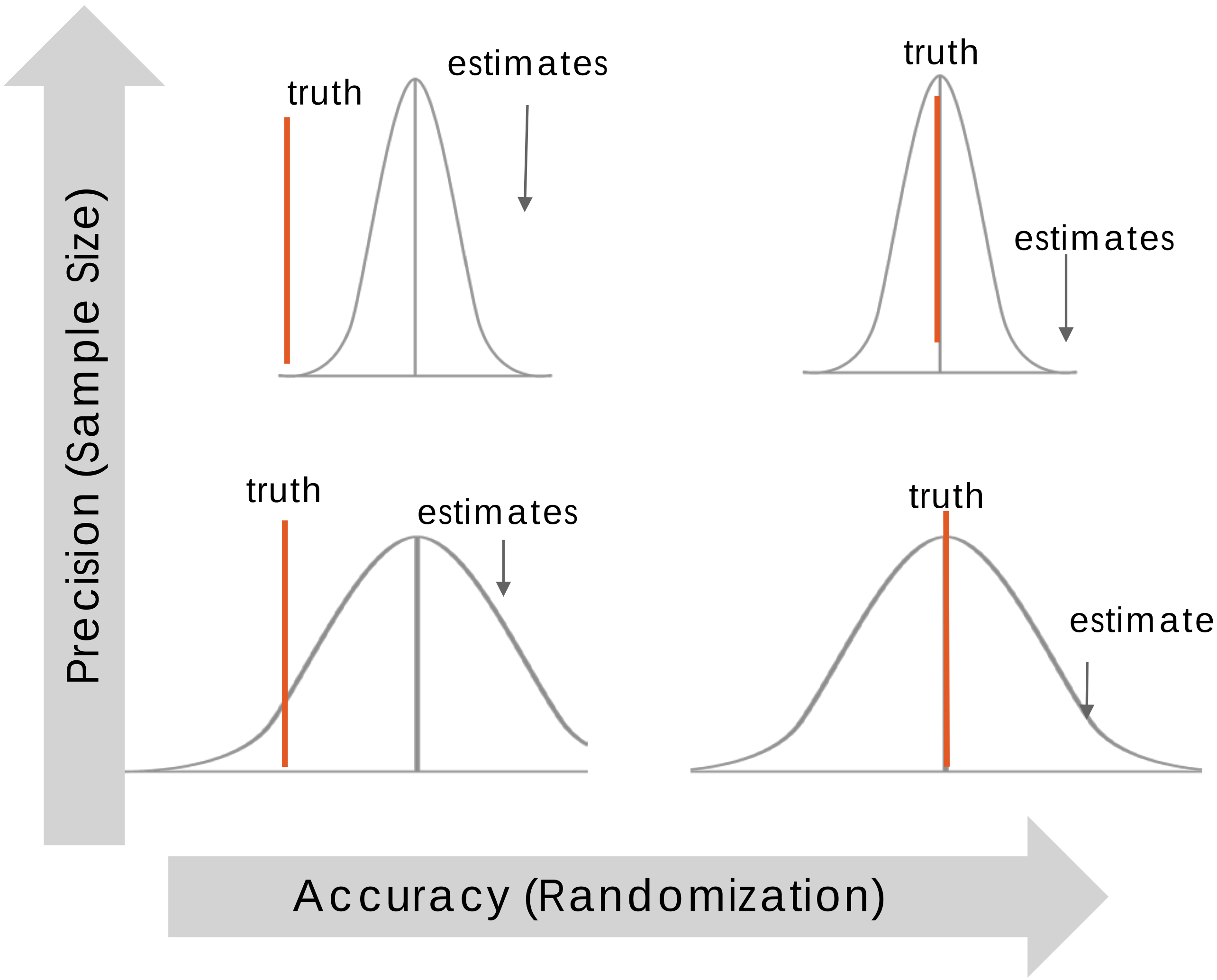
A. I.
B. II.
C. Don’t know



Accuracy versus Precision



Accuracy versus Precision



THE basic questions in statistics

- How confident can you be in your results?

- → How big does your sample need to be?

That was just the introduction

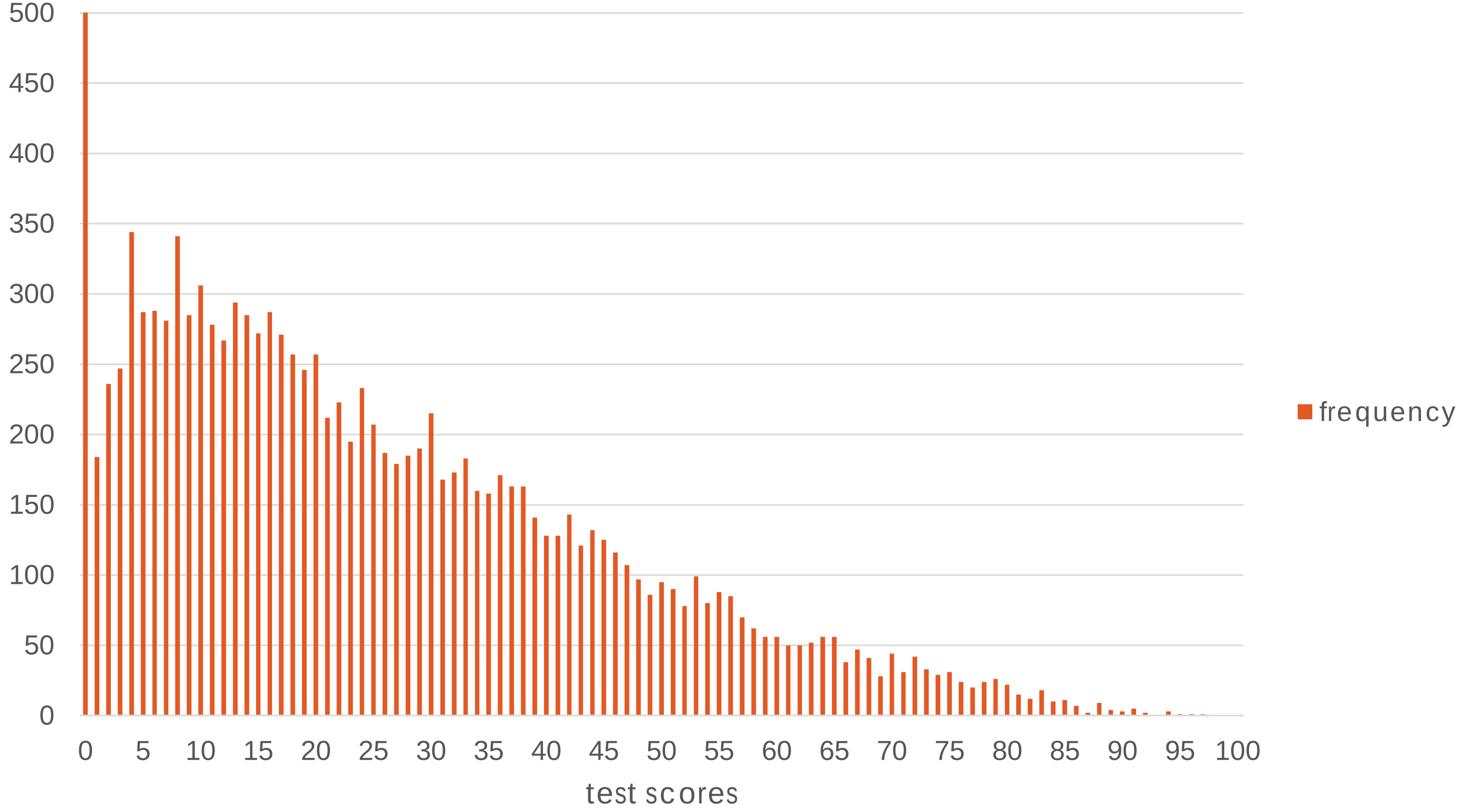
Outline

- Sampling distributions
  - population distribution
  - sampling distribution
  - law of large numbers/central limit theorem
  - standard deviation and standard error
- Detecting impact

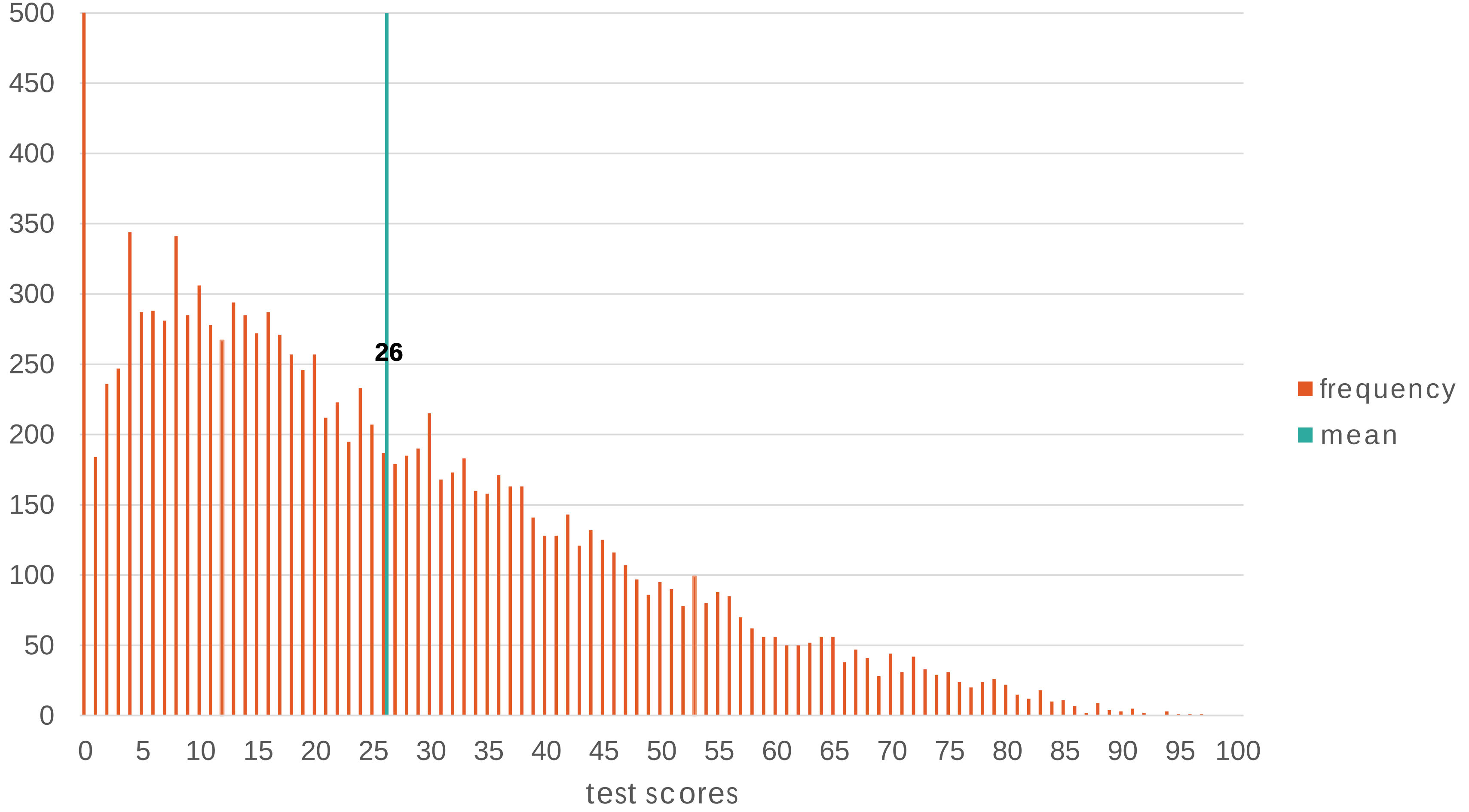
Outline

- Sampling distributions
  - population distribution
  - sampling distribution
  - law of large numbers/central limit theorem
  - standard deviation and standard error
- Detecting impact

Baseline test scores

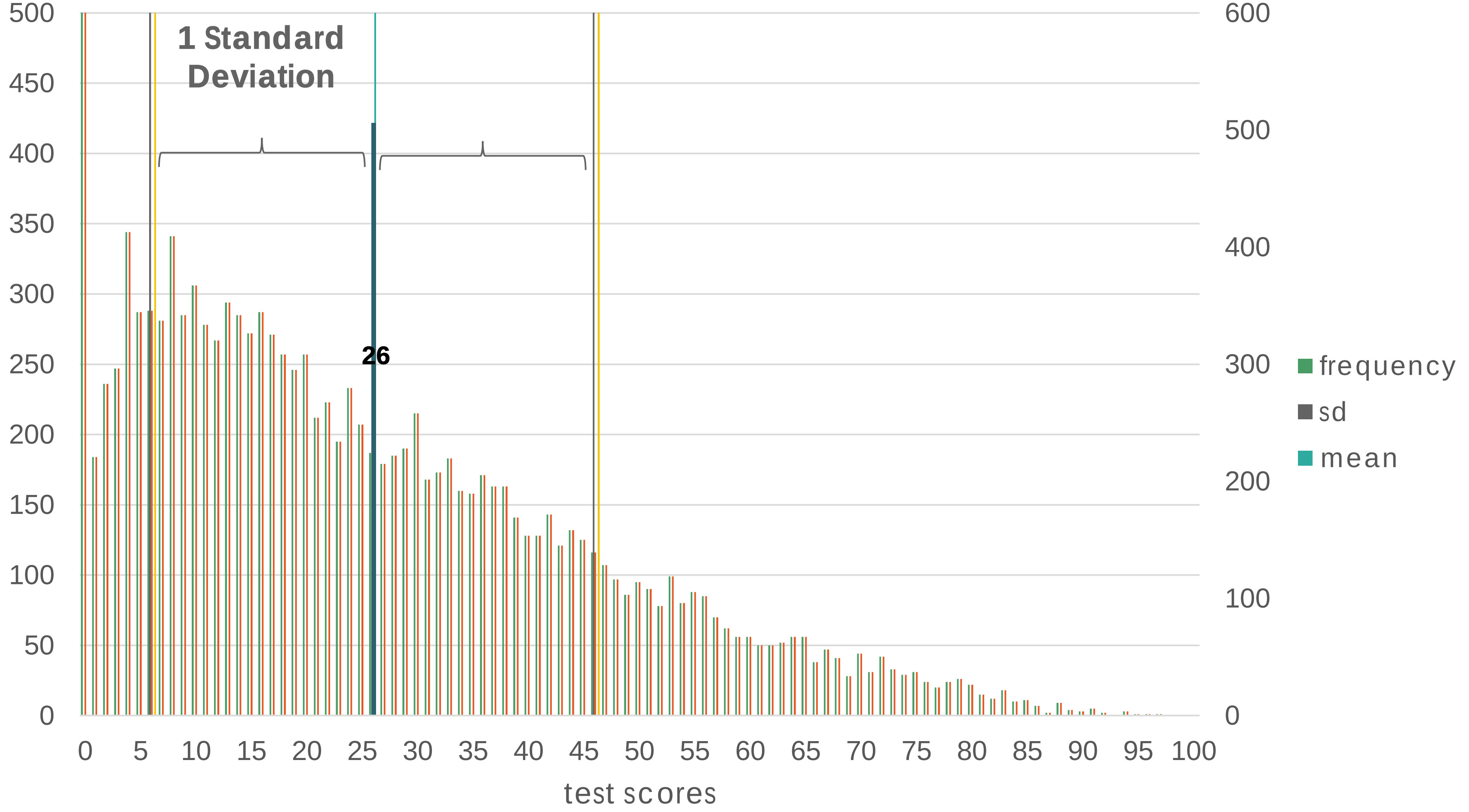


Mean = 26



500
450
400
350
300
250
200
150
100
50
0
26
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
test scores
frequency
mean

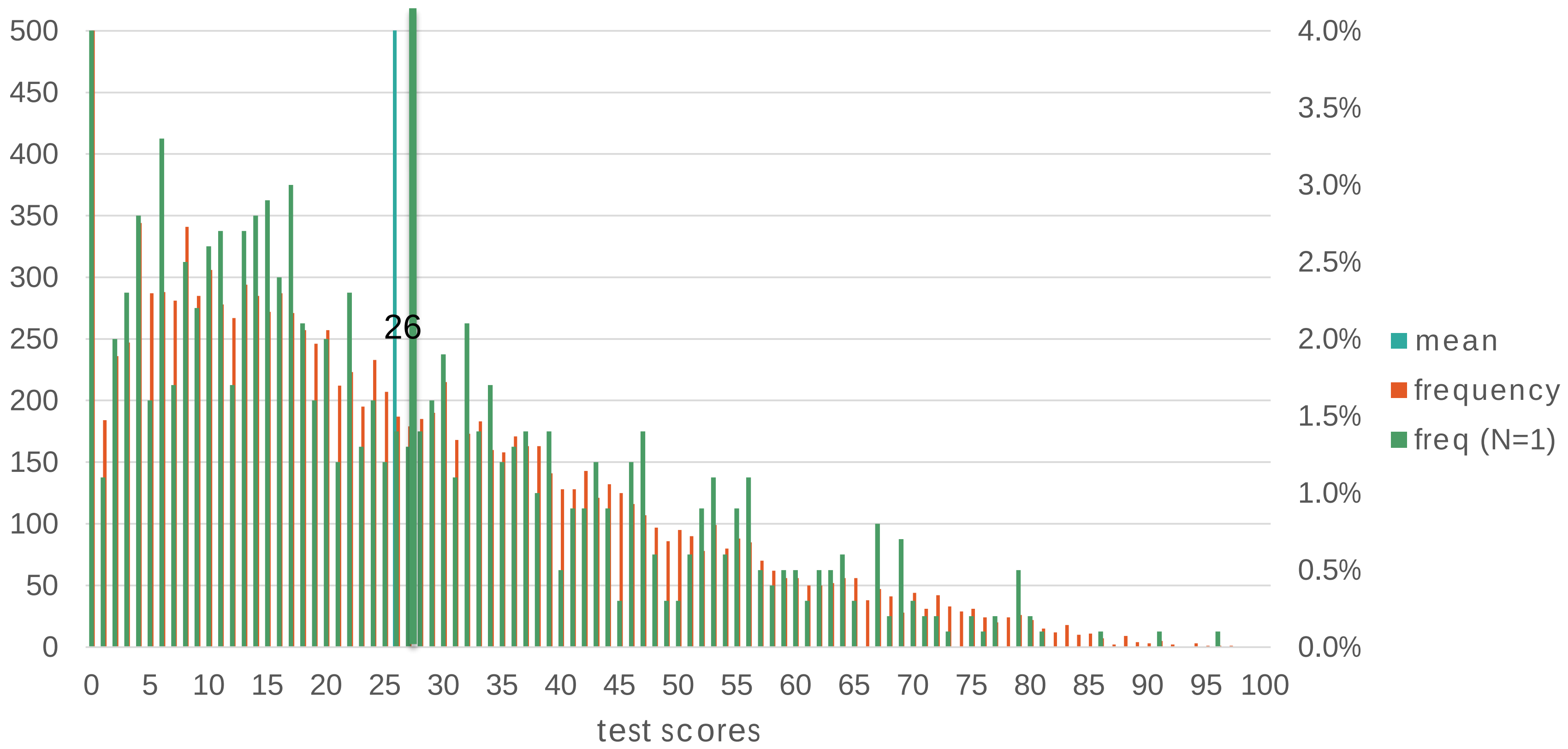
Standard Deviation = 20



Let’s do an experiment

- Take 1 Random test score from the pile of 16,000 tests
- Write down the value
- Put the test back
- Do these three steps again
- And again
- 8,000 times
- This is like a random sample of 8,000 (*with replacement*)

What can we say about this sample?



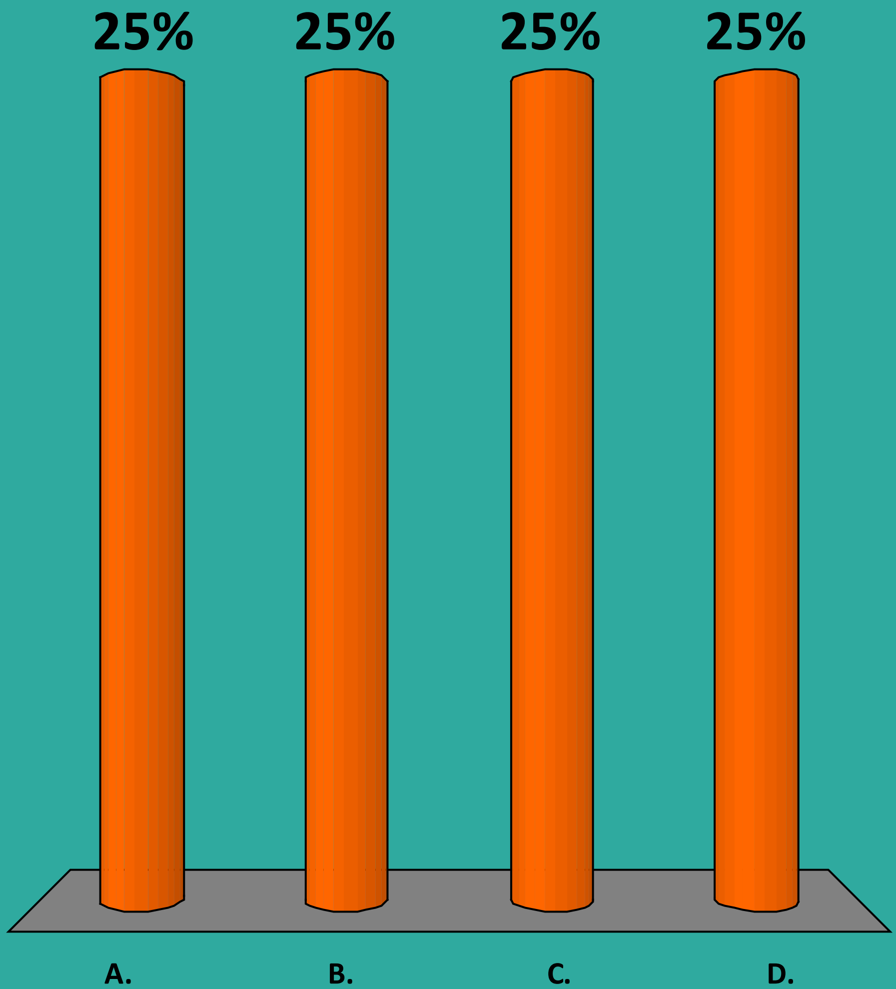
Good, the average of the sample is about 26...

But…

- … I remember that as my sample goes, up, isn't the sampling distribution supposed to turn into a bell curve?
- (Central Limit Theorem)
- Is it that my sample isn't large enough?

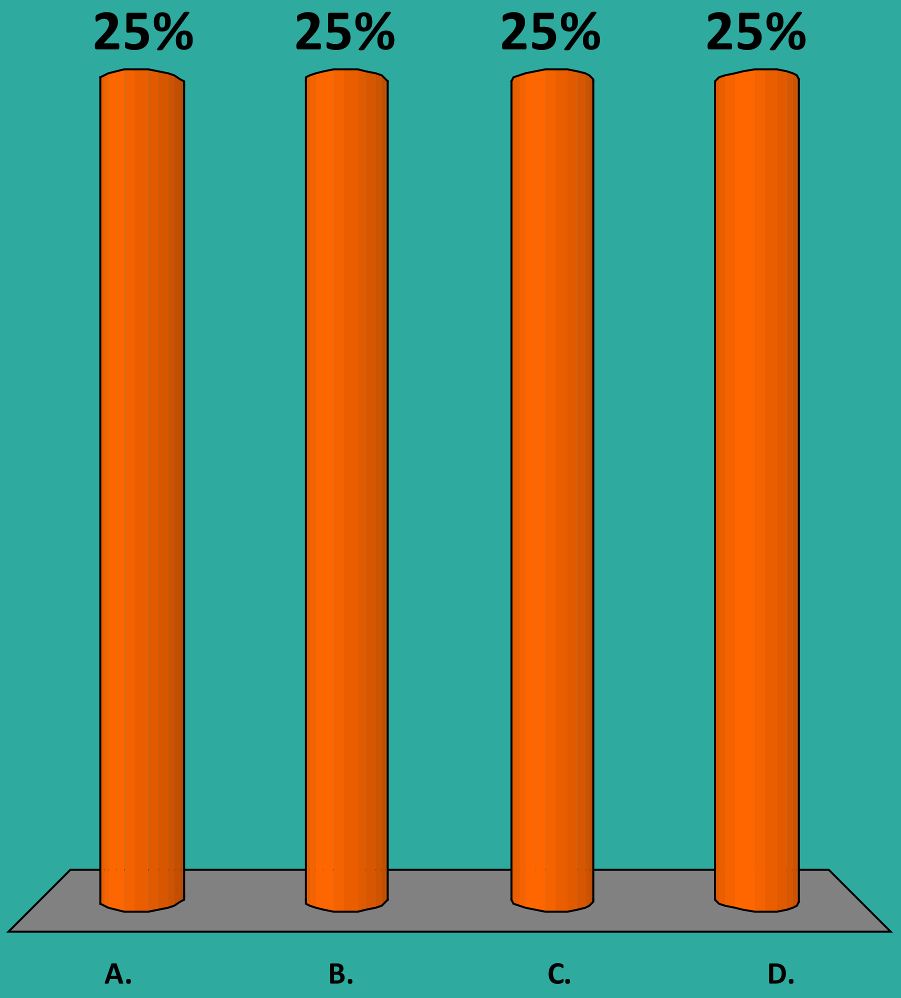
One limitation of statistical theory is that it assumes the population distribution is normally distributed

A. True
B. False
C. Depends
D. Don't know

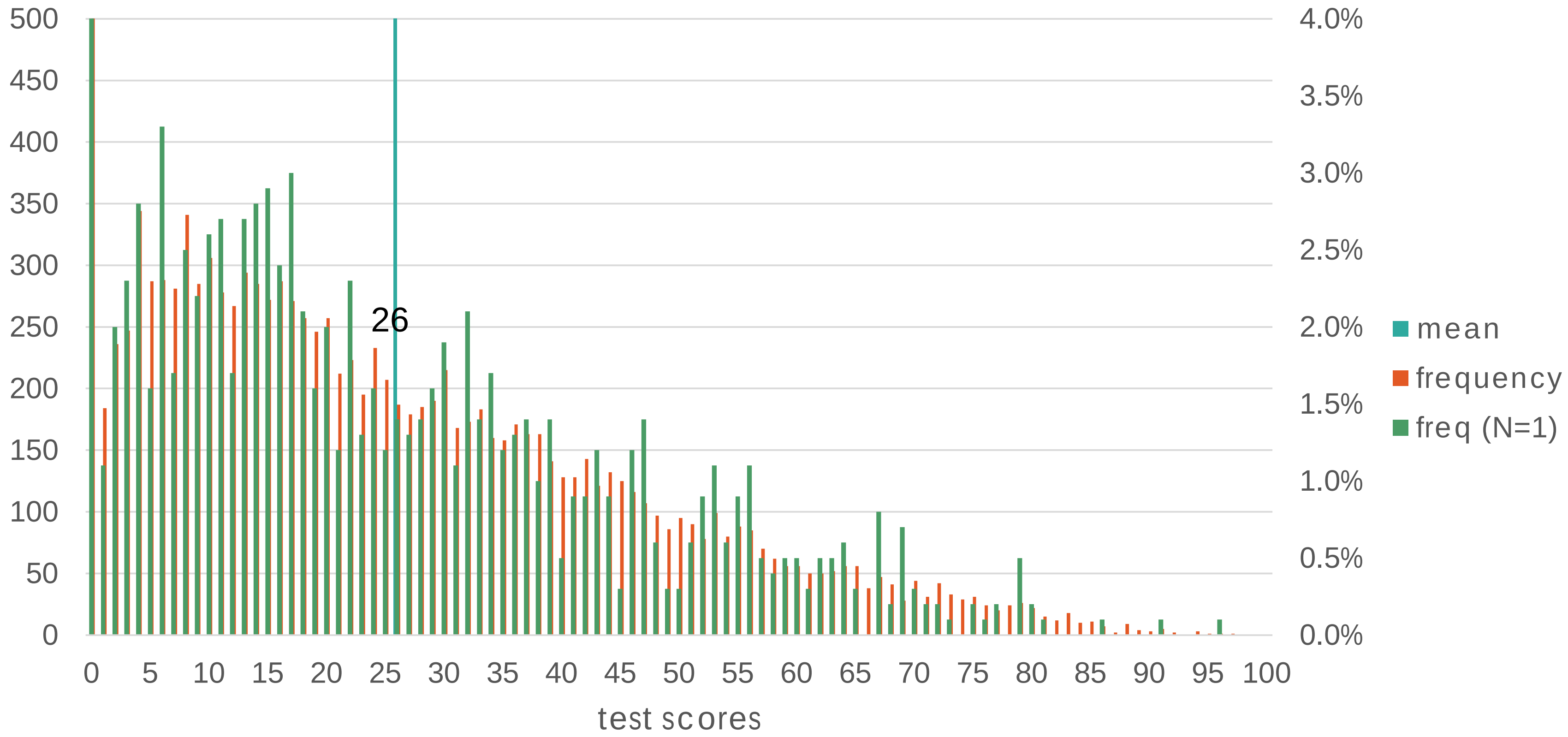


The sampling distribution may not be normal if the population distribution is skewed

A. True
B. False
C. Depends
D. Don’t know



Population v. sampling distribution

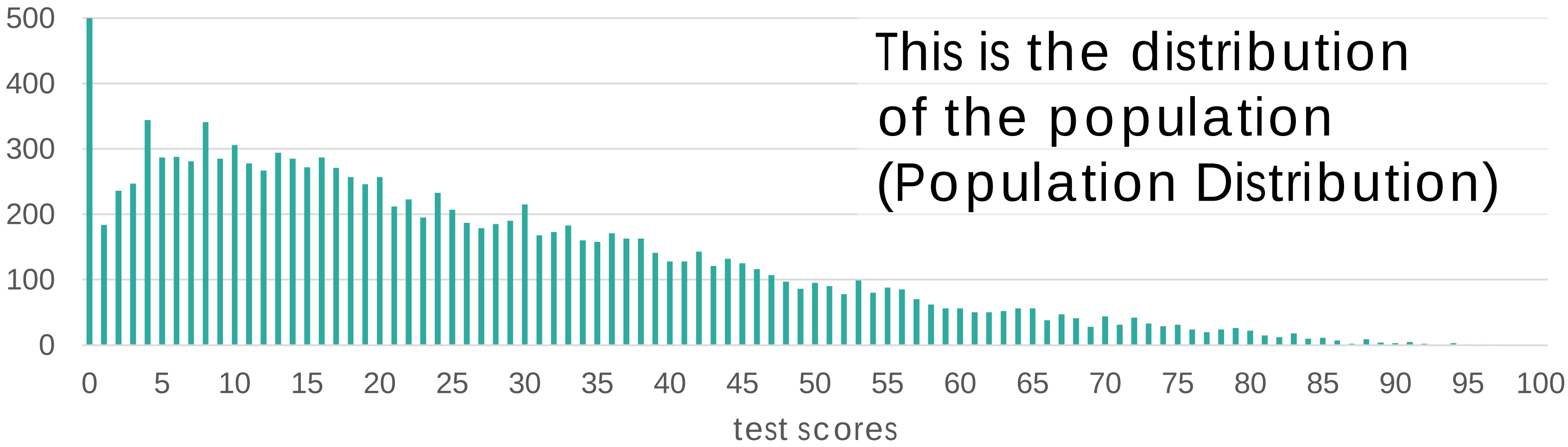


This is the distribution of my sample of 8,000 students

Outline

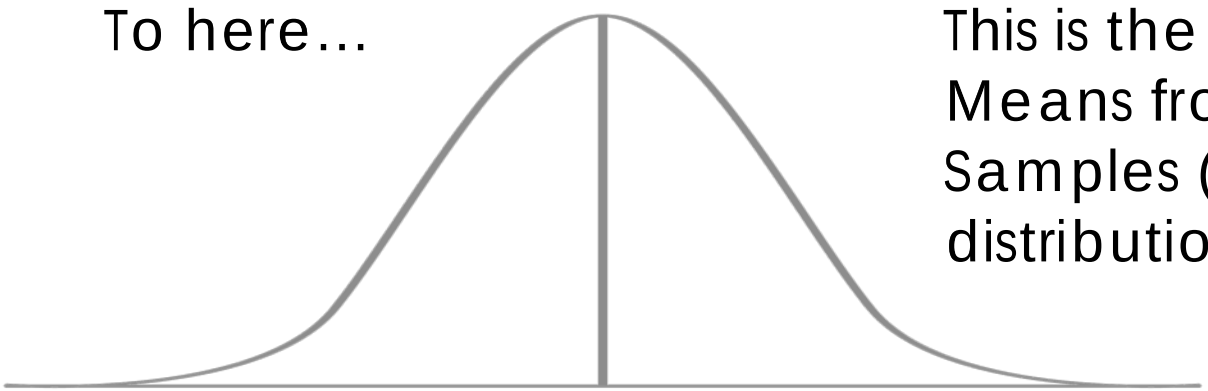
- Sampling distributions
  - population distribution
  - sampling distribution
  - law of large numbers/central limit theorem
  - standard deviation and standard error
- Detecting impact

How do we get from here...



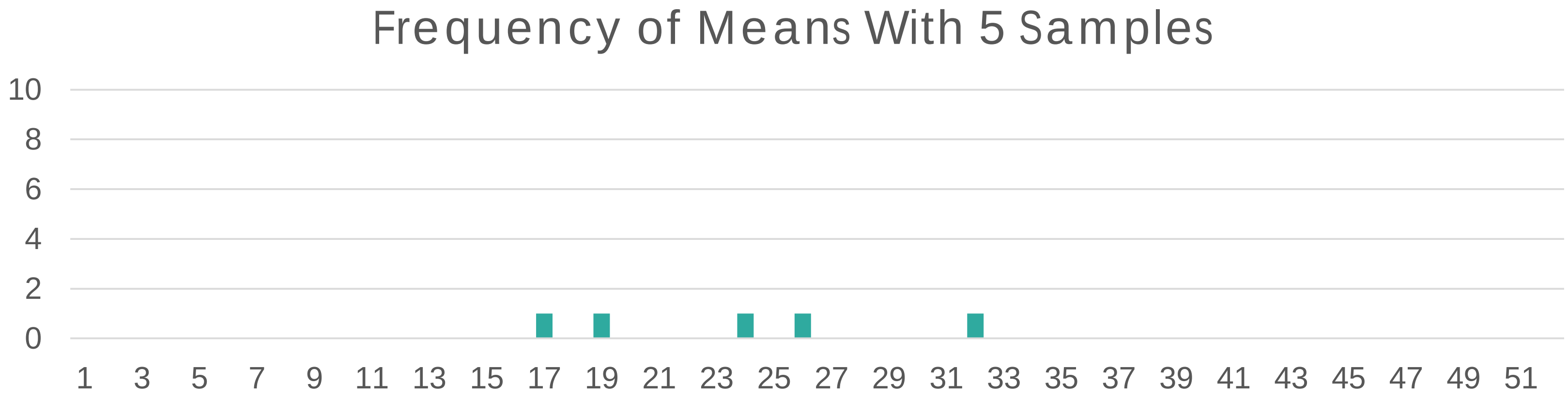
This is the distribution of the population (Population Distribution)

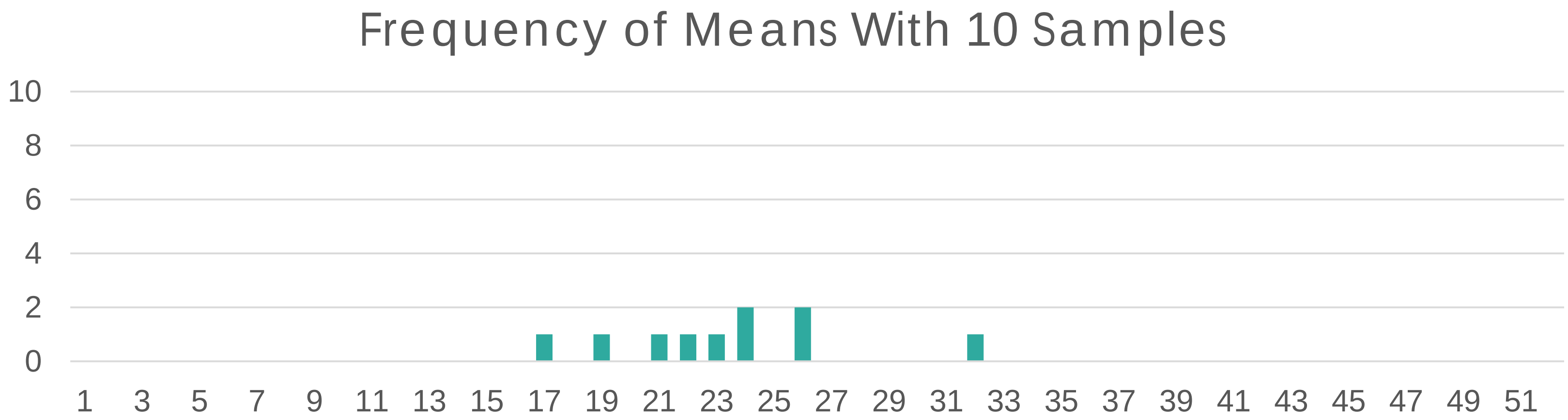
To here...



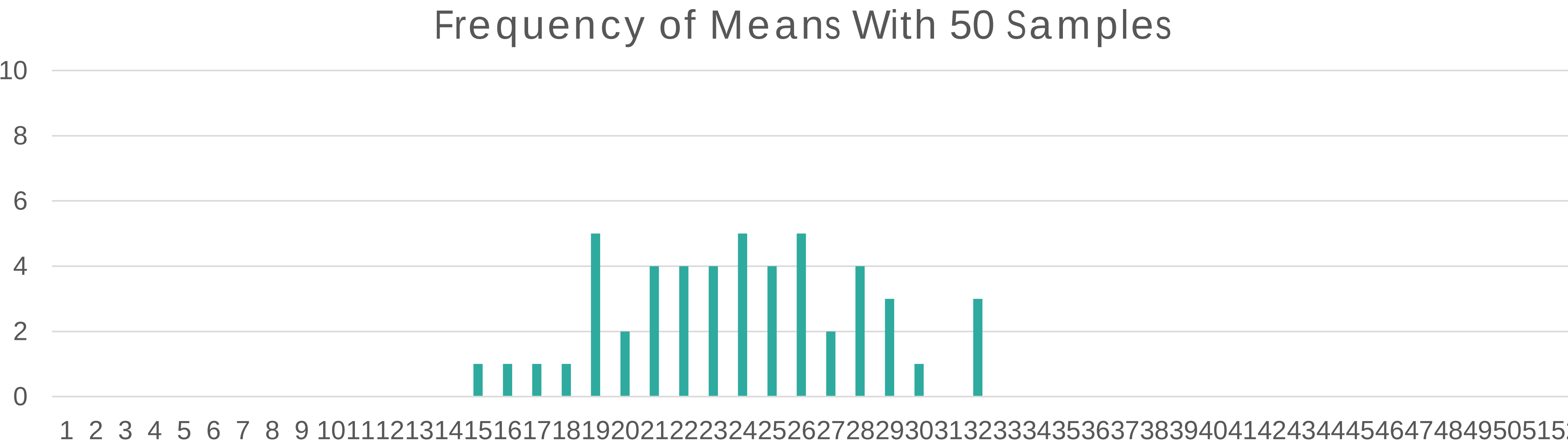
This is the distribution of Means from all Random Samples (Sampling distribution)

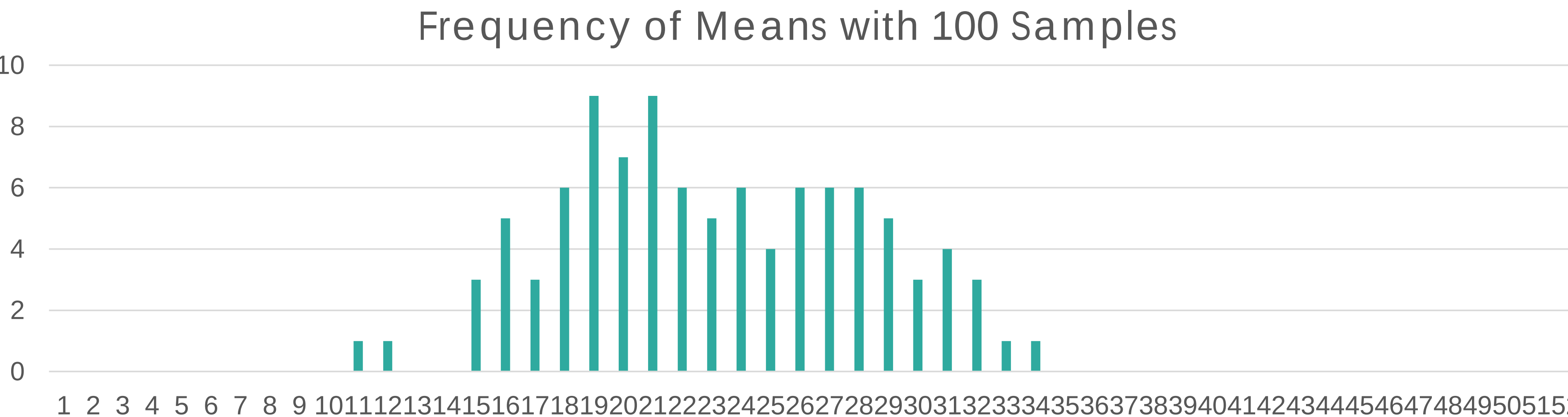
Draw 10 random students, take the average, plot it: Do this 5 & 10 times.



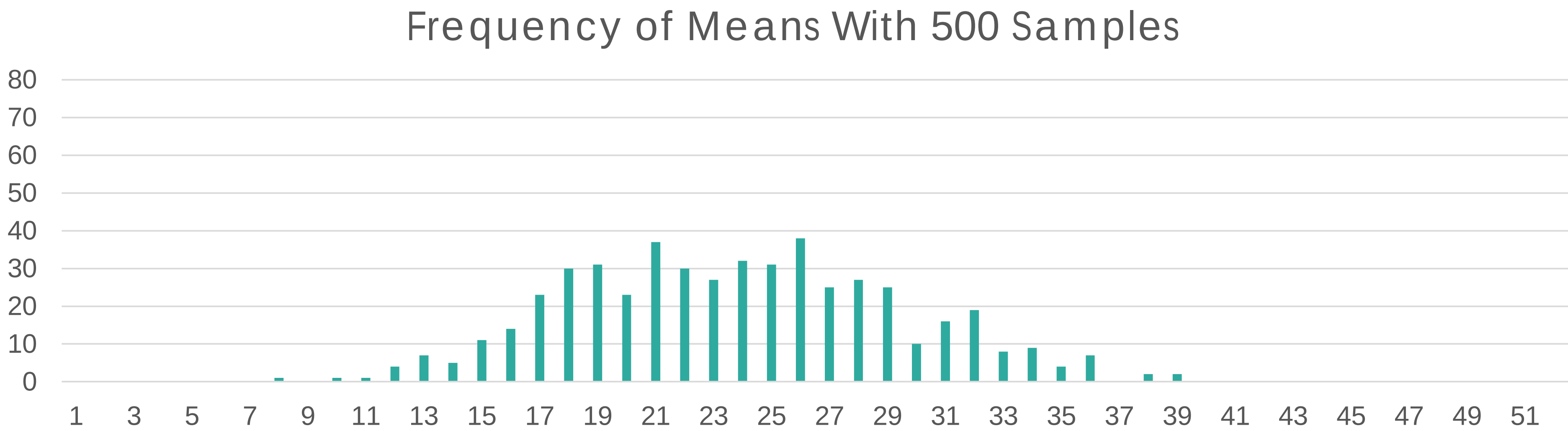


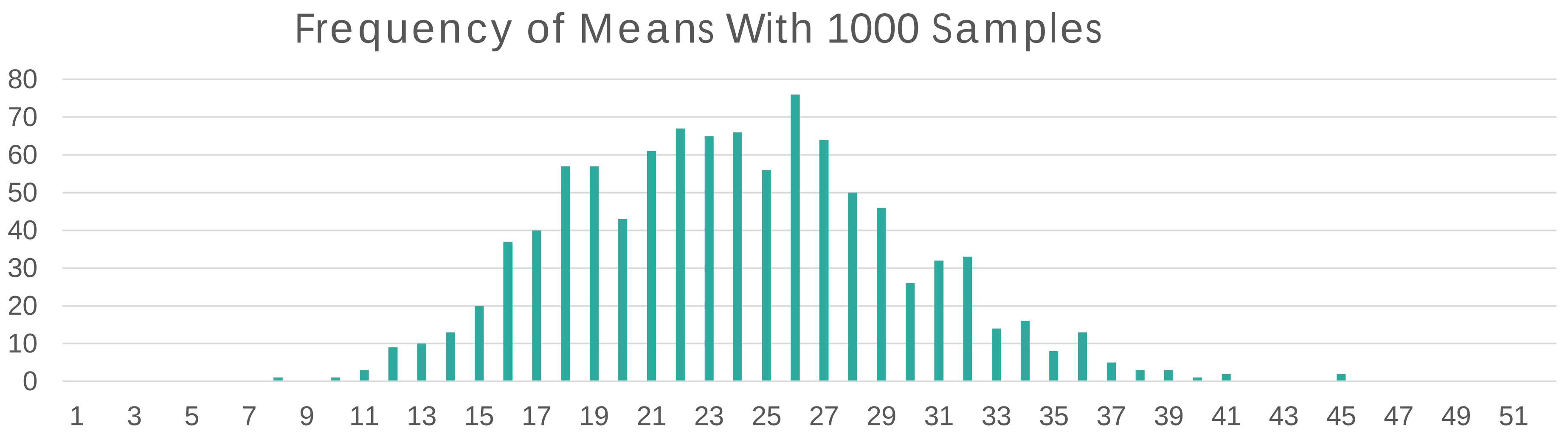
Draw 10 random students: 50 and 100 times





Draws 10 random students: 500 and 1000 times





Draw 10 Random students

- This is like a sample size of 10
- What happens if we take a sample size of 50?

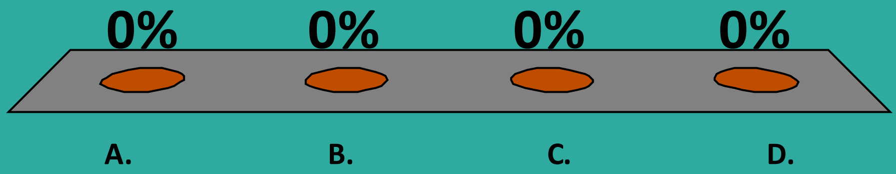
What happens to the sampling distribution if we draw a sample size of 50 instead of 10, and take the mean (thousands of times)?

A. We will approach a bell curve faster (than with a sample size of 10)

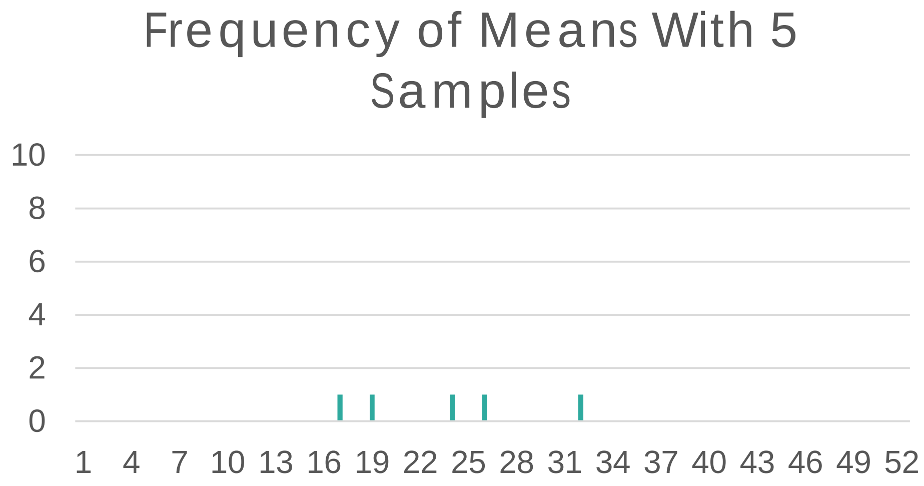
B. The bell curve will be narrower

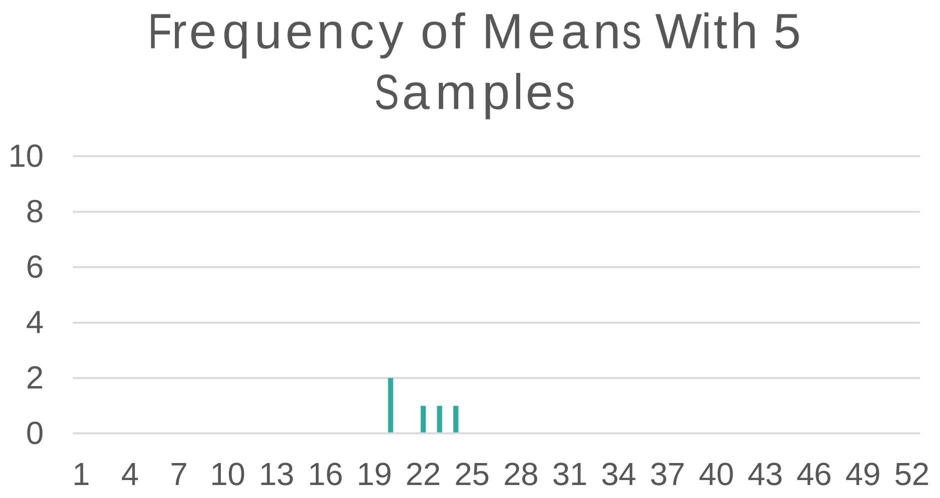
C. Both A & B

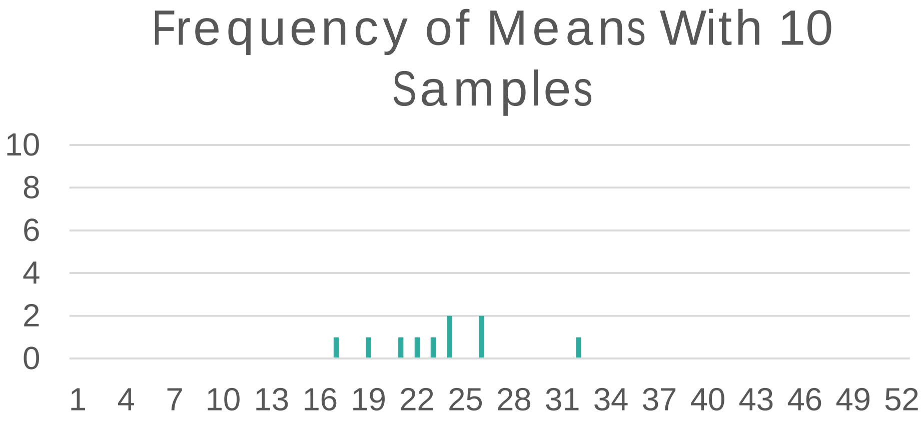
D. Neither. The underlying sampling distribution does not change.

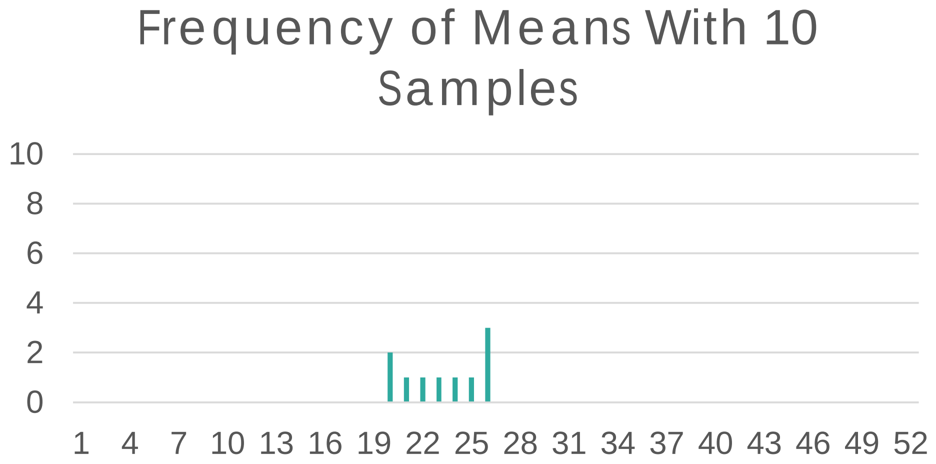


N = 10 N = 50

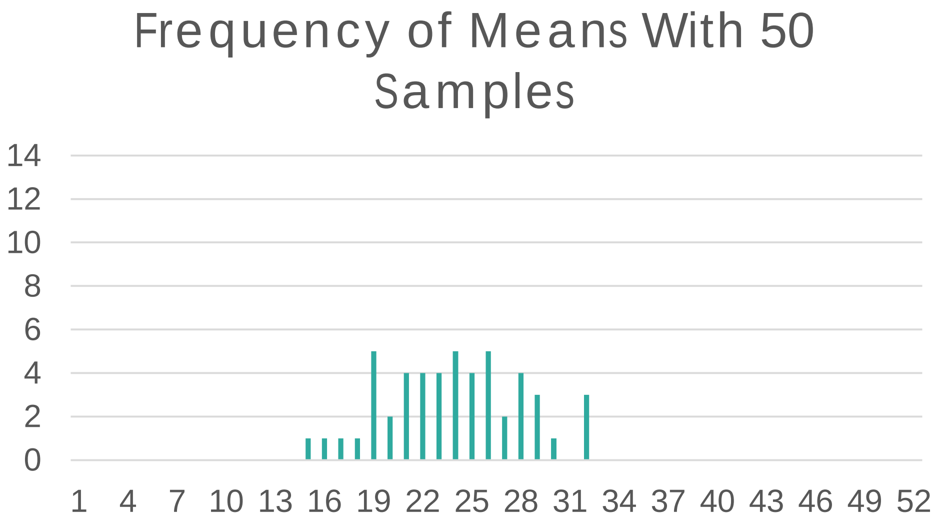

Frequency of Means With 5 Samples
10
8
6
4
2
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52

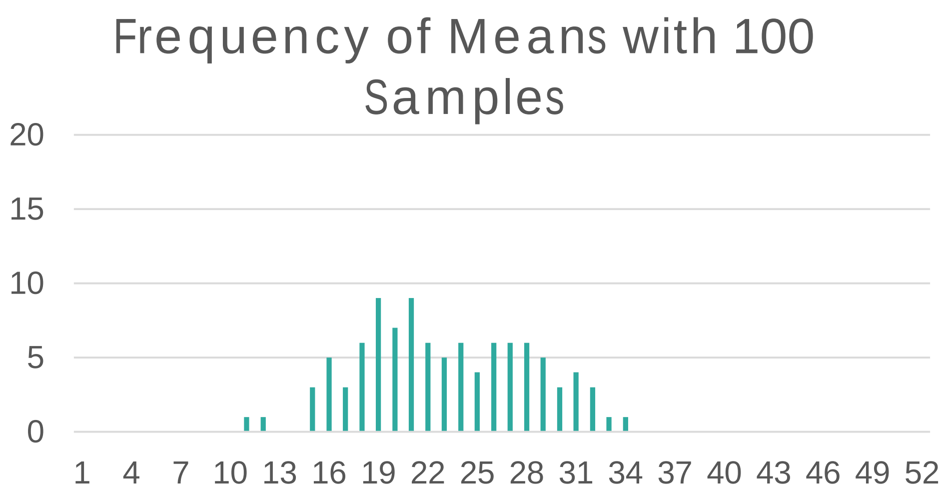

Frequency of Means With 5 Samples
10
8
6
4
2
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52


Frequency of Means With 10 Samples
10
8
6
4
2
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52

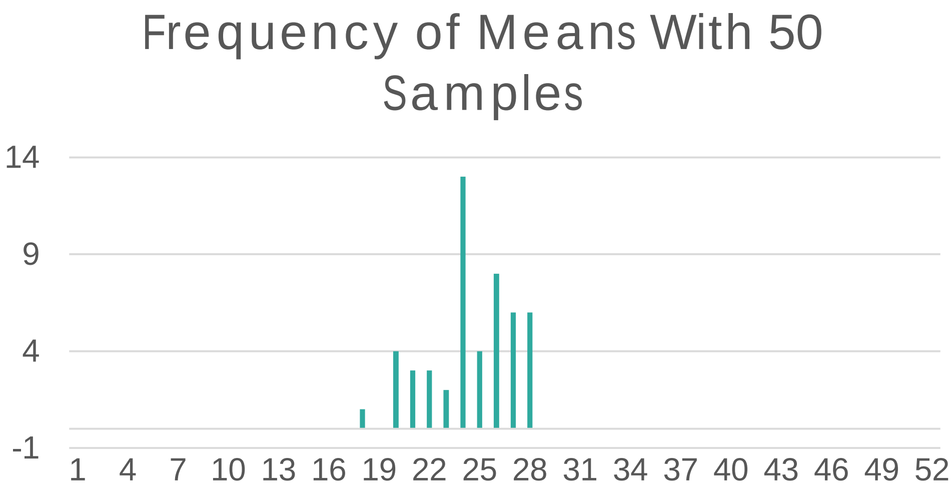

Frequency of Means With 10 Samples
10
8
6
4
2
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52

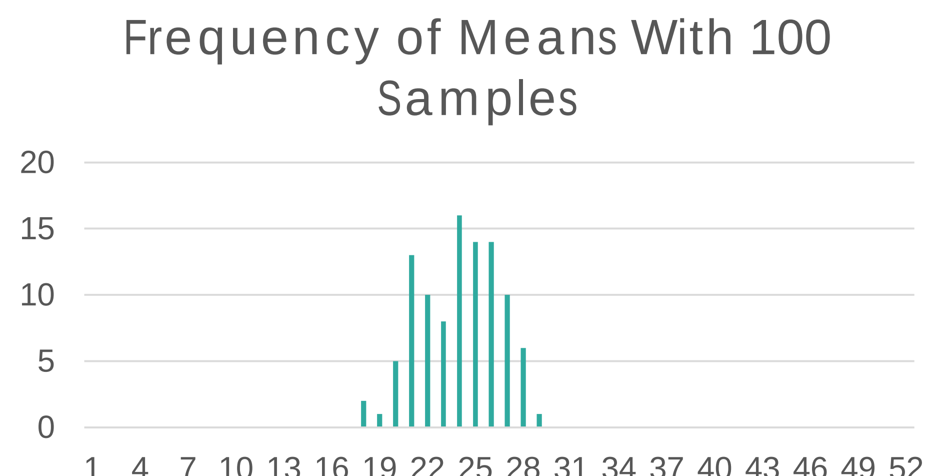
Draws of 10



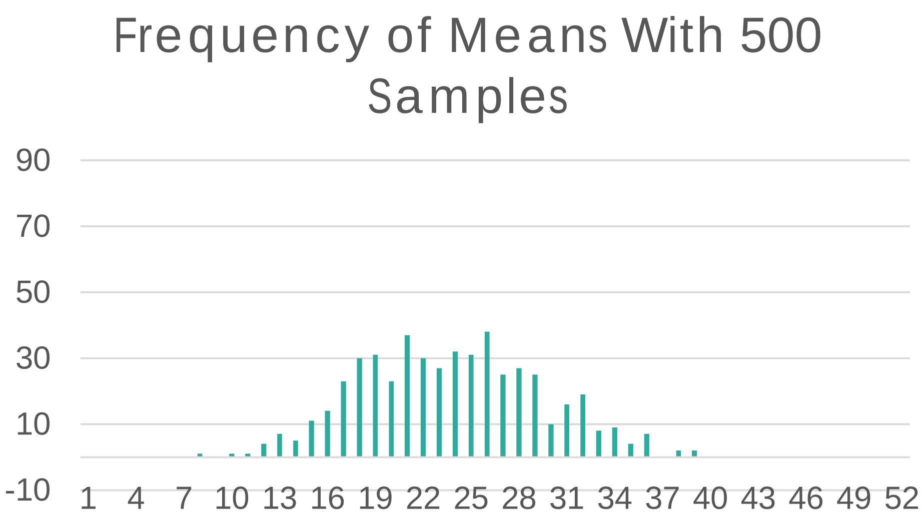


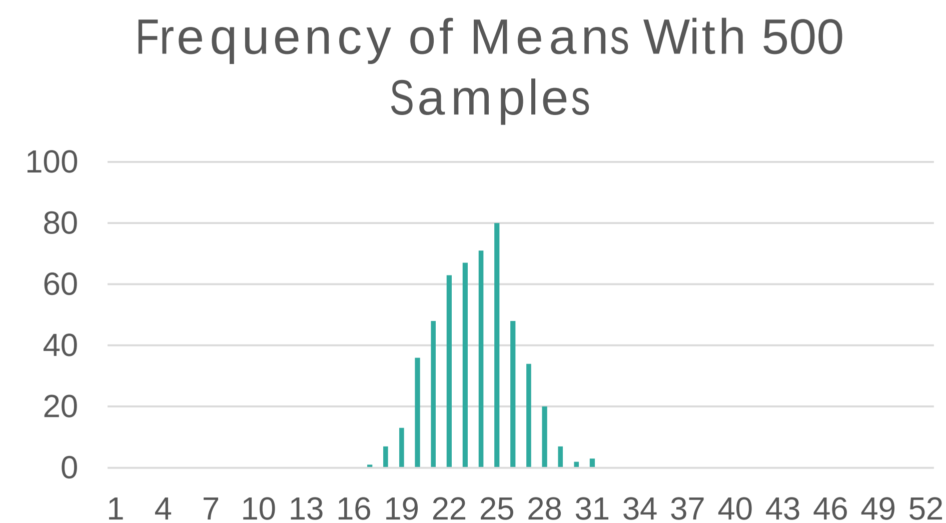
Draws of 50

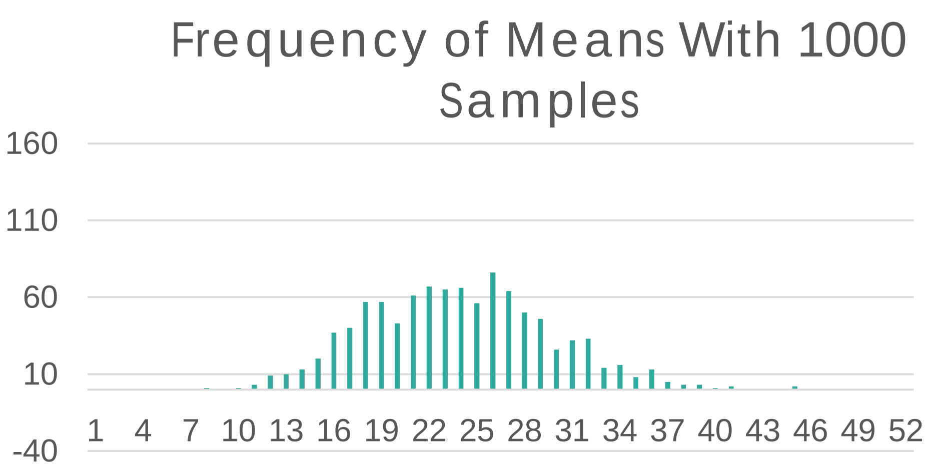


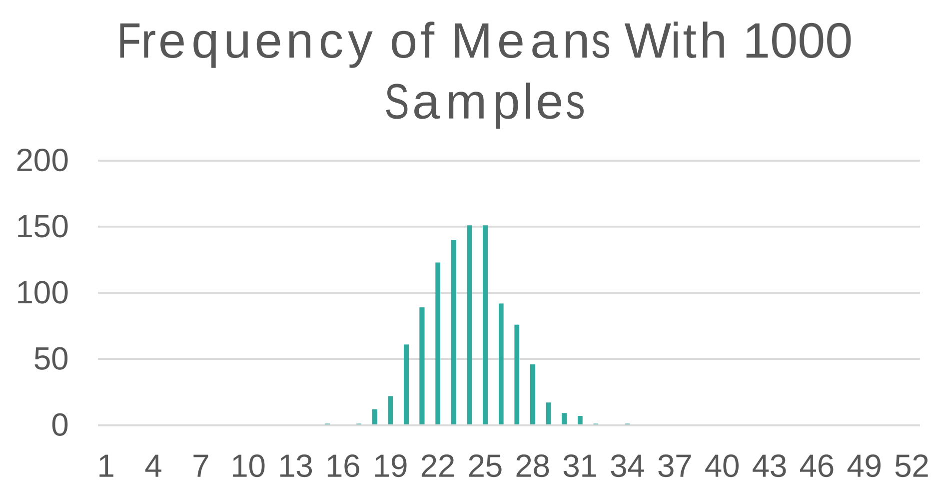


Draws of 10 | Draws of 50





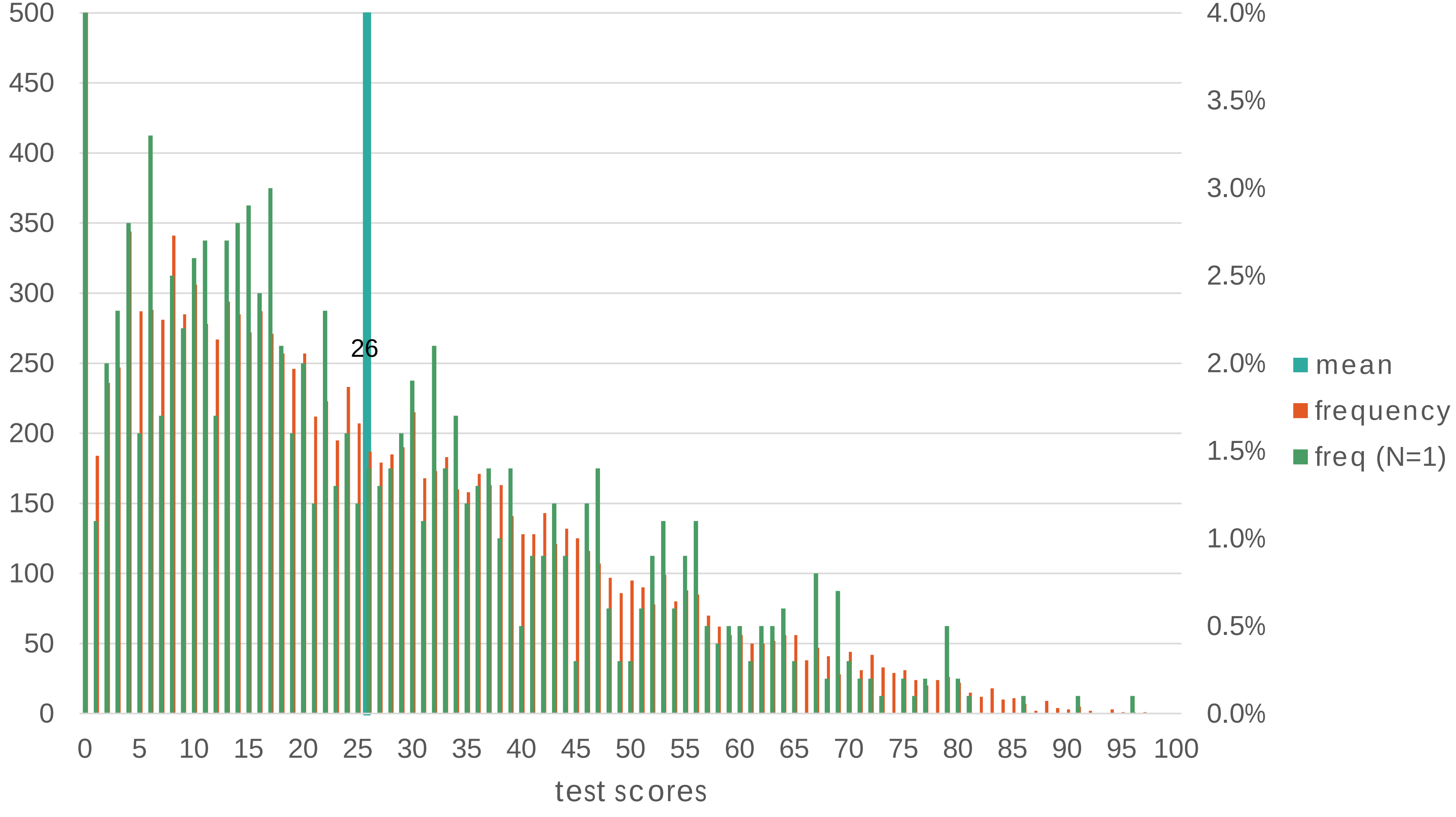




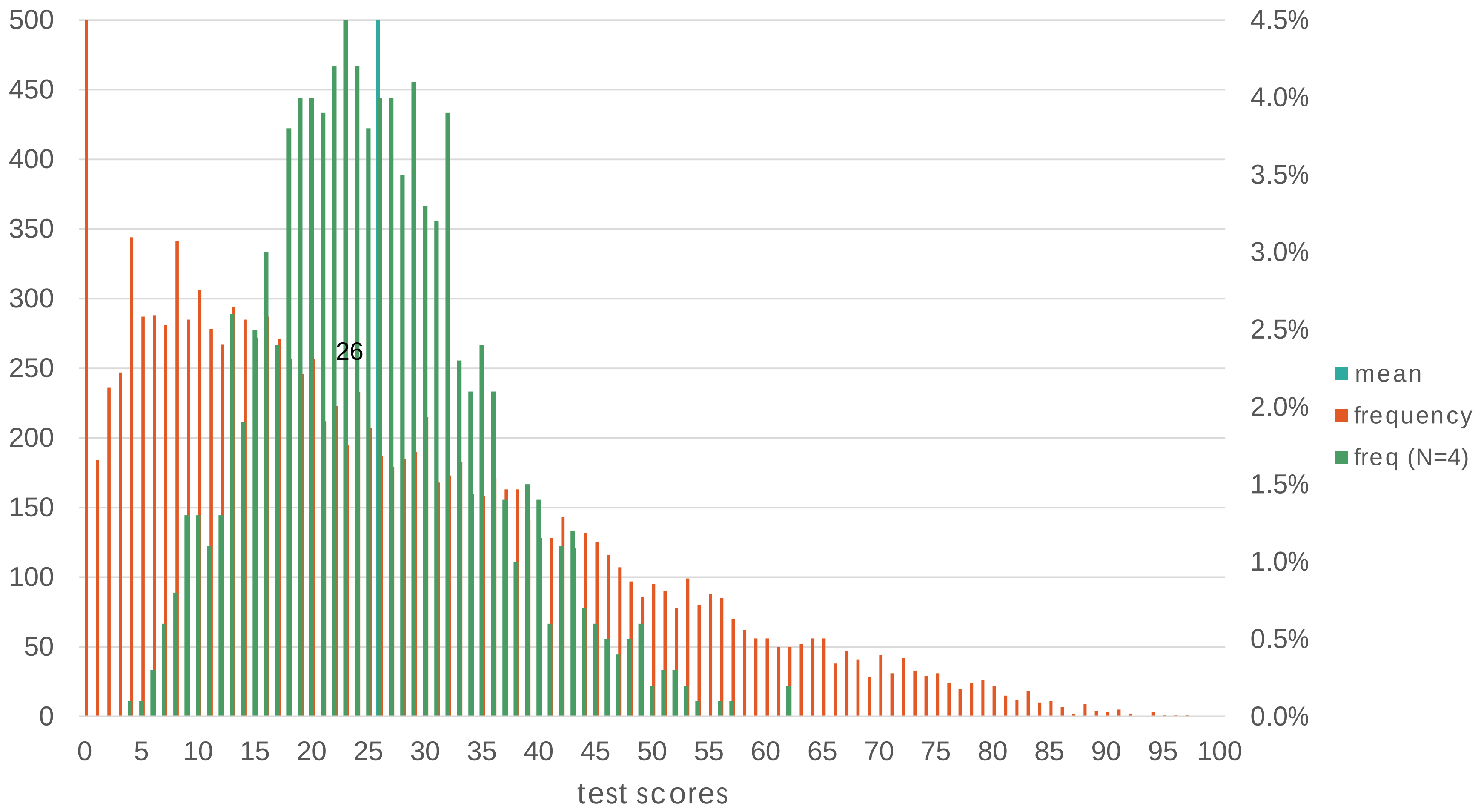
Outline

- Sampling distributions
  - population distribution
  - sampling distribution
  - law of large numbers/central limit theorem
  - standard deviation and standard error
- Detecting impact

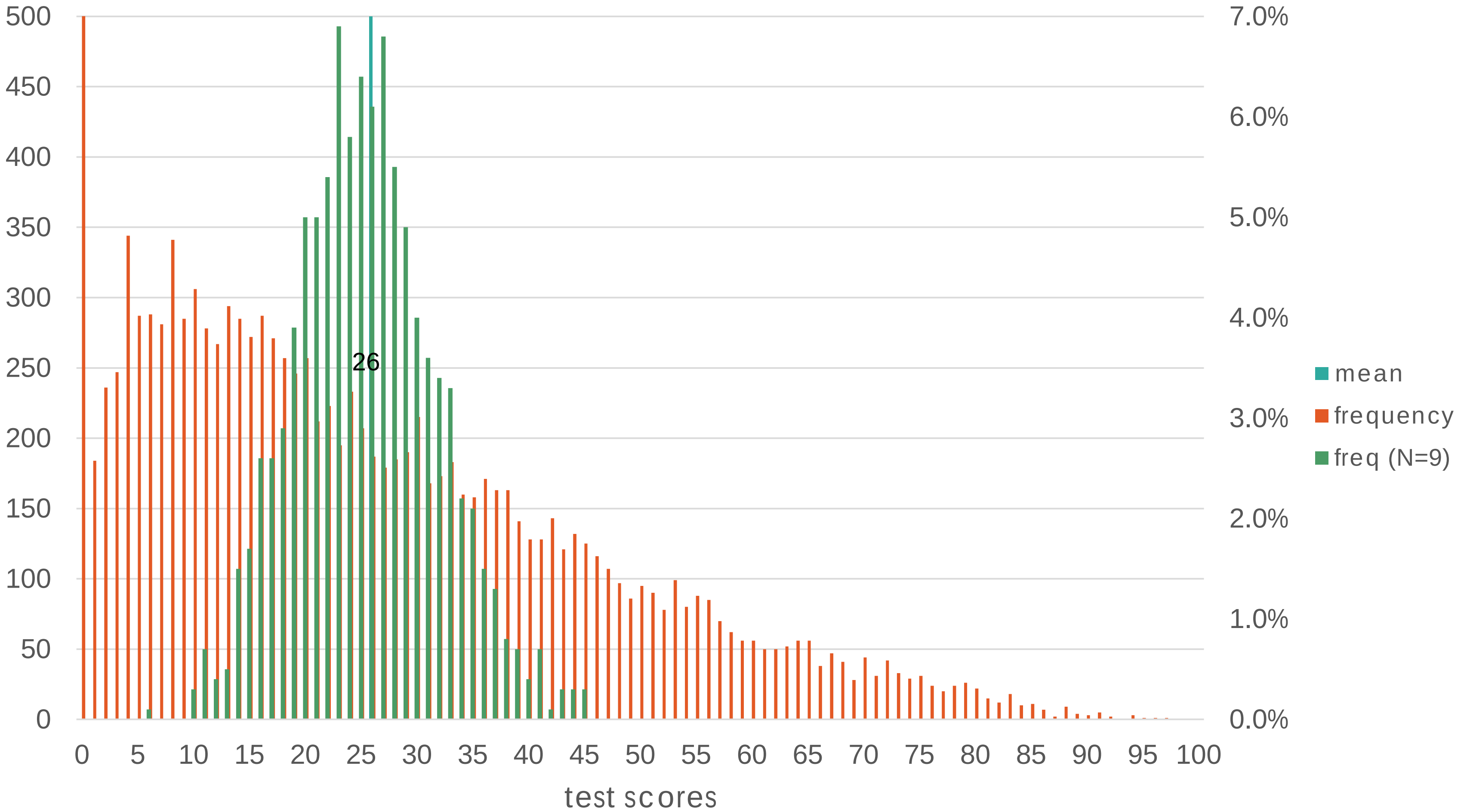
Population & sampling distribution: Draw 1 random student (from 8,000)



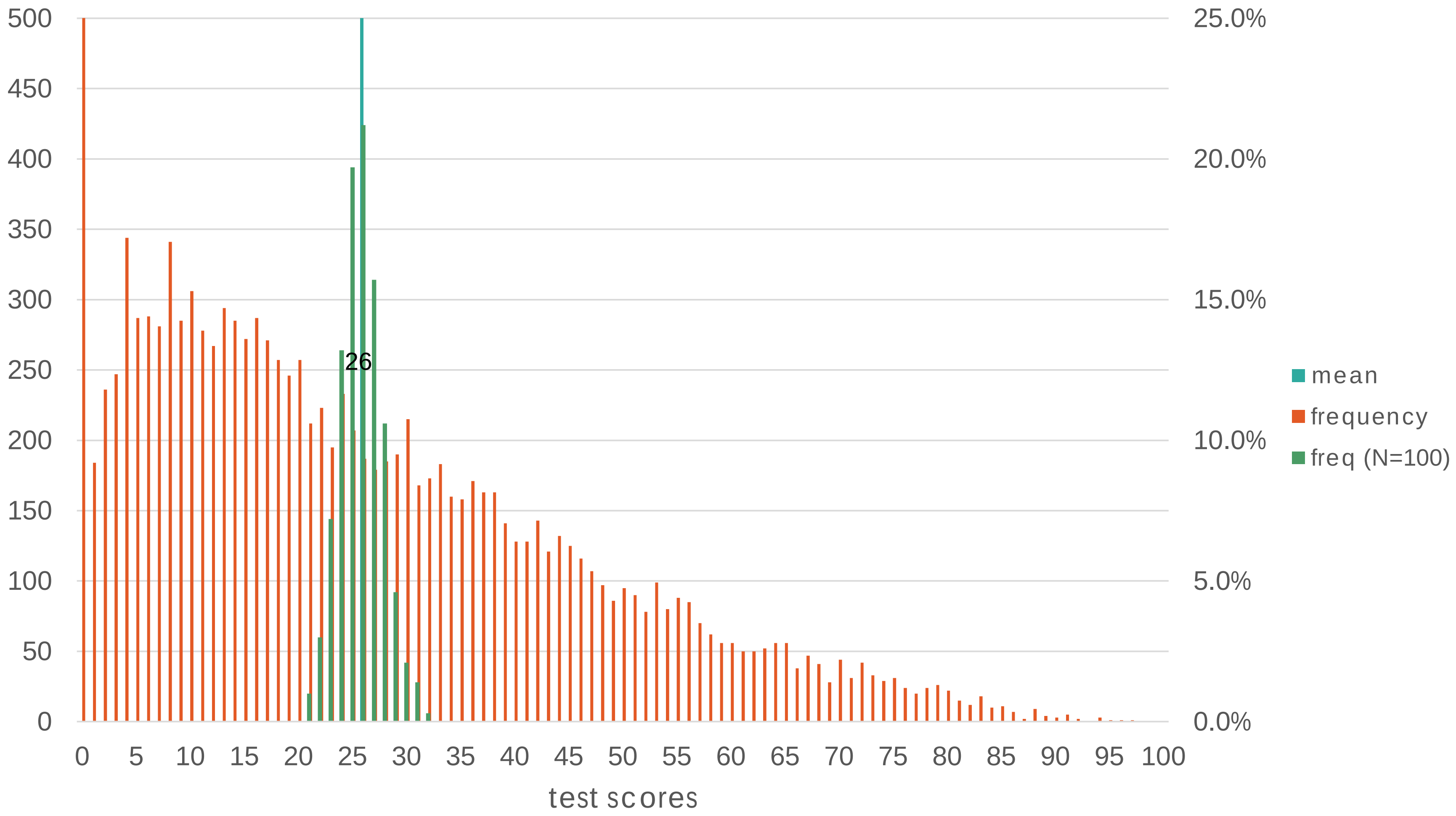
Sampling Distribution: Draw 4 random students (N=4)



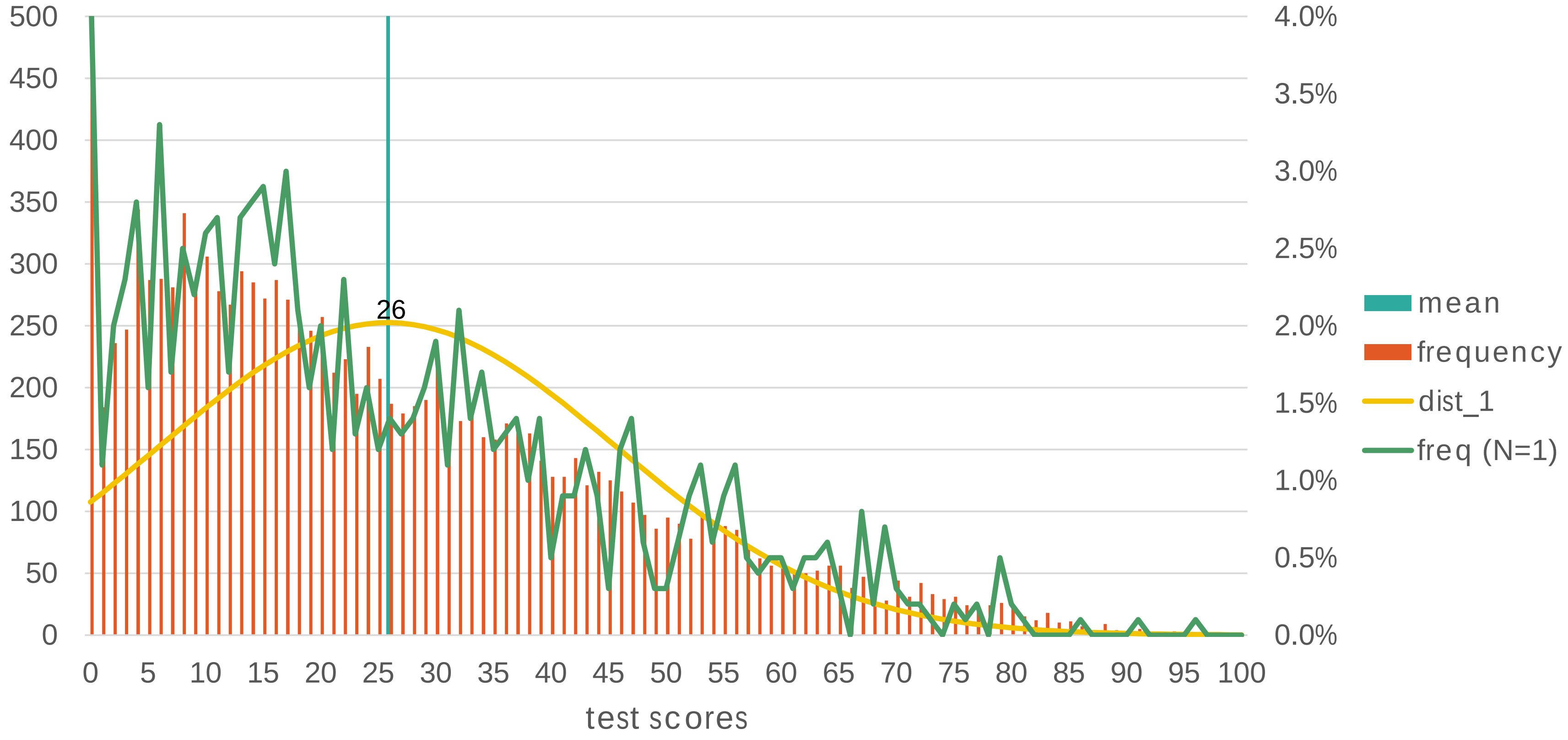
Law of Large Numbers : N=9



Law of Large Numbers: N =100

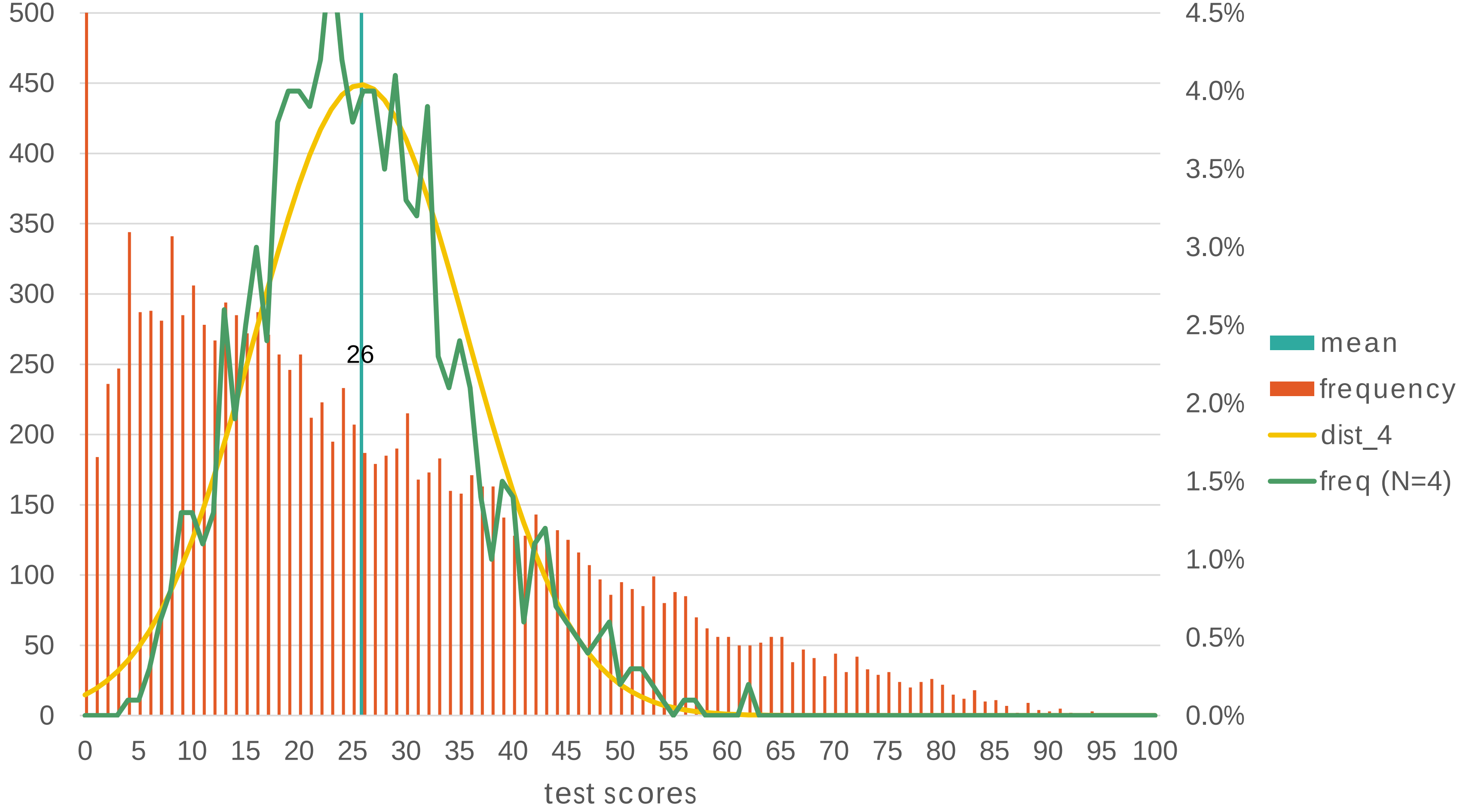


Central Limit Theorem: N=1

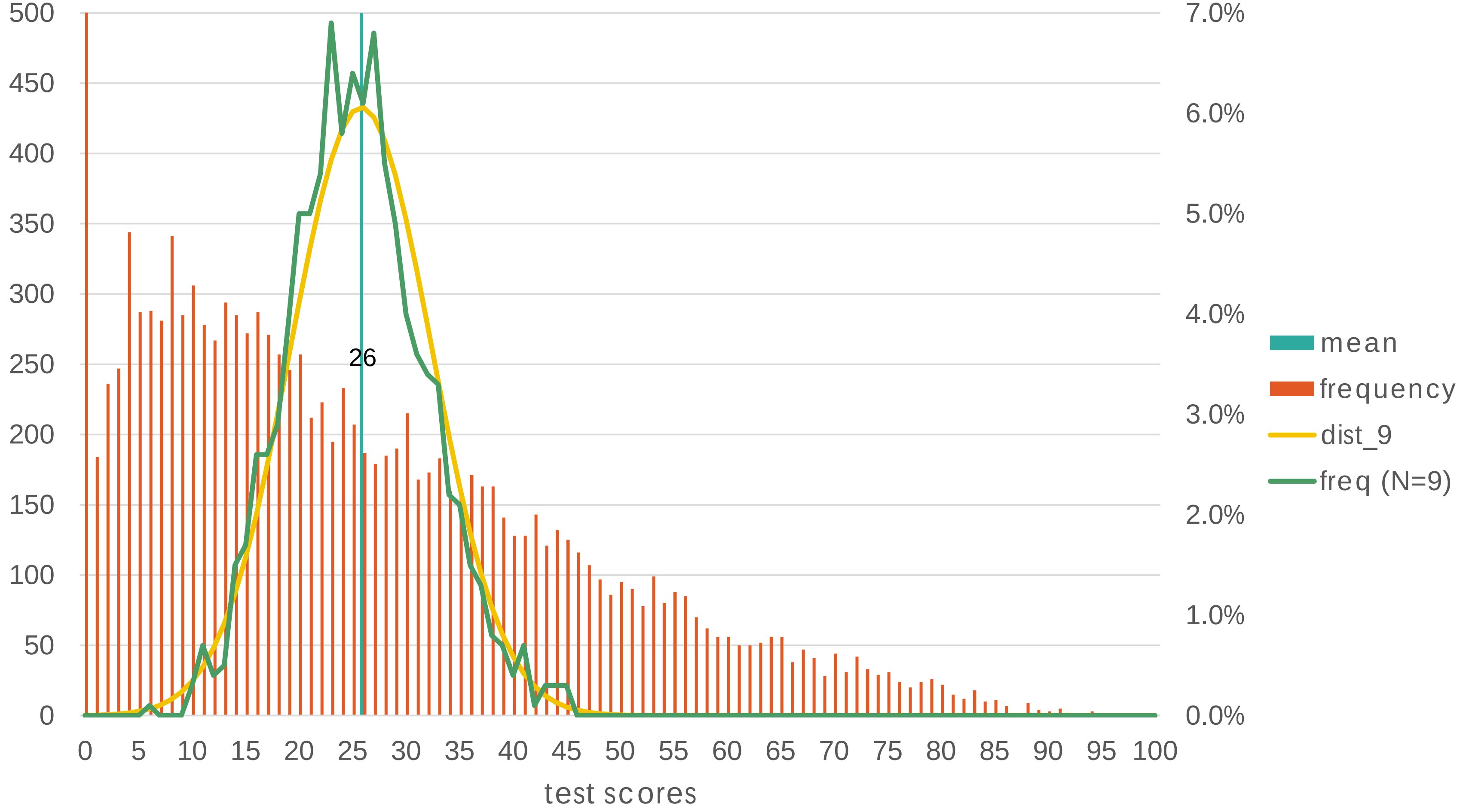


The yellow line is a theoretical distribution

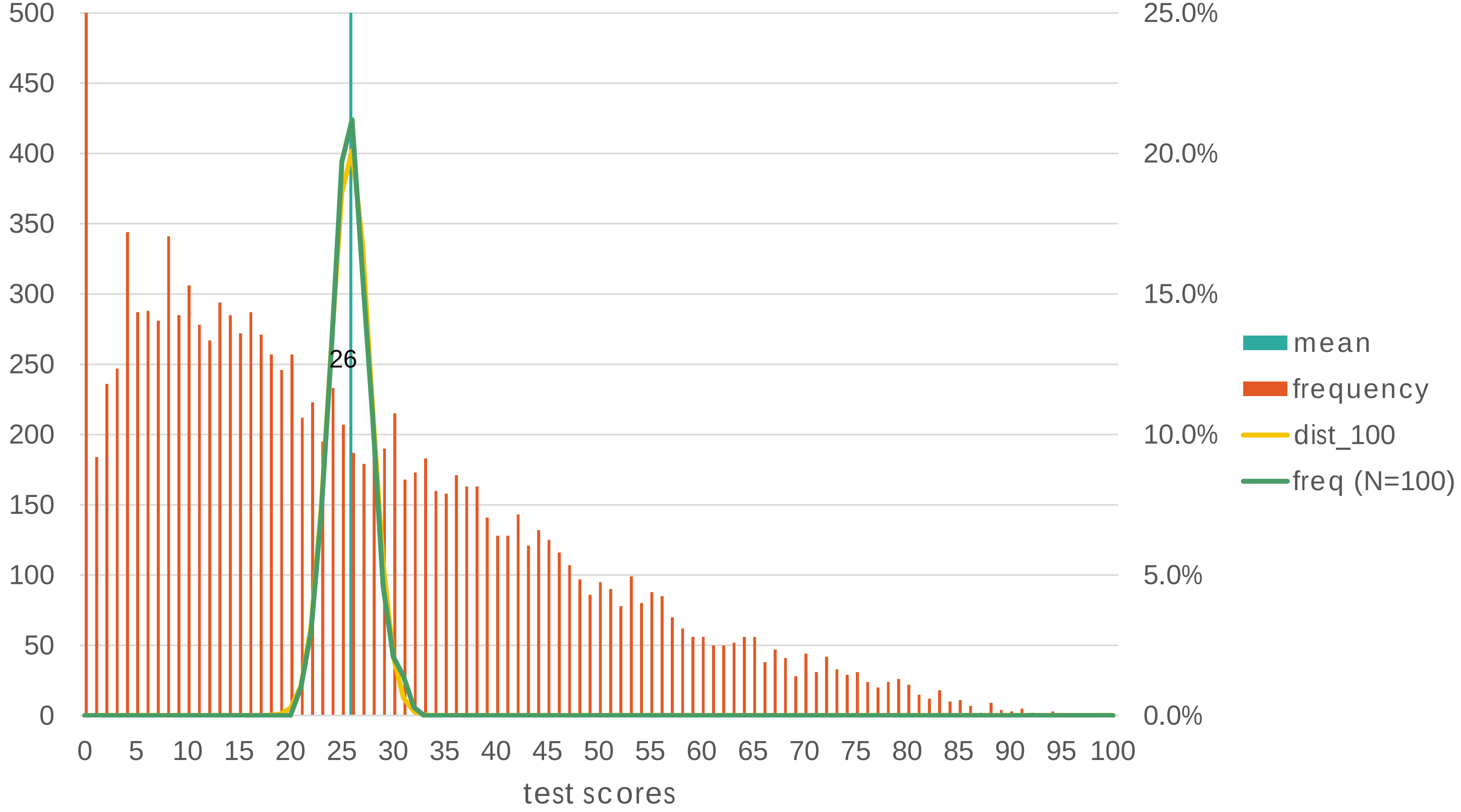
Central Limit Theorem : N=4



Central Limit Theorem : N=9



Central Limit Theorem : N =100



Outline

- Sampling distributions
  - population distribution
  - sampling distribution
  - law of large numbers/central limit theorem
  - standard deviation and standard error
- Detecting impact

Standard deviation/error

- What's the difference between the standard deviation and the standard error?
- The standard error = the standard deviation of the sampling distributions

Variance and Standard Deviation

- Variance = 400

$$\sigma^2 = \frac{\sum(Observation\,Value\, - Average)^2}{N}$$

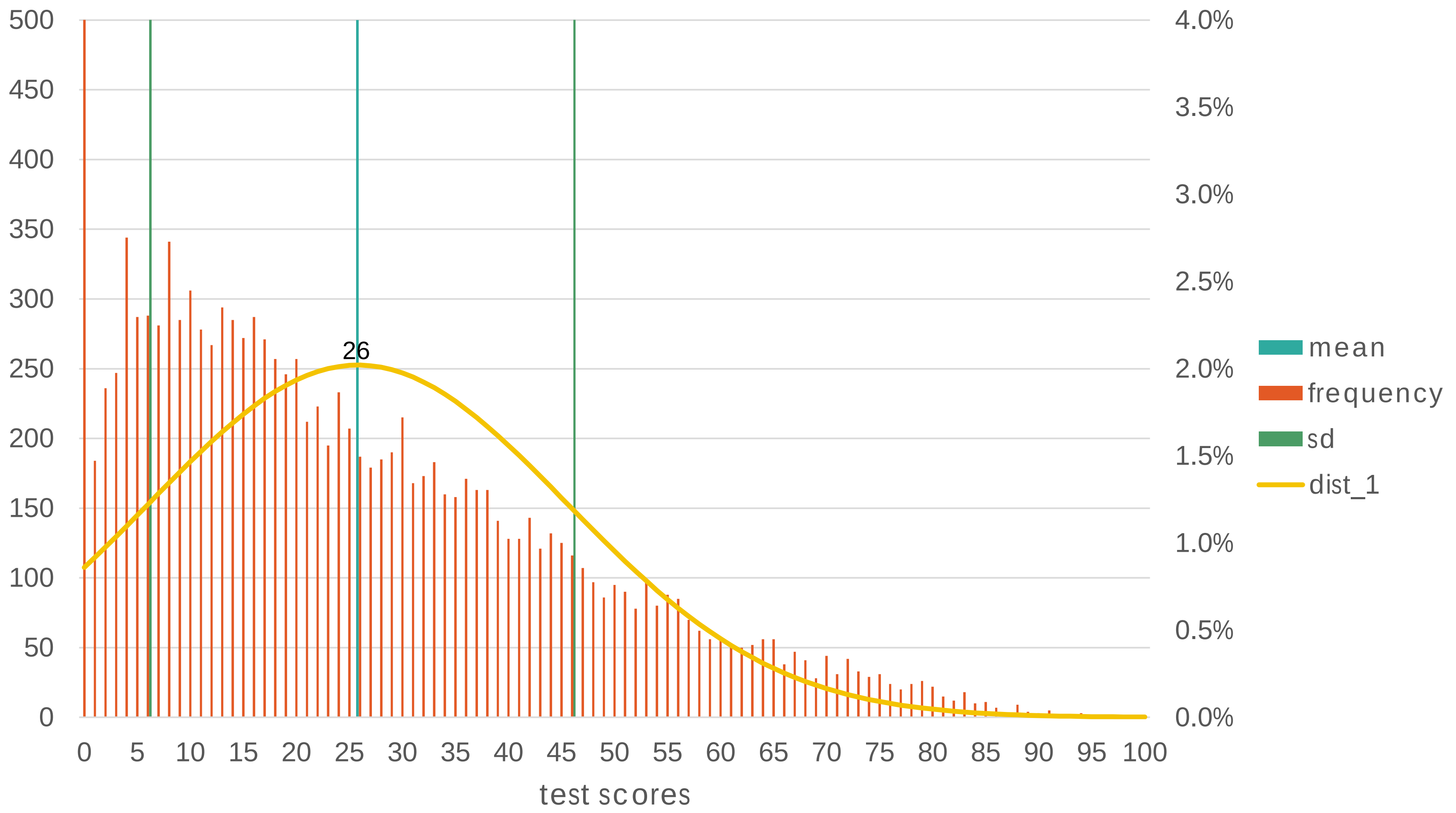
- Standard Deviation = 20

$$\sigma = \sqrt{Variance}$$

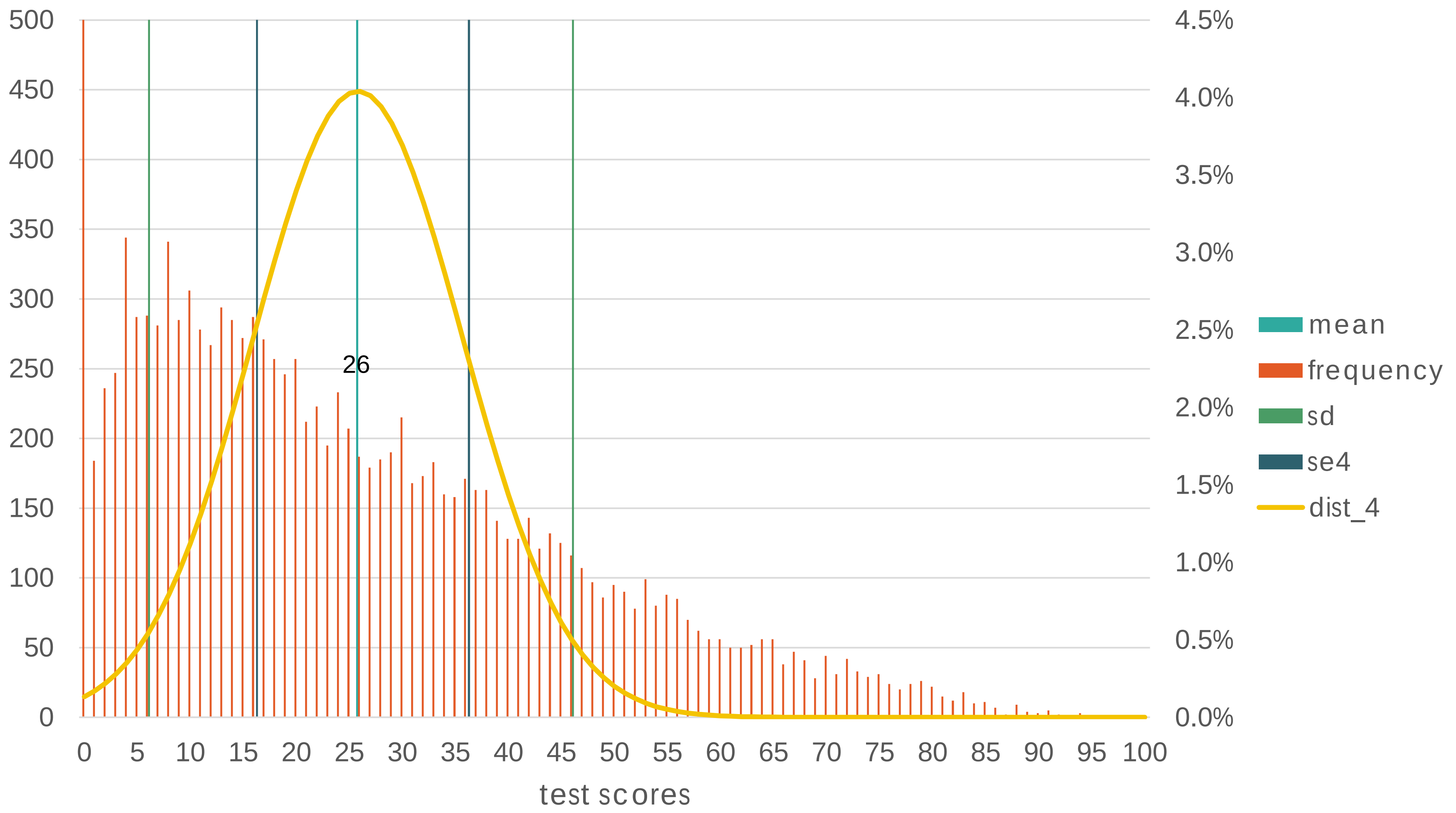
- Standard Error = $^{20}/_{\sqrt{N}}$

SE = $^{\sigma}/_{\sqrt{N}}$

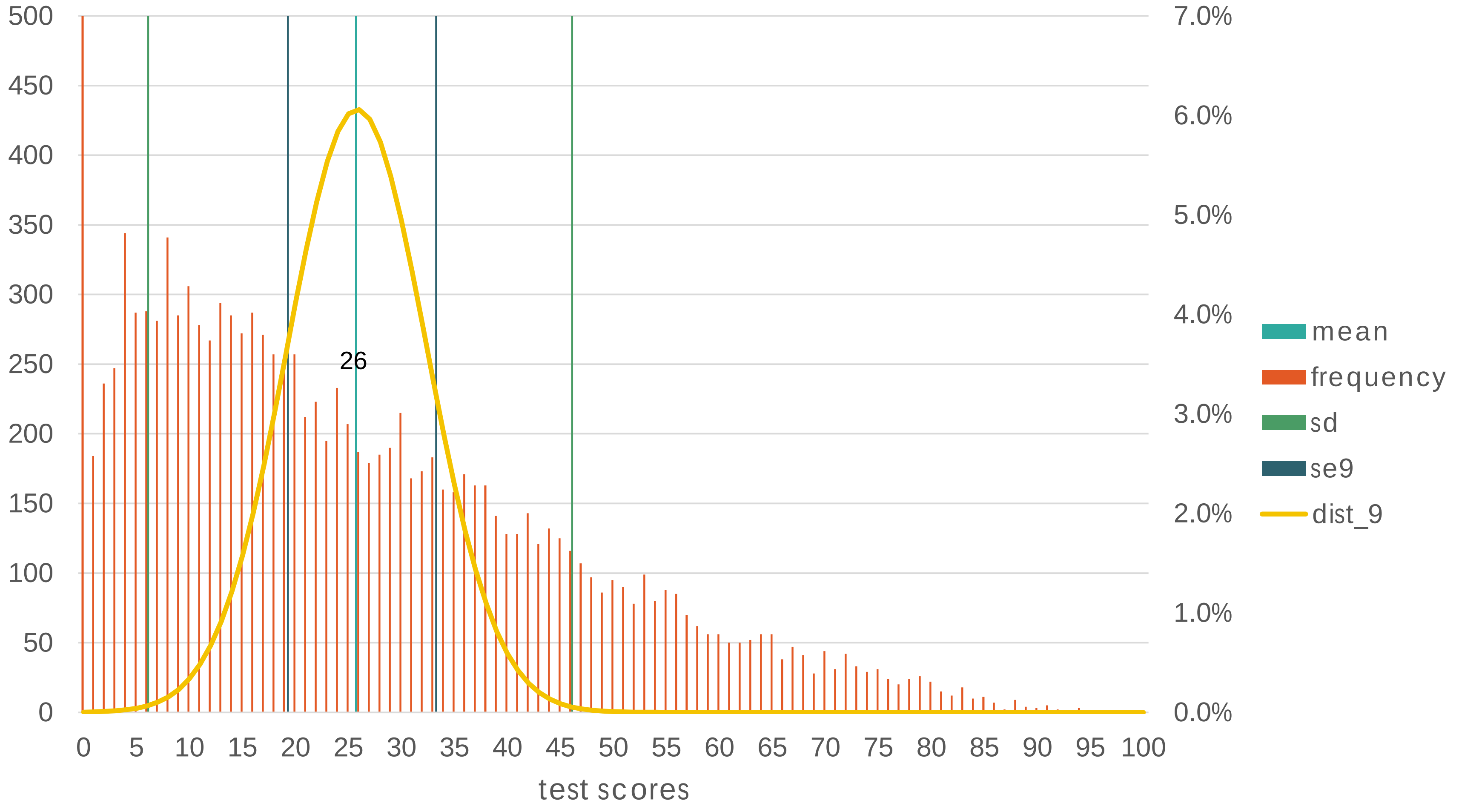
Standard Deviation/ Standard Error



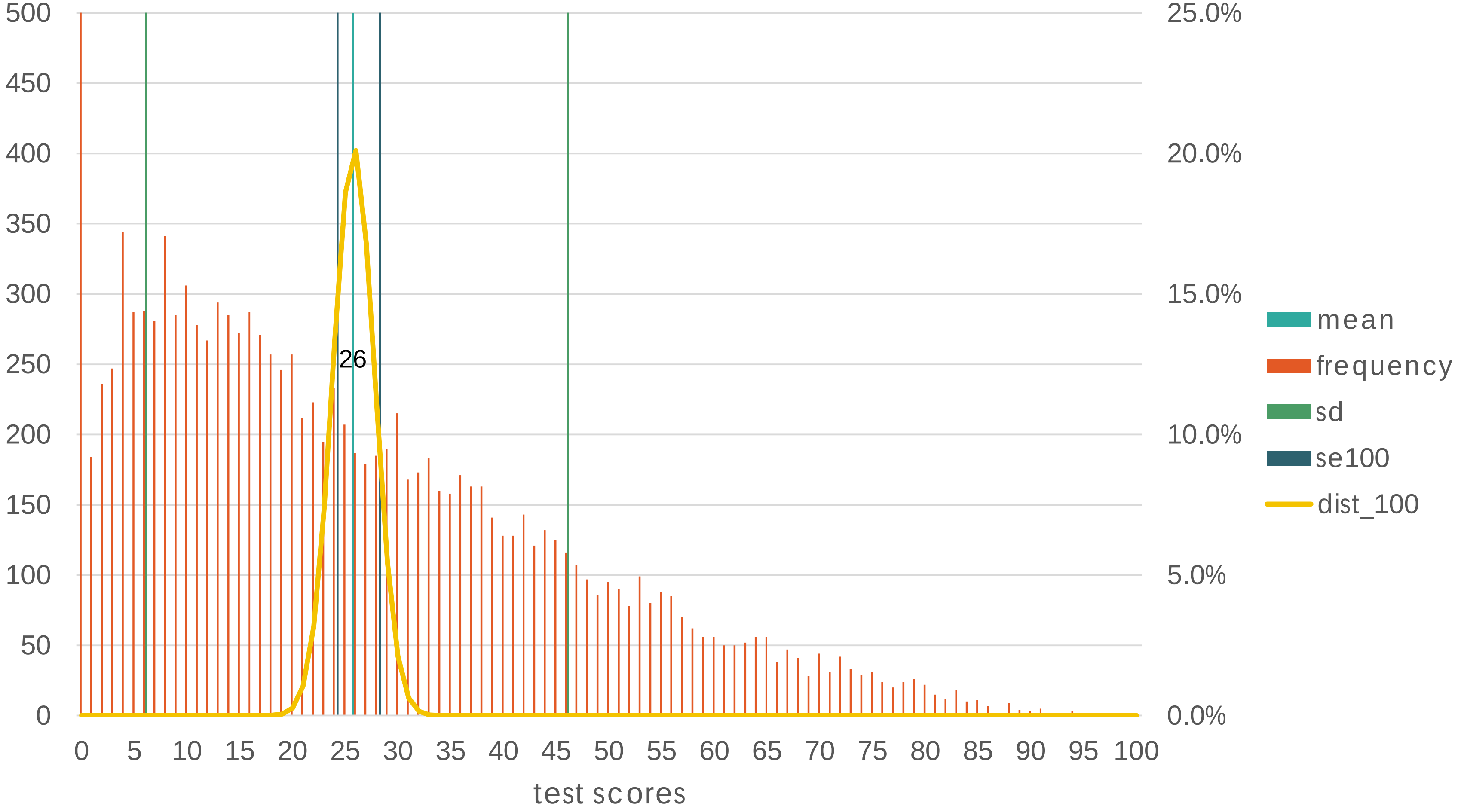
Sample size ↑ x4, SE ↓ ½



Sample size ↑ x9, SE ↓ ?



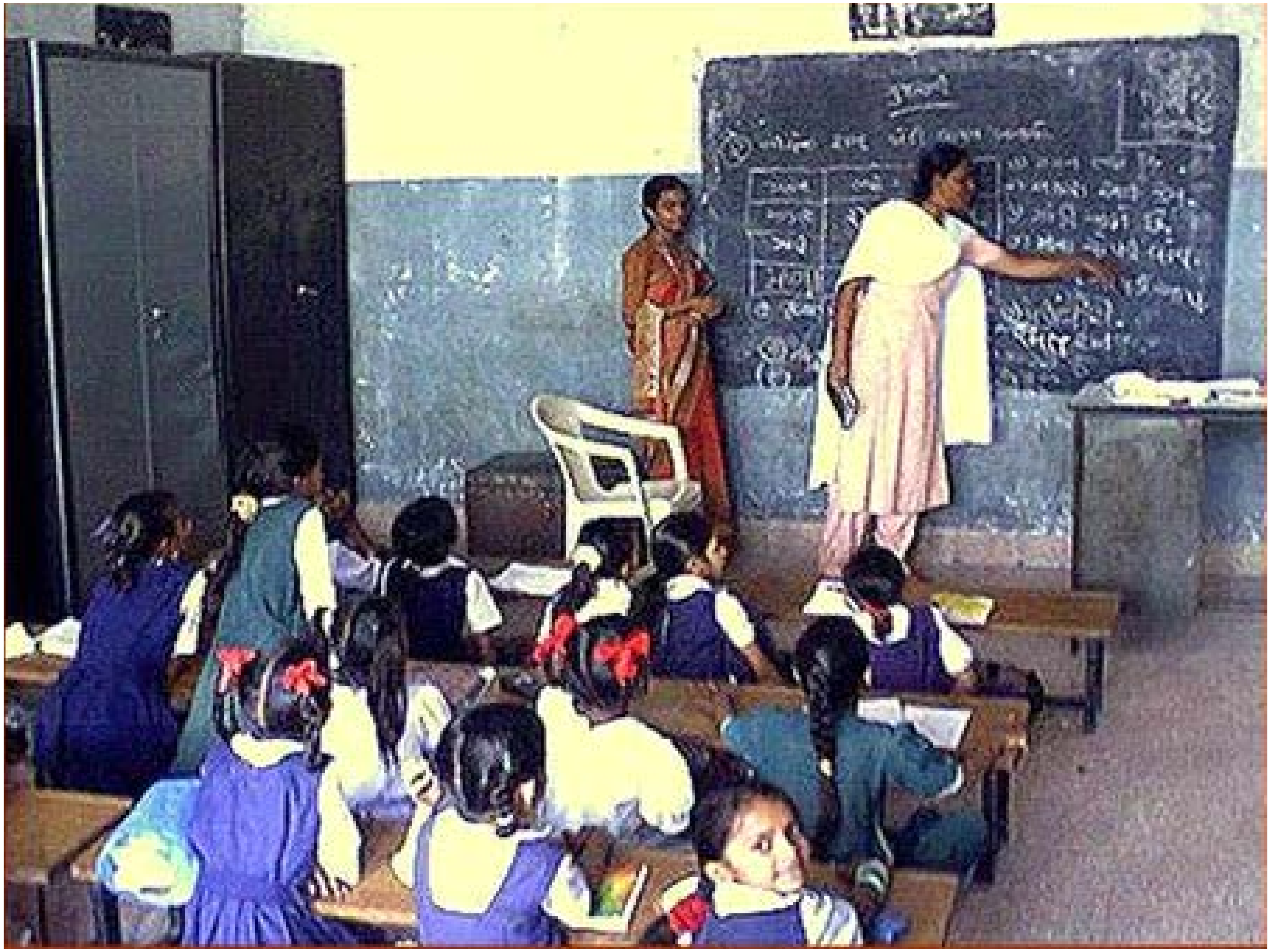
Sample size ↑ x100, SE ↓?



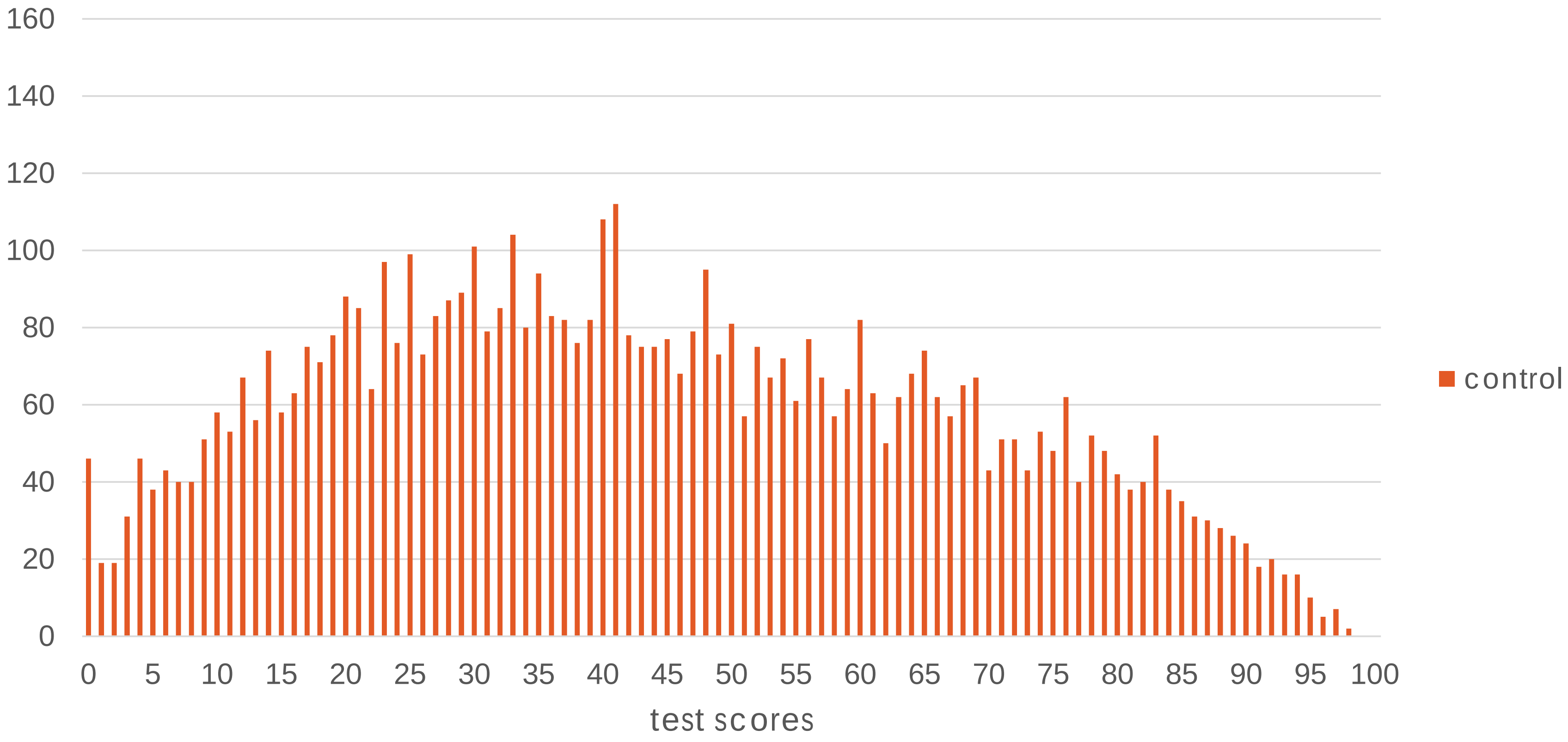
Outline

- Sampling distributions
- Detecting impact
  - significance
  - effect size
  - power
  - baseline and covariates
  - clustering
  - stratification

We implement the Balsakhi Program

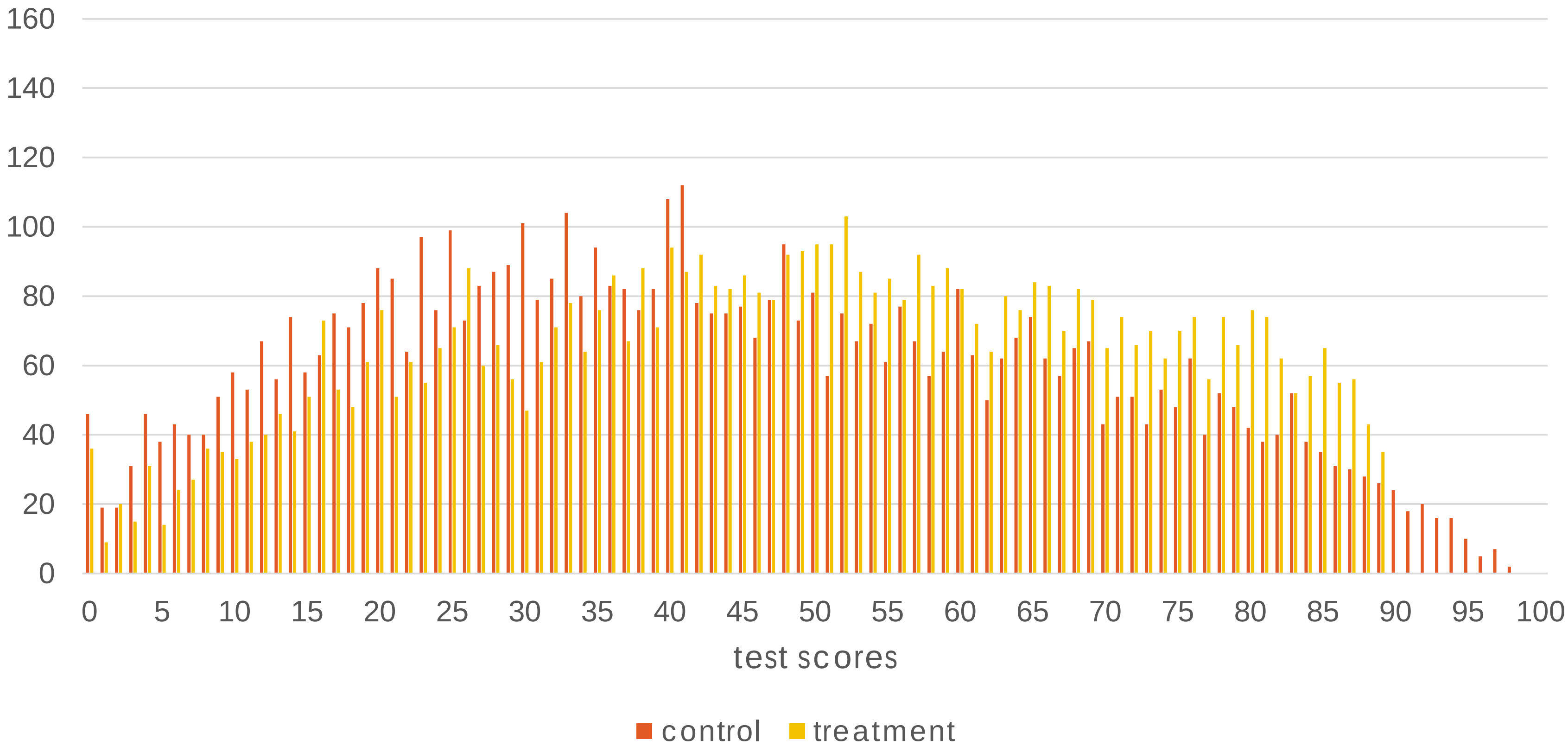


Control Group endline test scores



After the balsakhi programs, these are the endline test scores

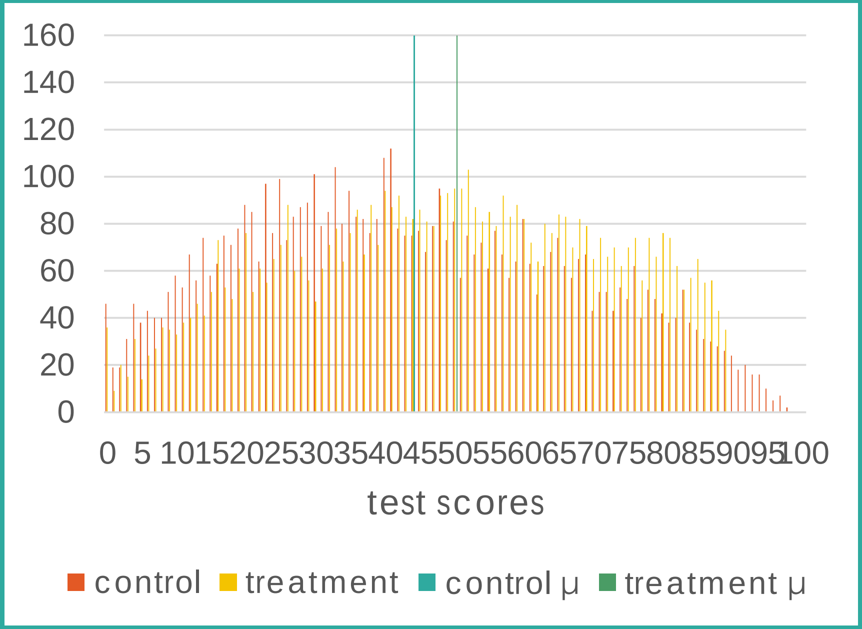
Endline test scores: control & treatment



Stop! That was the control group.
The treatment group is yellow.

Is this impact statistically significant?

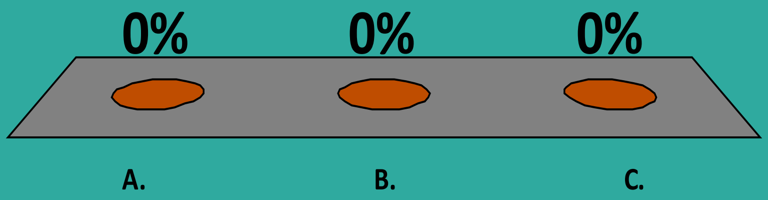
Average Difference = 6 points



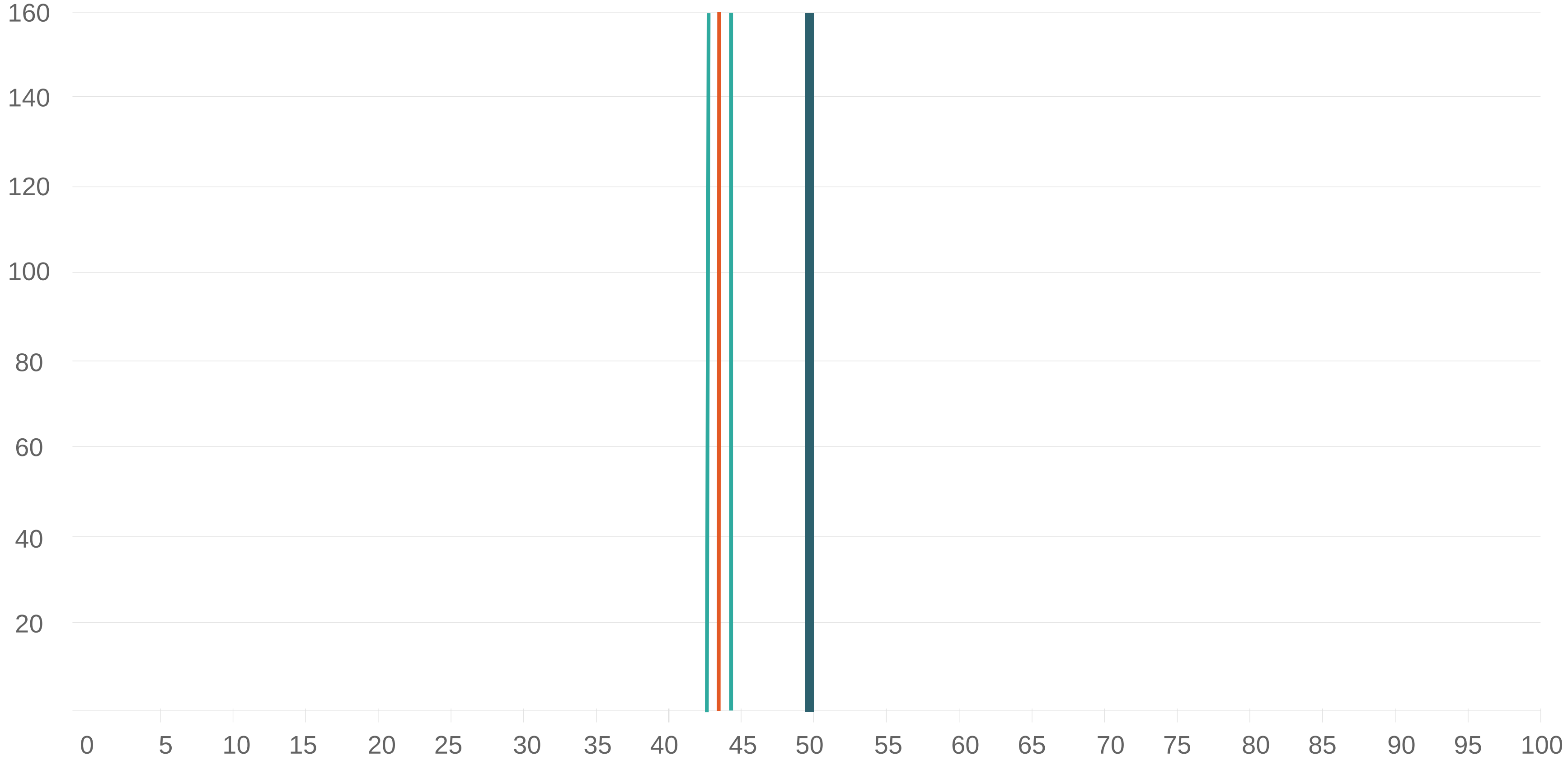
A. Yes

B. No

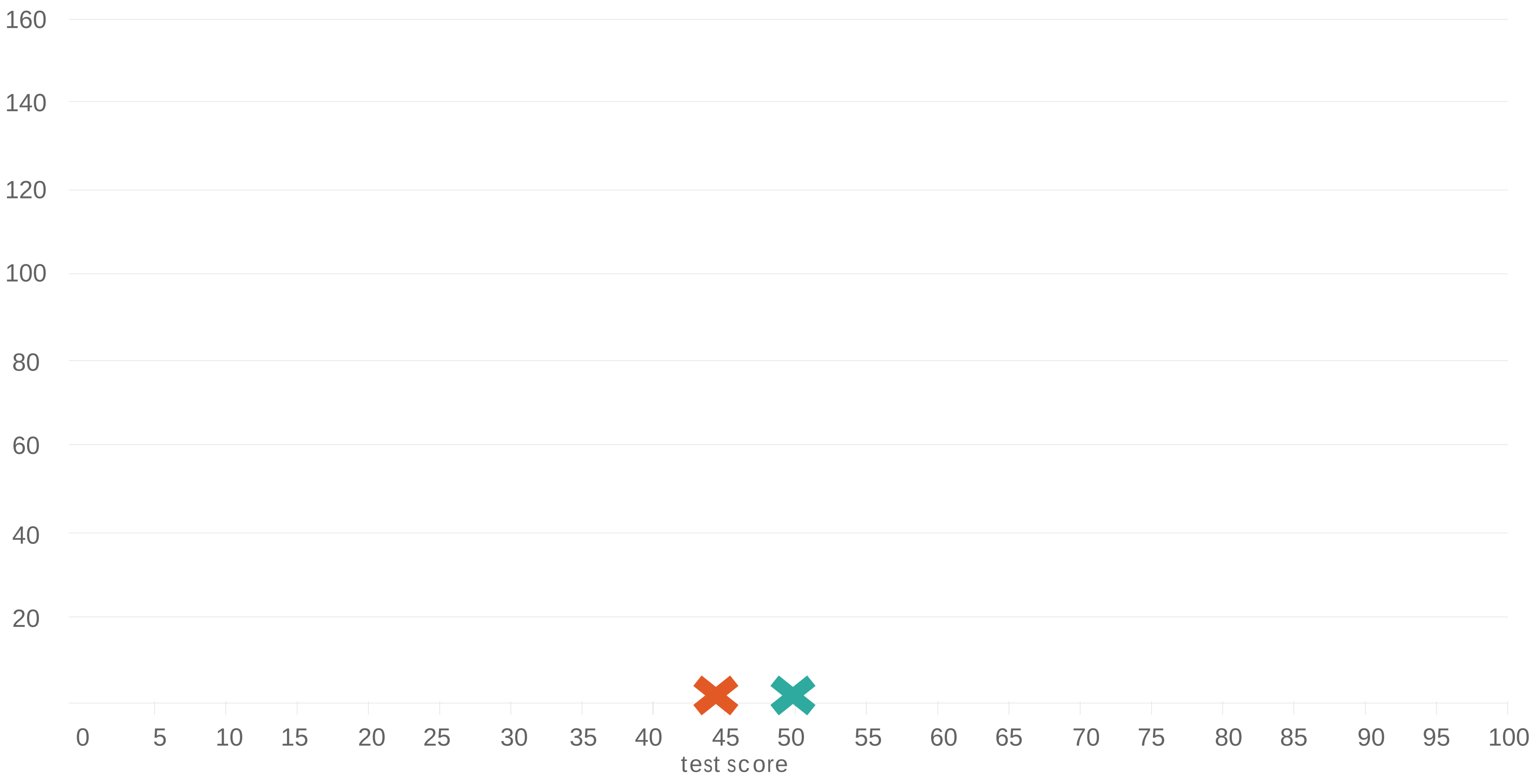
C. Don't know



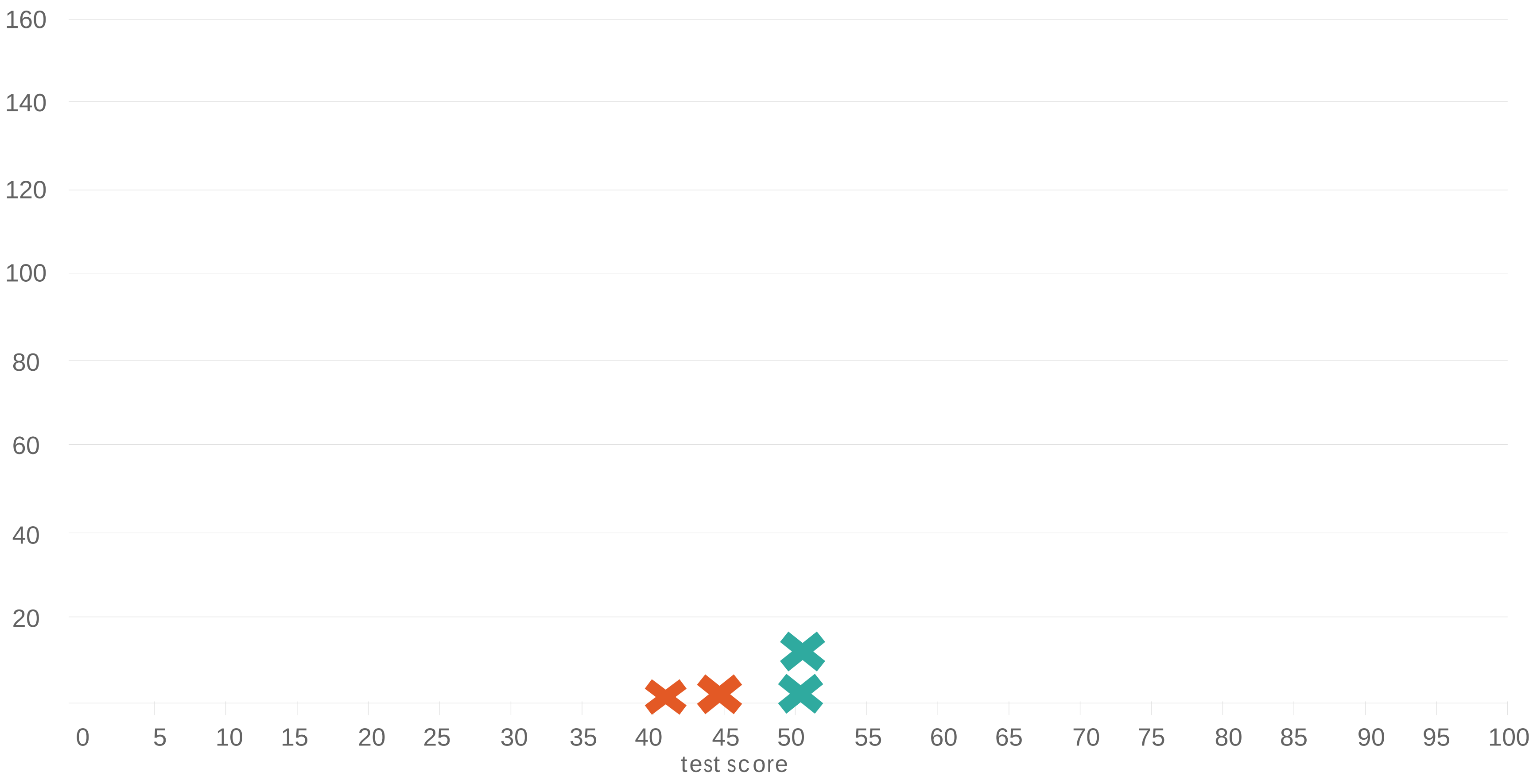
One experiment: 6 points



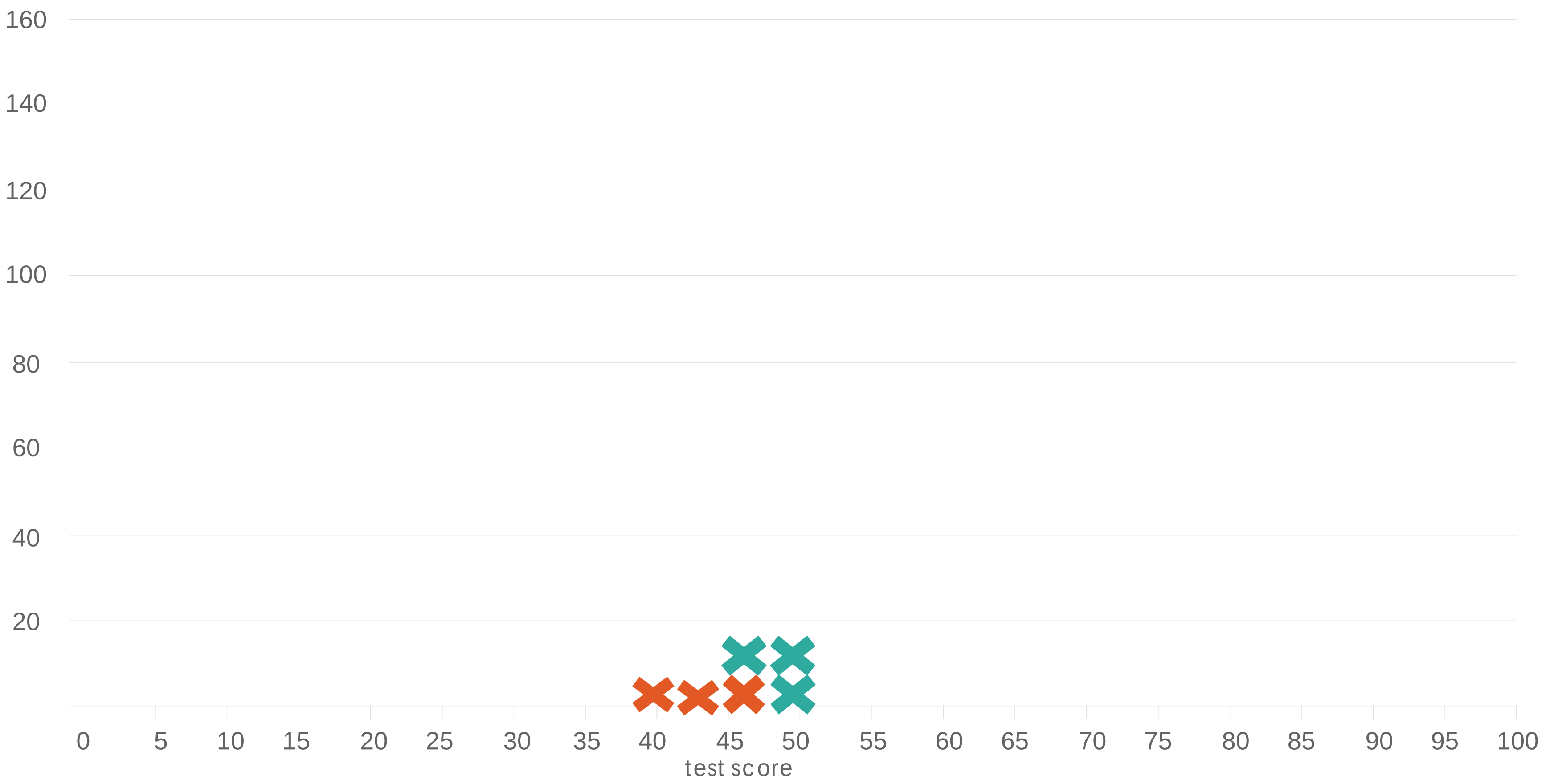
One experiment



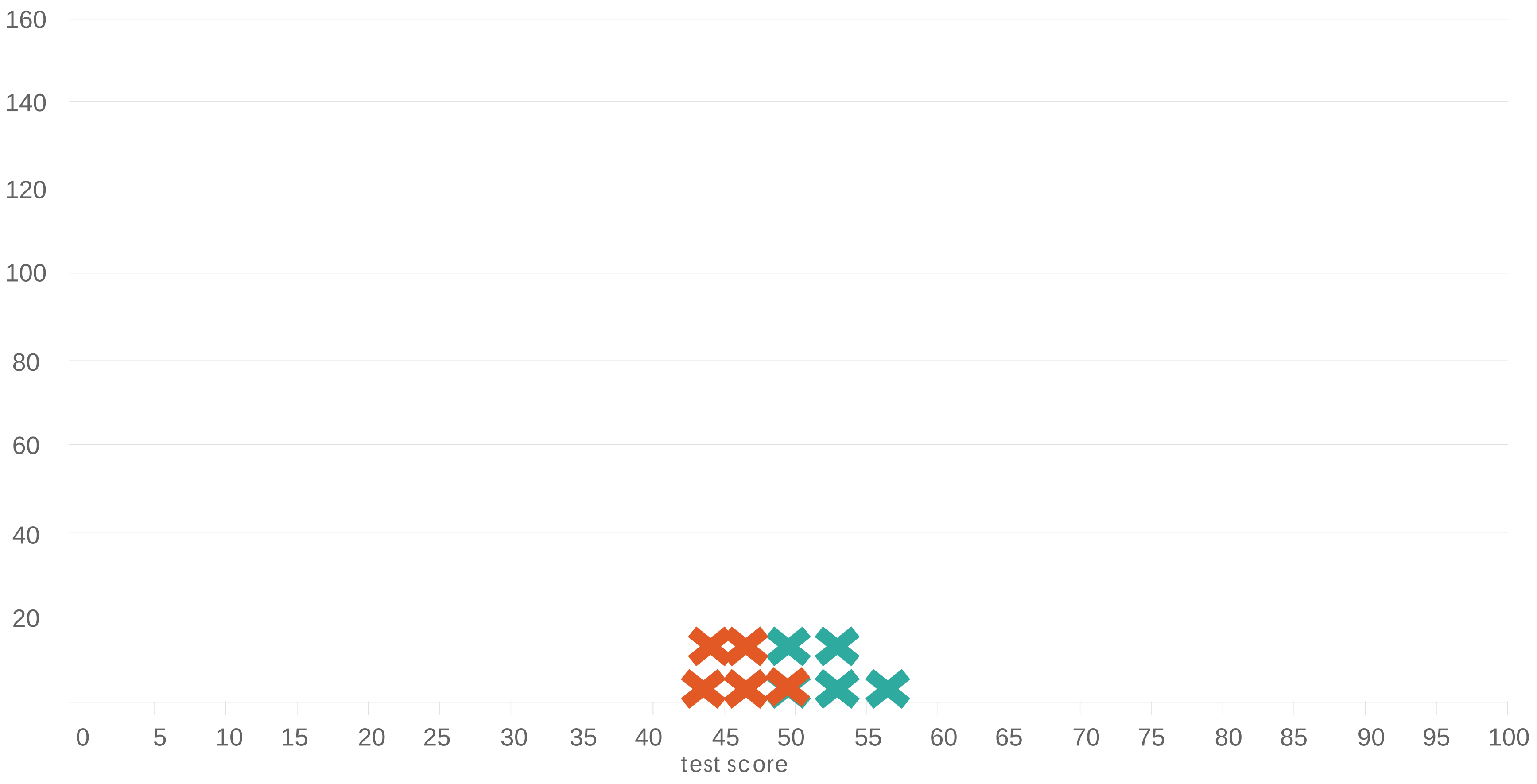
Two experiments



A few more…

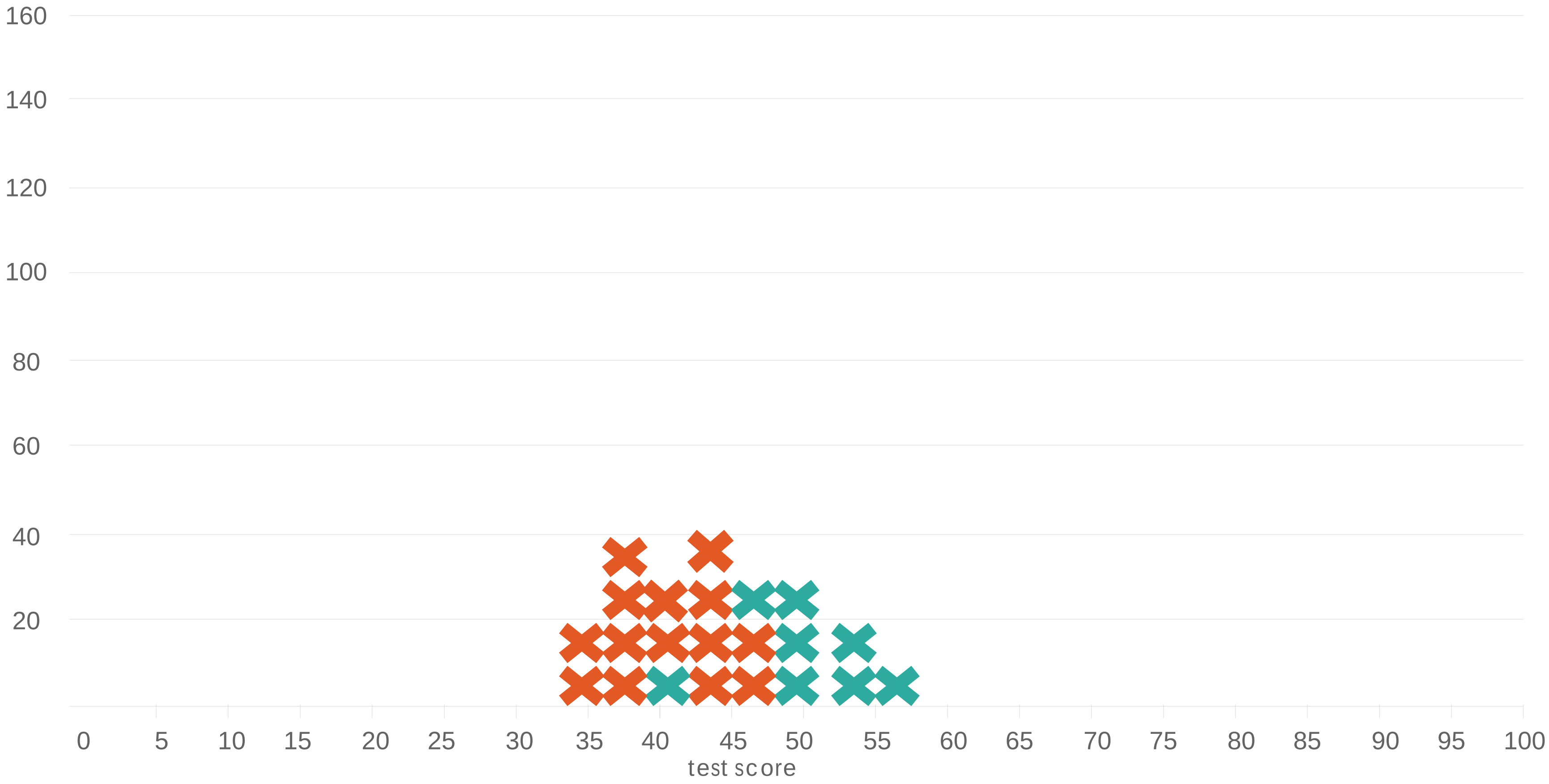

160
140
120
100
80
60
40
20
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
test score

A few more…

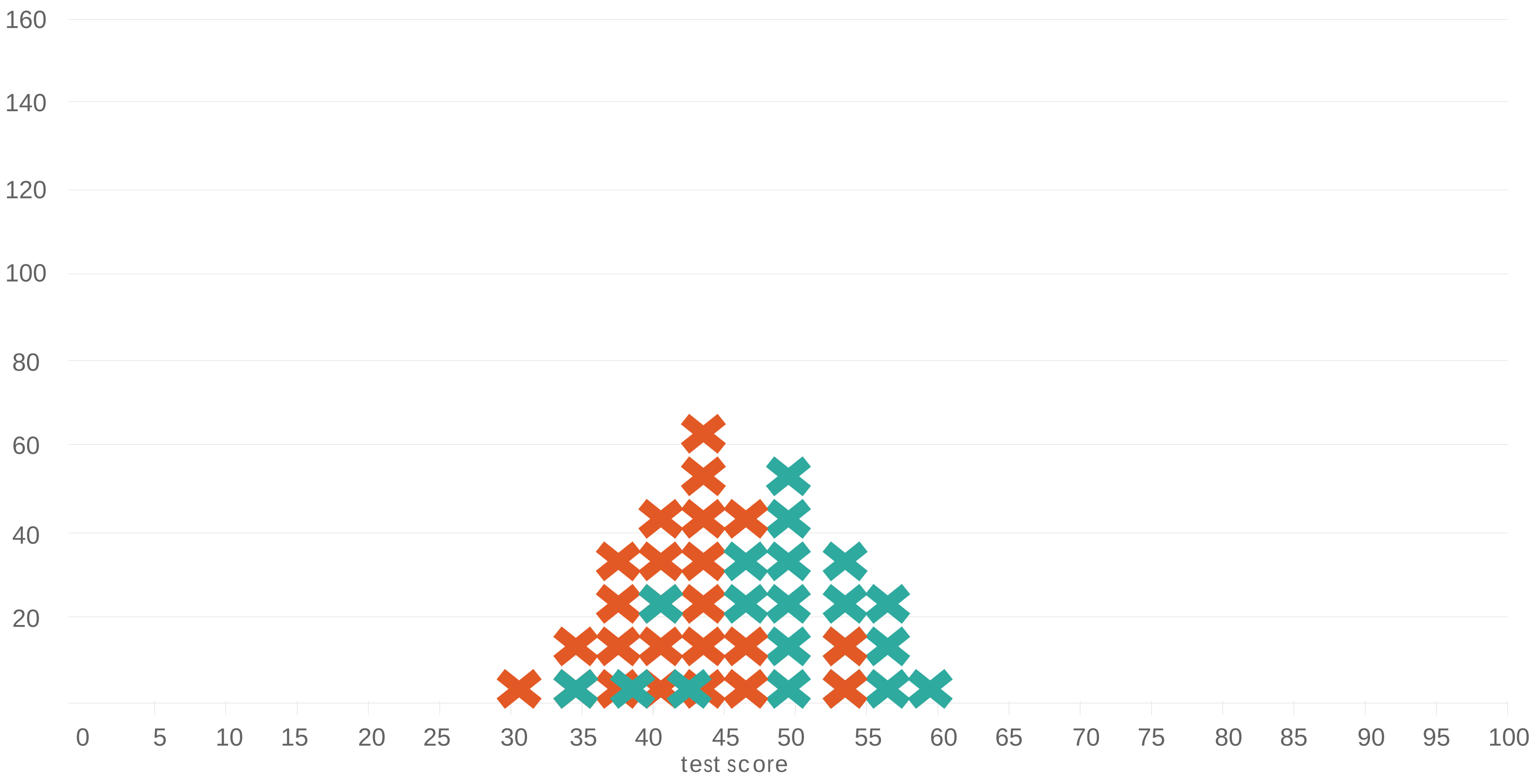


160
140
120
100
80
60
40
20
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
test score

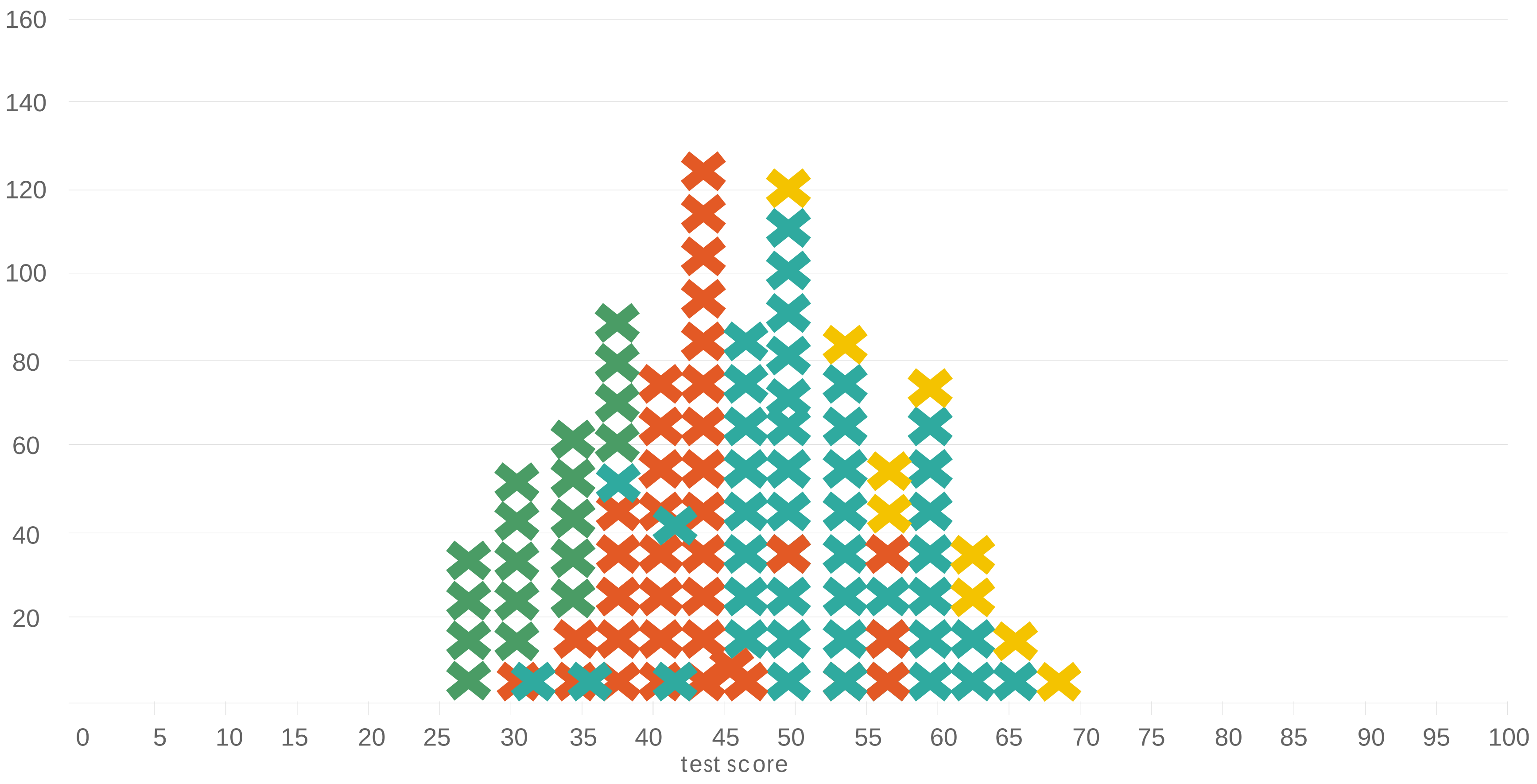
Many more...


160
140
120
100
80
60
40
20
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
test score

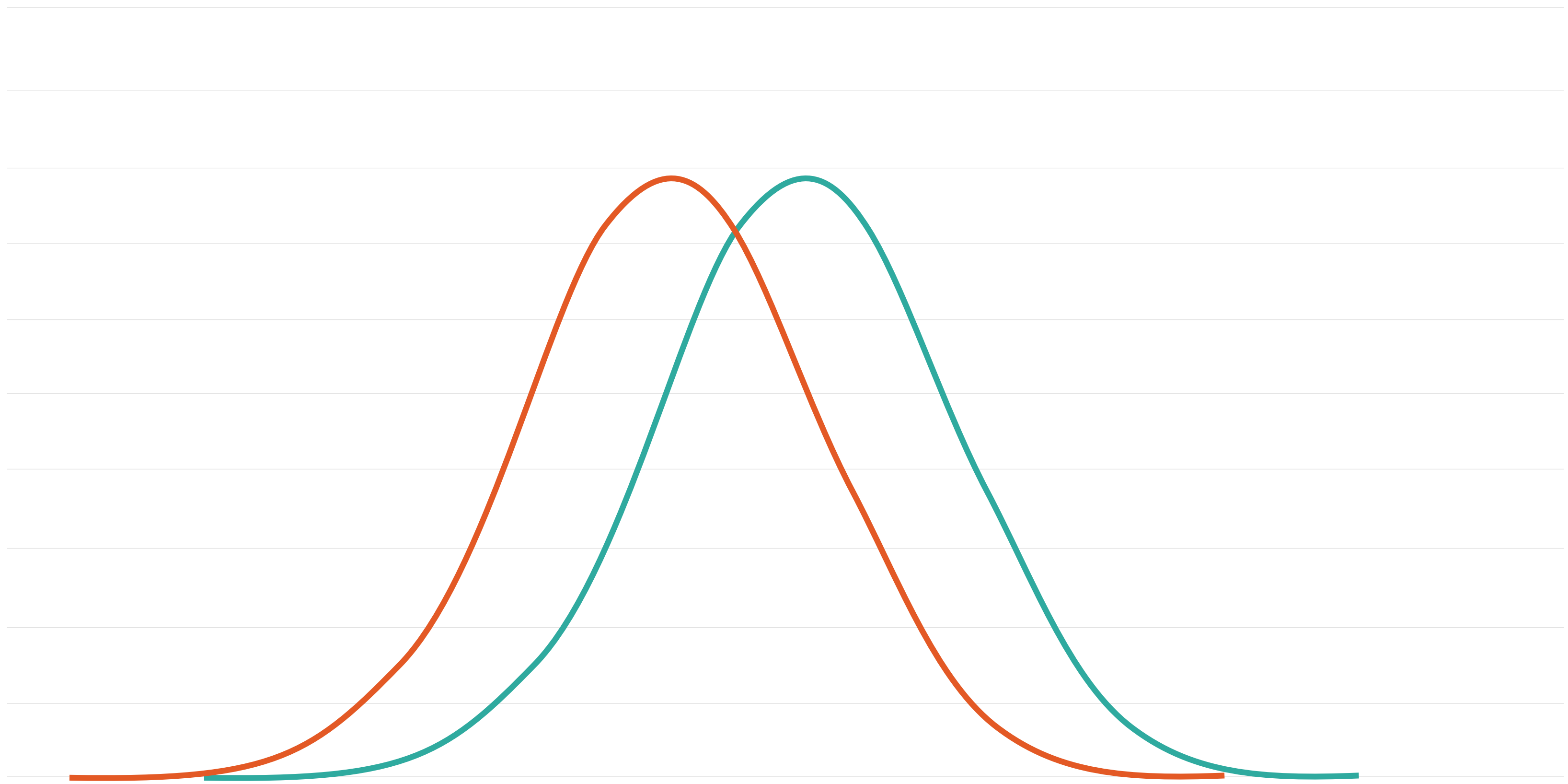
A whole lot more…



A whole.. lot more…



Running the experiment thousands of times...



By the Central Limit Theorem, these are normally distributed

Hypothesis Testing

- In criminal law, most institutions follow the rule: “innocent until proven guilty”
- The presumption is that the accused is innocent and the burden is on the prosecutor to show guilt
  - The jury or judge starts with the “null hypothesis” that the accused person is innocent
  - The prosecutor has a hypothesis that the accused person is guilty

Hypothesis Testing

- In program evaluation, instead of "presumption of innocence," the rule is: "presumption of insignificance"
- The "Null hypothesis" (H0) is that there was no (zero) impact of the program
- The burden of proof is on the evaluator to show a significant effect of the program

Hypothesis Testing: Conclusions

- If it is very unlikely (less than a 5% probability) that the difference is solely due to chance:
  - We “reject our null hypothesis”
- We may now say:
  - “our program has a statistically significant impact”

Hypothesis Testing: Steps

1. Determine the (size of the) sampling distribution around the null hypothesis H_0 by calculating the standard error
2. Choose the confidence interval, e.g. 95% (or significance level: α) (α=5%)
3. Identify the critical value (boundary of the confidence interval)
4. If our observation falls in the critical region we can reject the null hypothesis

What is the significance level?

- Type I error: rejecting the null hypothesis even though it is true (false positive)

- Significance level: *The probability* that we will reject the null hypothesis even though it is true

Hypothesis testing: 95% confidence

		You Conclude	
		Effective	*No Effect*
The Truth	*Effective*	☺	Type II Error (low power) ☹
	No Effect	Type I Error (5% of the time) ☹	☺

What is Power?

- Type II Error: Failing to reject the null hypothesis (concluding there is no difference), when indeed the null hypothesis is false.

- Power: If there is a measureable effect of our intervention (the null hypothesis is false), the probability that we will detect an effect (reject the null hypothesis)

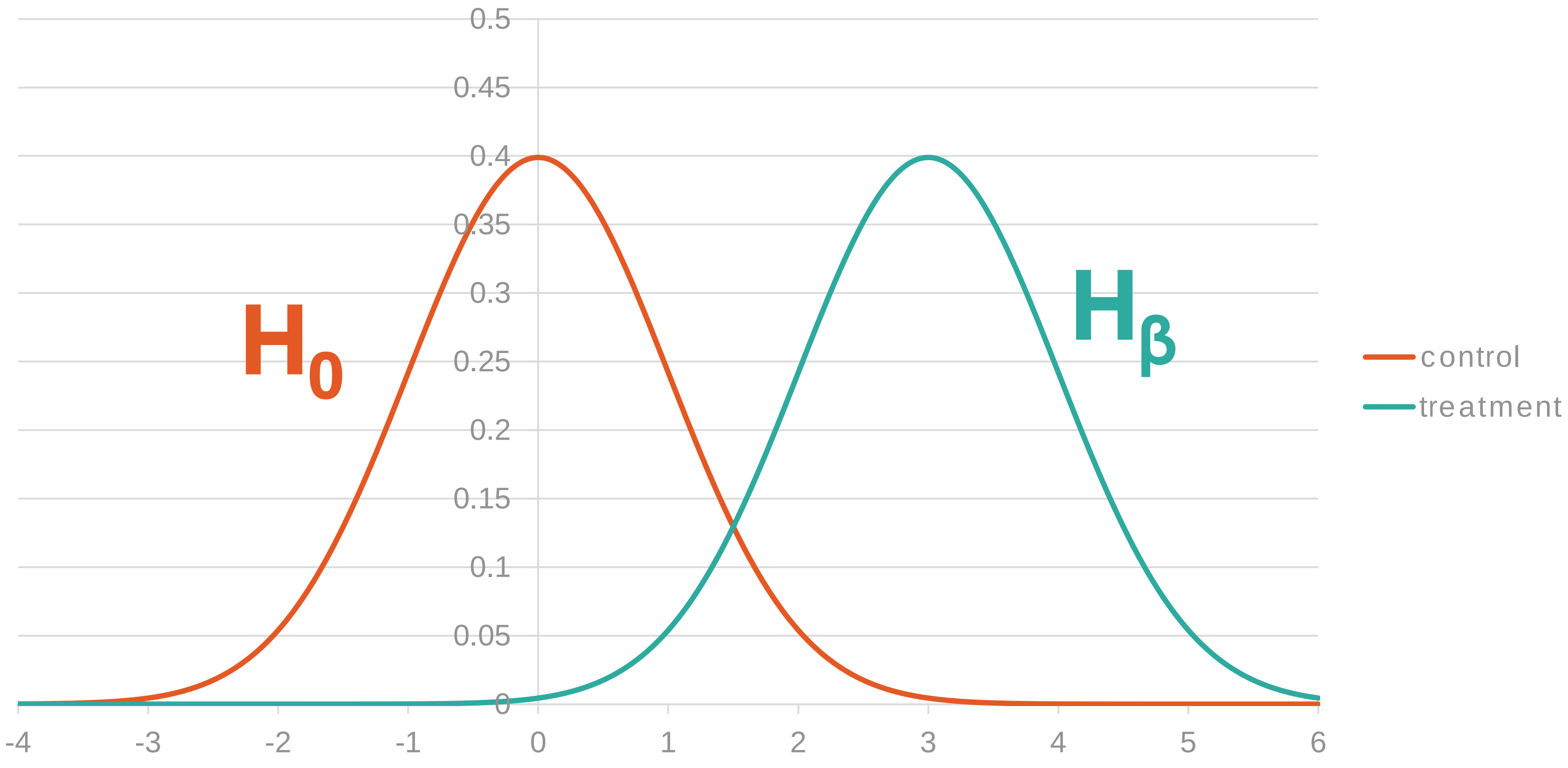
Hypothesis Testing: Steps

1. Determine the (size of the) sampling distribution around the null hypothesis H_0 by calculating the standard error
2. Choose the confidence interval, e.g. 95% (or significance level: α) (α=5%)
3. Identify the critical value (boundary of the confidence interval)
4. If our observation falls in the critical region we can reject the null hypothesis

Determining Power: Steps

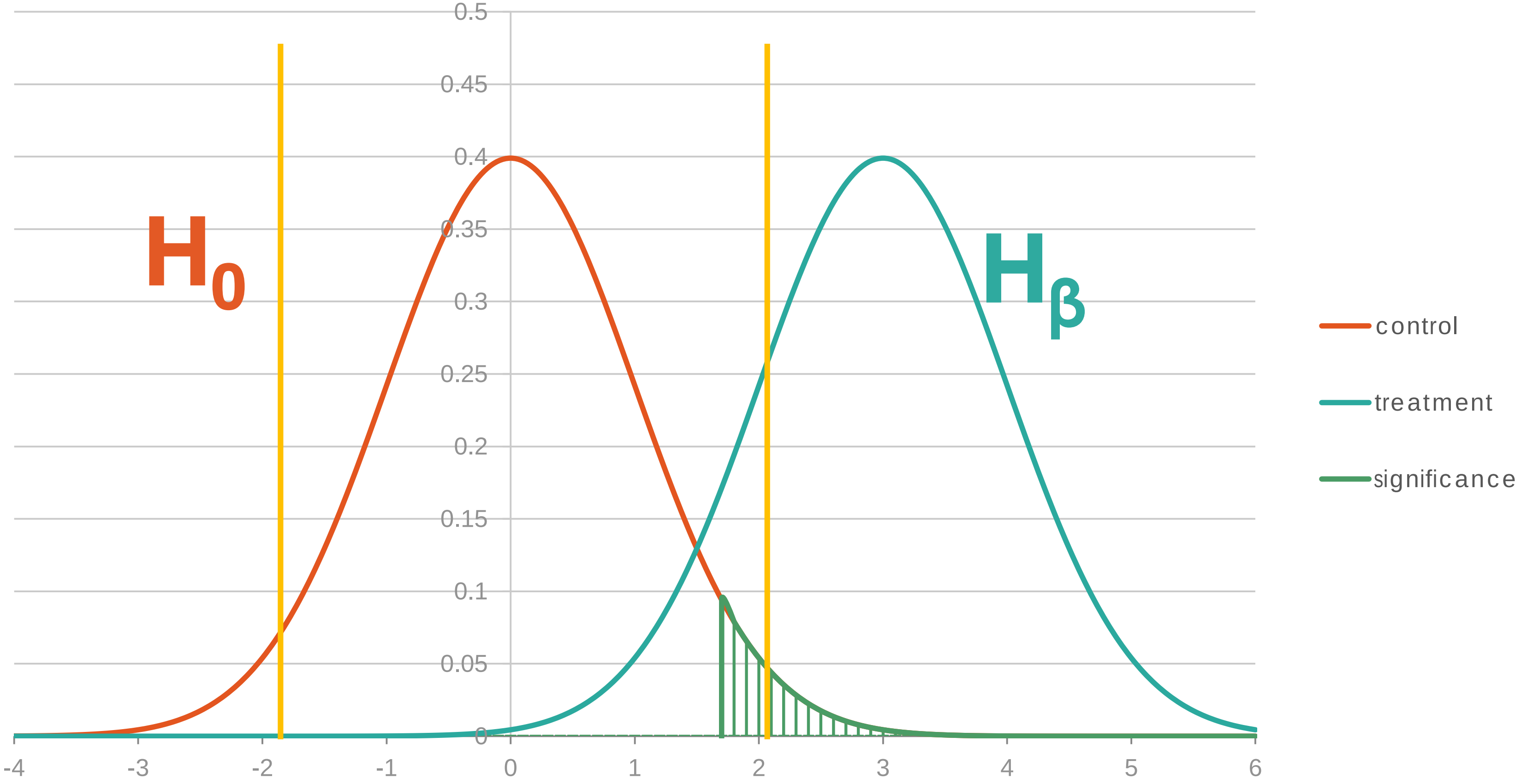
1. Determine the (size of the) sampling distribution around the null hypothesis H_0 by calculating the standard error
2. Hypothesize an effect size H_β
3. Determine the (size of the) sampling distribution around the alternate hypothesis
4. Choose the confidence interval, e.g. 95%
   (or significance level: α) (α=5%)
5. Identify the critical value (boundary of the confidence interval)
6. Determine where in the H_β sampling distribution, the critical value lies.
7. Calculate the proportion of the mass under the H_β sampling distribution that lies on the other side of the critical value (away from the null hypothesis)

Before the experiment



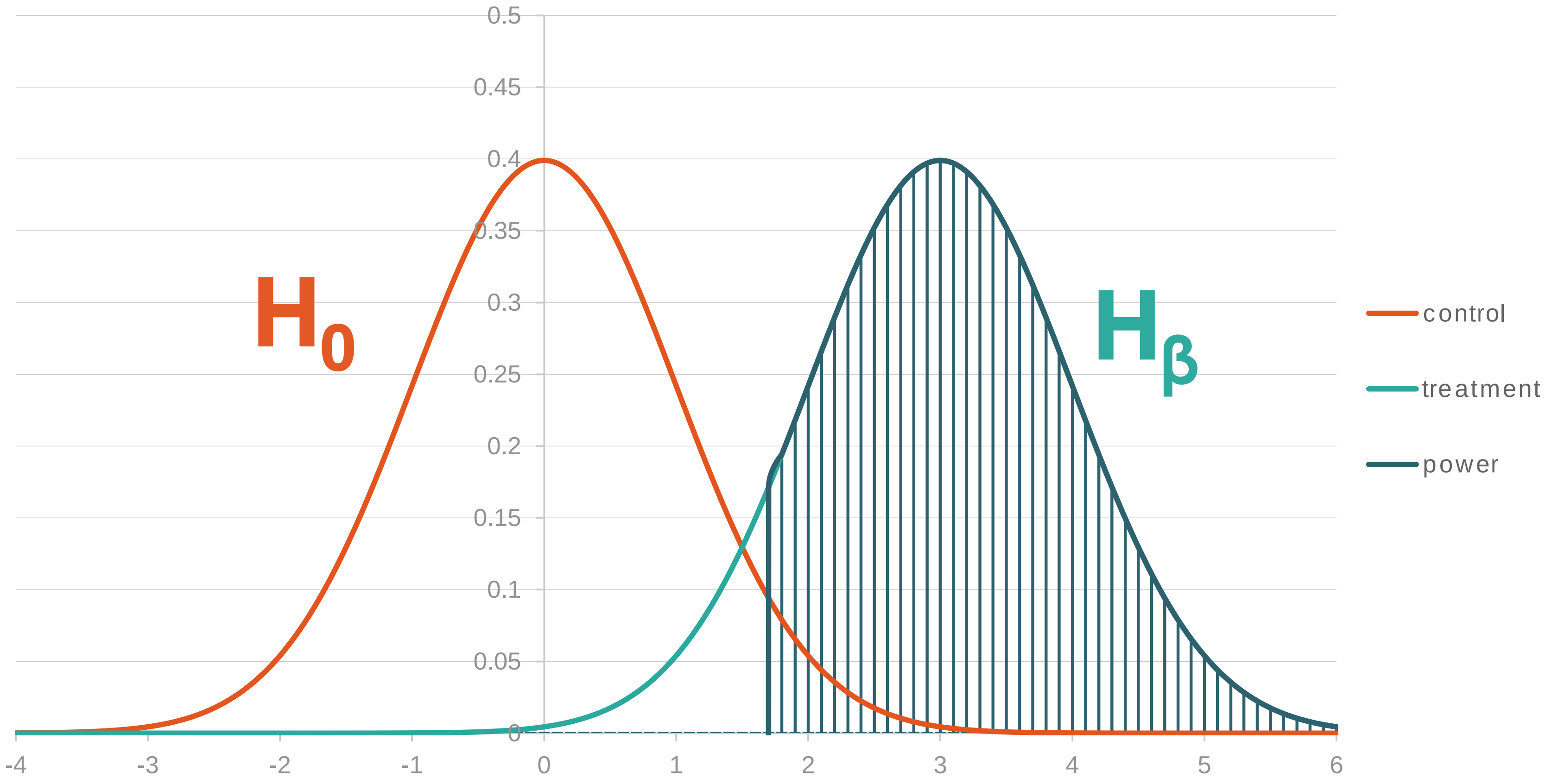
ASSUME TWO EFFECTS: no effect and treatment effect β

Impose significance level of 5%



Anything between lines cannot be distinguished from 0

Can we distinguish Hβ from H0 ?



Shaded area shows % of time we would find Hβ true if it was

What influences power?

- What are the factors that change the proportion of the research hypothesis that is shaded—i.e. the proportion that falls to the right (or left) of the null hypothesis curve?
- Understanding this helps us design more powerful experiments

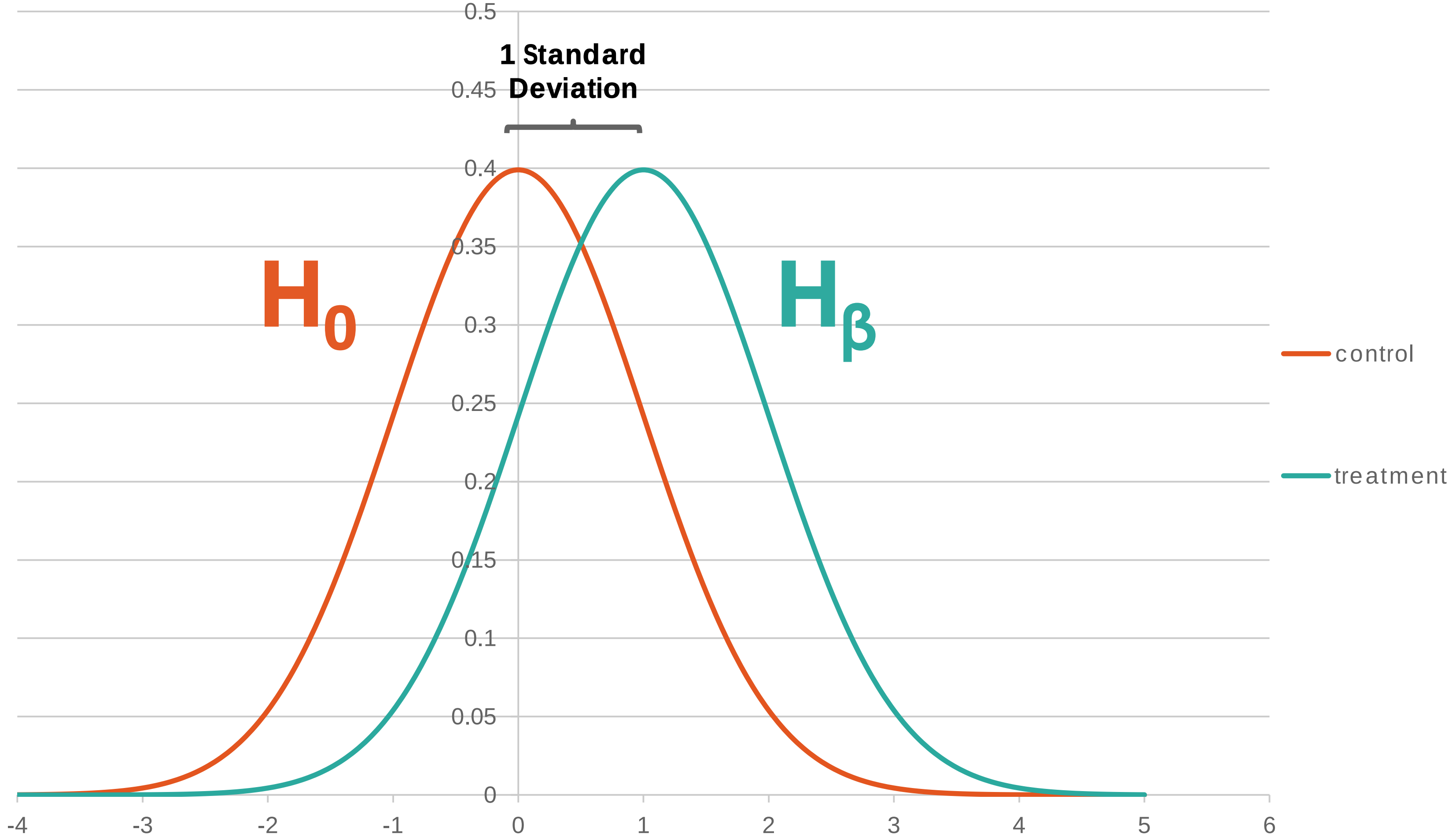
Power: main ingredients

1. Effect Size
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C
5. Clustering

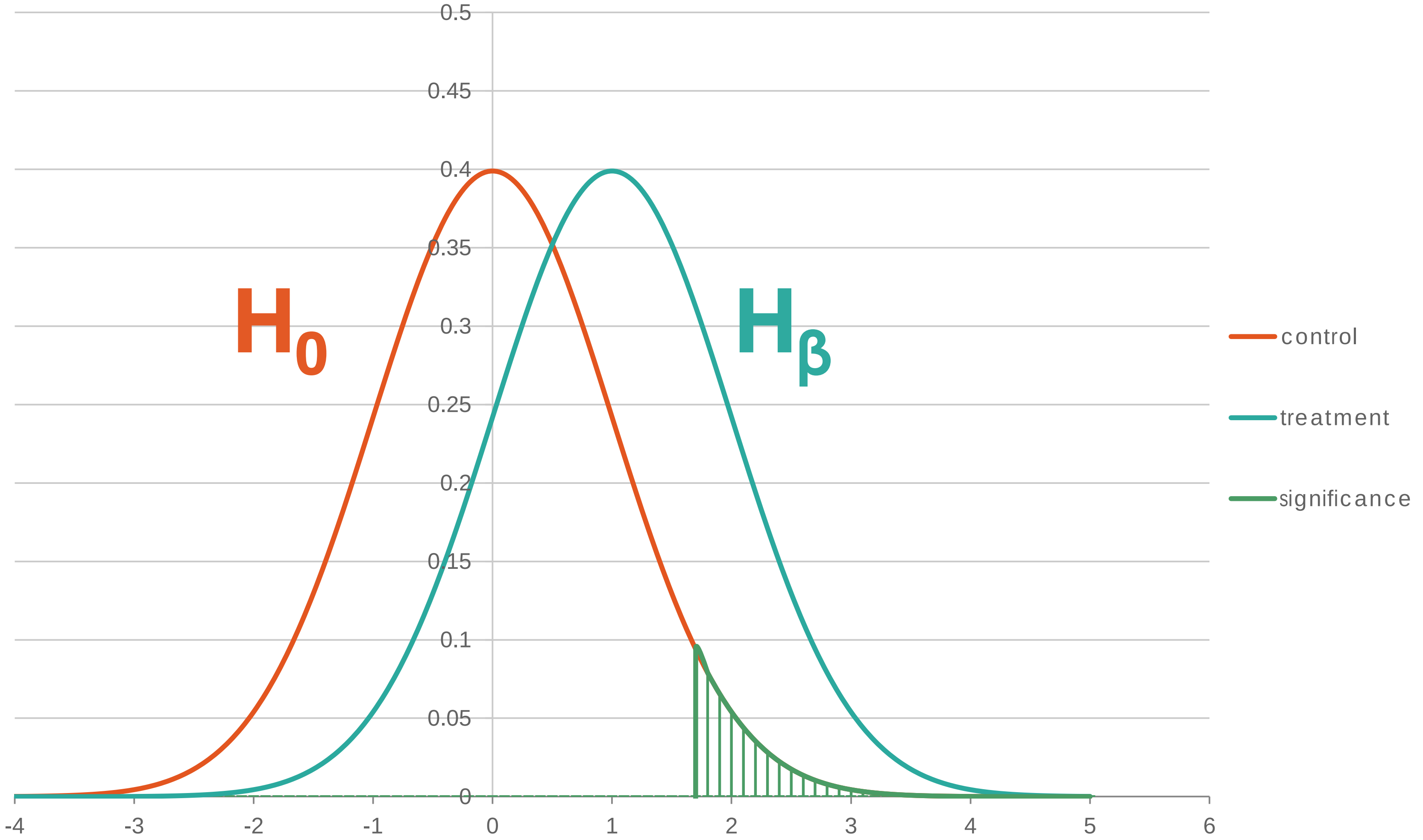
Power: main ingredients

1. Effect Size
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C
5. Clustering

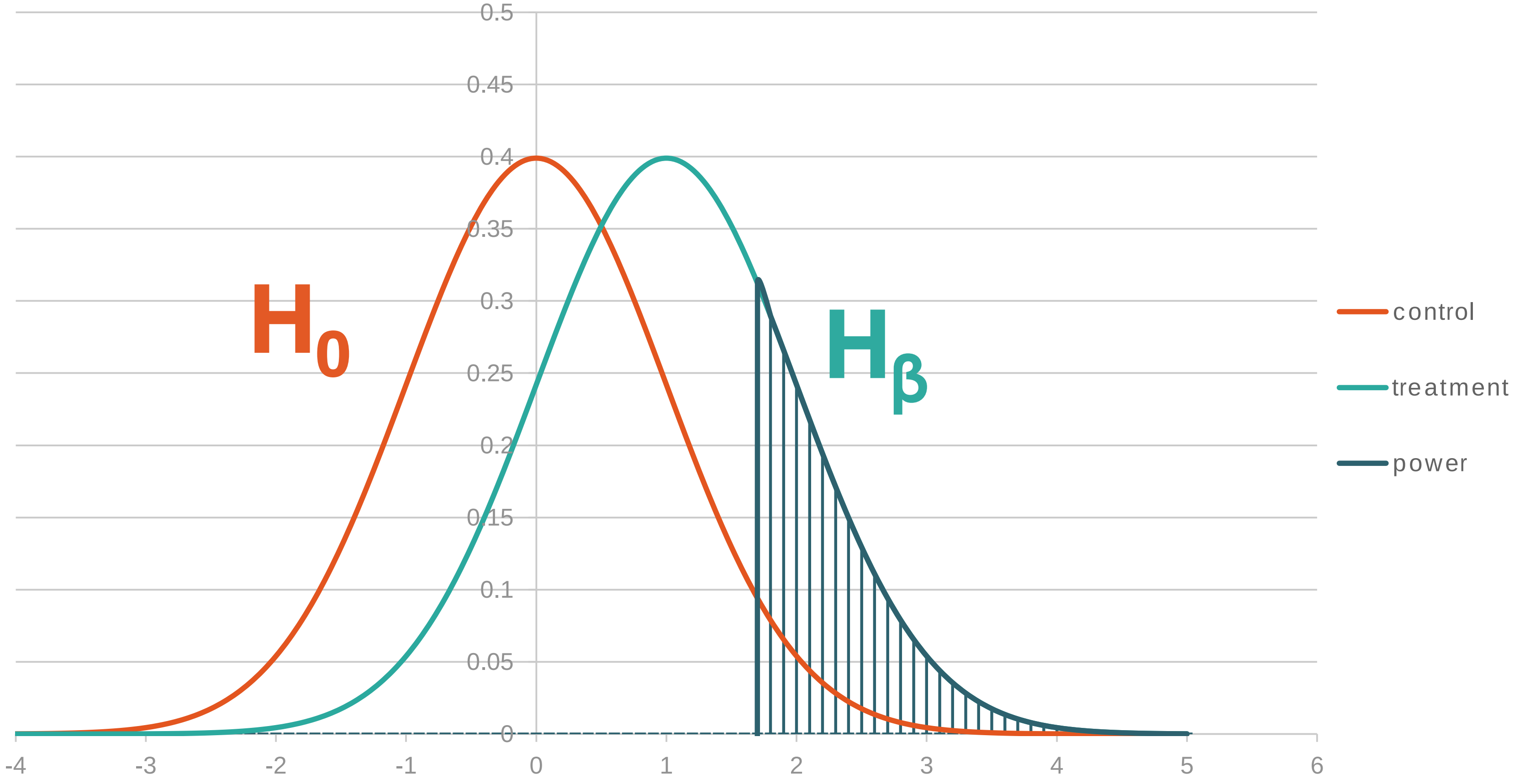
Effect Size: 1*SE



Effect Size = 1*SE

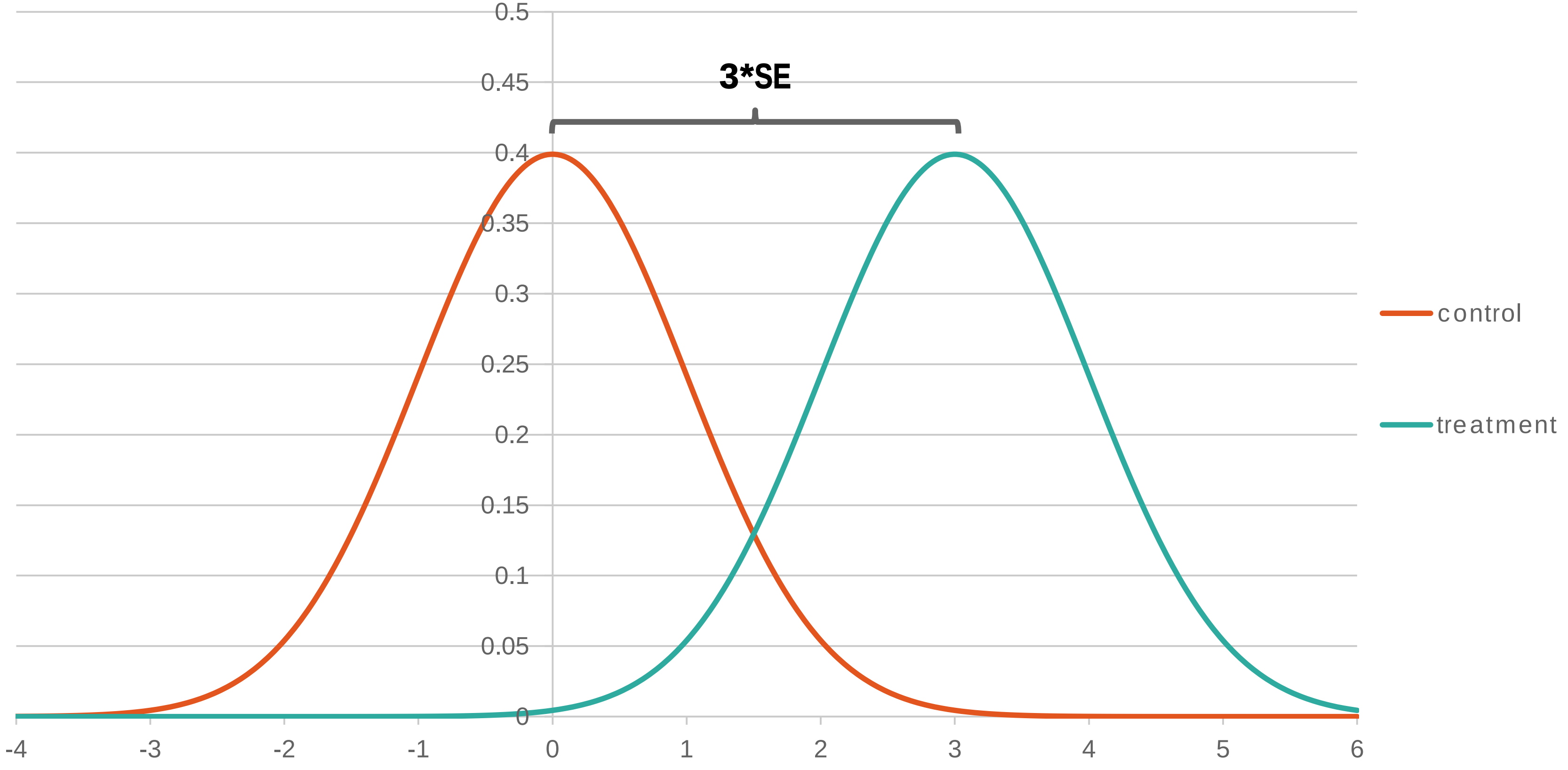


Power: 26%
If the true impact was 1*SE…



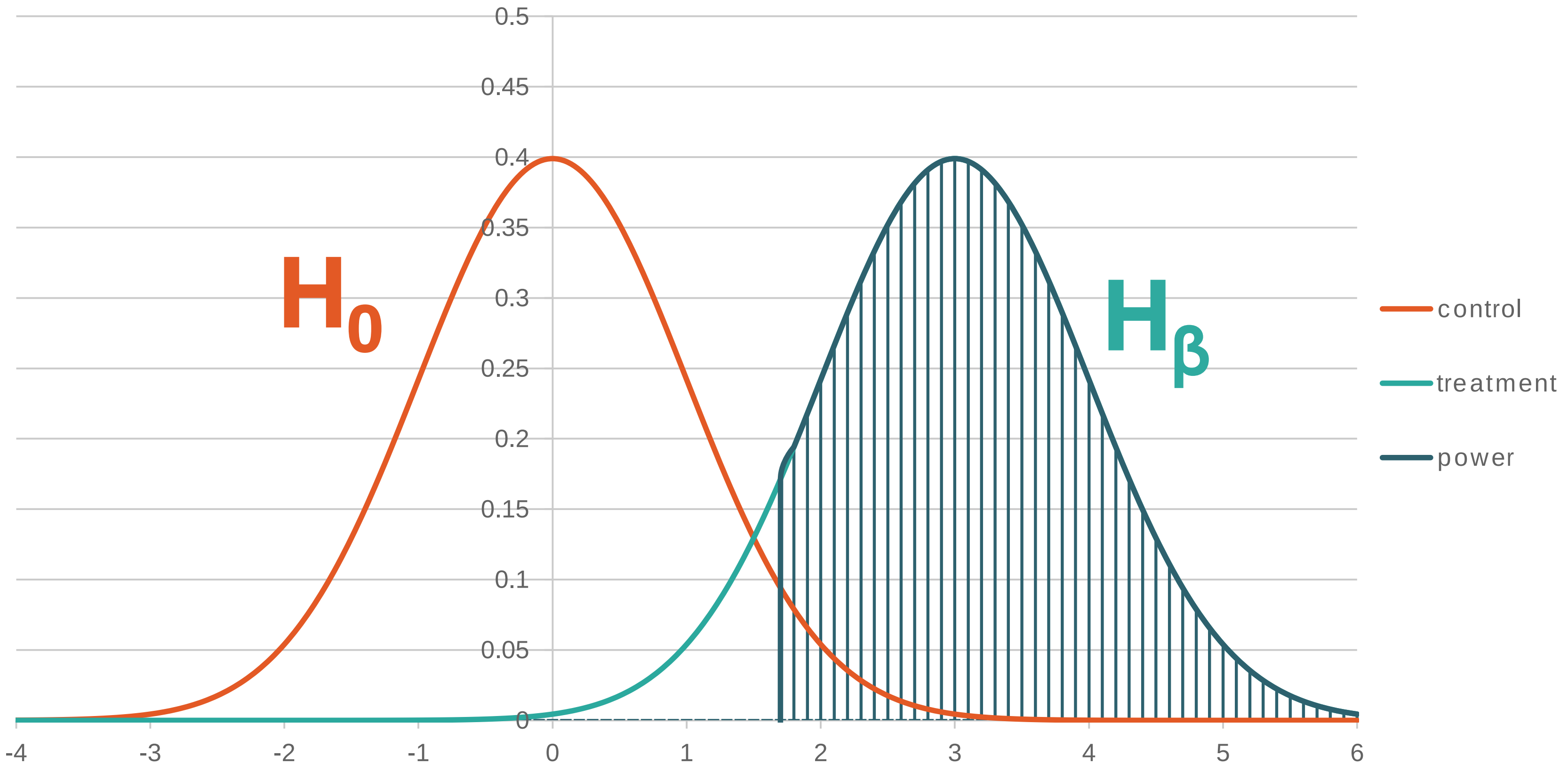
The Null Hypothesis would be rejected only 26% of the time

Effect Size: 3*SE



Bigger hypothesized effect size→ distributions farther apart

Effect size 3*SE: Power= 91%



Bigger Effect size means more power

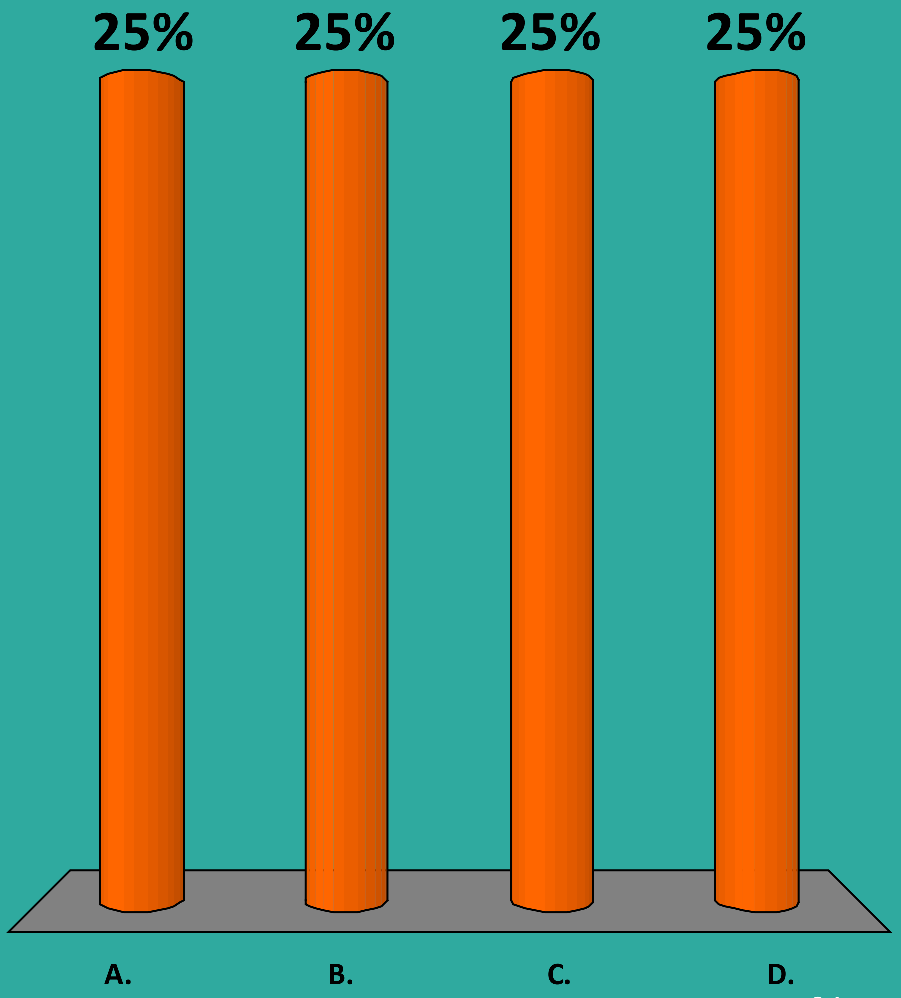
What effect size should you use when designing your experiment?

A. Smallest effect size that is still cost effective

B. Largest effect size you expect your program to produce

C. Both

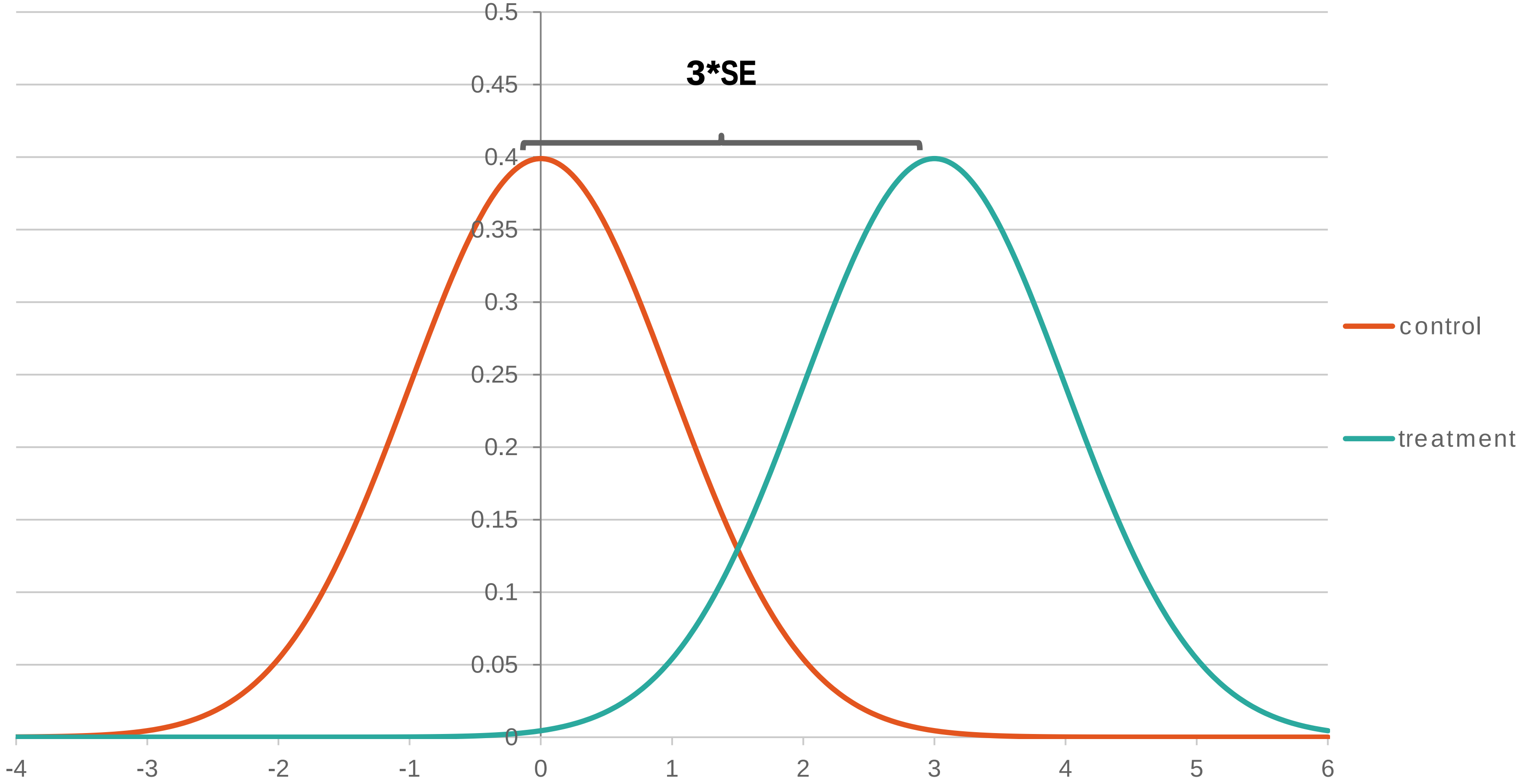
D. Neither



Effect size and take-up

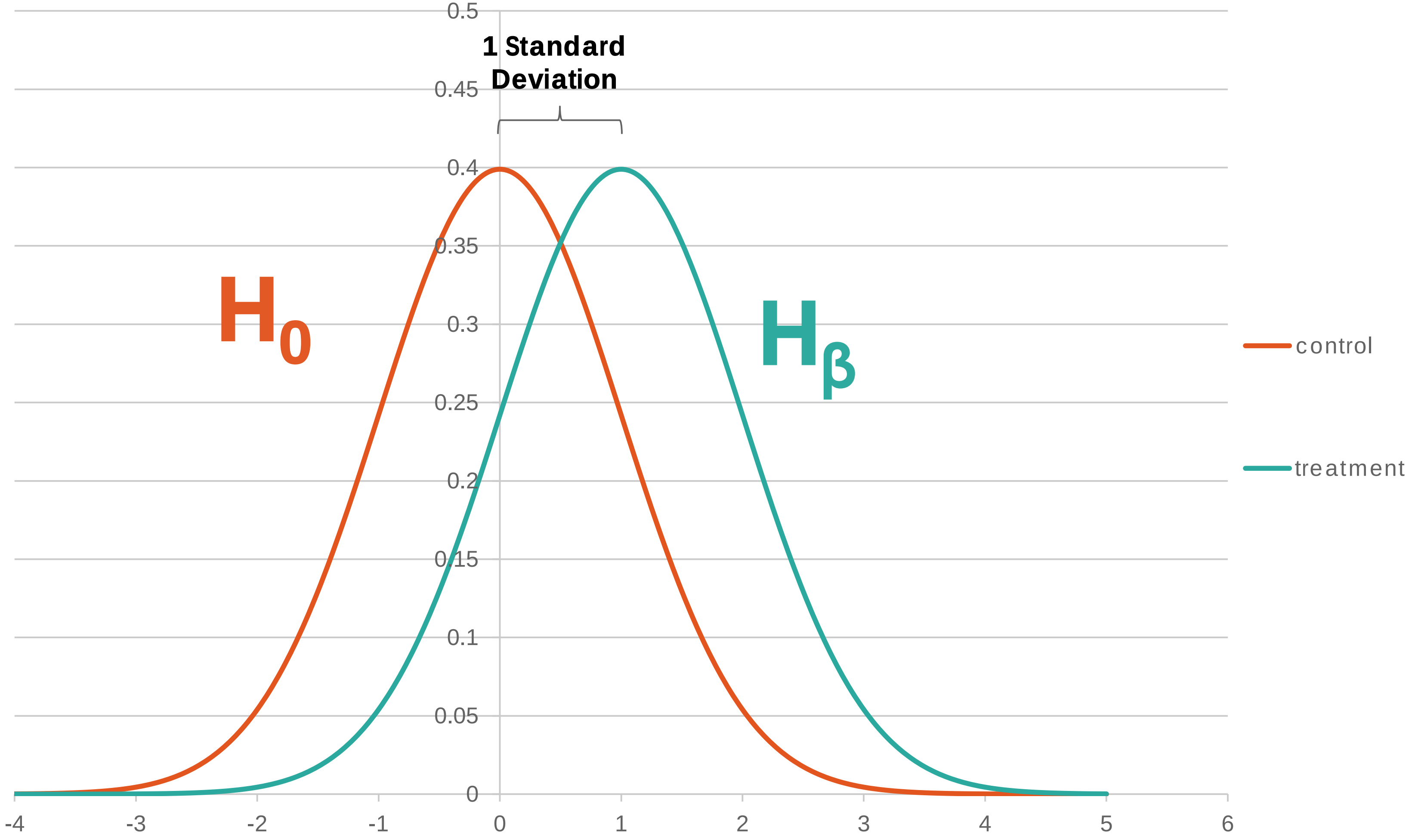
- Let's say we believe the impact on our participants is "3"
- What happens if take up is 1/3?
- Let's show this graphically

Effect Size: 3*SE

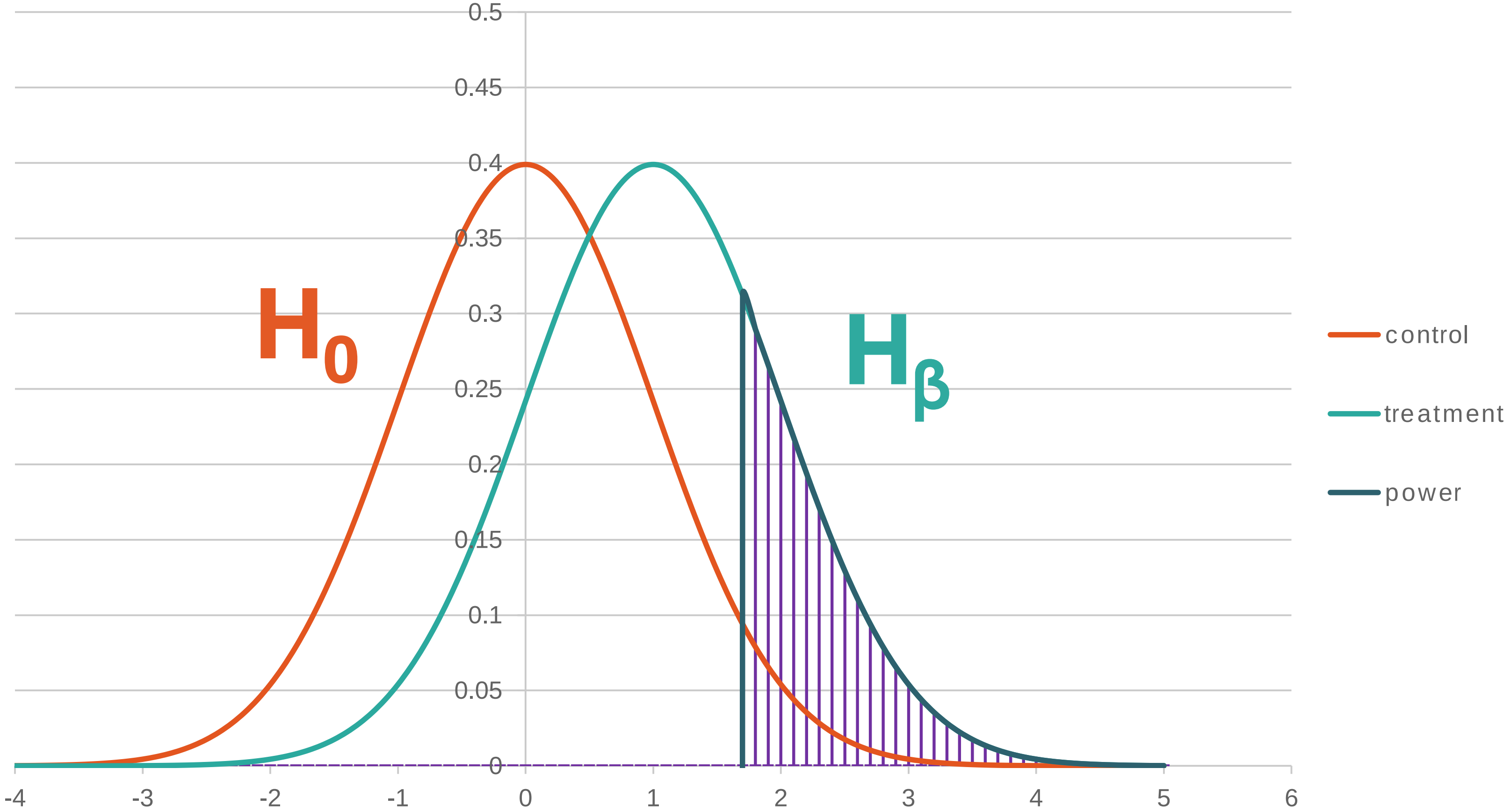


Let's say we believe the impact on our participants is "3"

Take up is 33%. Effect size is 1/3rd



Back to: Power = 26%

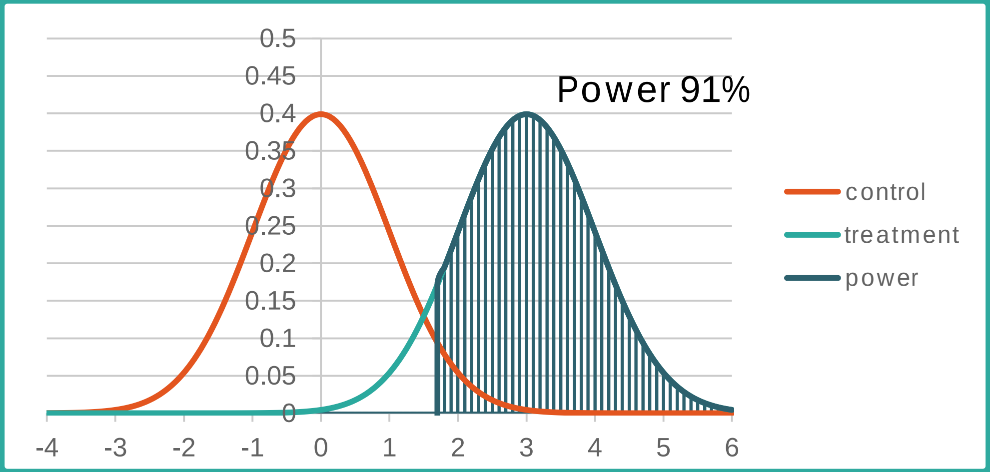


Take-up is reflected in the effect size

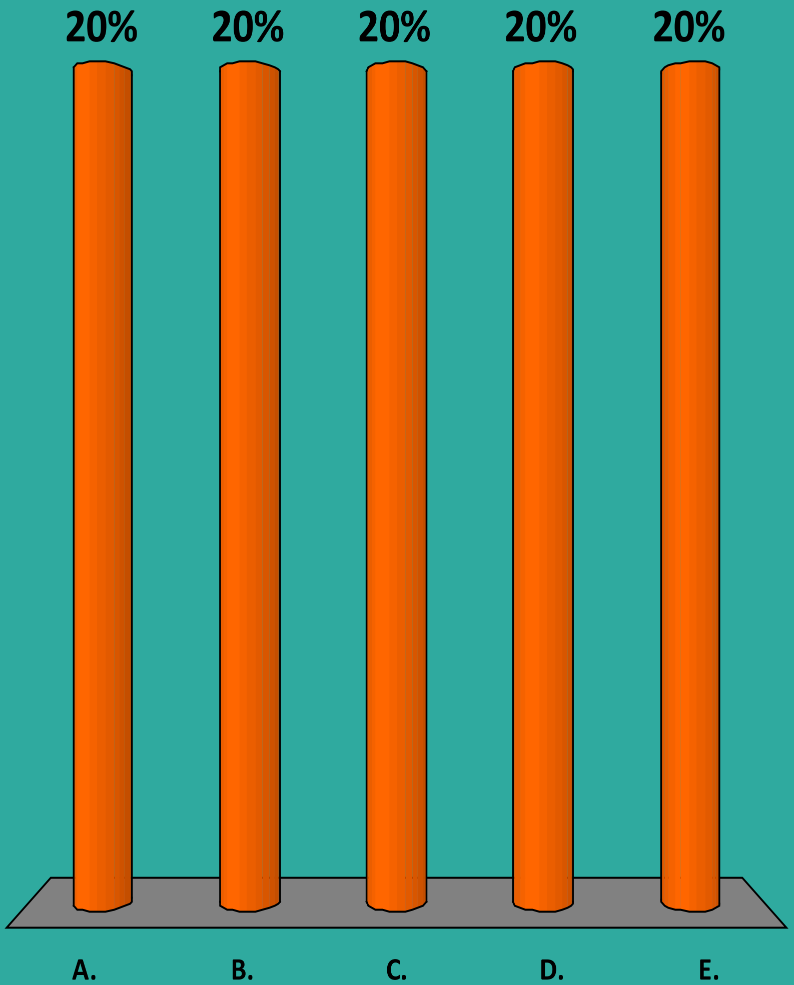
Power: main ingredients

1. Effect Size
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C
5. Clustering

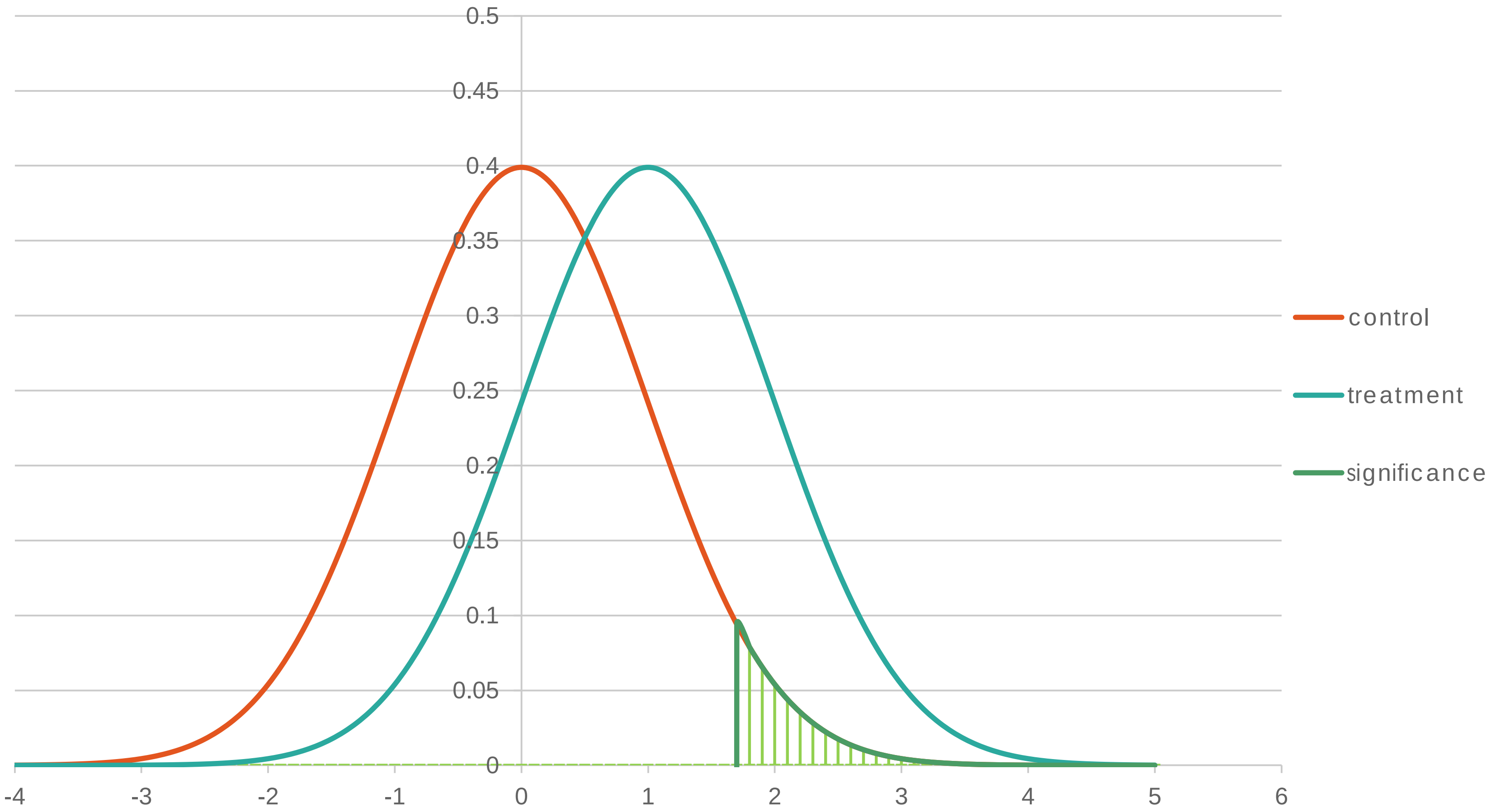
By increasing sample size you increase…



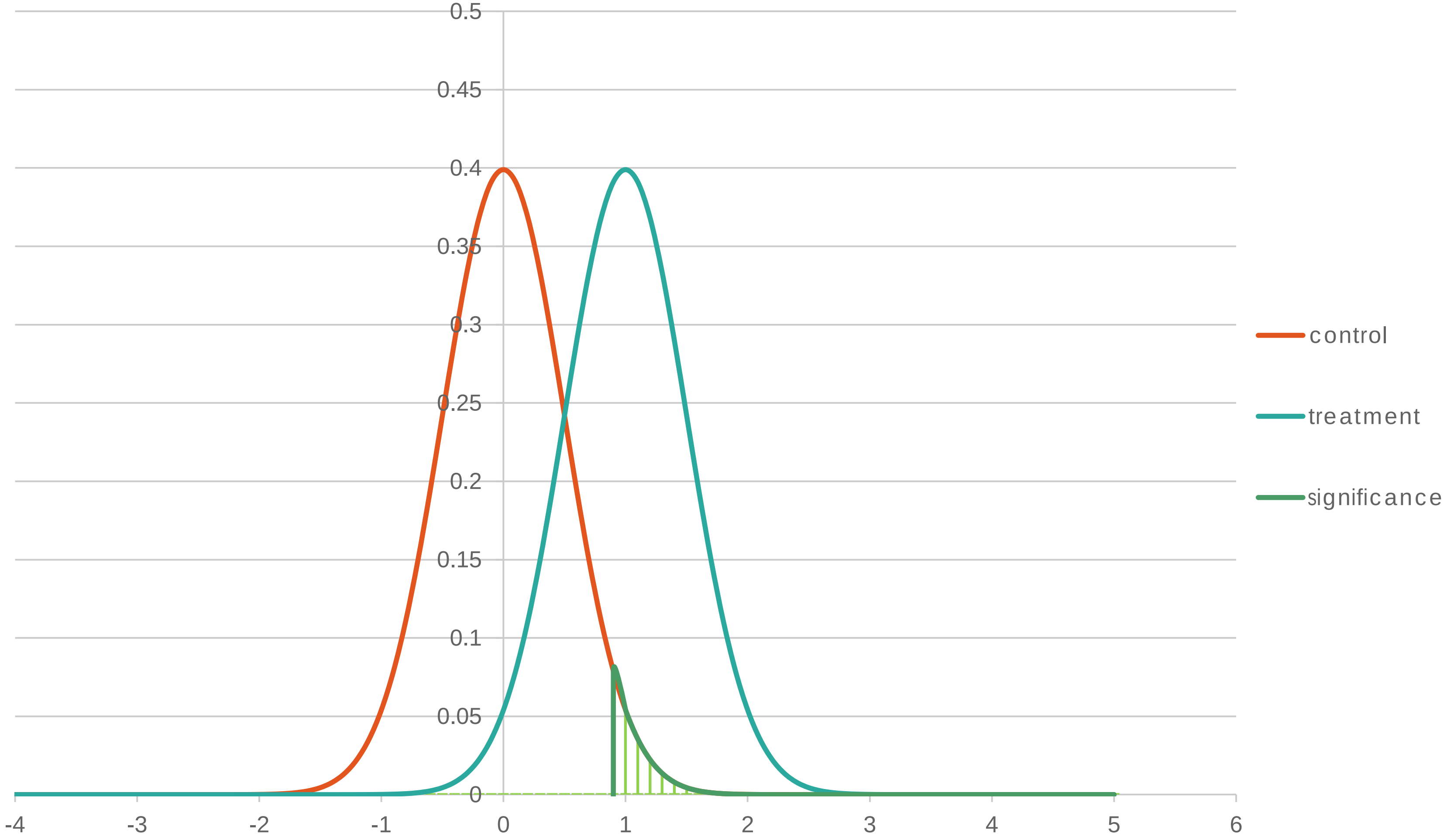
A. Accuracy
B. Precision
C. Both
D. Neither
E. Don’t know



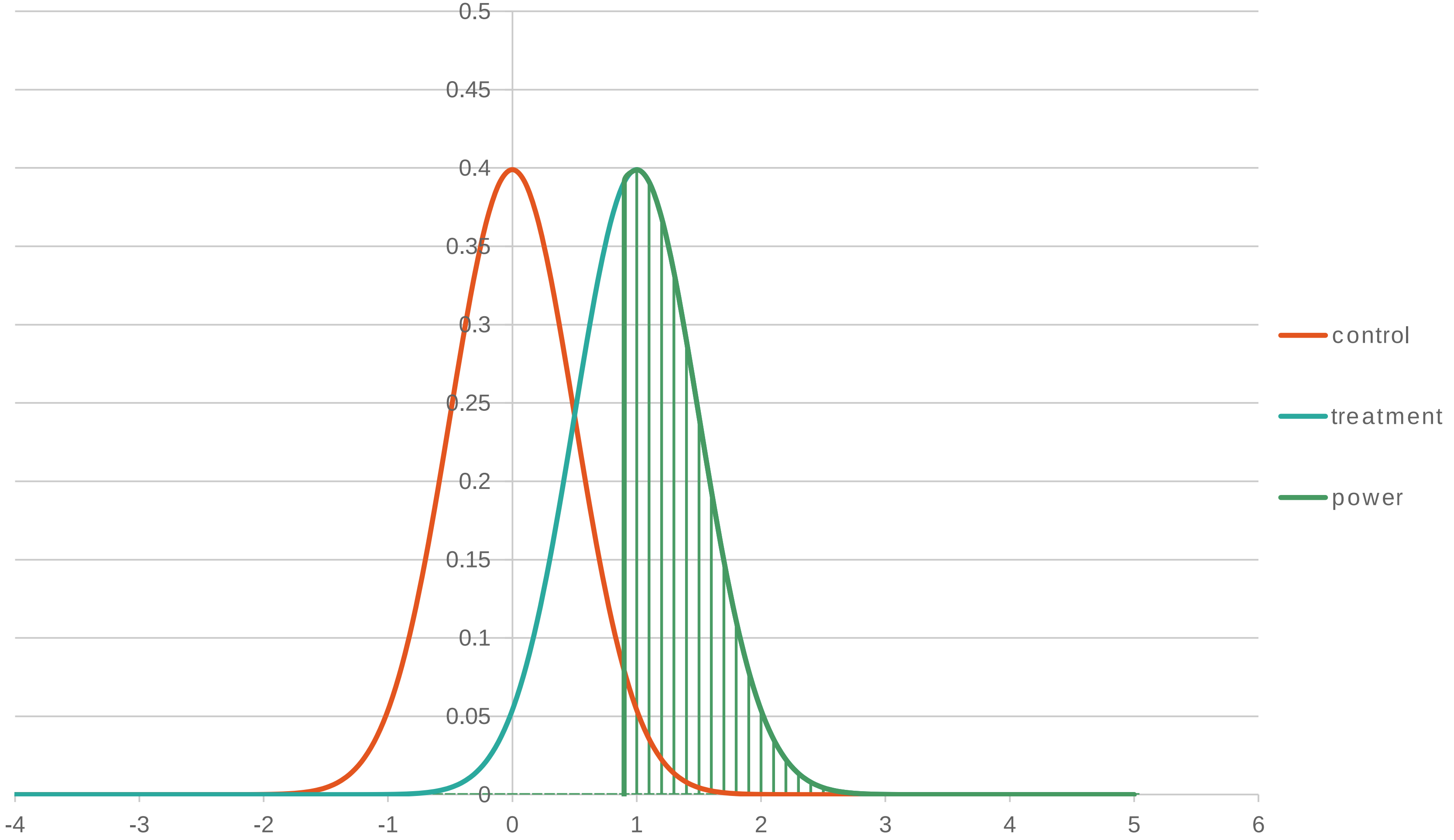
Power: Effect size = 1SD, Sample size = N



Power: Sample size = 4N

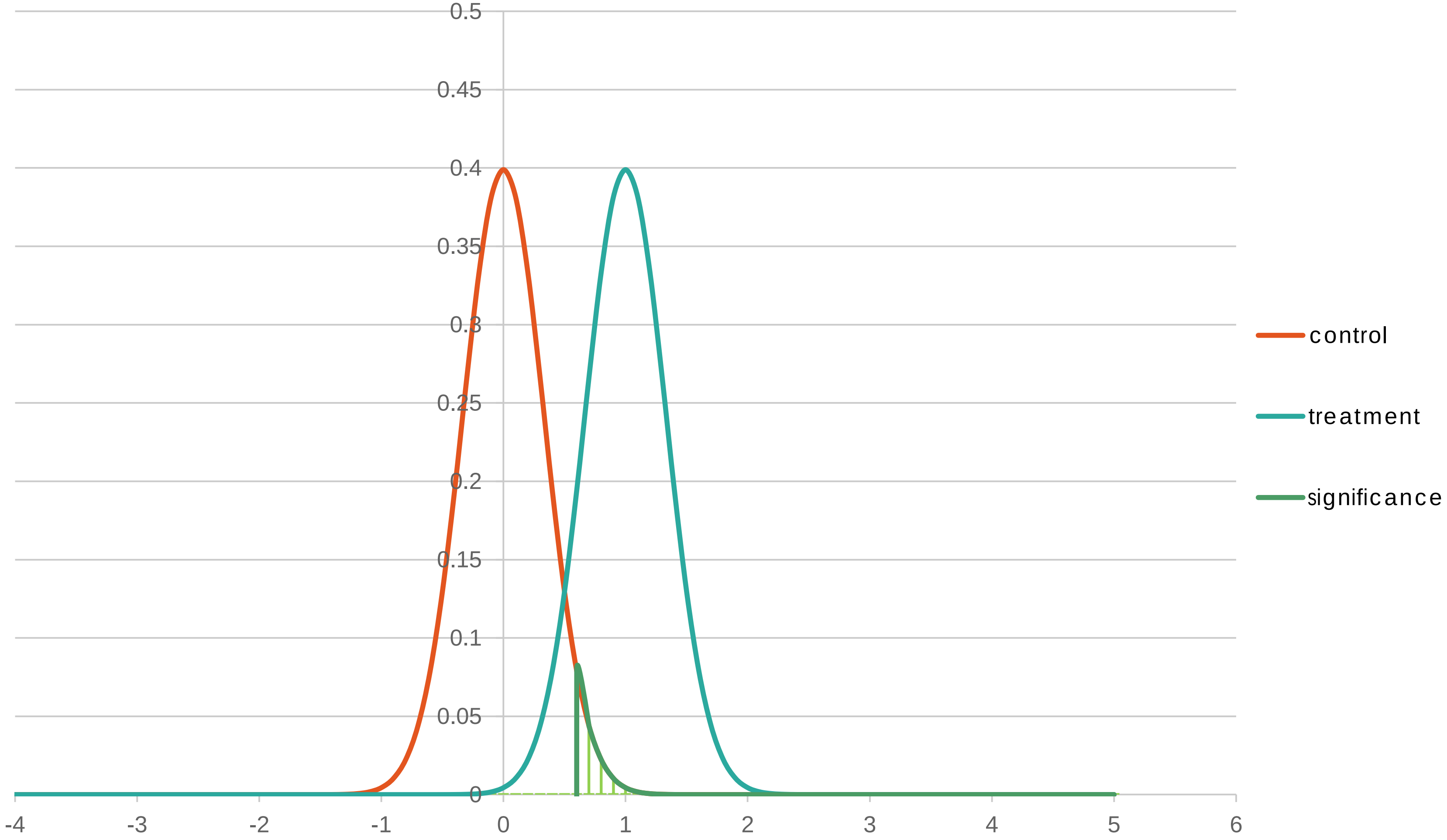


Power: 64%

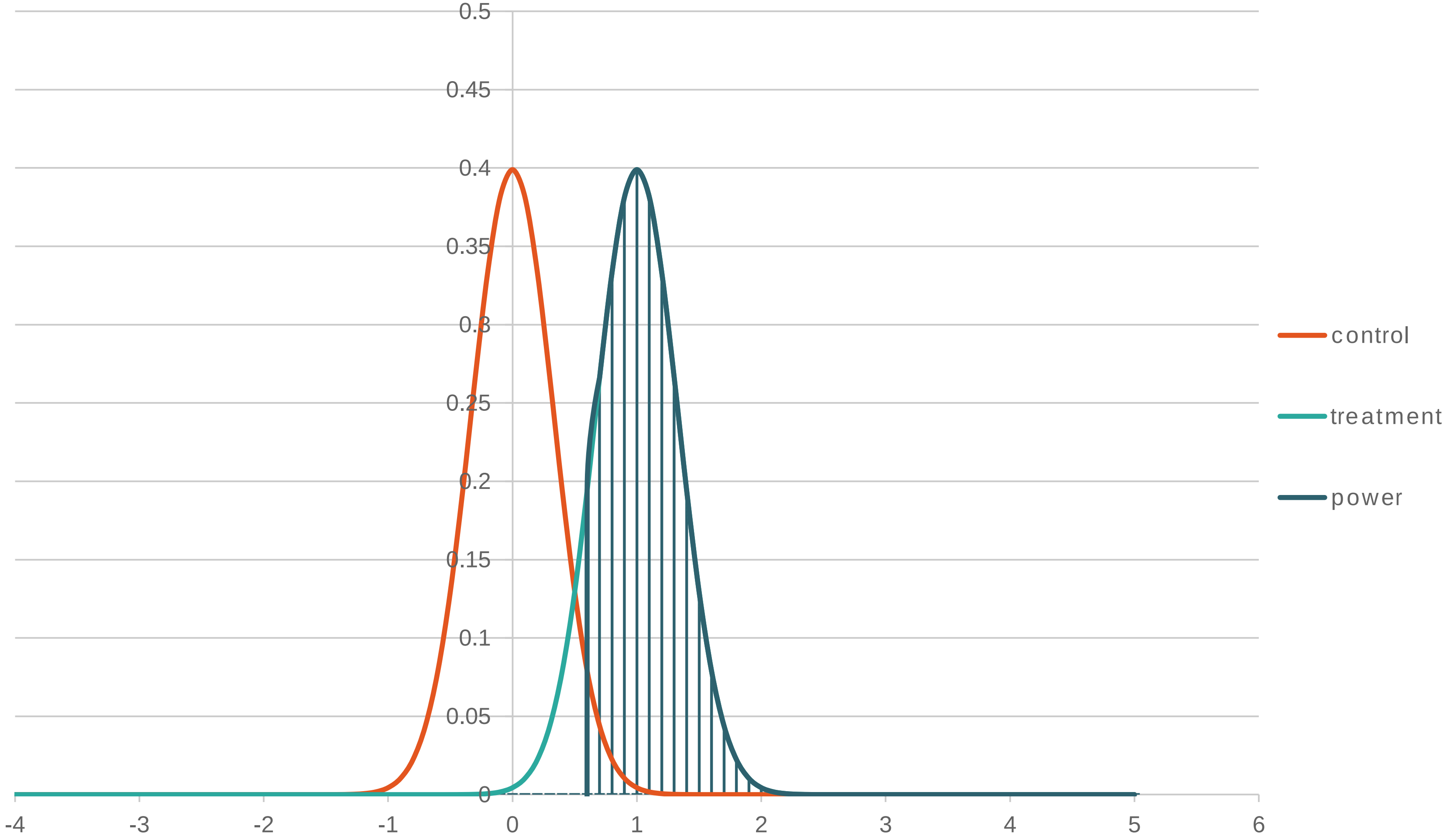


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6
control
treatment
power

Power: Sample size = 9N



Power: 91%

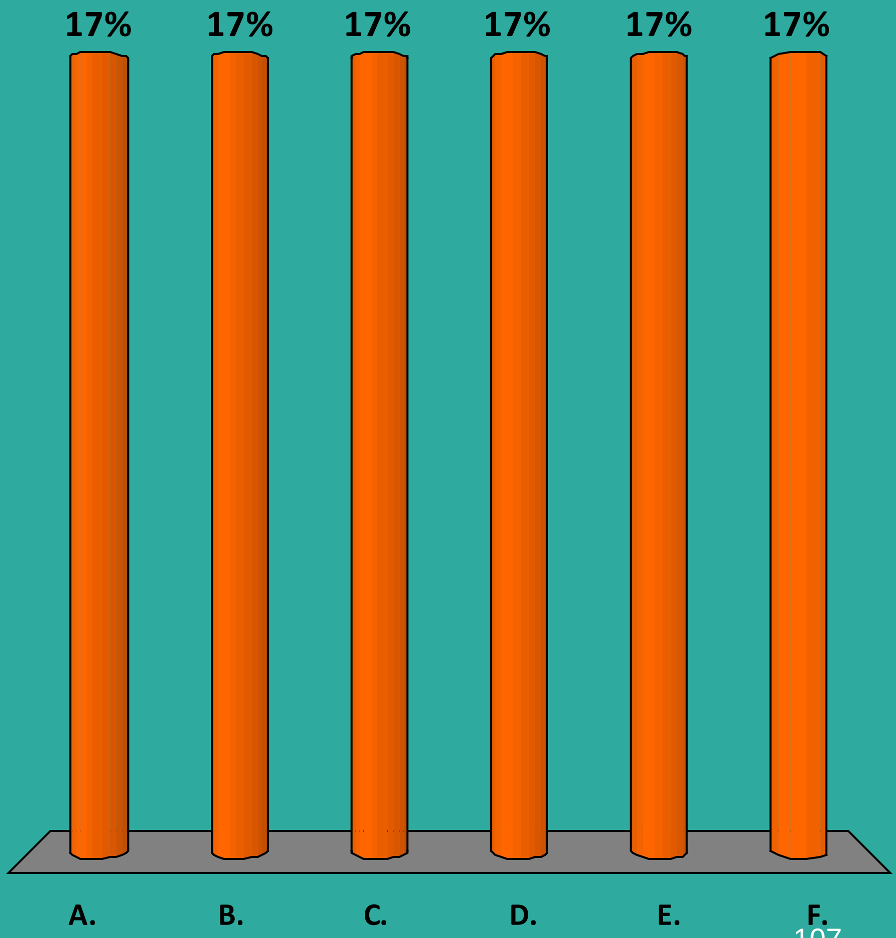

0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6
control
treatment
power

Power: main ingredients

1. Effect Size
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C
5. Clustering

What are typical ways to reduce the underlying (population) variance

A. Include covariates
B. Increase the sample
C. Do a baseline survey
D. All of the above
E. A and B
F. A and C



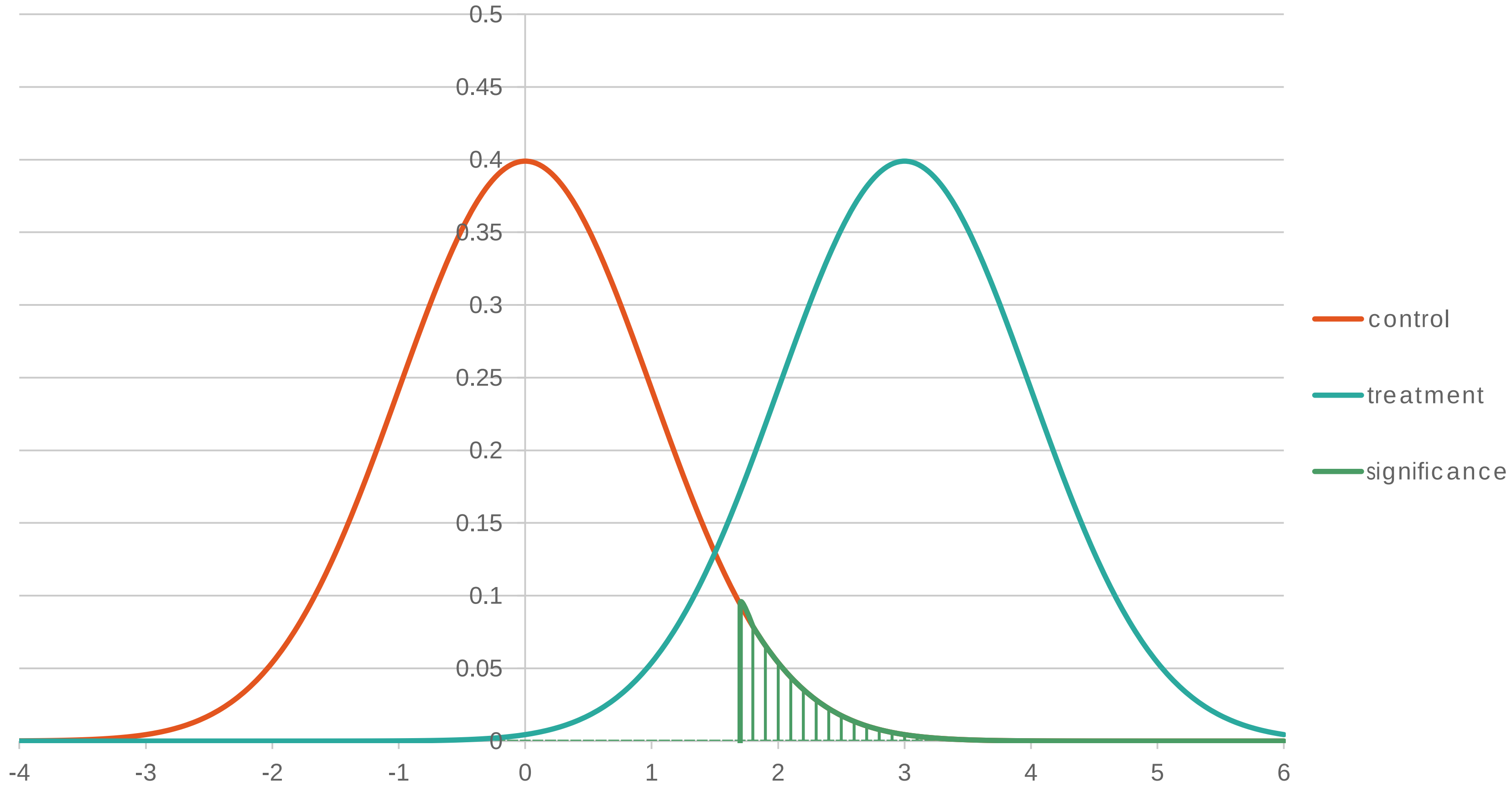
Variance

- There is sometimes very little we can do to reduce the noise
- The underlying variance is what it is
- We can try to “absorb” variance:
  - using a baseline
  - controlling for other variables
    - In practice, controlling for other variables (besides the baseline outcome) buys you very little

Power: main ingredients

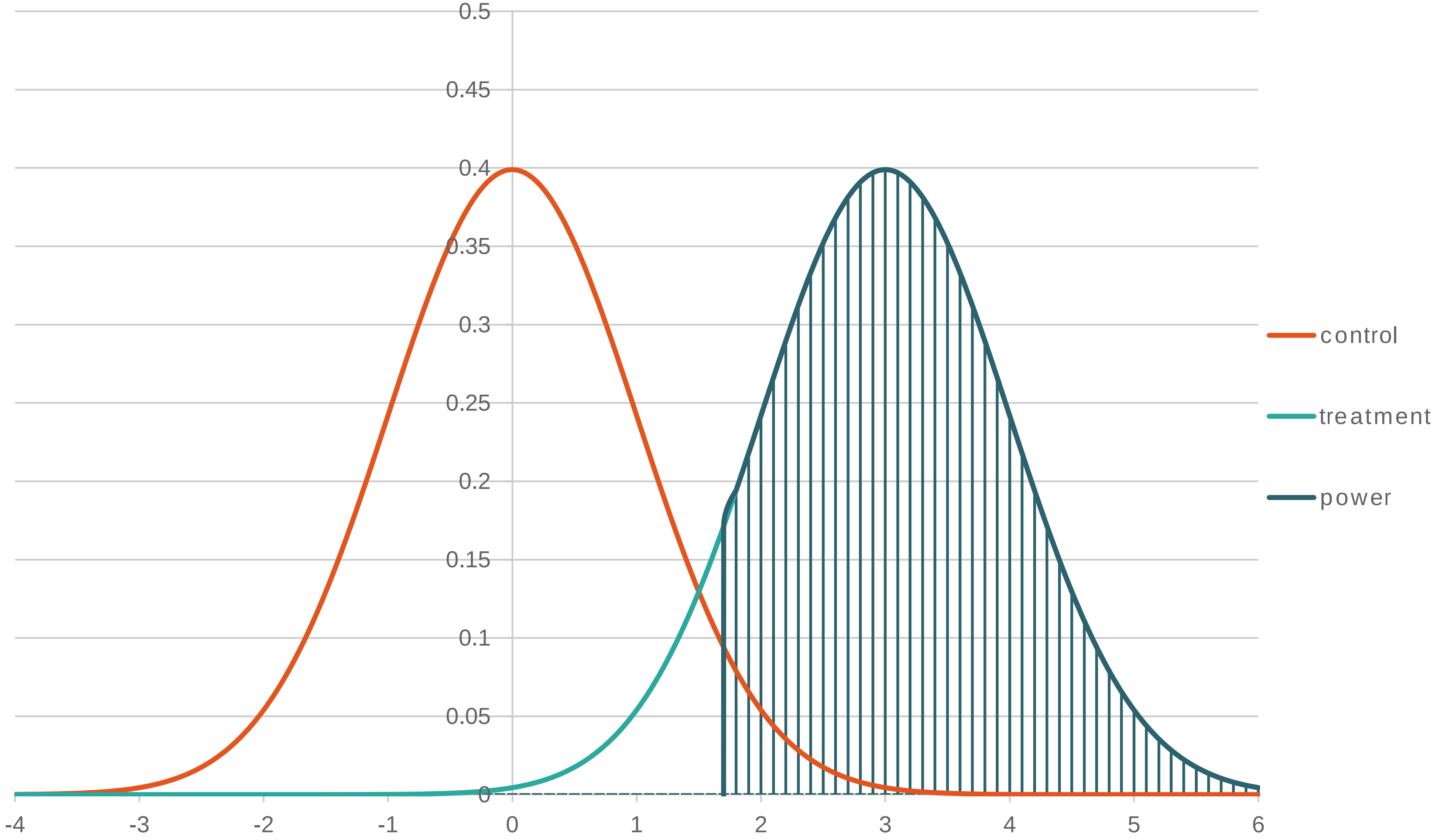
1. Effect Size
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C
5. Clustering

Sample split: 50% C, 50% T



Equal split gives distributions that are the same “fatness”

Power: 91%

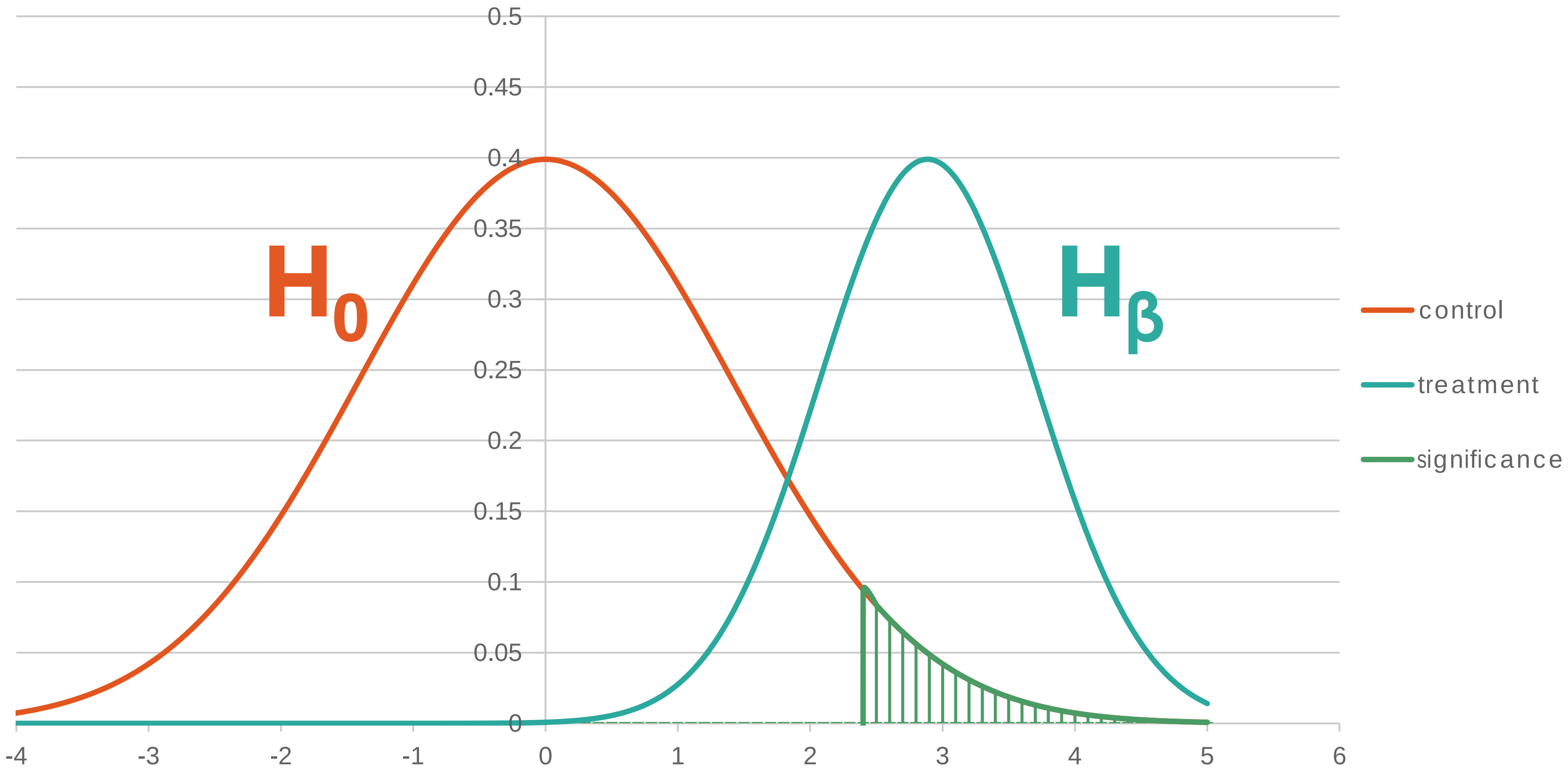


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6
control
treatment
power

If it’s not 50-50 split?

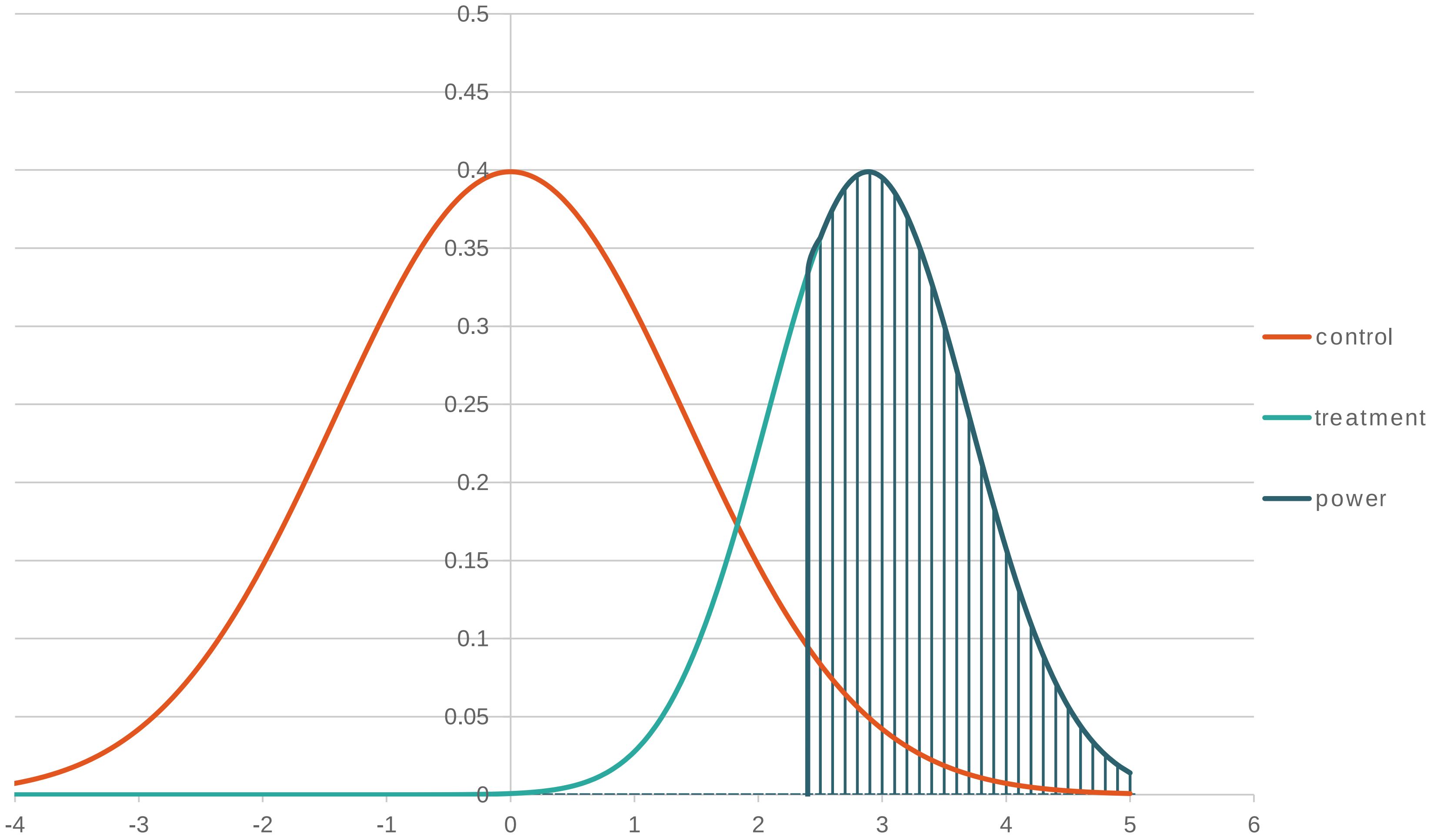
- What happens to the relative fatness if the split is not 50-50.
- Say 25-75?

Sample split: 25% C, 75% T



Uneven distributions, not efficient, i.e. less power

Power: 83%



0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6
control
treatment
power

Allocation to T v C

$$sd(X_1 - X_2) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$sd(X_1 - X_2) = \sqrt{\frac{1}{2} + \frac{1}{2}} = \sqrt{\frac{2}{2}} = 1$$

$$sd(X_1 - X_2) = \sqrt{\frac{1}{3} + \frac{1}{1}} = \sqrt{\frac{4}{3}} = 1.15$$

Power: main ingredients

1. Effect Size
2. Sample Size
3. Variance
4. Proportion of sample in T vs. C
5. Clustering

Clustered design: definition

- In sampling:
  - When clusters of individuals (e.g. schools, communities, etc.) are randomly selected from the population, before selecting individuals for observation
- In randomized evaluation:
  - When clusters of individuals are randomly assigned to different treatment groups

Clustered design: intuition

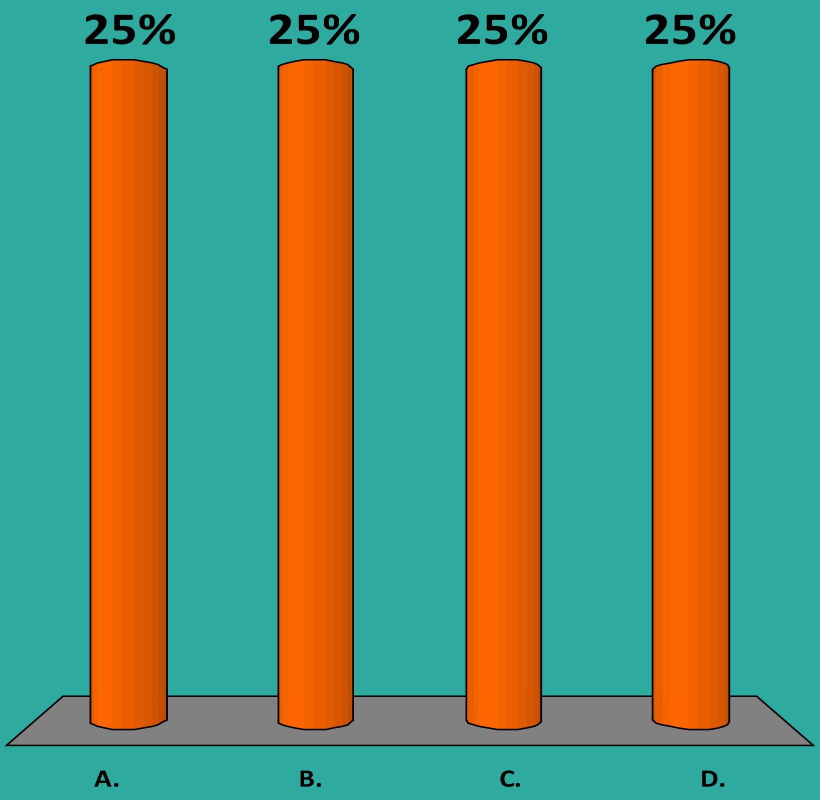
- You want to know how close the upcoming national elections will be
- Method 1: Randomly select 50 people from entire Indian population
- Method 2: Randomly select 5 families, and ask ten members of each family their opinion

Low intra-cluster correlation (ICC) aka ρ (rho)

HIGH intra-cluster correlation (ρ)

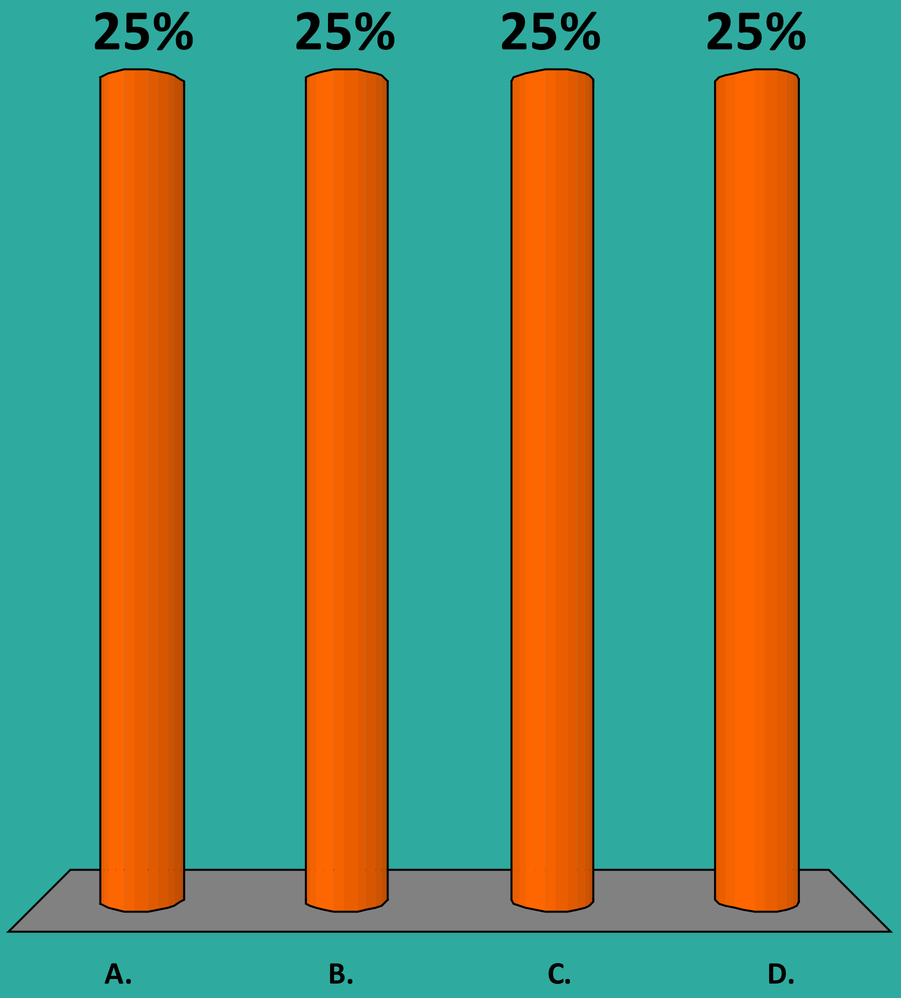
All uneducated people live in one village. People with only primary education live in another. College grads live in a third, etc. ICC (ρ) on education will be..

A. High
B. Low
C. No effect on rho
D. Don't know



If ICC (ρ) is high, what is a more efficient way of increasing power?

A. Include more clusters in the sample
B. Include more people in clusters
C. Both
D. Don't know



Testing multiple treatments

50 ←0.15 SD→ 150 / 250

Control Group **Balsakhi**

↑ ↖ ↗ ↑

0.15 SD 0.05 SD / 0.25 SD 0.10 SD

↓ ↙ ↘ ↓

500 / 100 / 200 ←0.10 SD→ 100

CAL program Balsakhi + CAL

END!