ABDUL LATIF JAMEEL
Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION

Sampling and Sample Size

Rohit Naimpally
J-PAL

Course Overview

1. What is Evaluation?
2. Outcomes, Impact, and Indicators
3. Why Randomize and Common Critiques
4. How to Randomize
5. **Sampling and Sample Size**
6. Threats and Analysis
7. Project from Start to Finish
8. Cost-Effectiveness Analysis and Scaling Up

Framing the discussion…

"Trevor was a painter. Indeed, few people escape that nowadays. But he was also an artist, and artists are rather rare."

- Oscar Wilde

"Power is as much an art as a science."

- Unknown (probably not Oscar Wilde)

Learning Objectives

At the end of the presentation, you will:

1. Know the **Central Limit Theorem** and the **Law of Large Numbers**, and why they matter.

2. Know the difference between a ***Type I*** and a ***Type II*** error.

3. Know what the "**power**" of a study is and why you should care.

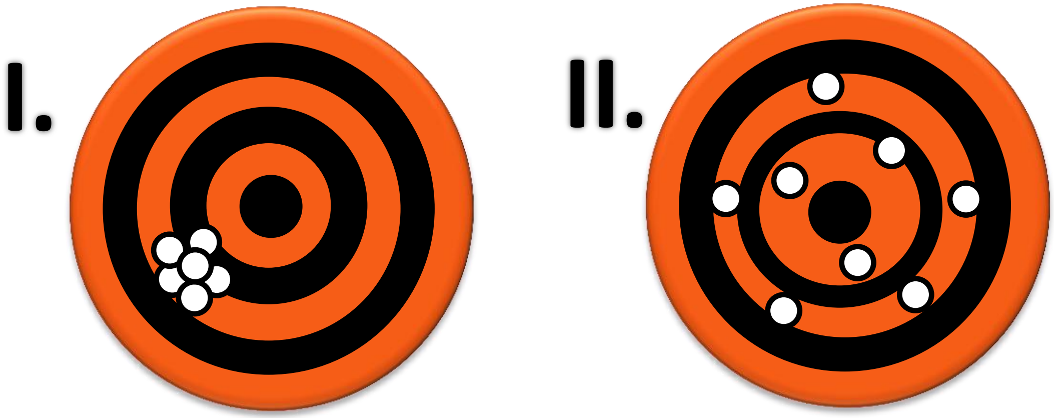
4. Be ready to tackle the power exercise in the next session!

<u>THE</u> basic questions in statistics

- How confident can you be in your results?
  - This is given by the **significance** level of your results (remember the "asterisks"?)

- How big does your sample need to be?
  - This is given by the **power** of your design.

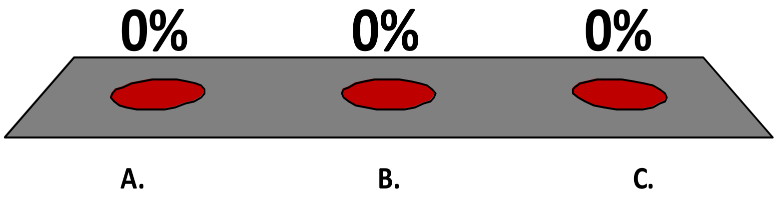
Recap: Which of these is more accurate?



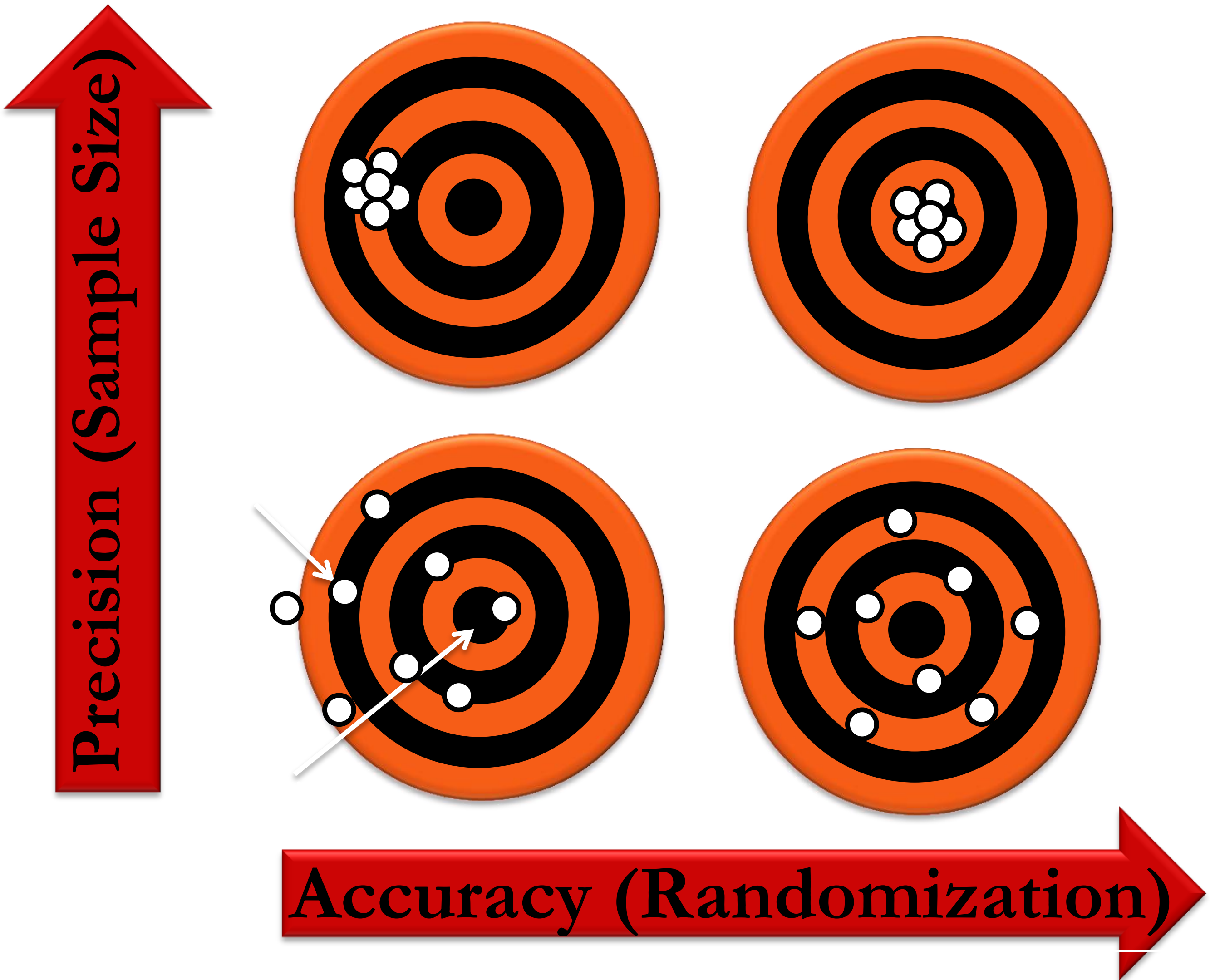
A. I.

B. II.

C. Don't know

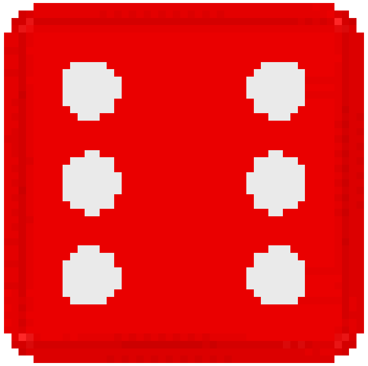


Recap: Accuracy versus Precision

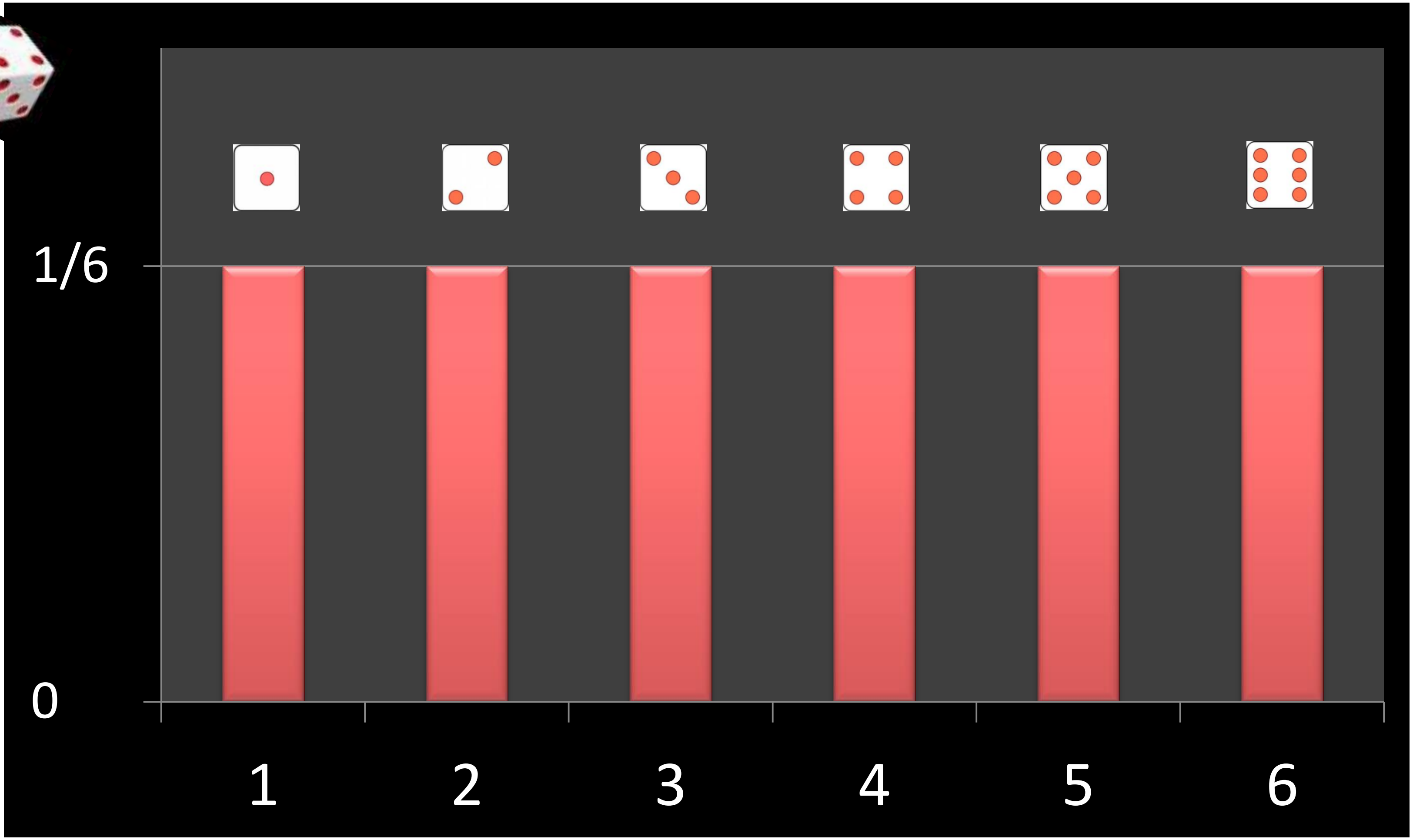


What's the average result?

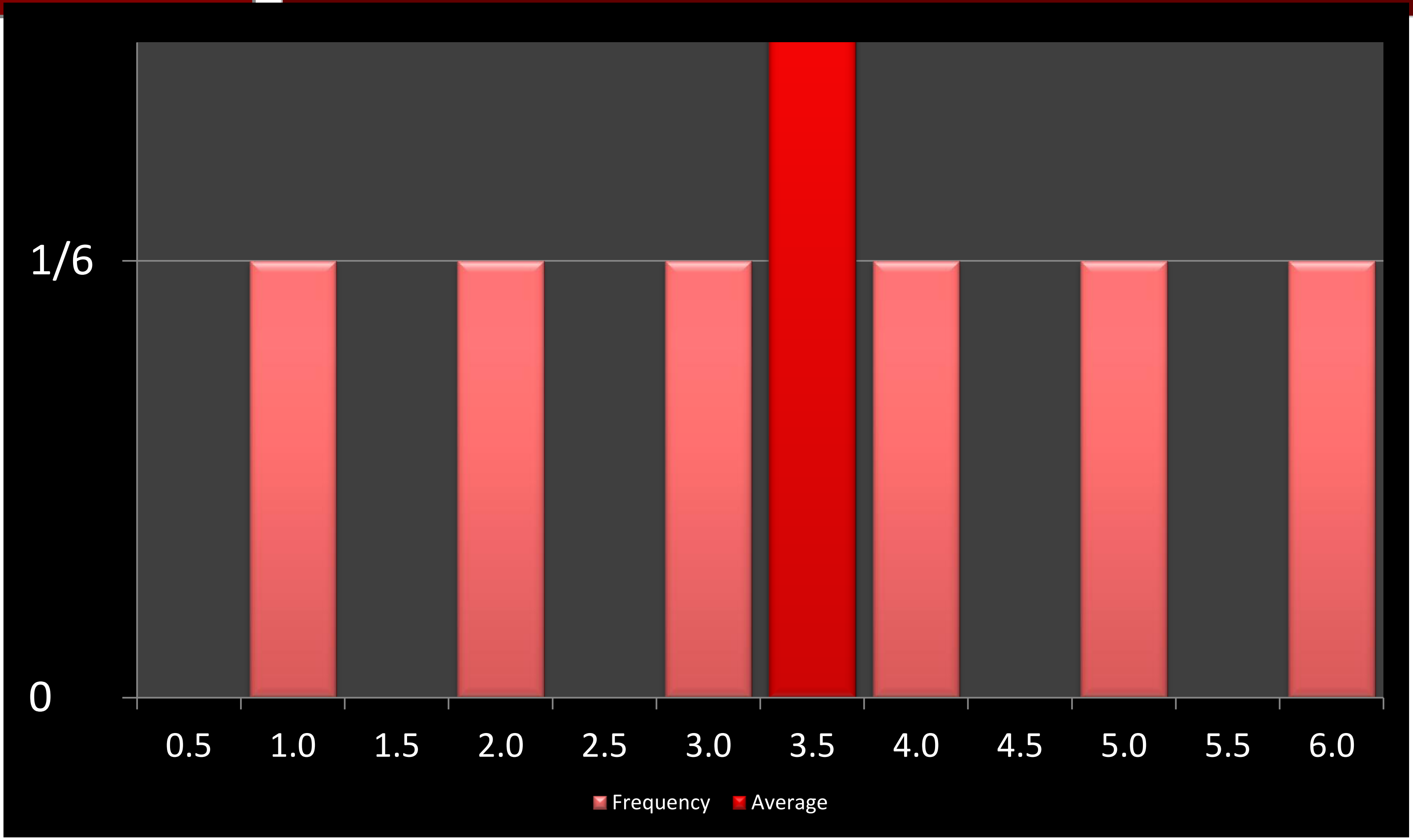
- If you were to roll a die once, what is the "expected result"? (i.e. the average)



Possible results & probability: 1 die

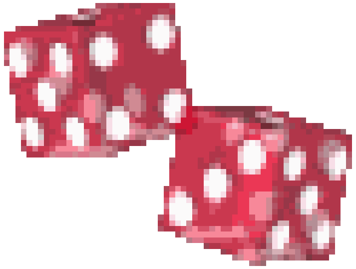


Rolling 1 die: possible results & average

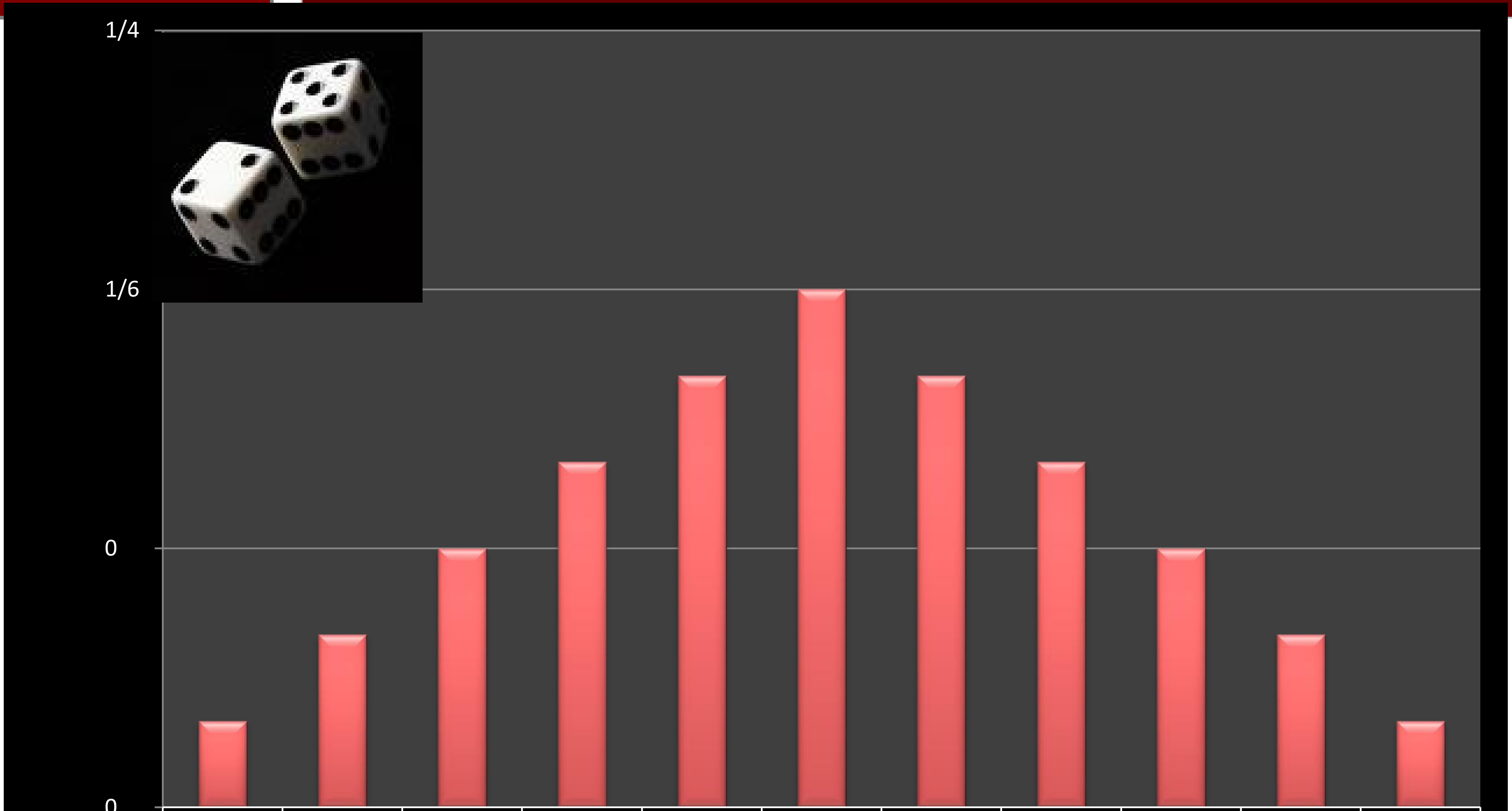


What's the average result?

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



Rolling 2 dice:
Possible totals & likelihood

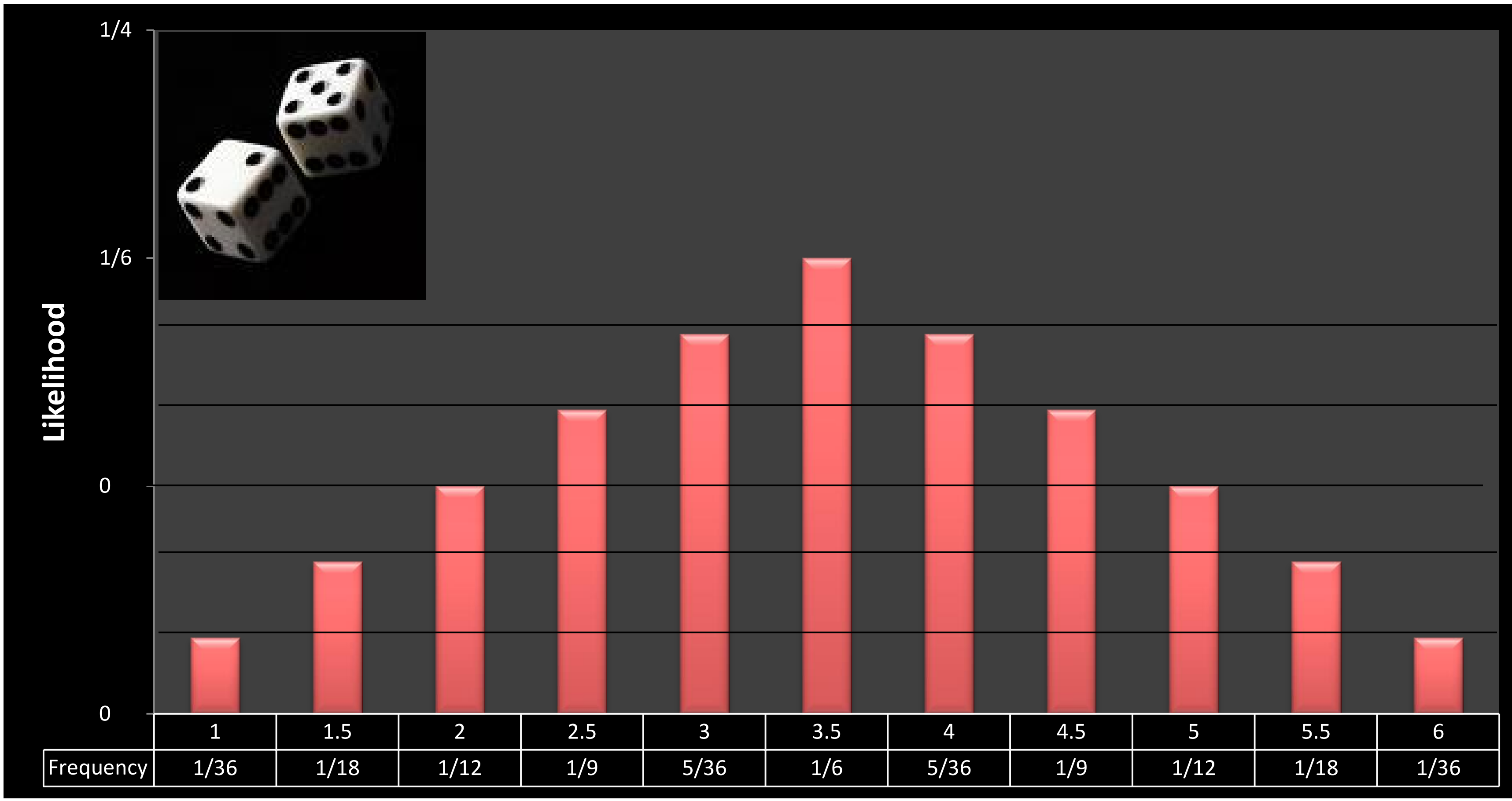


	2	3	4	5	6	7	8	9	10	11	12
Frequency	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

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

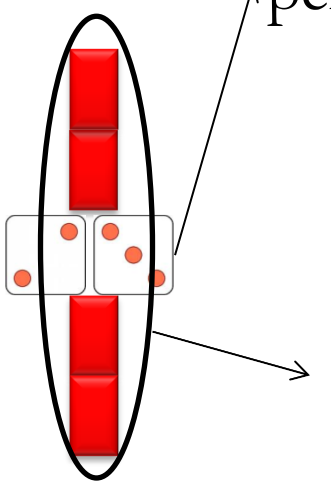


Outcomes and Permutations

Putting together permutations, you get:

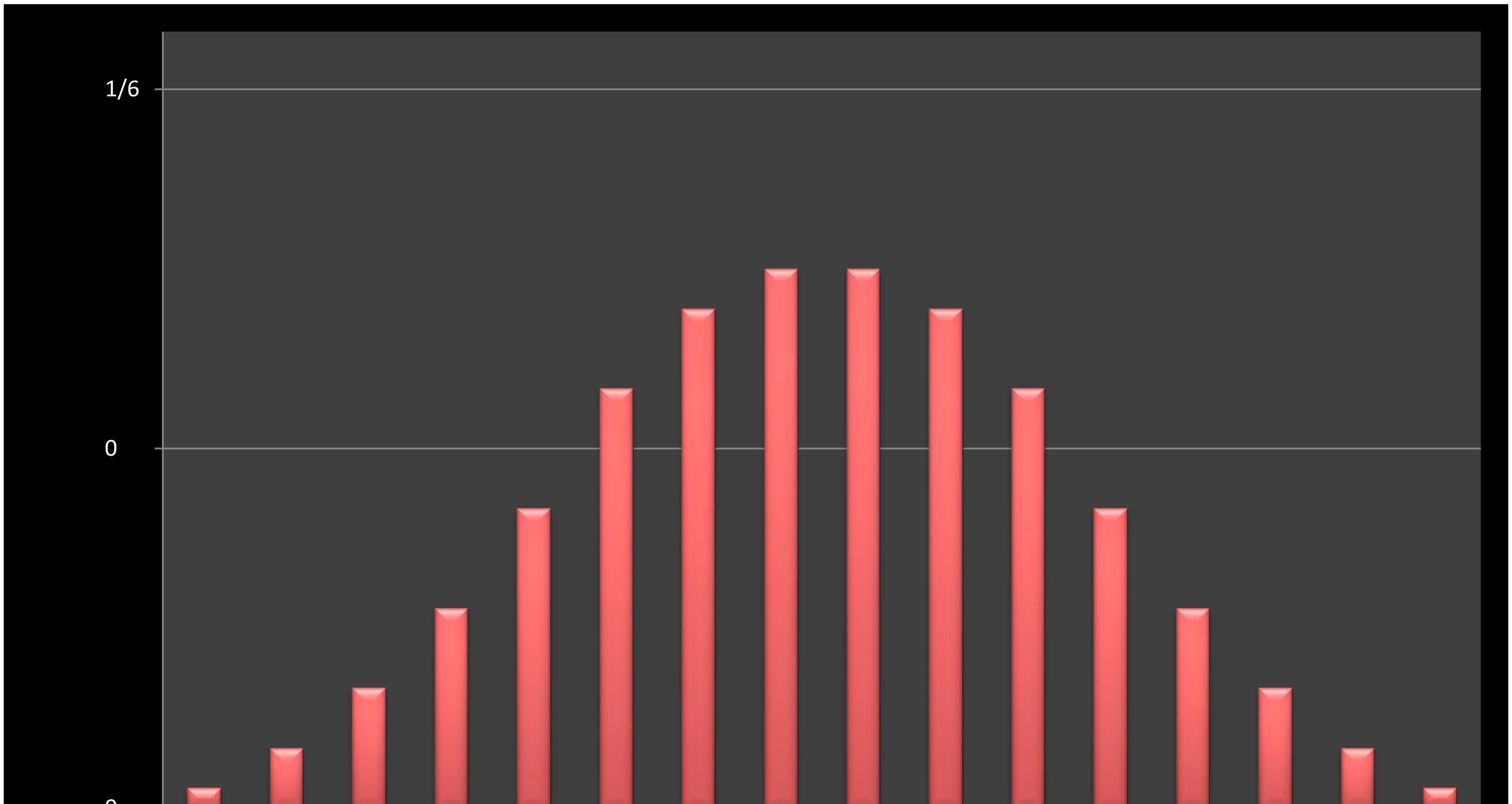
1. All possible outcomes
2. The likelihood of each of those outcomes

Each block within a column represents one possible permutation (to obtain that average)



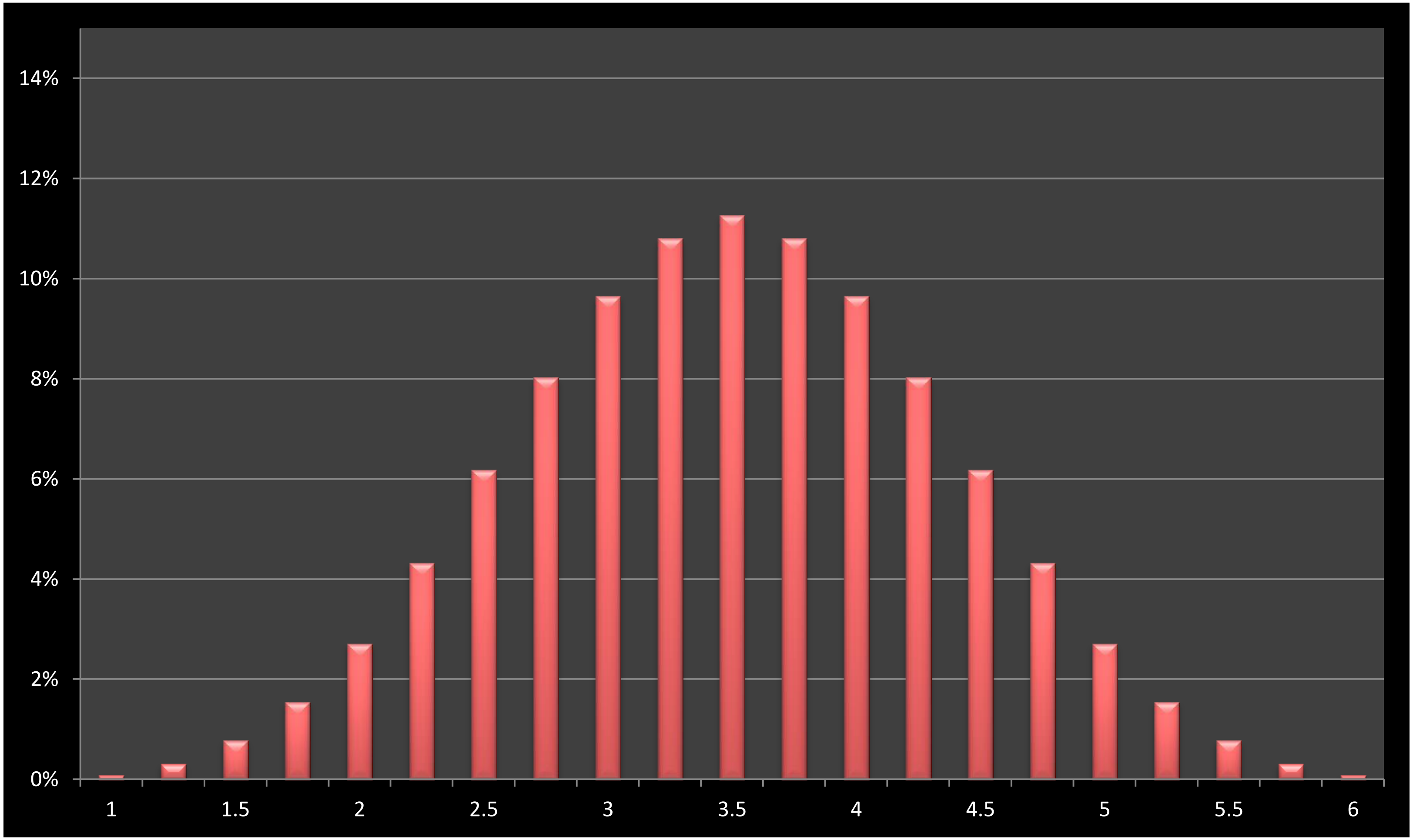
Each column represents one possible outcome (average result)

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

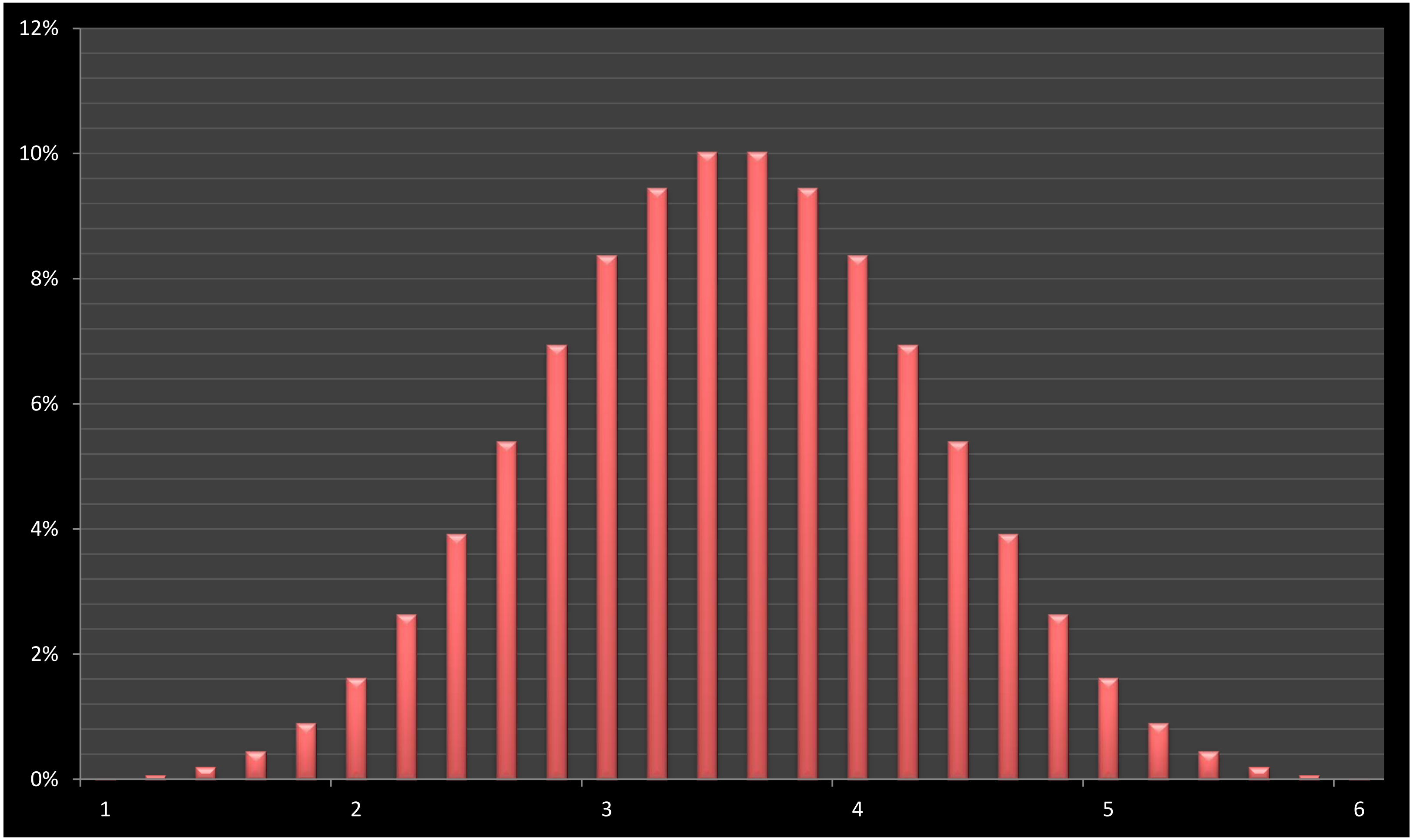


	1	1 1/3	1 2/3	2	2 1/3	2 2/3	3	3 1/3	3 2/3	4	4 1/3	4 2/3	5	5 1/3	5 2/3	6
Frequency	0%	1%	3%	5%	7%	10%	12%	13%	13%	12%	10%	7%	5%	3%	1%	0%

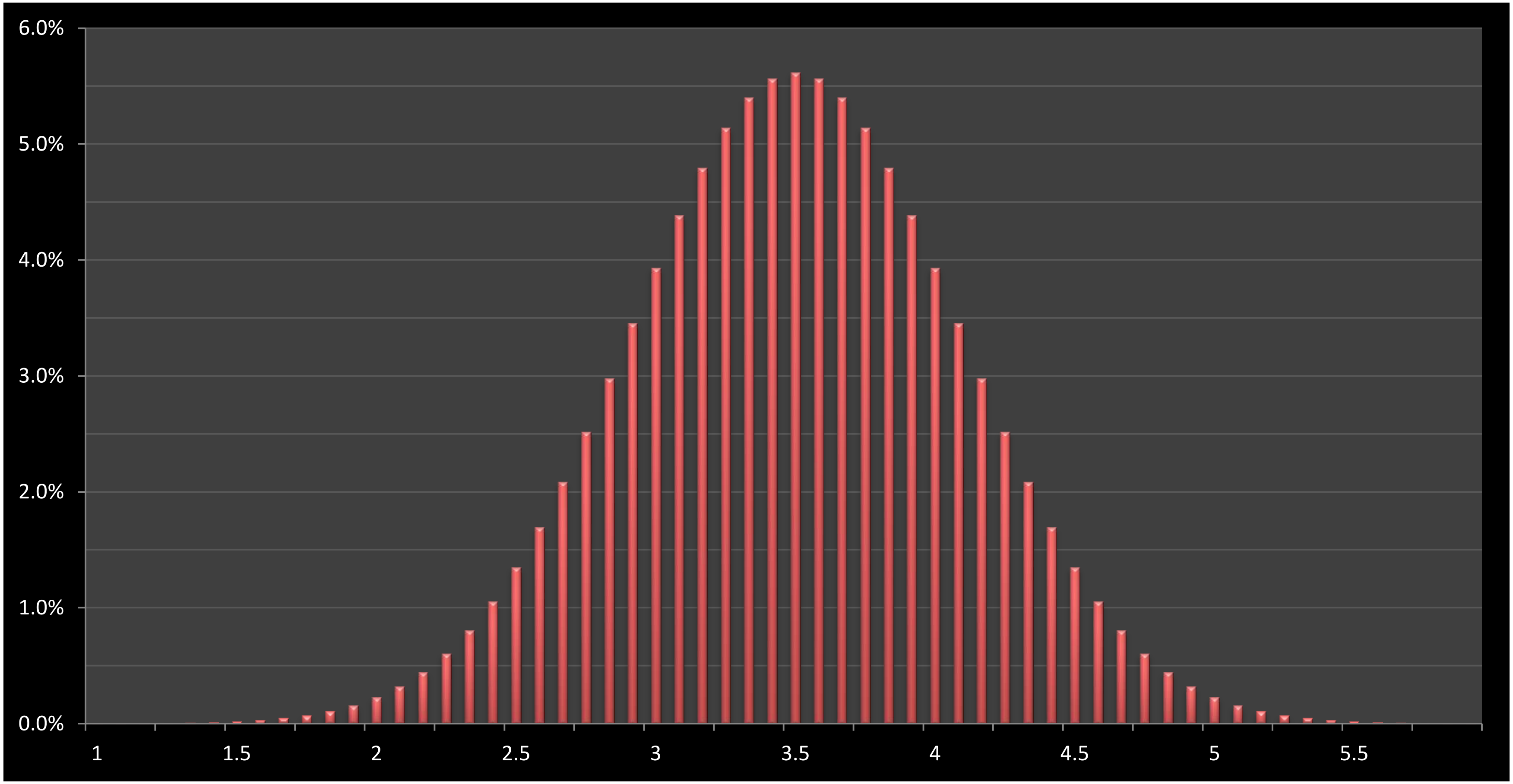
Rolling 4 dice:
21 results, 1296 permutations



Rolling 5 dice:
26 results, 7776 permutations

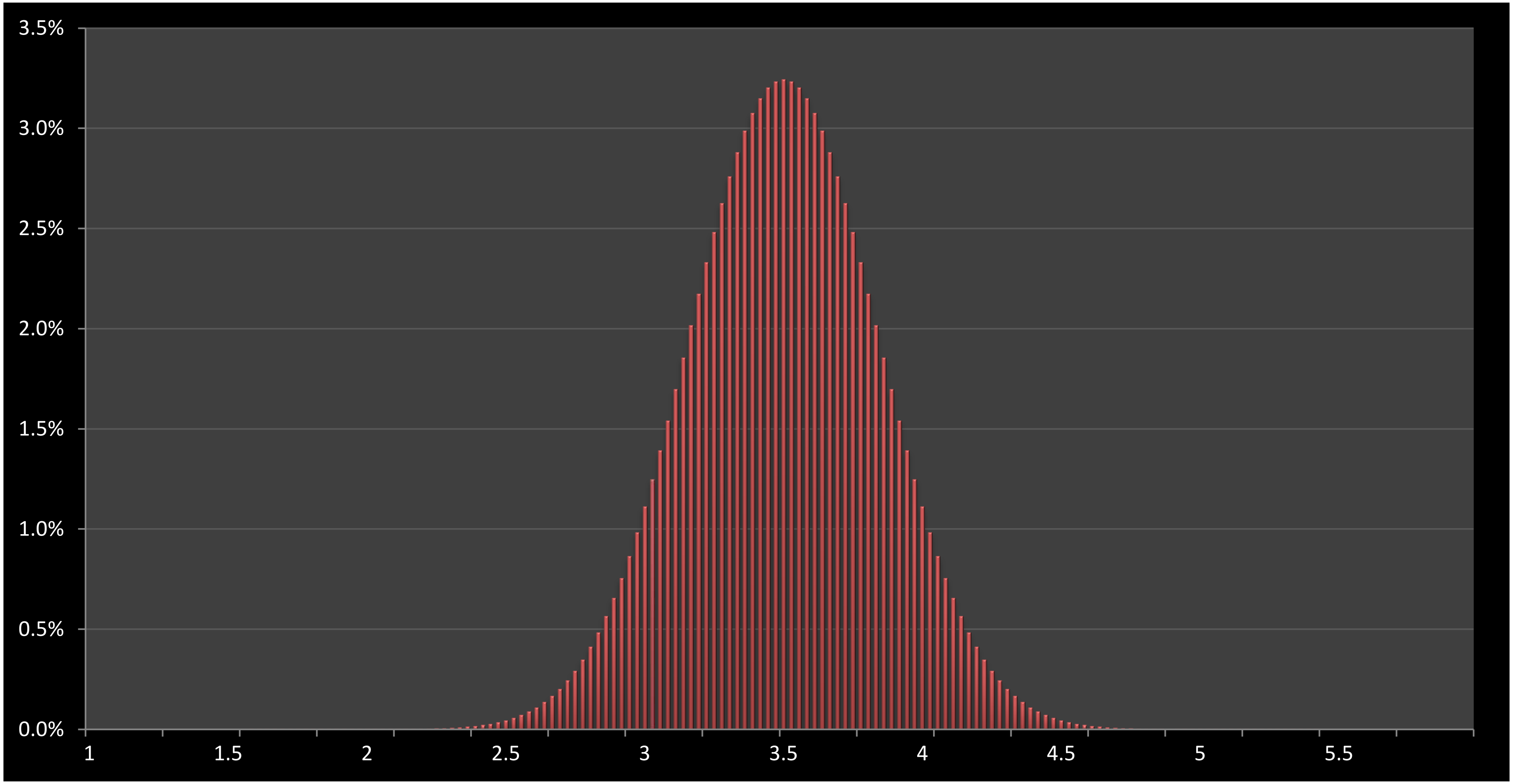


Rolling 10 dice:
50 results, >60 million permutations



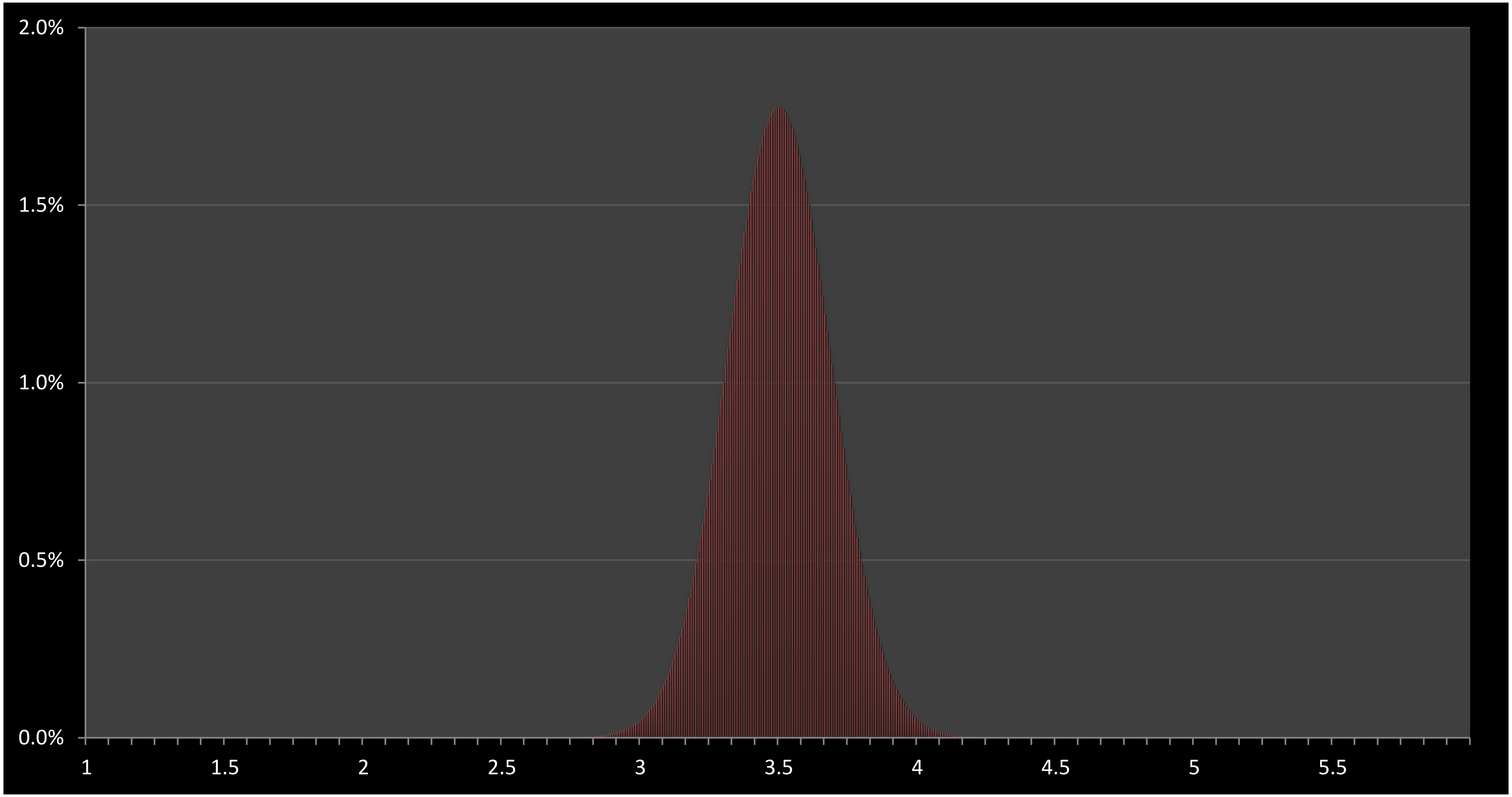
Looks like a bell curve, or a *normal* distribution

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

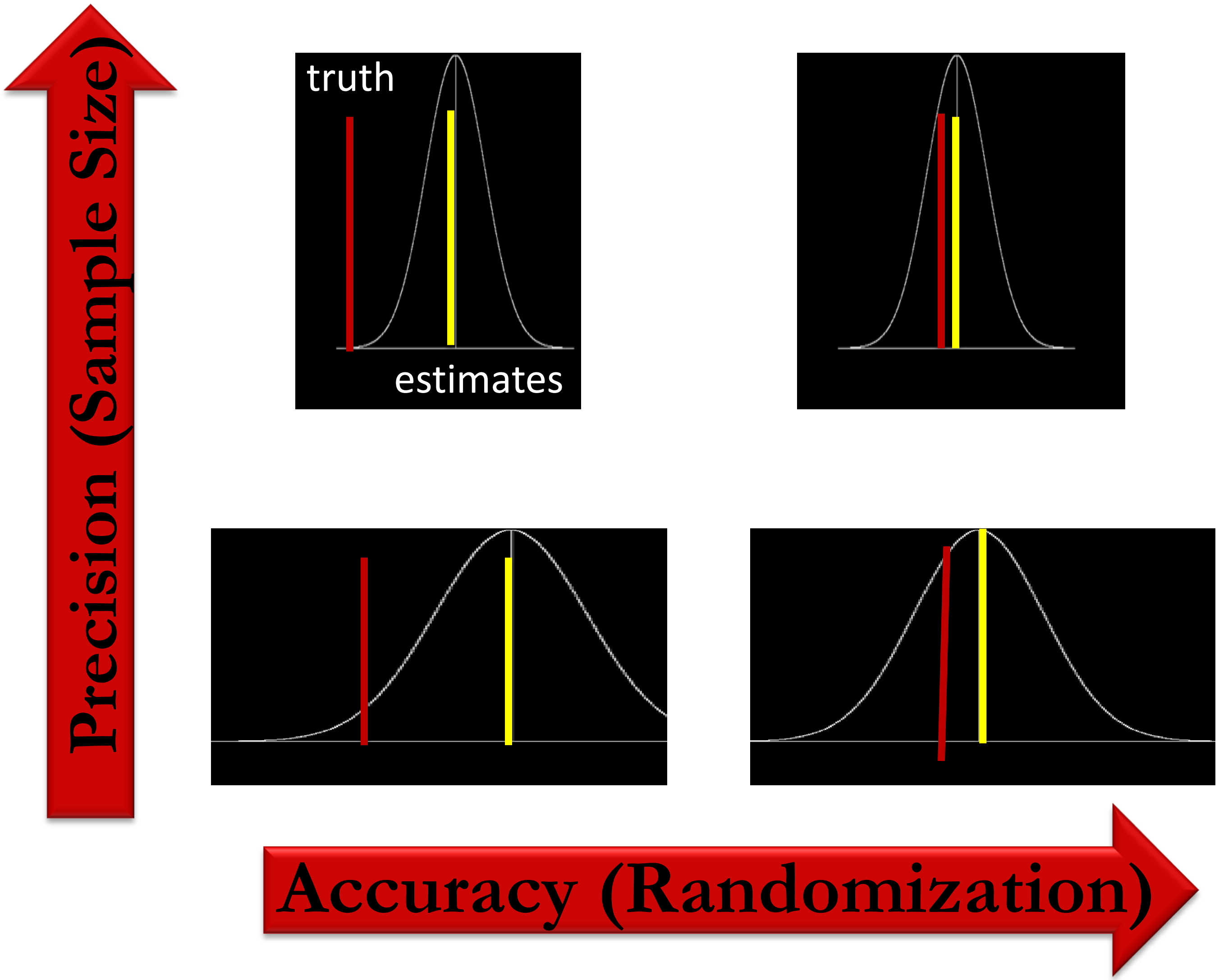
1. The more dice you roll, the closer most averages are to the <u>true</u> average (the distribution gets "tighter")

-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-

Accuracy versus Precision



THAT WAS JUST THE INTRODUCTION

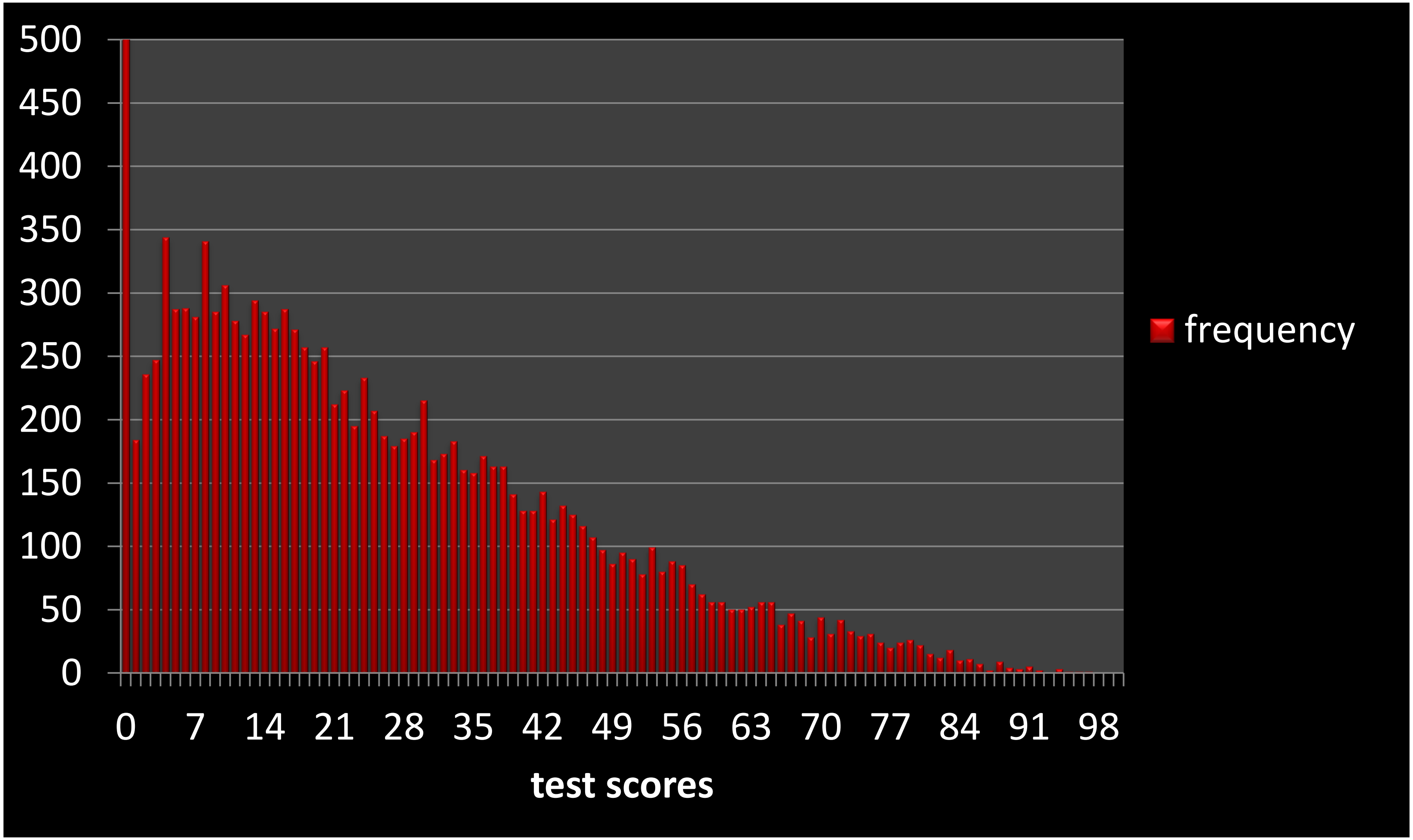
Outline

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

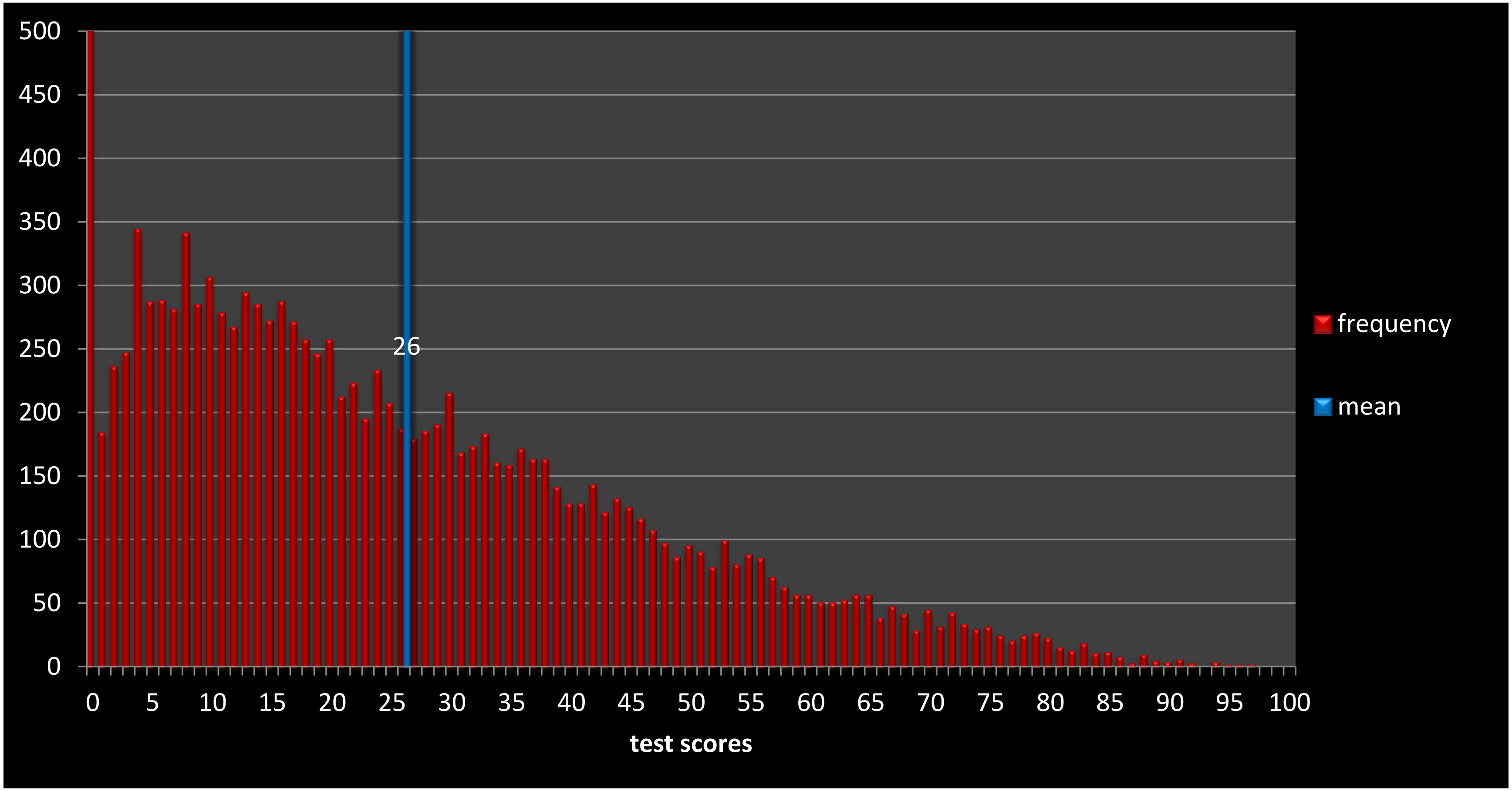
Outline

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

Baseline test scores

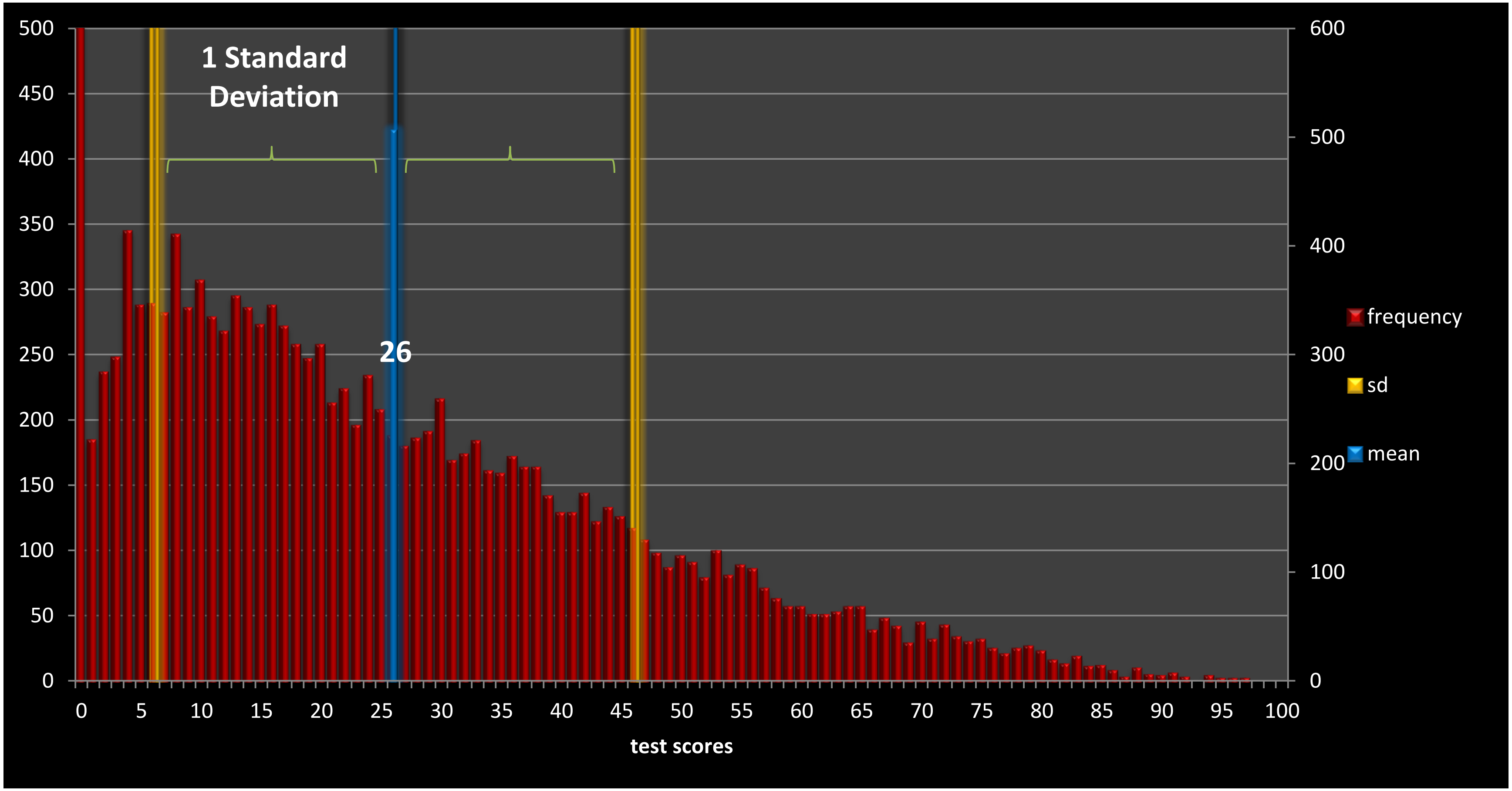


Mean = 26



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

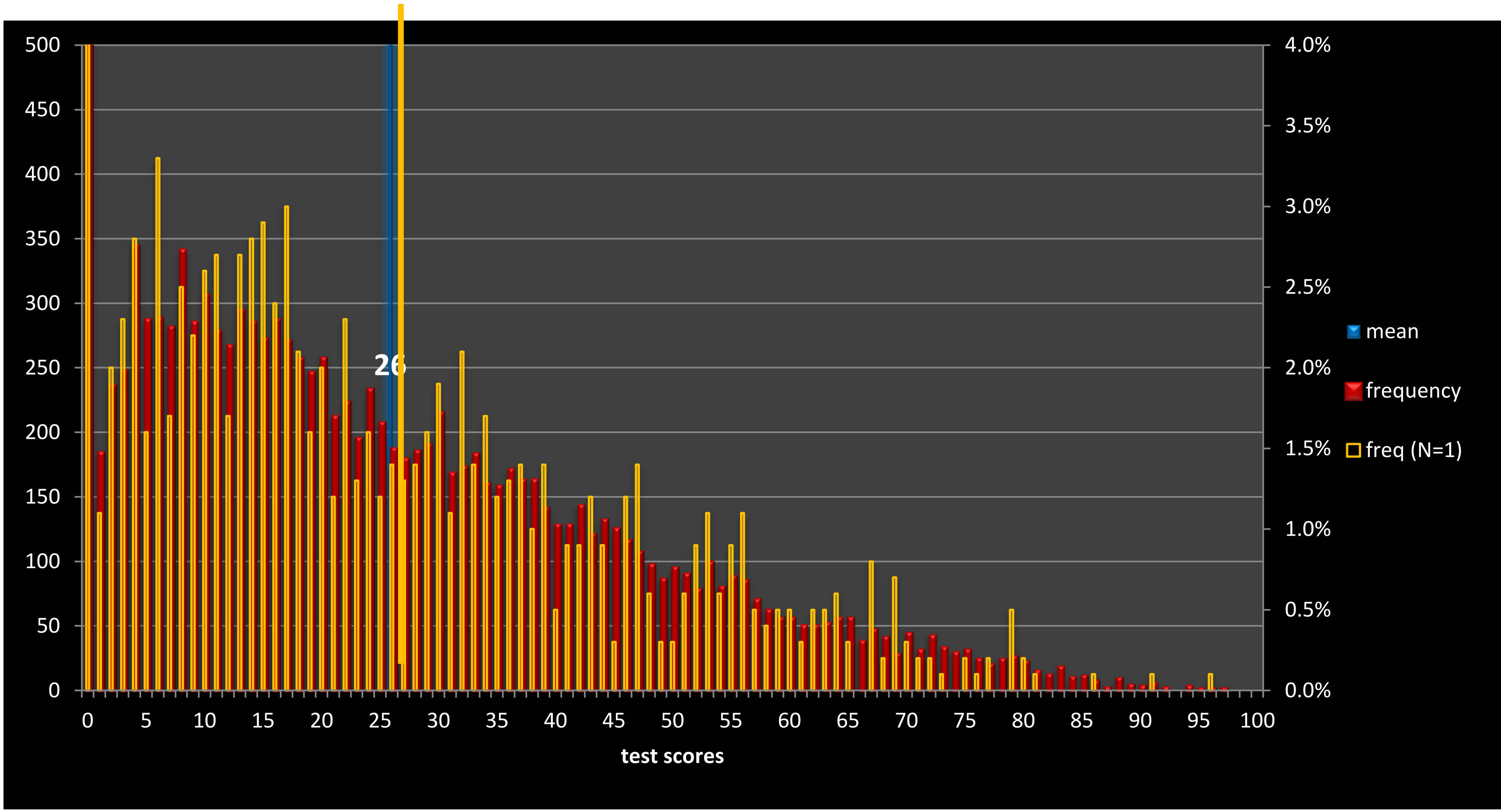
Standard Deviation = 20



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?

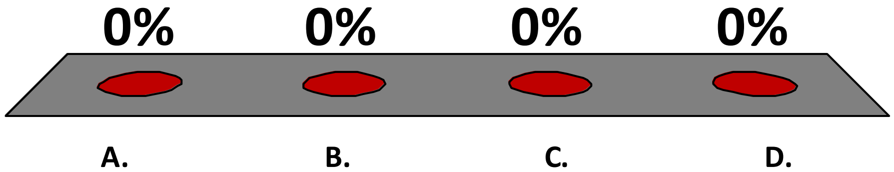
One limitation of statistical theory is that it assumes the population distribution is *normally distributed*

A. True

B. False

C. Depends

D. Don't know



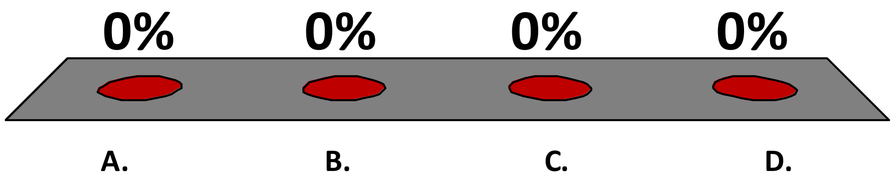
The sampling distribution may not be normal if the population distribution *is skewed*

A. True

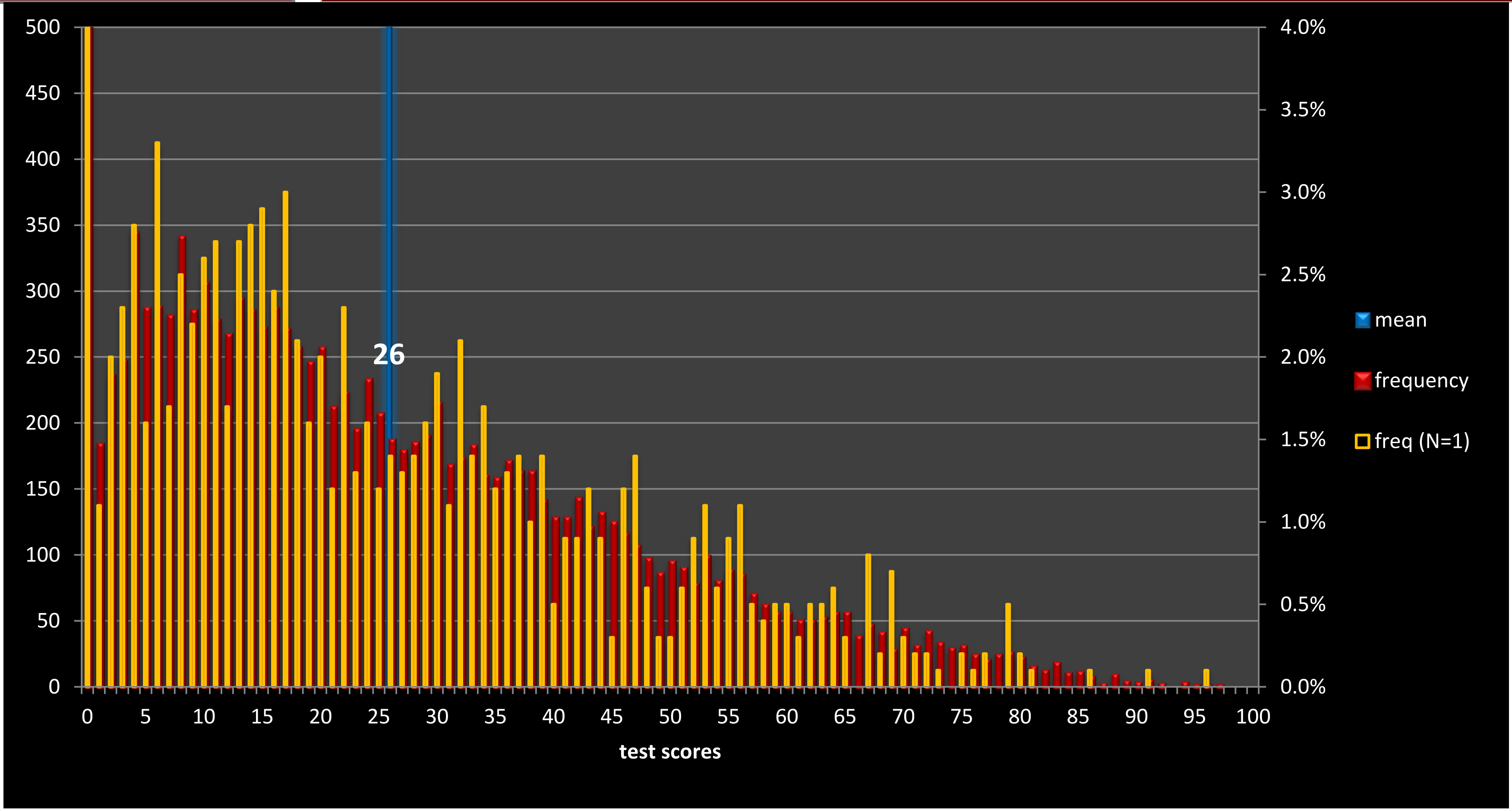
B. False

C. Depends

D. Don’t know



Population vs. sampling distribution

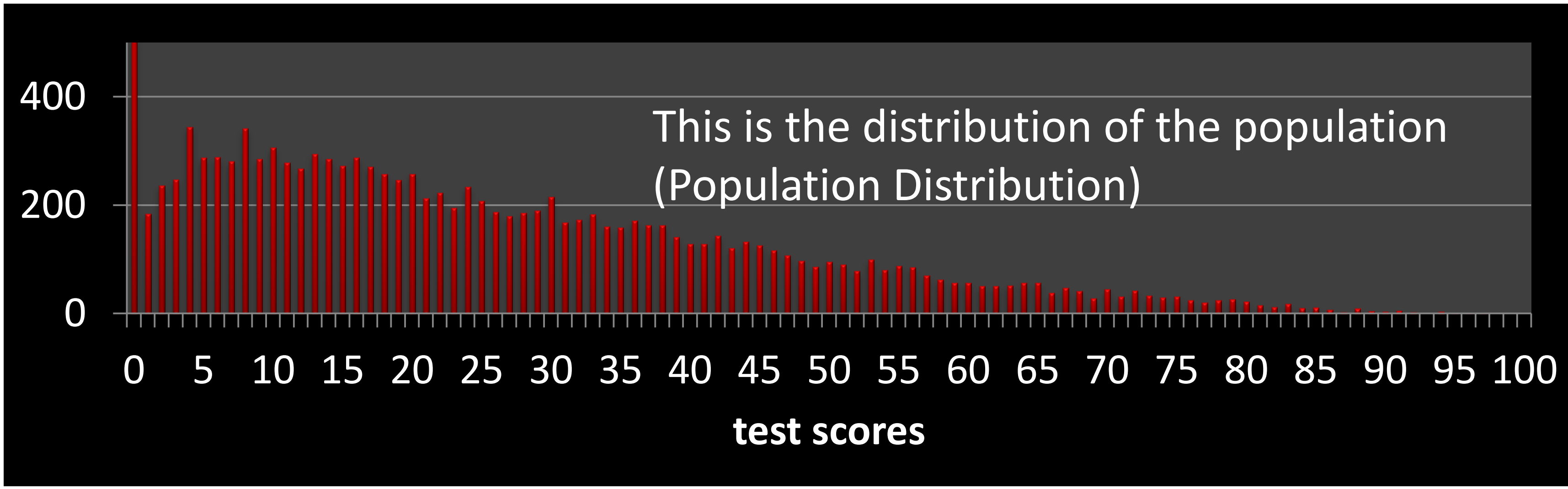


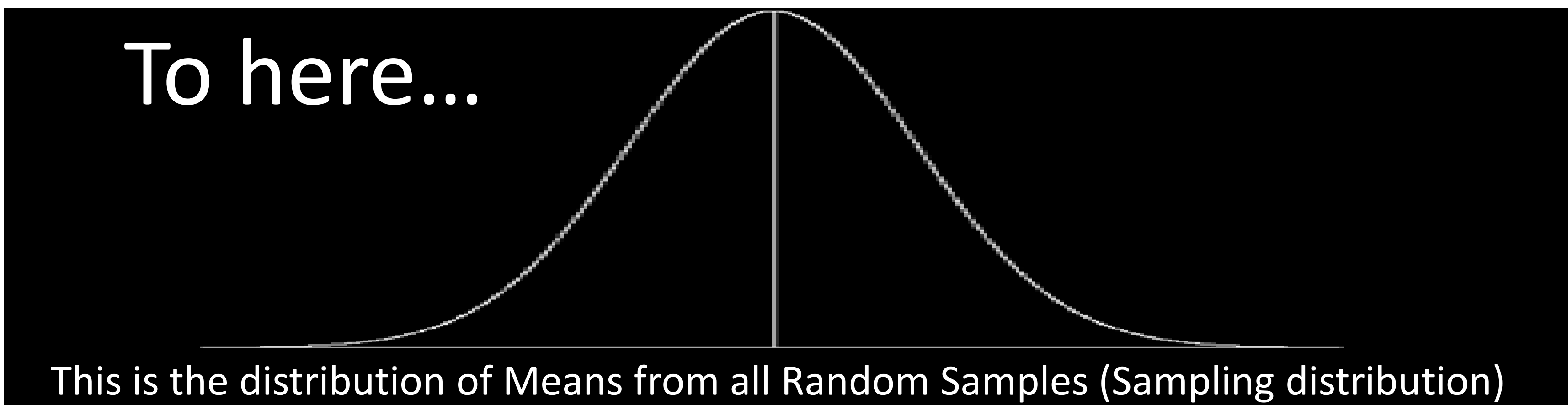
This is the distribution of my sample of 8,000 students!

Outline

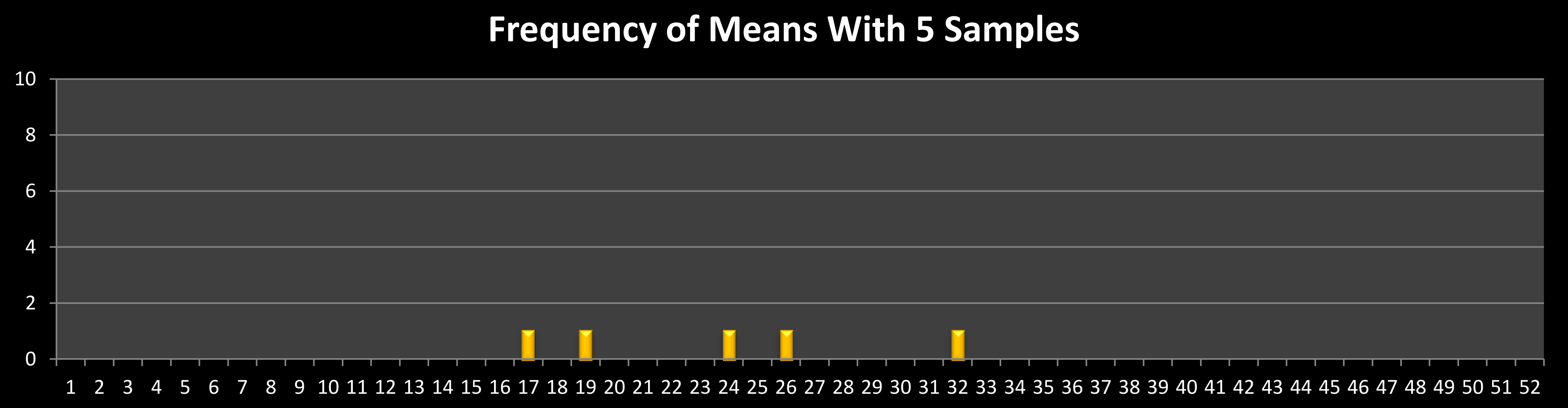
- Sampling distributions
  - population distribution
  - **sampling distribution**
  - law of large numbers/central limit theorem
  - standard deviation and standard error
- Detecting impact

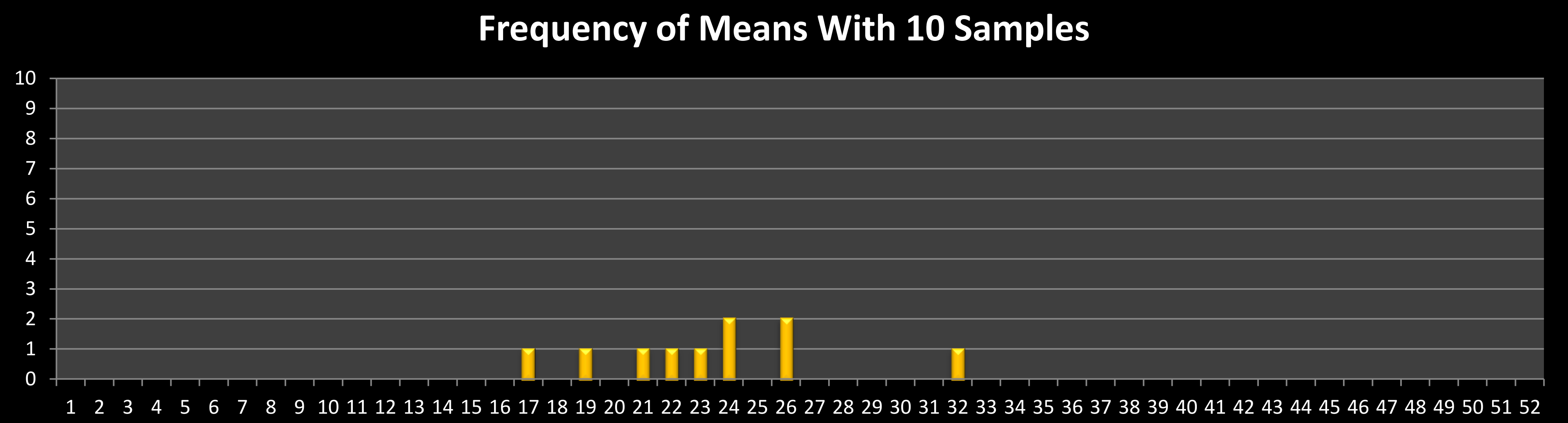
How do we get from here…



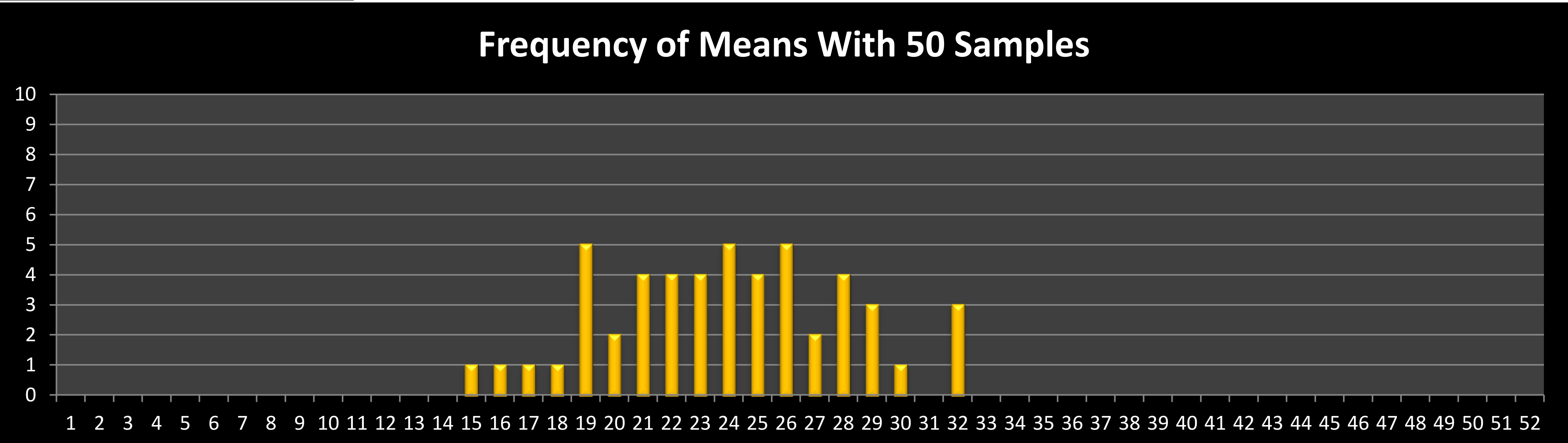


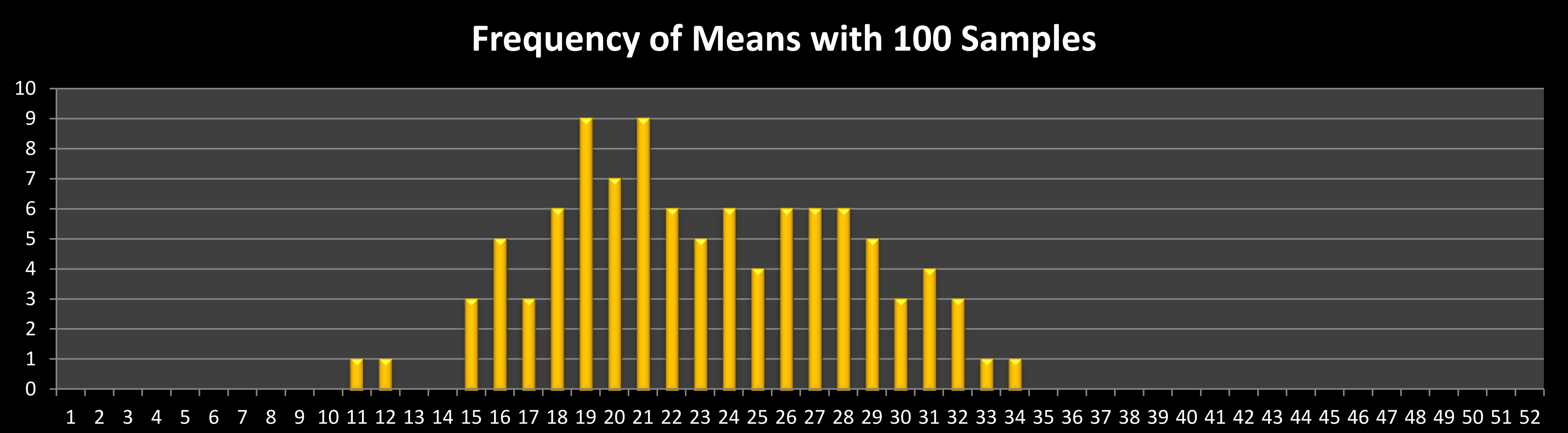
Draw 10 random students, take the average, plot it: Do this 5 & 10 times.



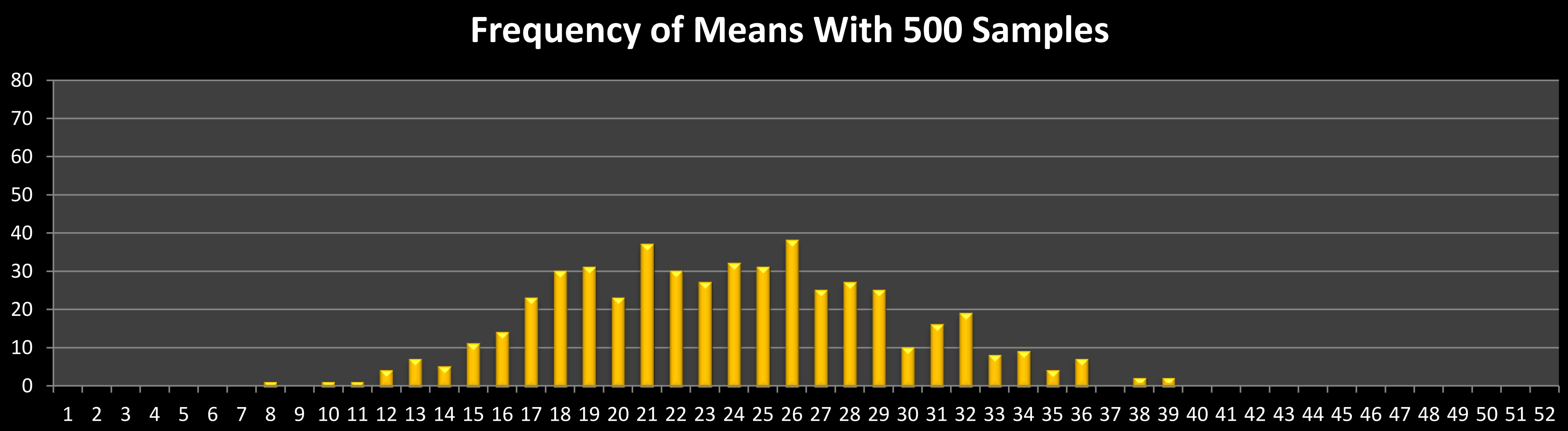


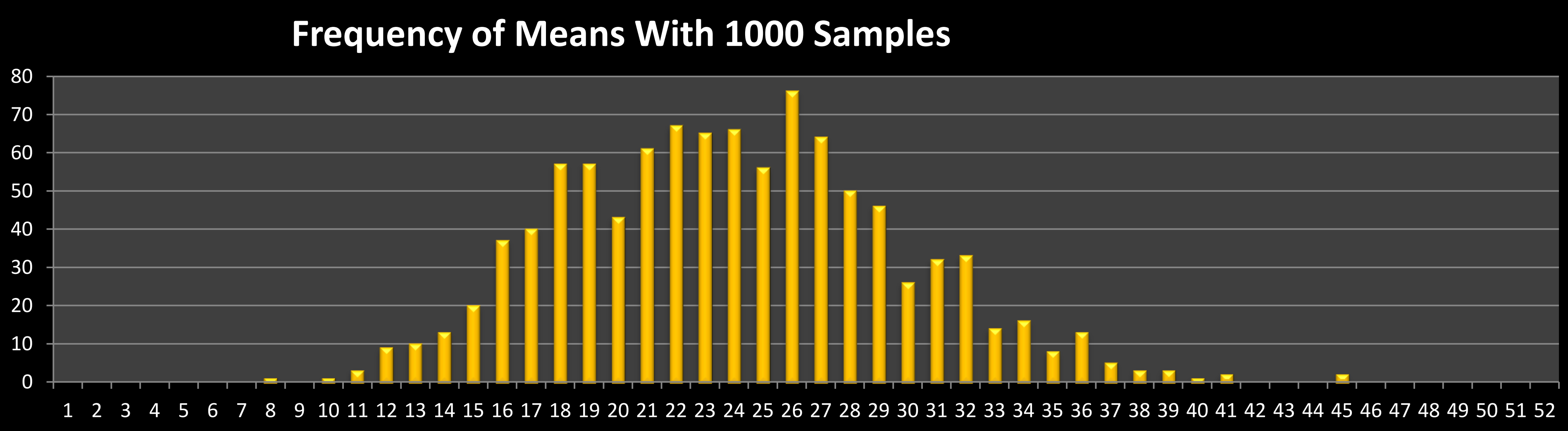
Draw 10 random students: 50 and 100 times





Draws 10 random students: 500 and 1000 times



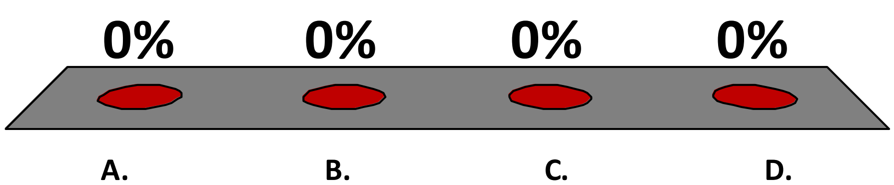


Draw 10 Random students

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

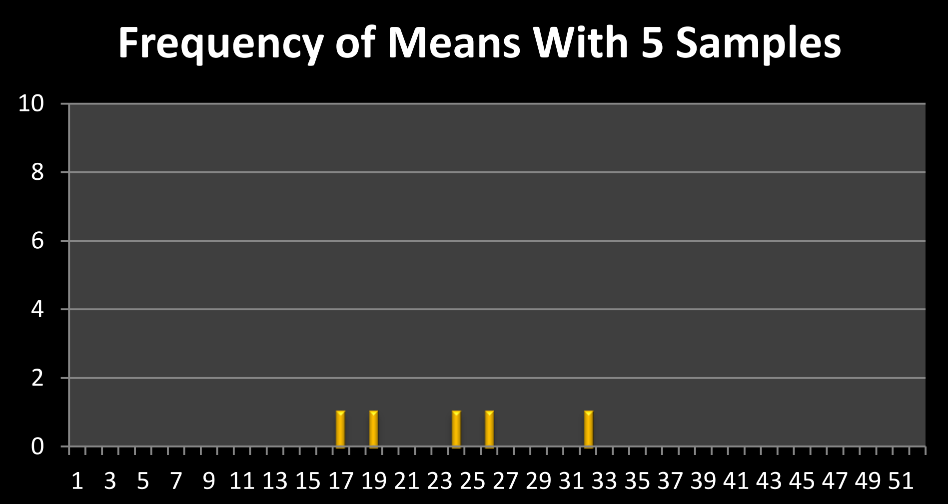
What happens to the sampling distribution if we draw a sample size of 50 instead of 10, and take the mean (thousands of times)?

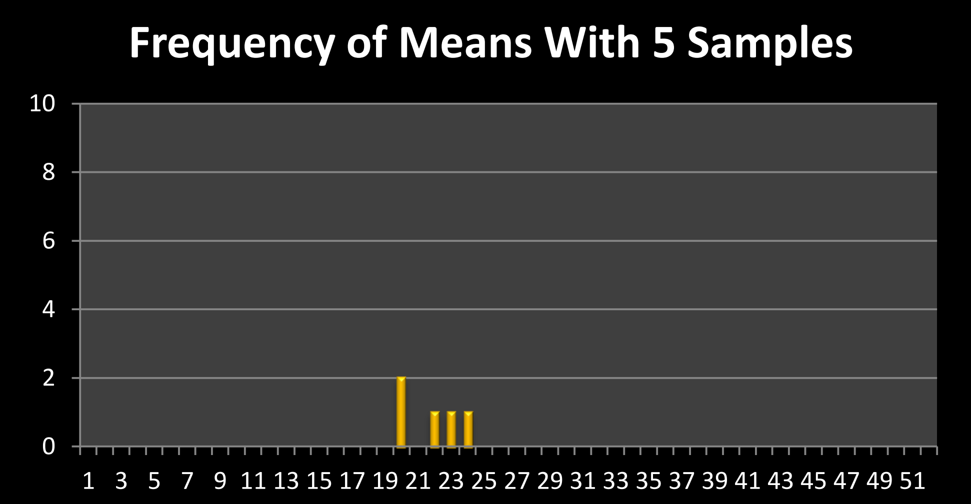
A. We will approach a bell curve faster (than with a sample size of 10)
B. The bell curve will be narrower
C. Both A & B
D. Neither. The underlying sampling distribution does not change.

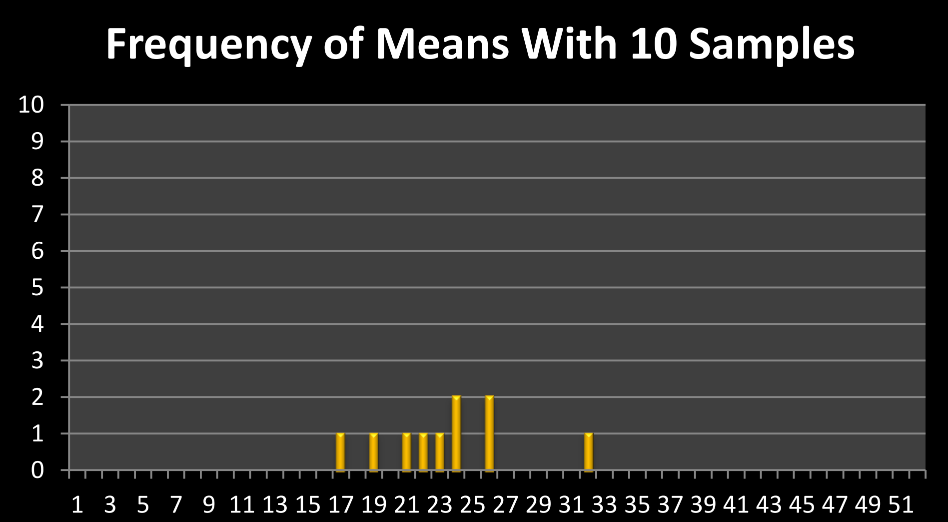


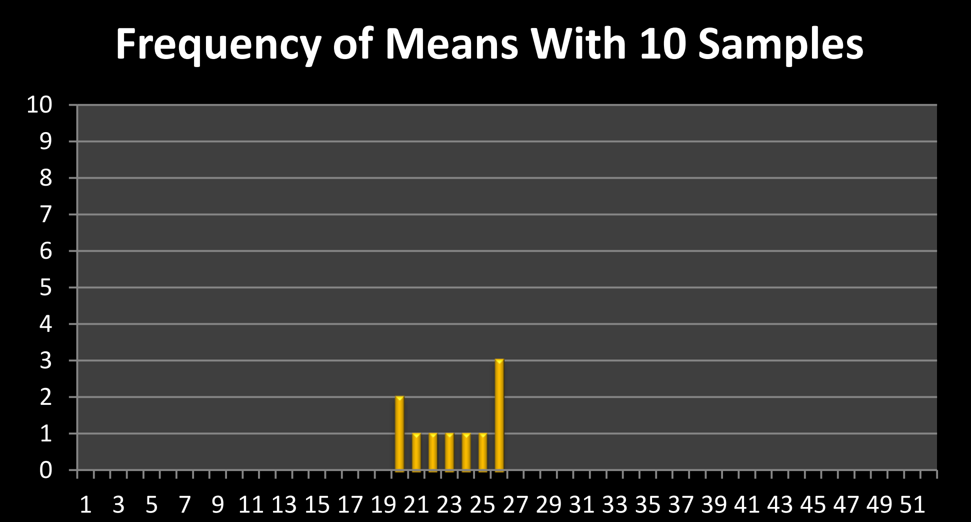
N = 10

N = 50

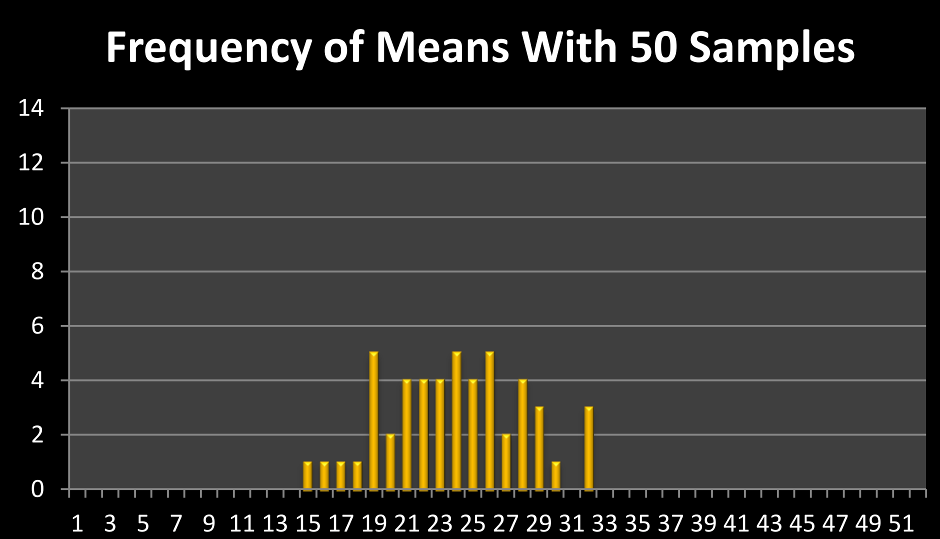

Frequency of Means With 5 Samples
10
8
6
4
2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

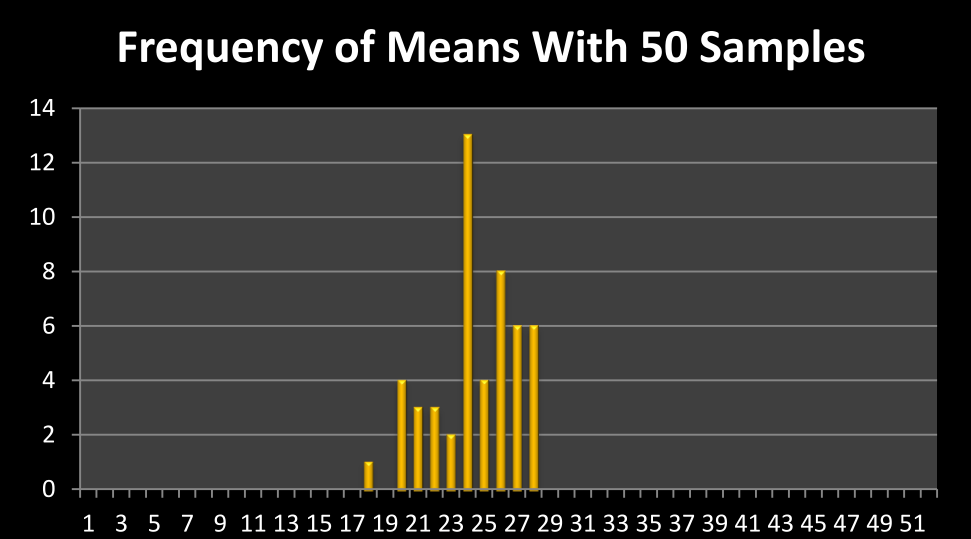

Frequency of Means With 5 Samples
10
8
6
4
2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

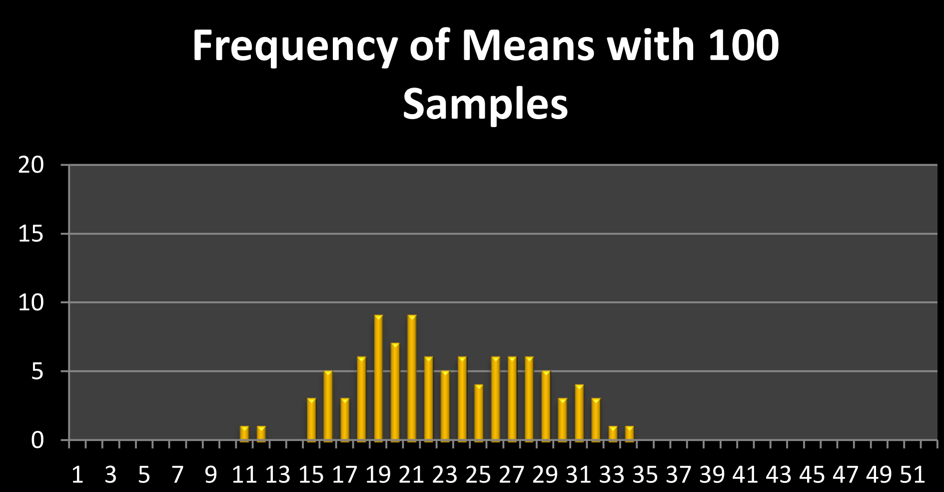

Frequency of Means With 10 Samples
10
9
8
7
6
5
4
3
2
1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

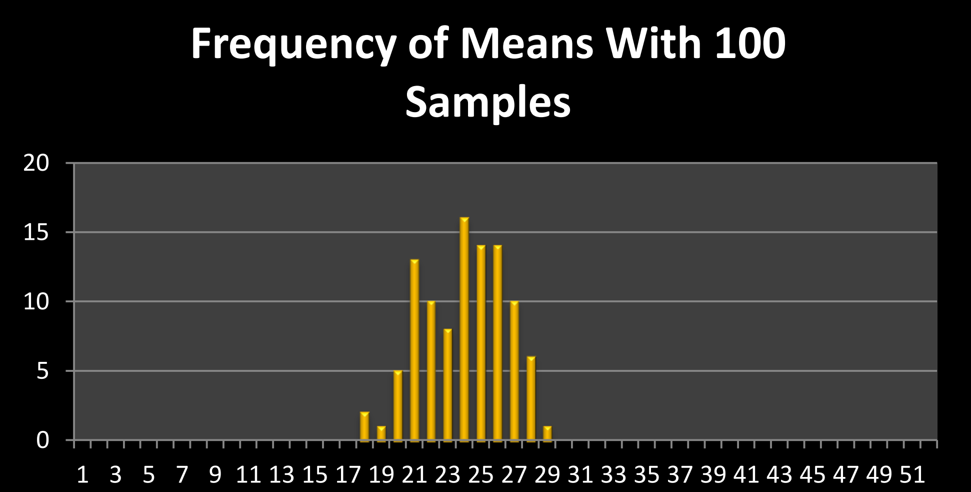

Frequency of Means With 10 Samples
10
9
8
7
6
5
4
3
2
1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

N = 10 N = 50

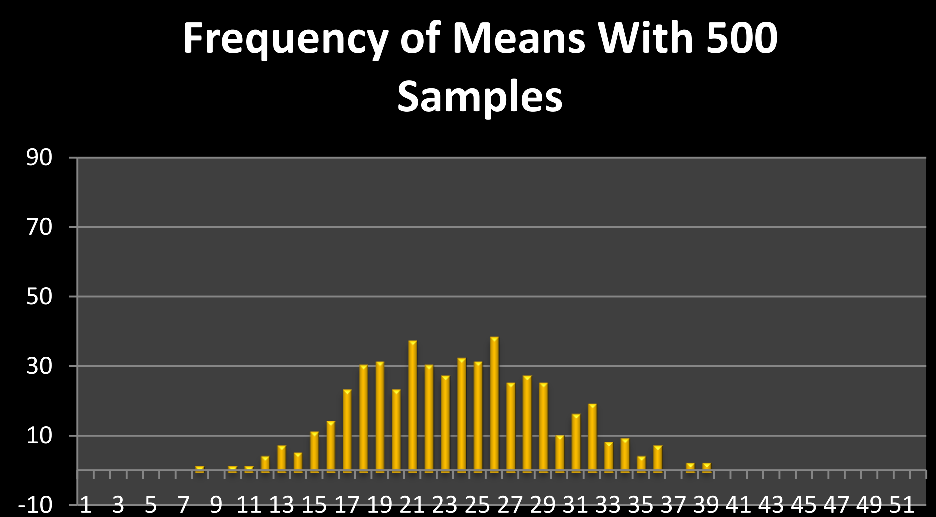

Frequency of Means With 50 Samples
0 2 4 6 8 10 12 14
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

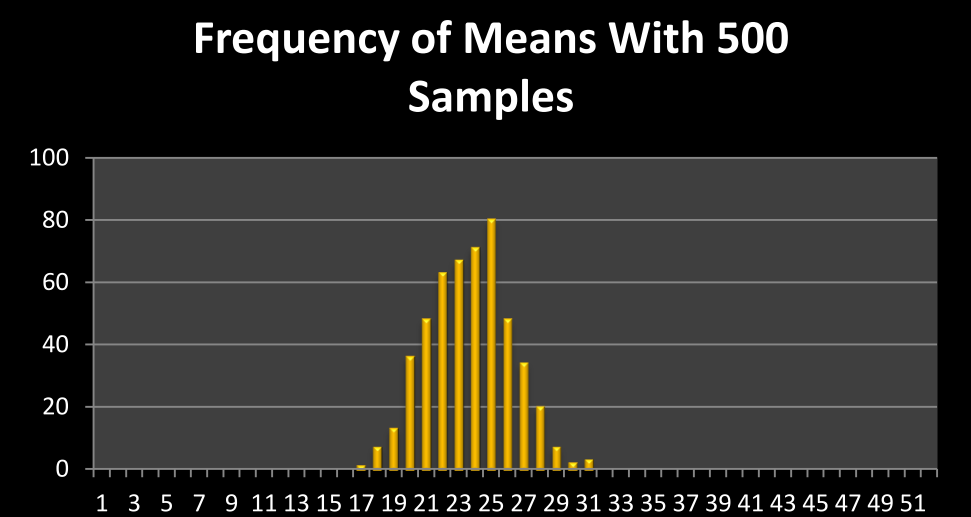

Frequency of Means With 50 Samples
0 2 4 6 8 10 12 14
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

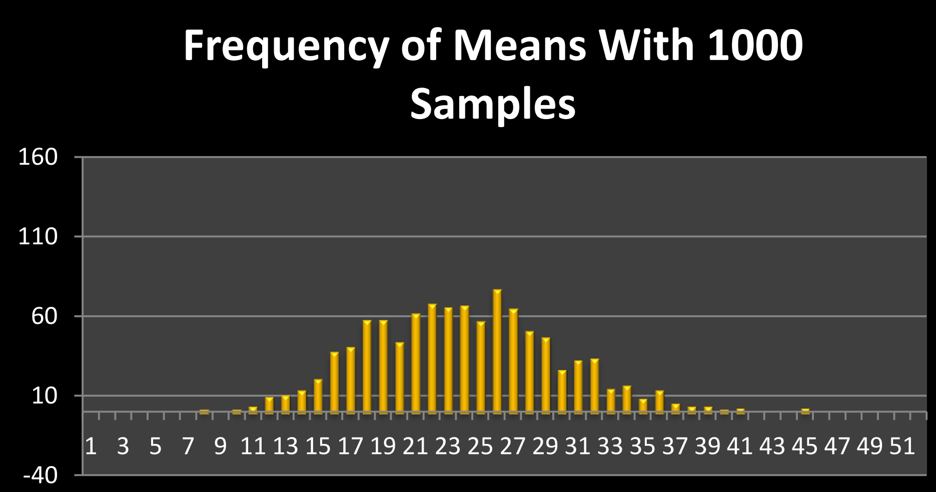

Frequency of Means with 100 Samples
0 5 10 15 20
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

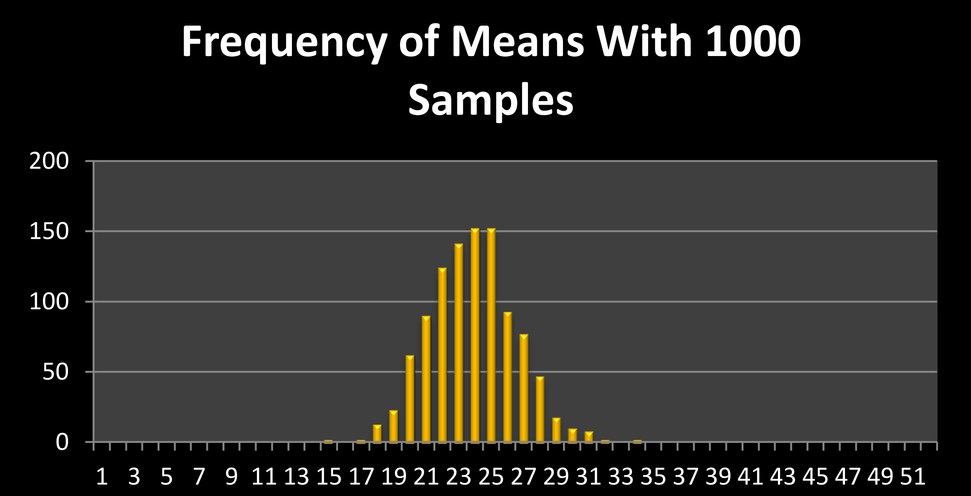

Frequency of Means With 100 Samples
0 5 10 15 20
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

N = 10 N = 50


Frequency of Means With 500 Samples
90
70
50
30
10
-10
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51


Frequency of Means With 500 Samples
100
80
60
40
20
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

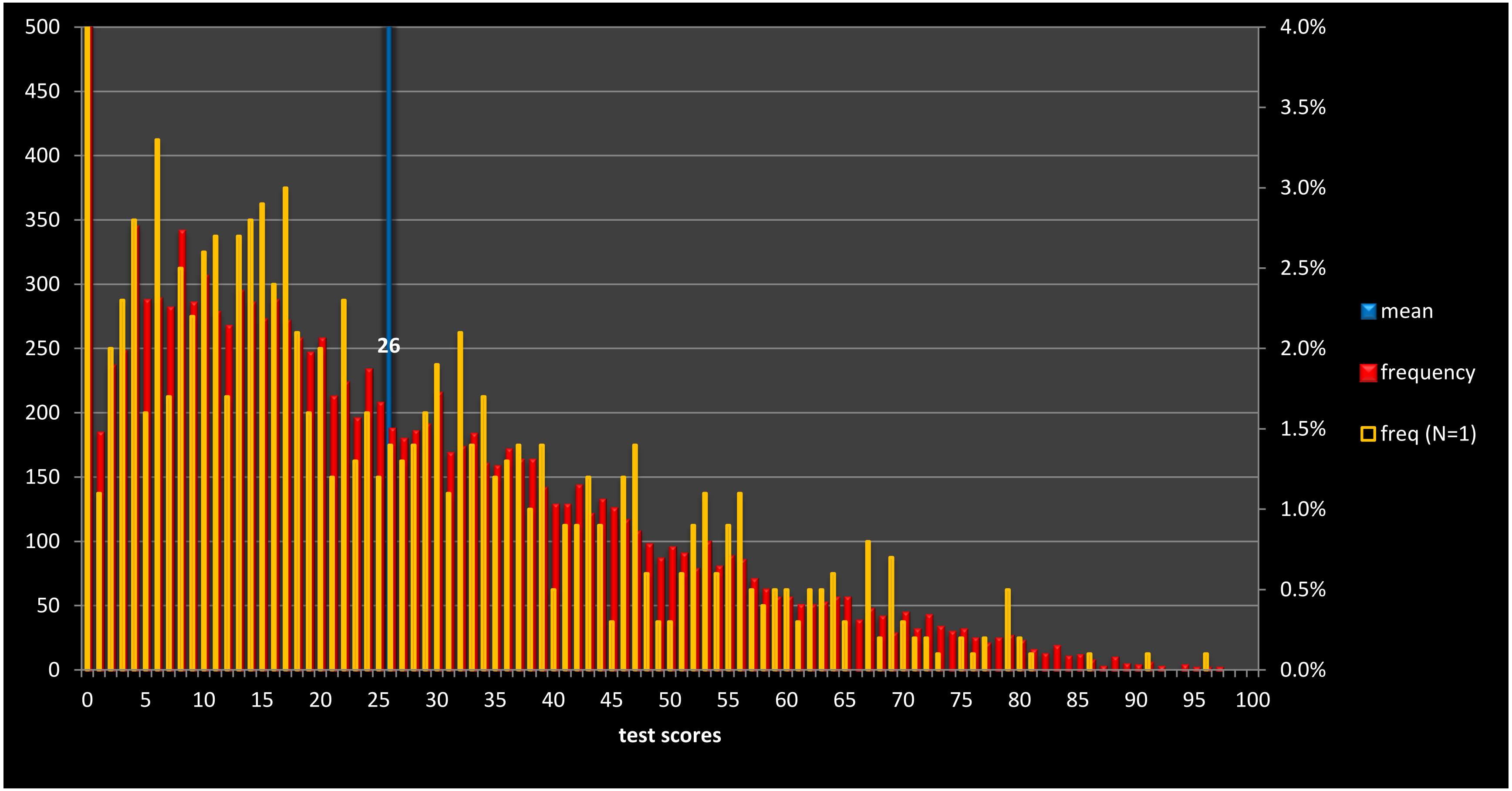

Frequency of Means With 1000 Samples
160
110
60
10
-40
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51


Frequency of Means With 1000 Samples
200
150
100
50
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51

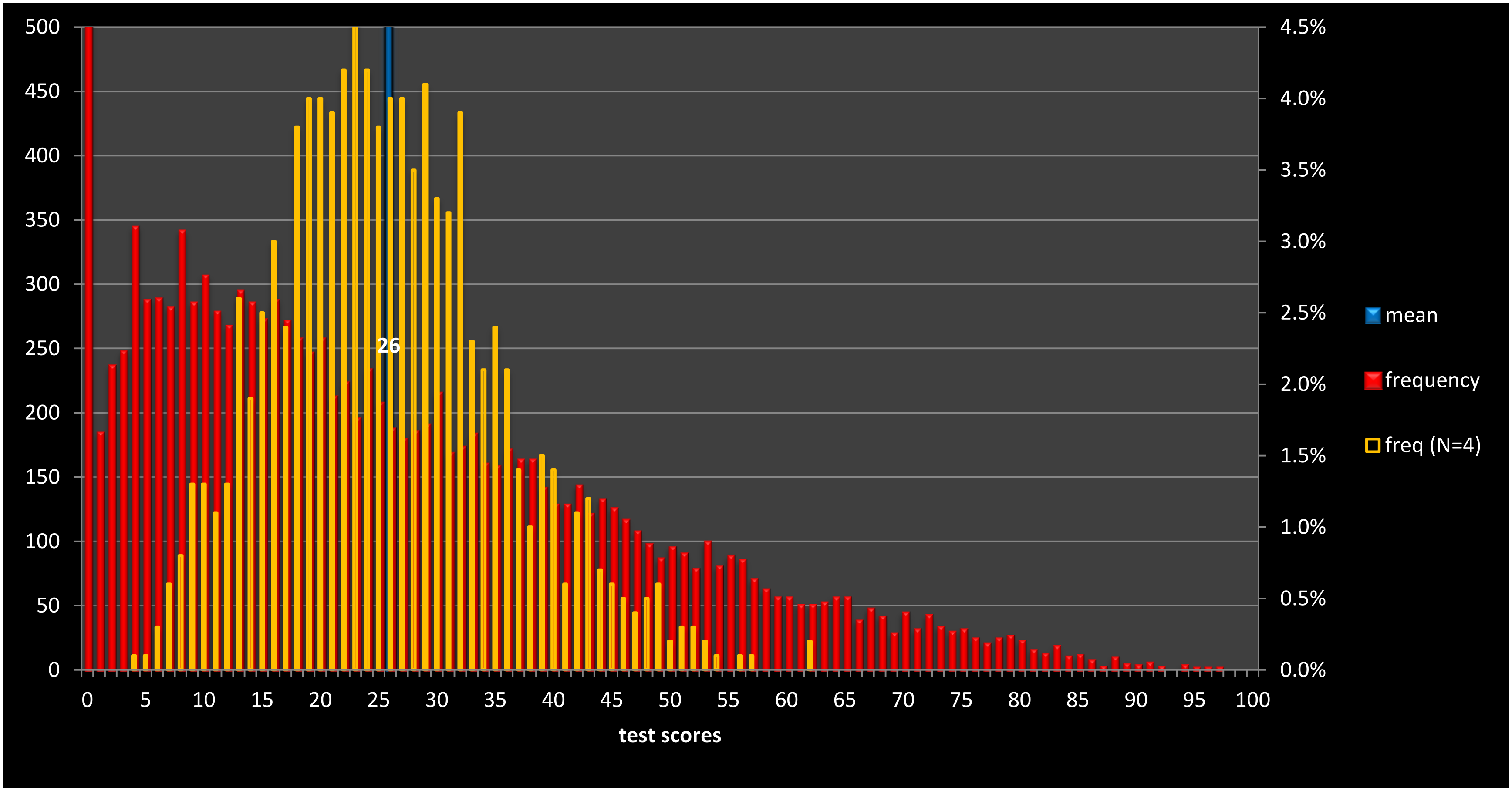
Outline

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

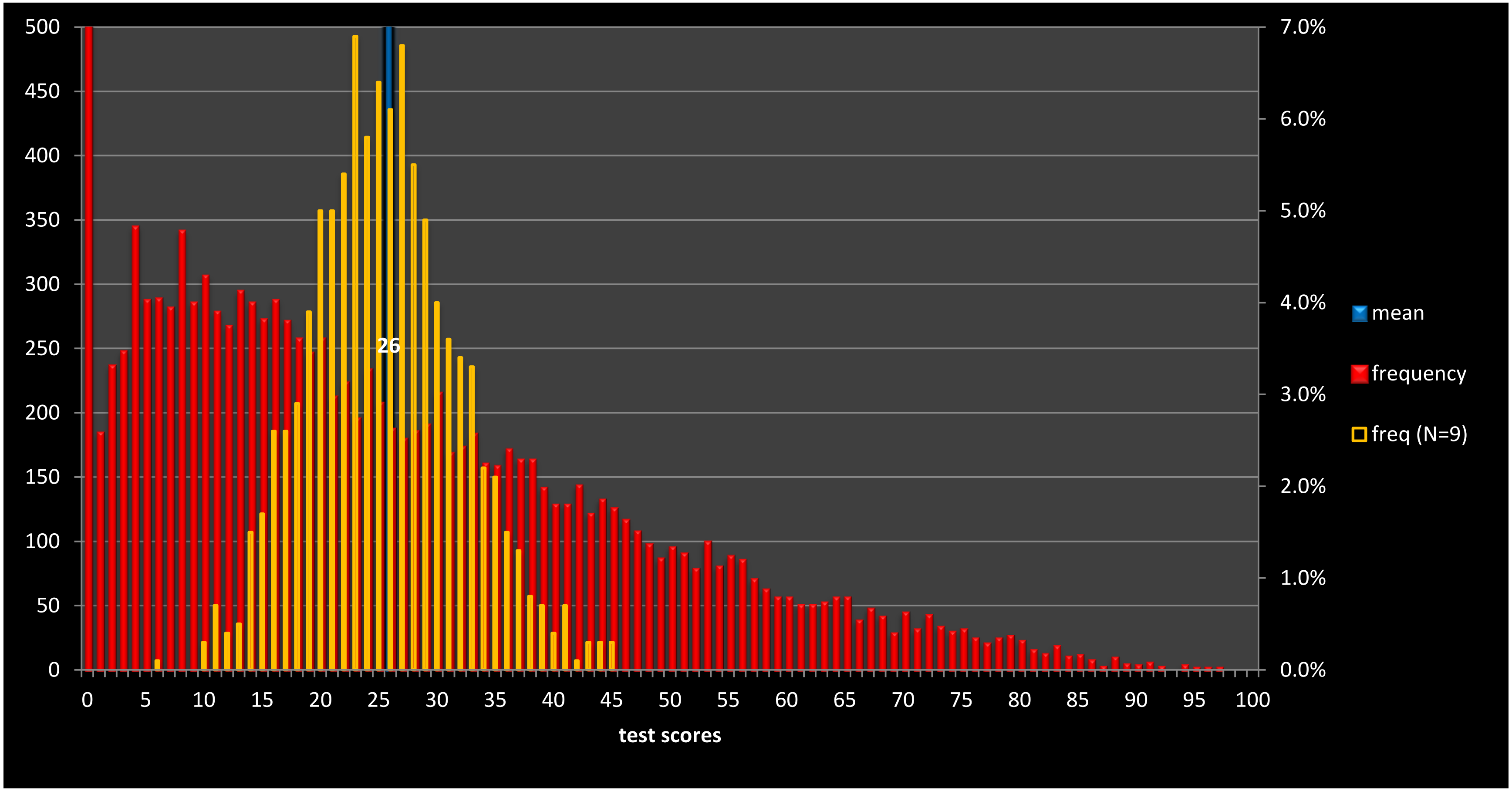
Population & sampling distribution: Draw 1 random student (from 8,000)



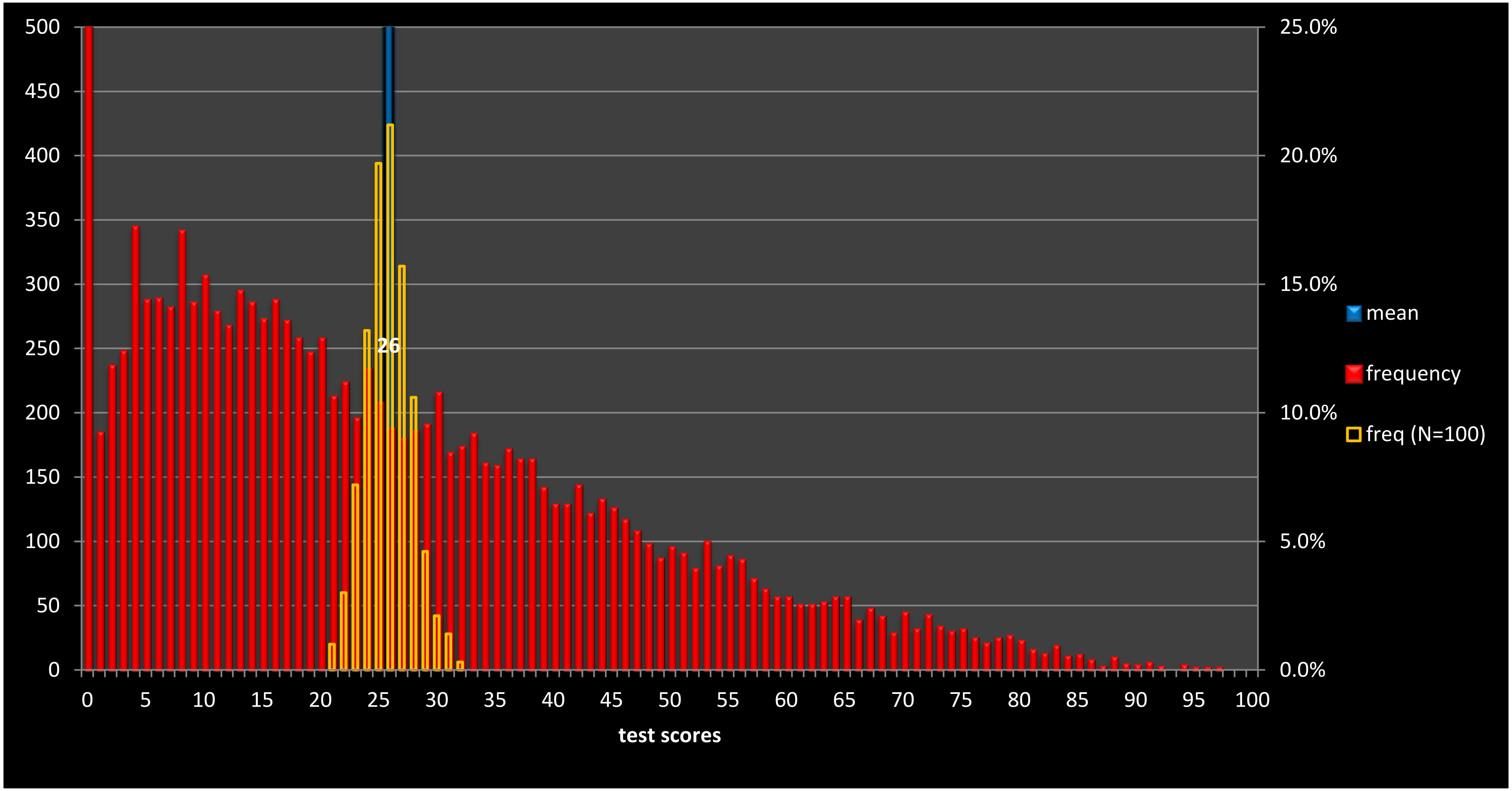
Sampling Distribution:
Draw 4 random students (N=4)



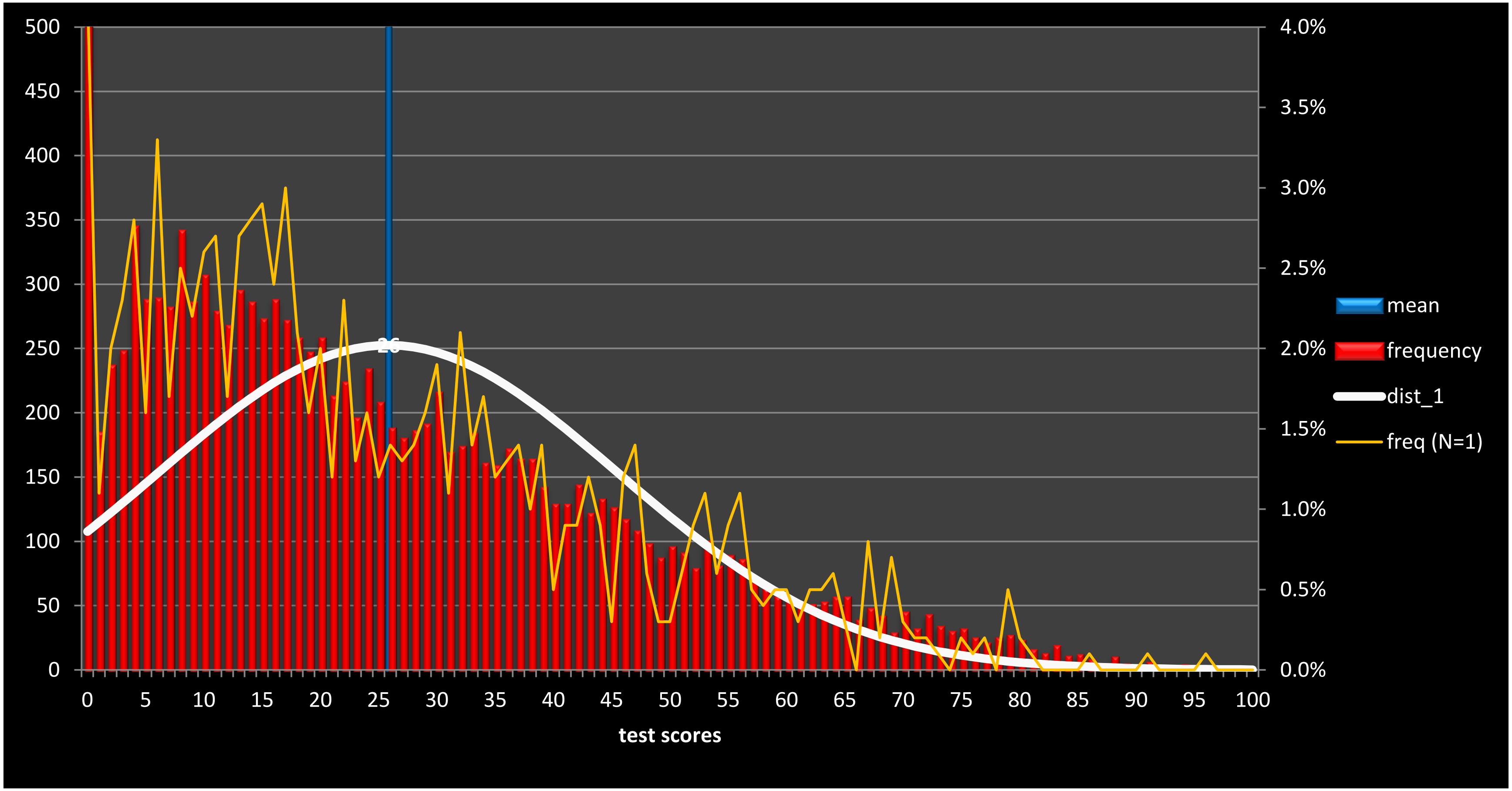
Law of Large Numbers : N=9



Law of Large Numbers: N =100

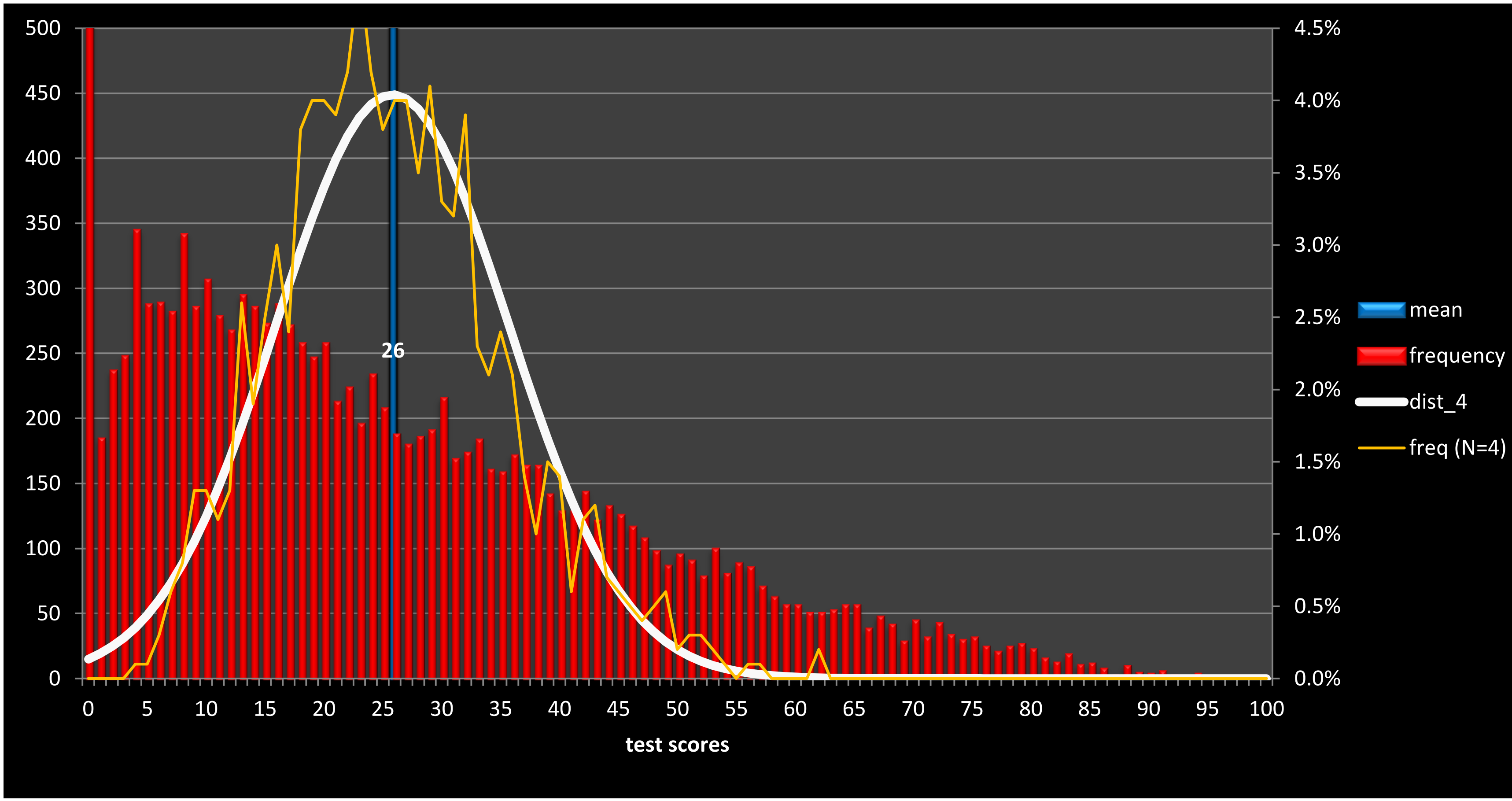


Central Limit Theorem: N=1

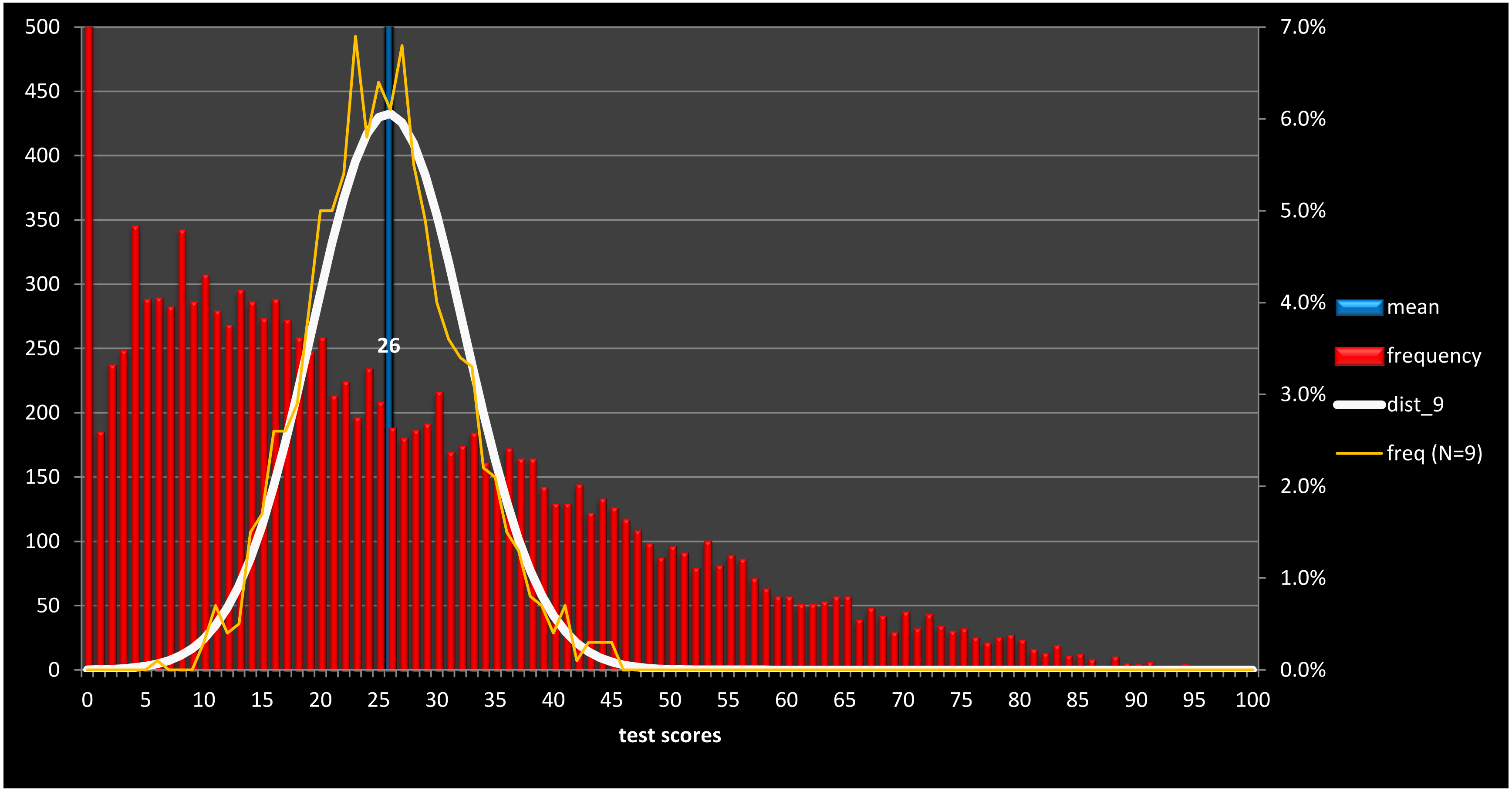


The white line is a theoretical distribution

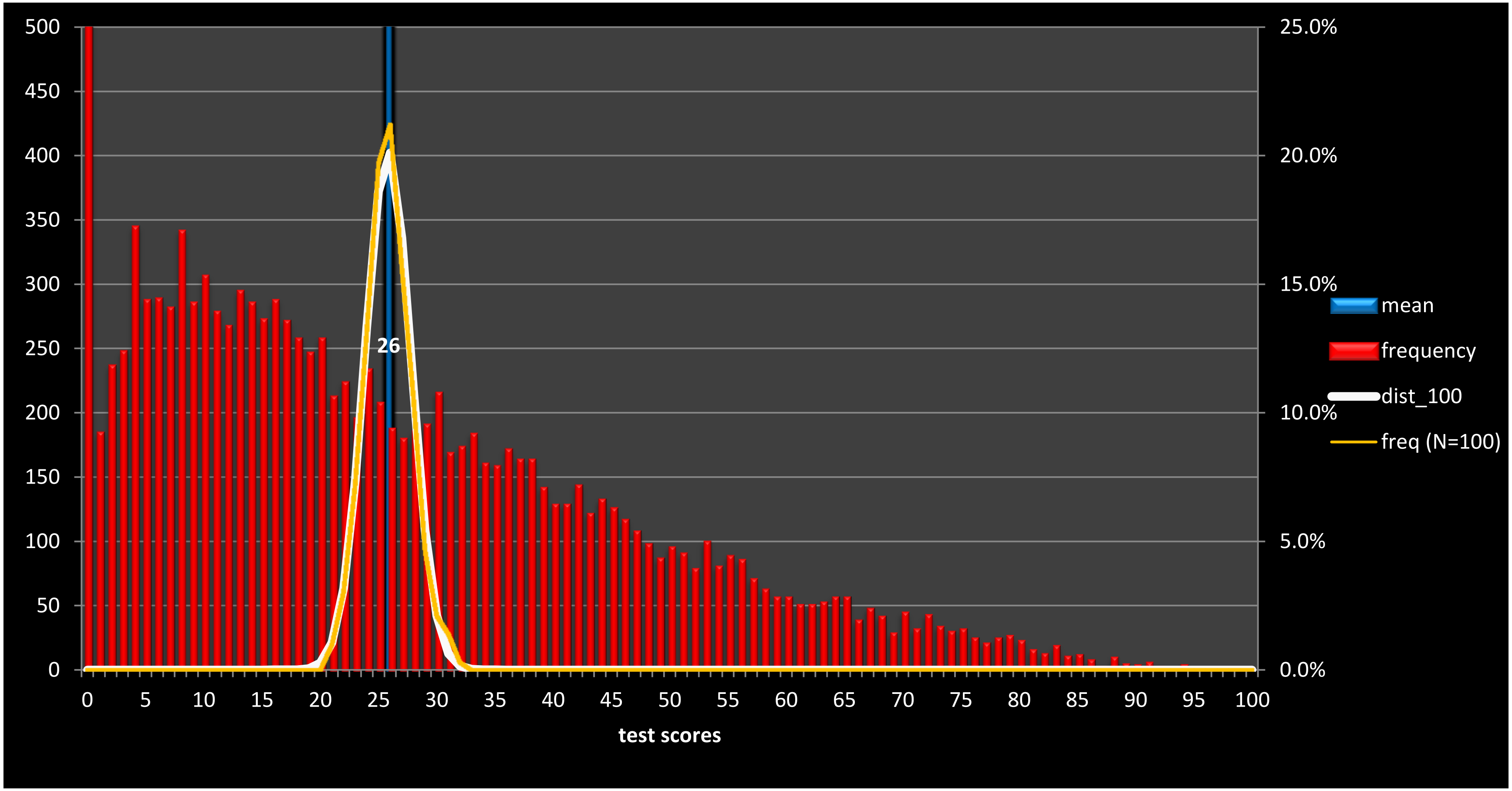
Central Limit Theorem : N=4



Central Limit Theorem : N=9



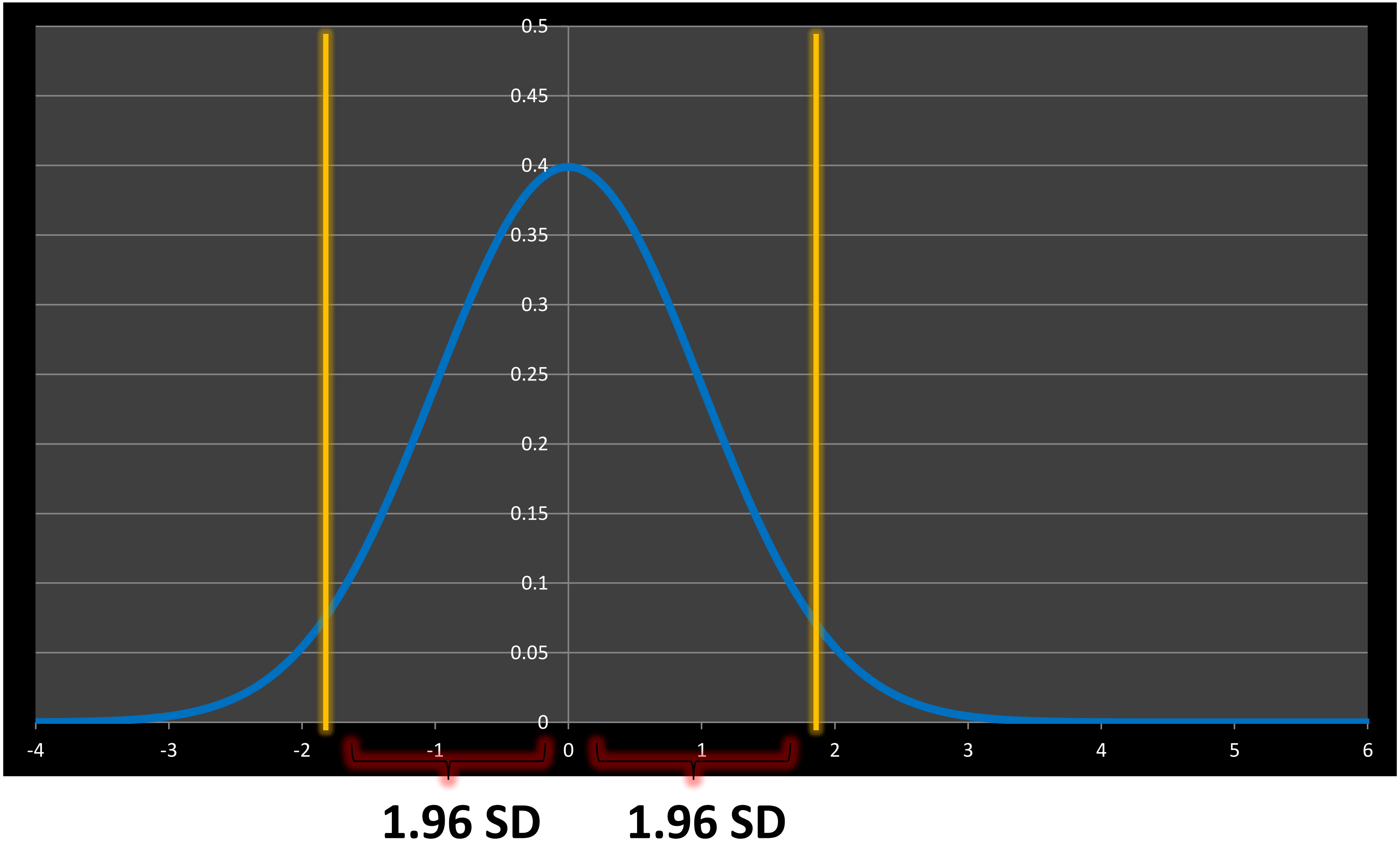
Central Limit Theorem : N =100



So Why Do We Care?

- Sampling distribution is a probability distribution
- Sampling Distribution is a bell curve (*irrespective* of what the underlying distribution is)
- Why does it matter?
- Why do we care if the probability distribution looks like a bell curve?
- **Because we know how to calculate the area underneath!**

95% Confidence Interval



Outline

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

Standard deviation/error

- But wait! The regression results that I have seen typically report the standard **error**, not the standard **deviation**.

- What's the difference between the standard deviation and the standard error?

The standard error = the standard deviation of the sampling distribution

Variance and Standard Deviation

- Variance = 400

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

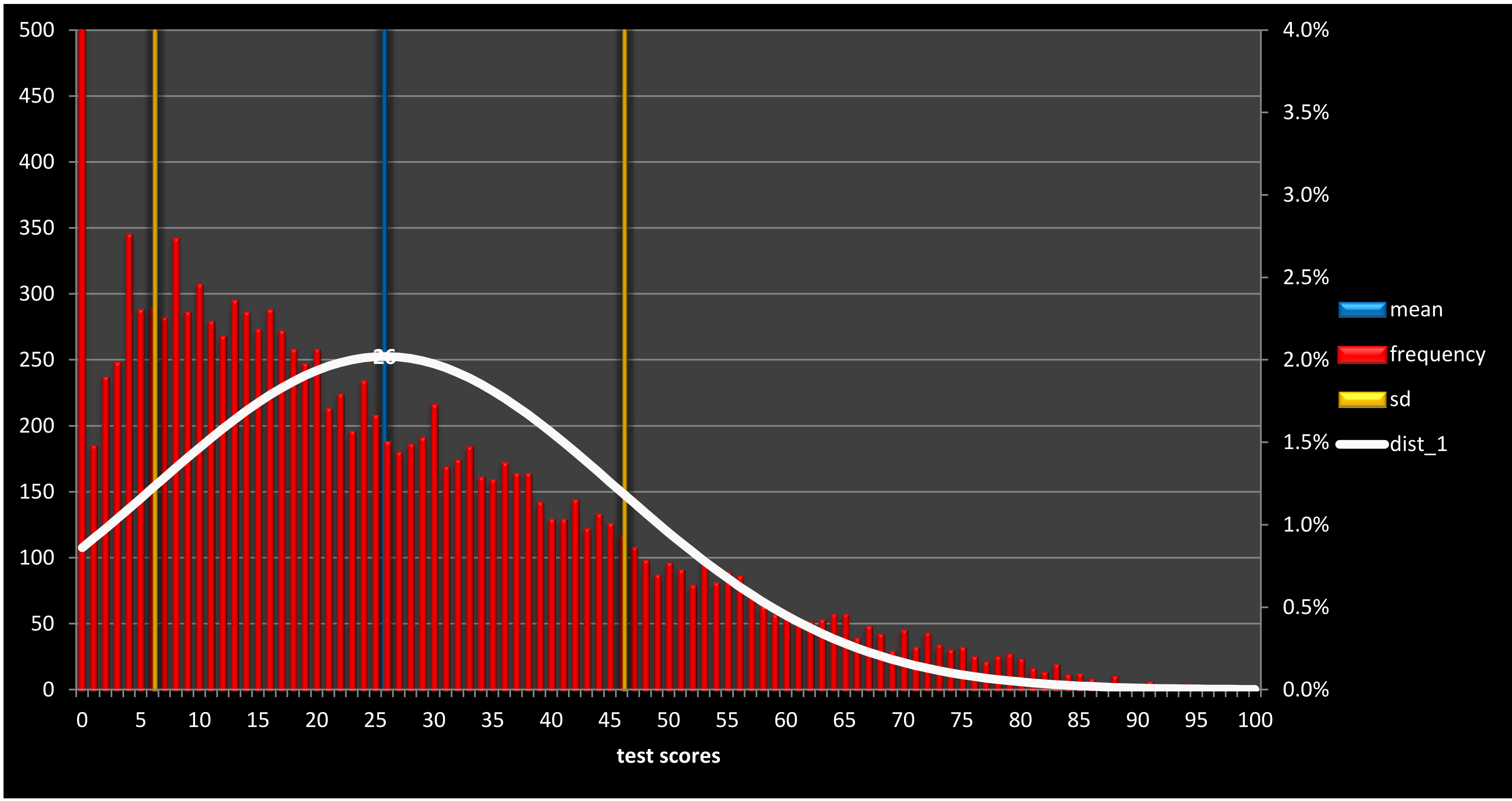
- Standard Deviation = 20

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

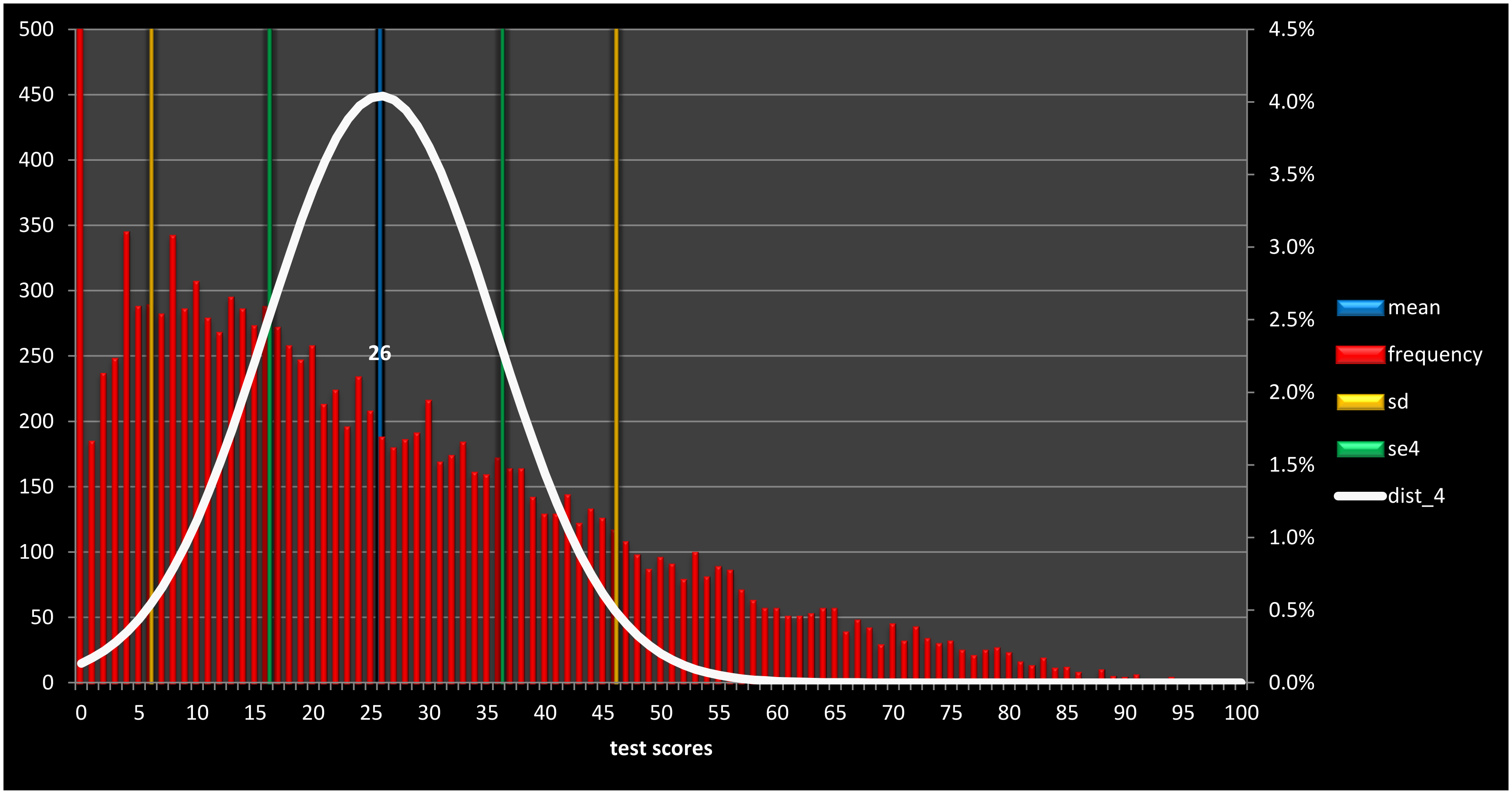
- Standard Error = ${}^{20}/_{\sqrt{N}}$

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

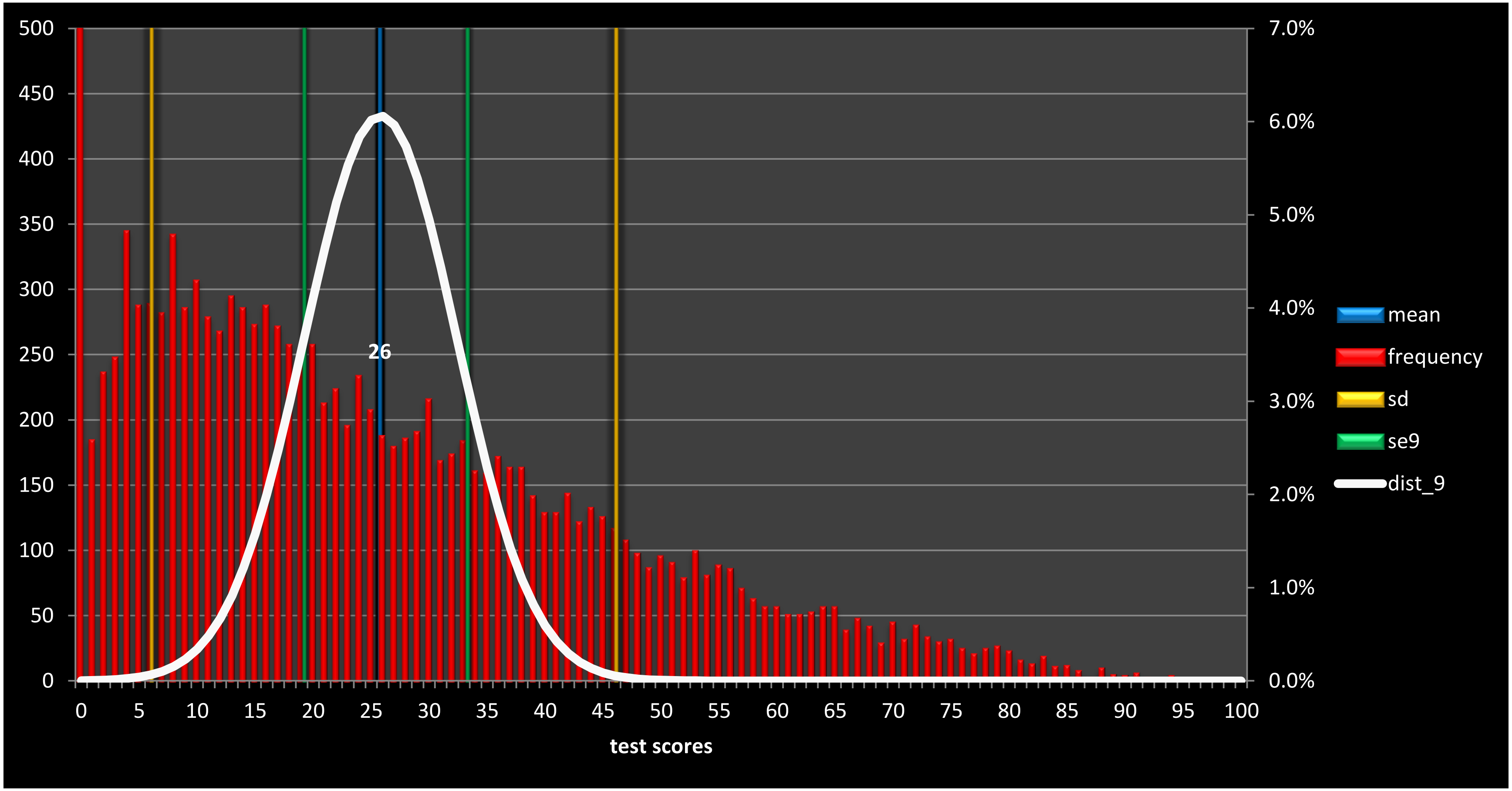
Standard Deviation/ Standard Error



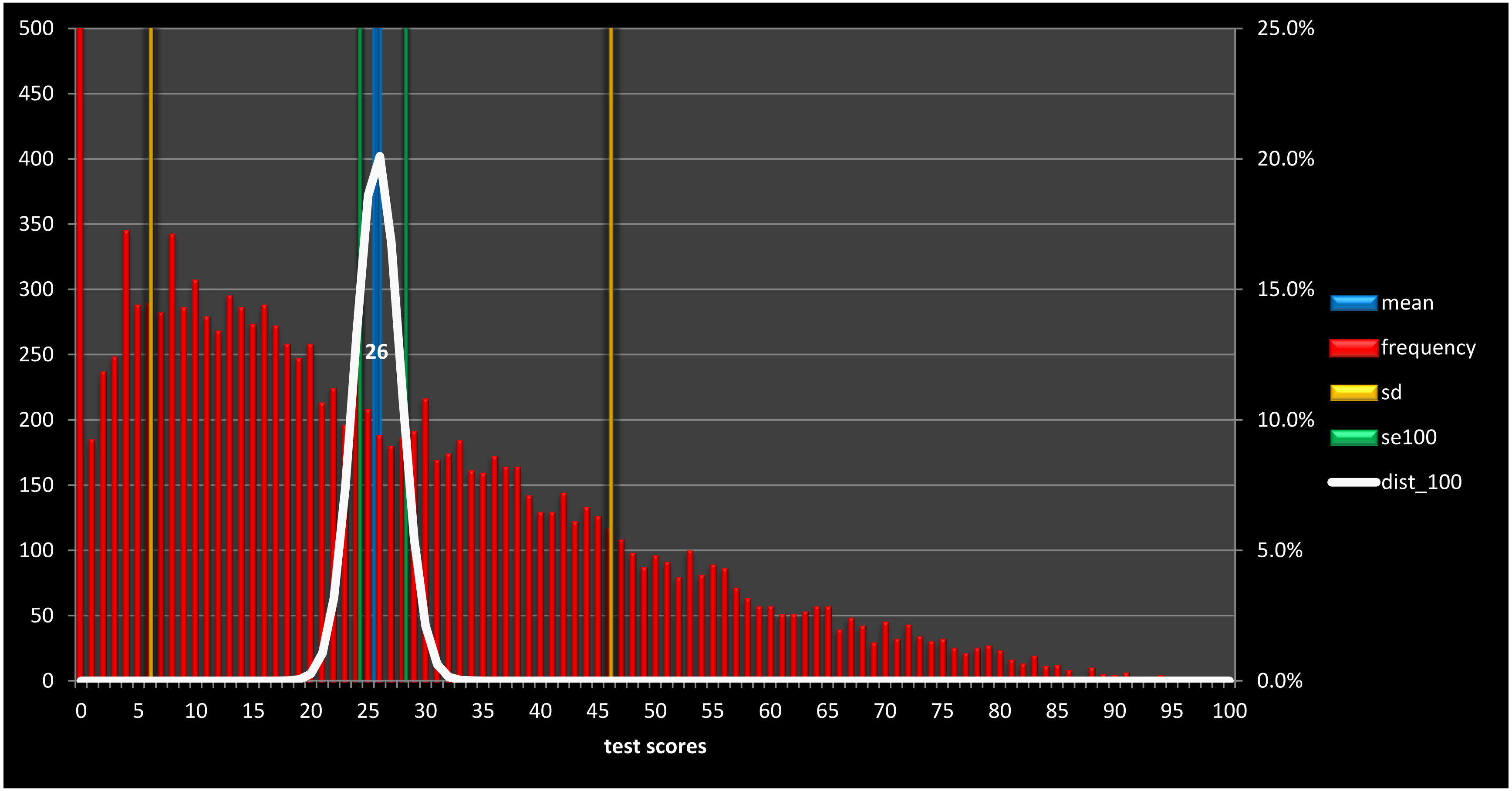
Sample size ↑ x4, SE ↓ ½



Sample size ↑ x9, SE ↓ ?



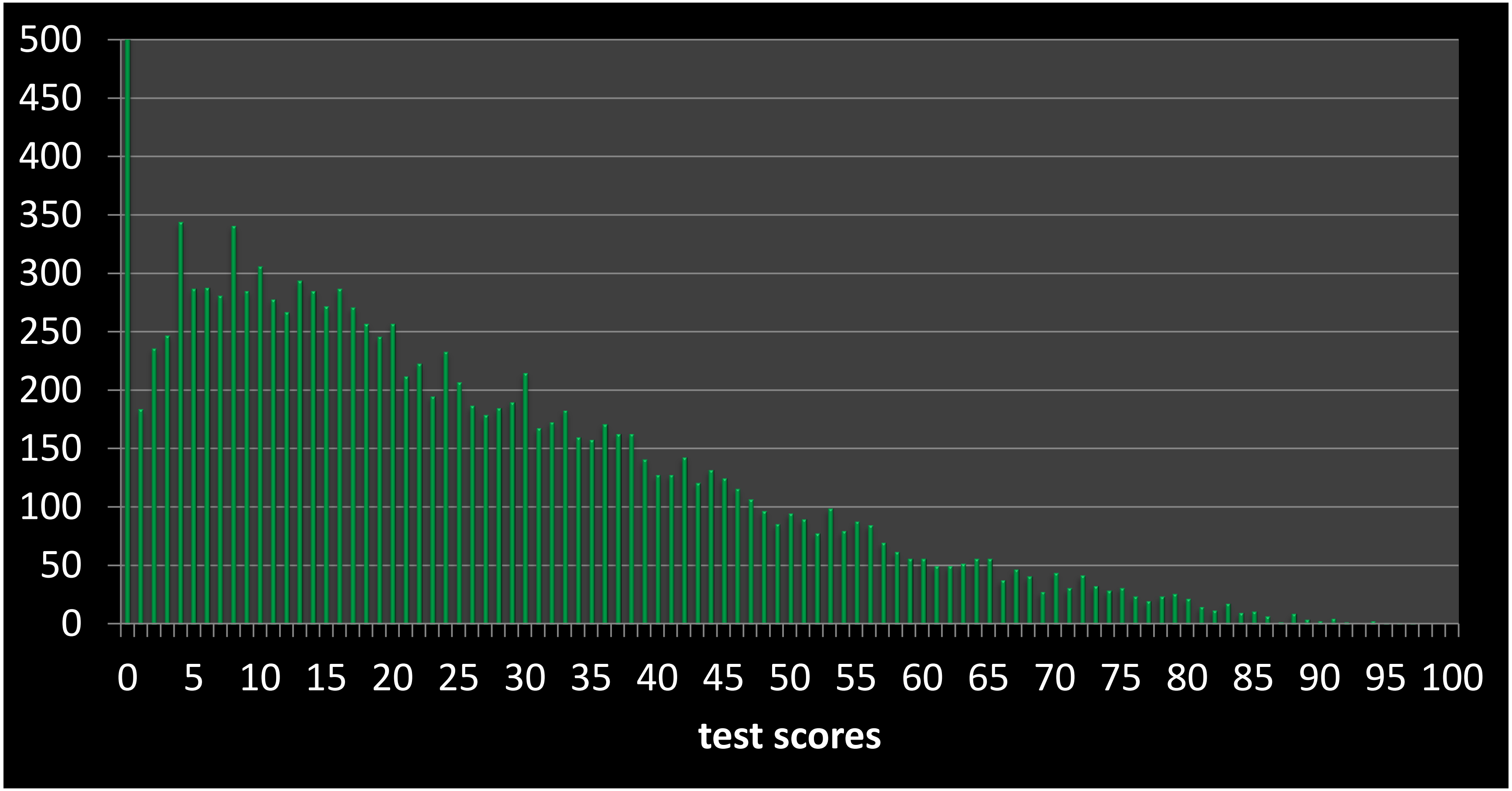
Sample size ↑ x100, SE ↓?



Outline

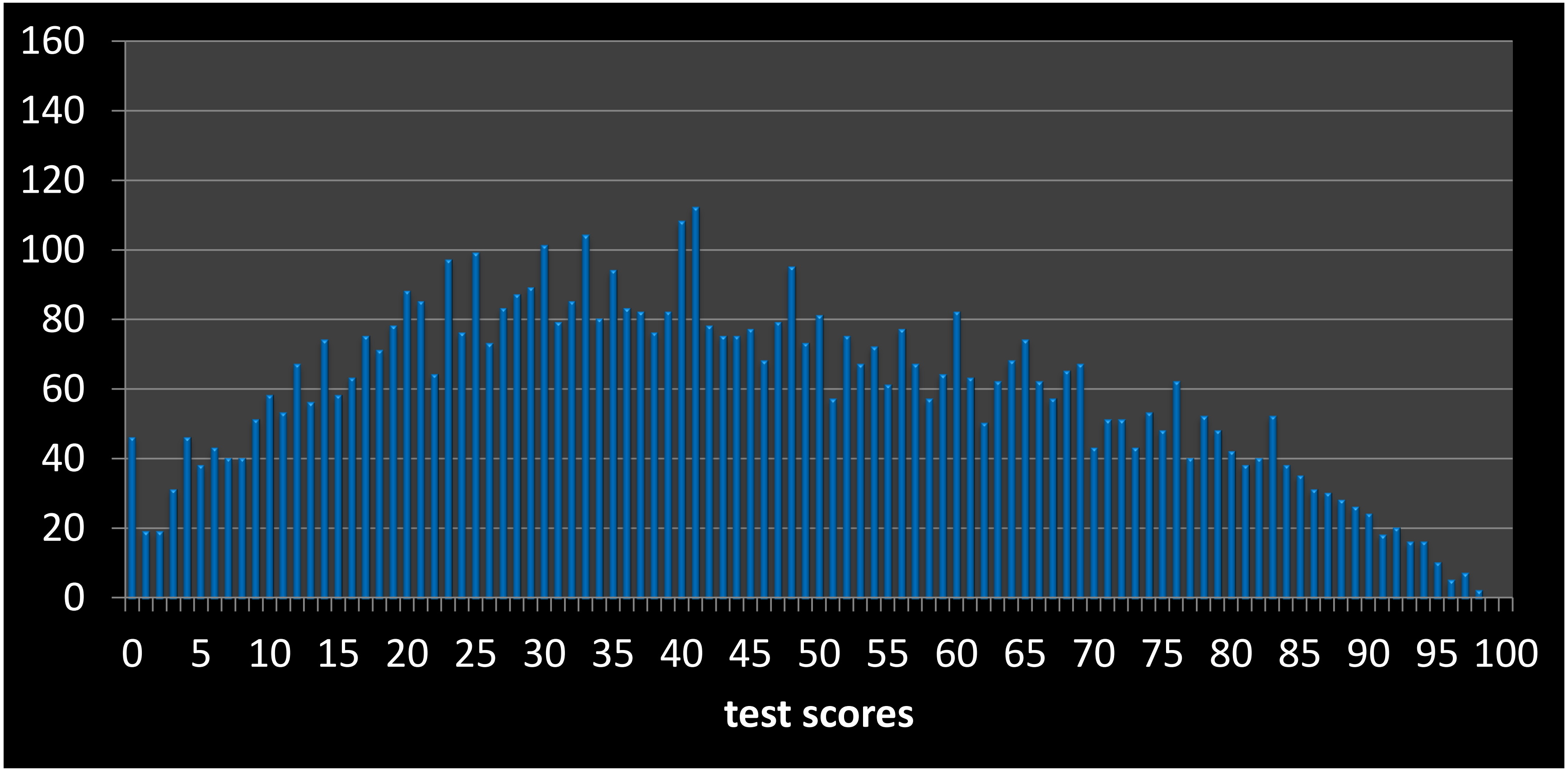
- Sampling distributions
- Detecting impact
  - significance
  - effect size
  - power
  - baseline and covariates
  - clustering
  - stratification

Baseline test scores



We implement the Balsakhi Program

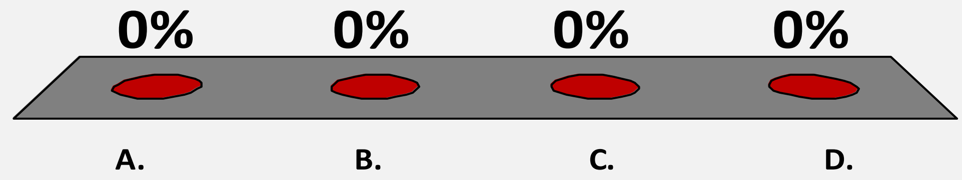
Endline test scores



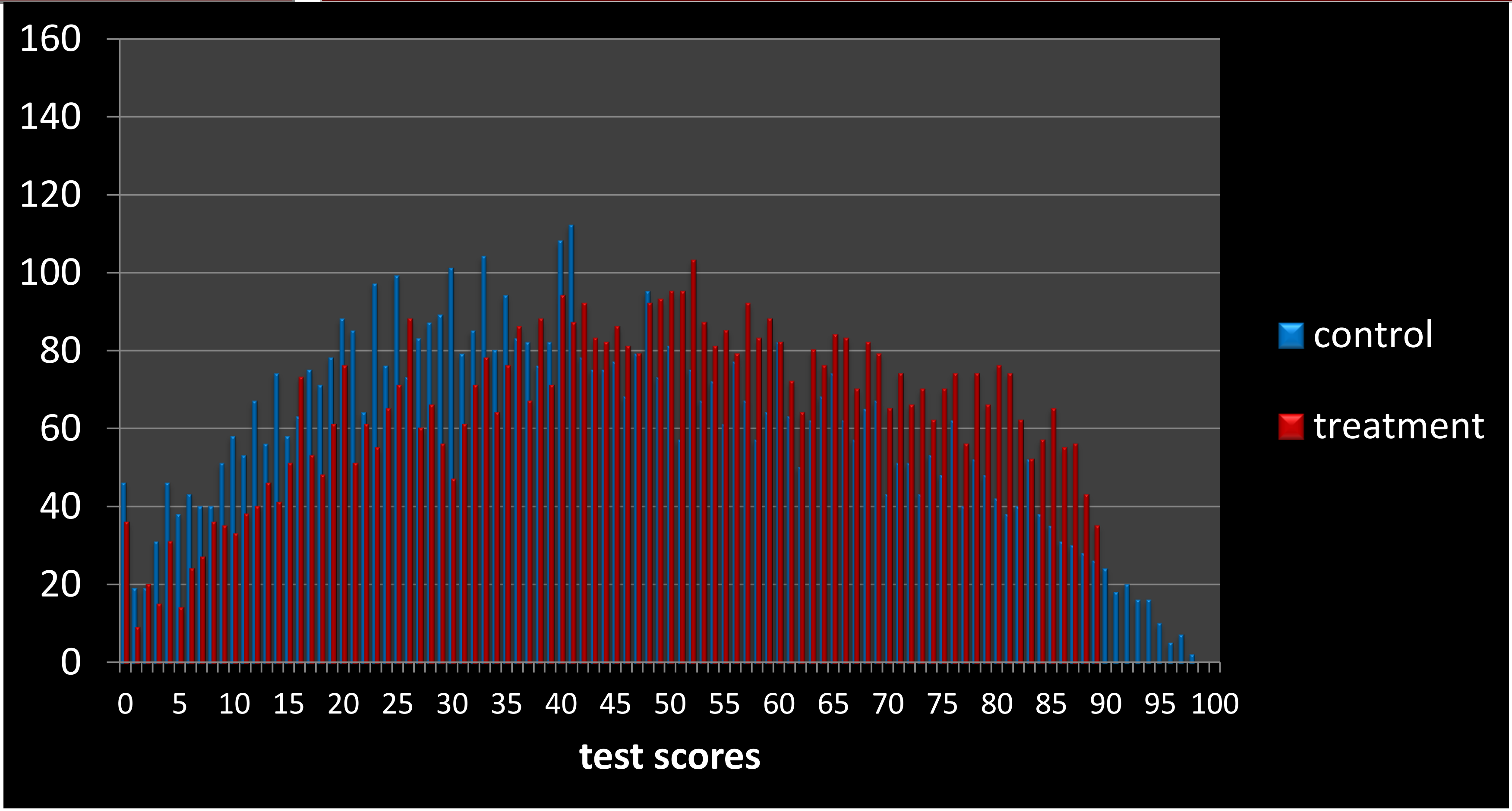
After the balsakhi programs, these are the endline test scores

The impact appears to be?

A. Positive
B. Negative
C. No impact
D. Don’t know

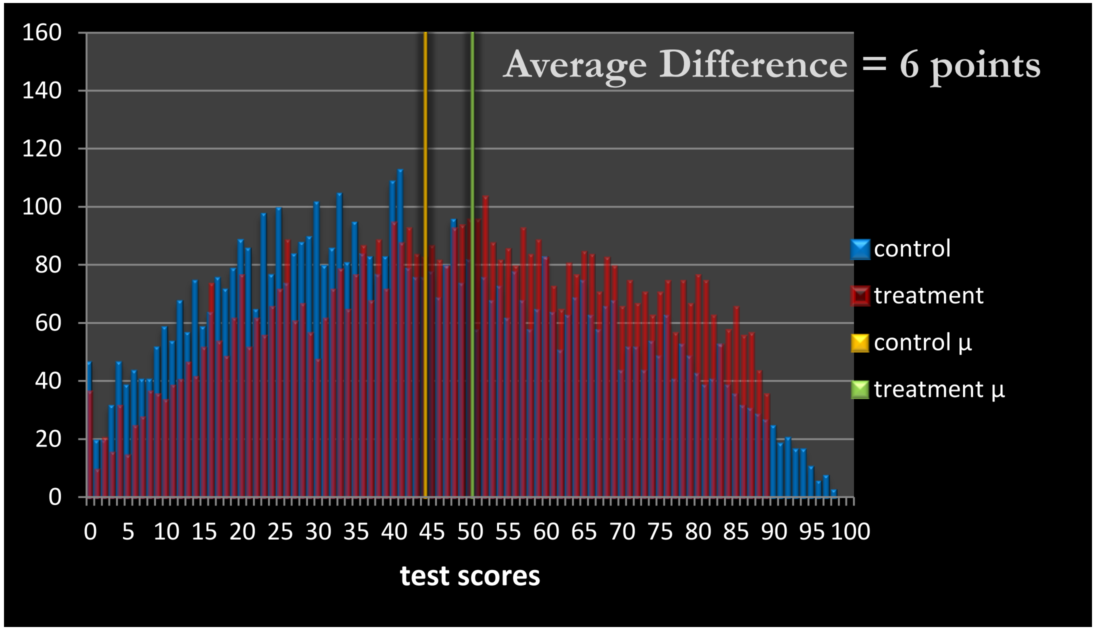


Post-test: control & treatment



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

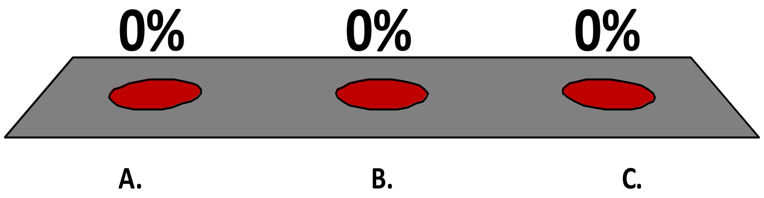
Is this impact statistically significant?



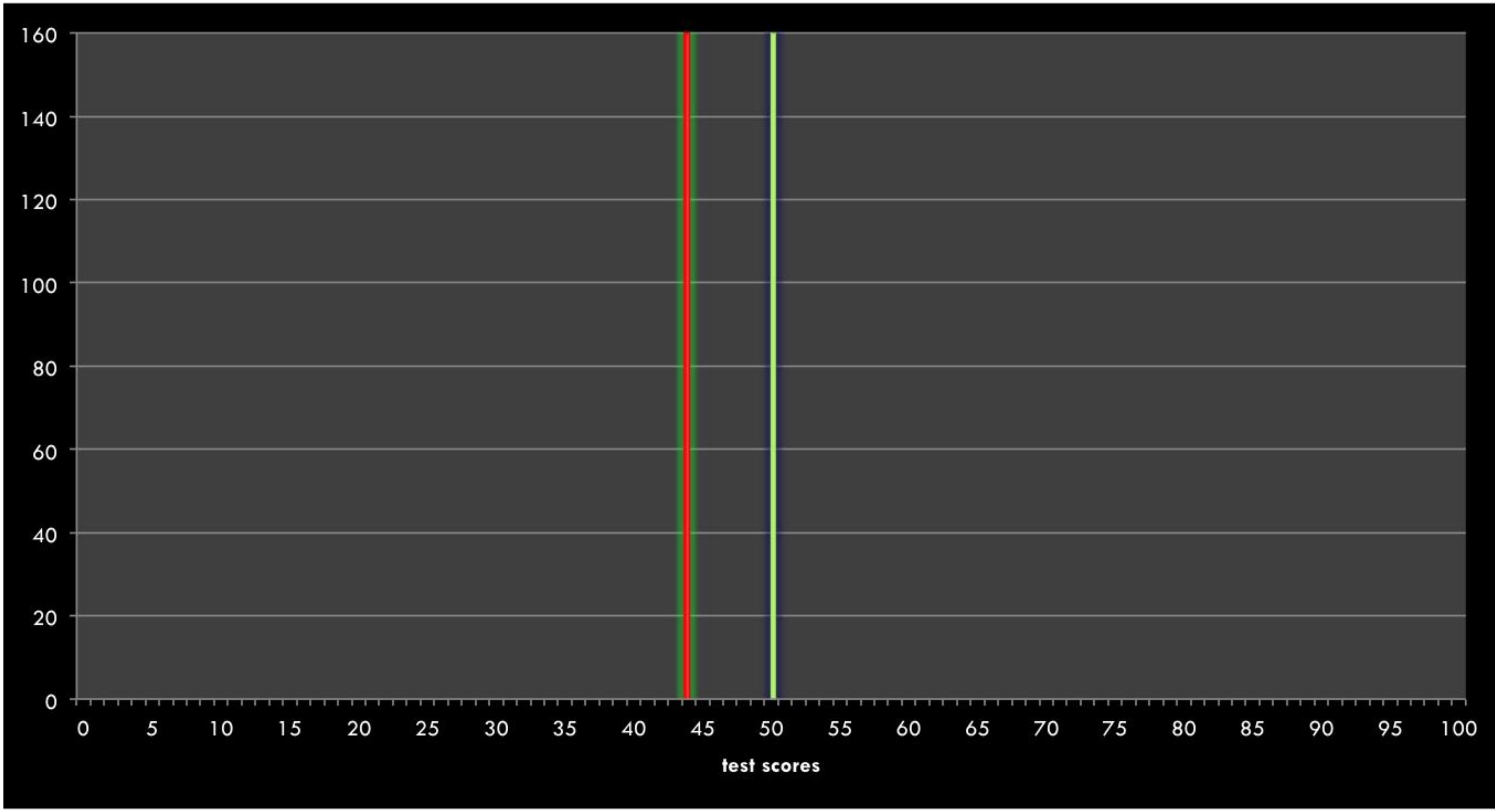
A. Yes

B. No

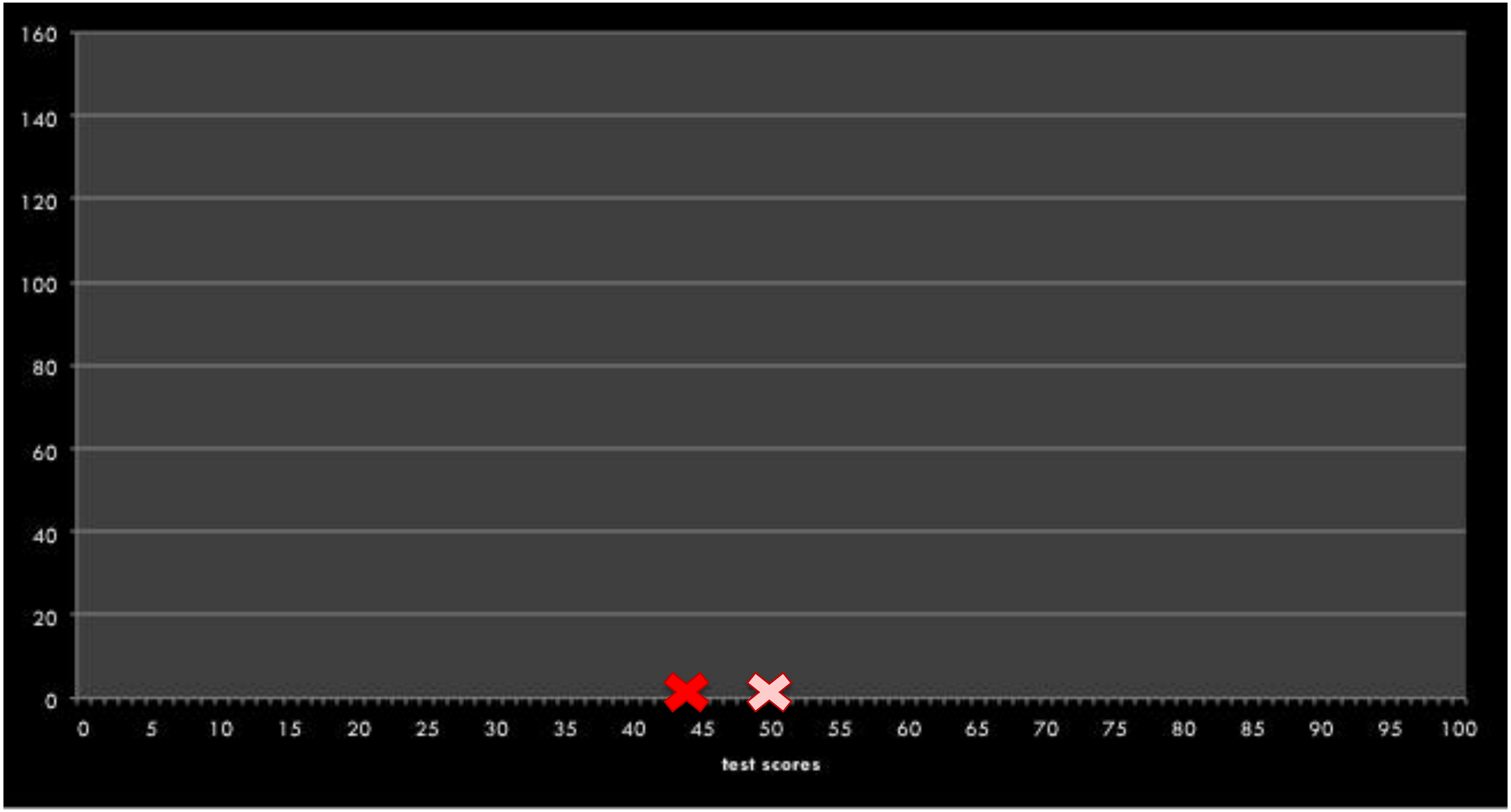
C. Don't know



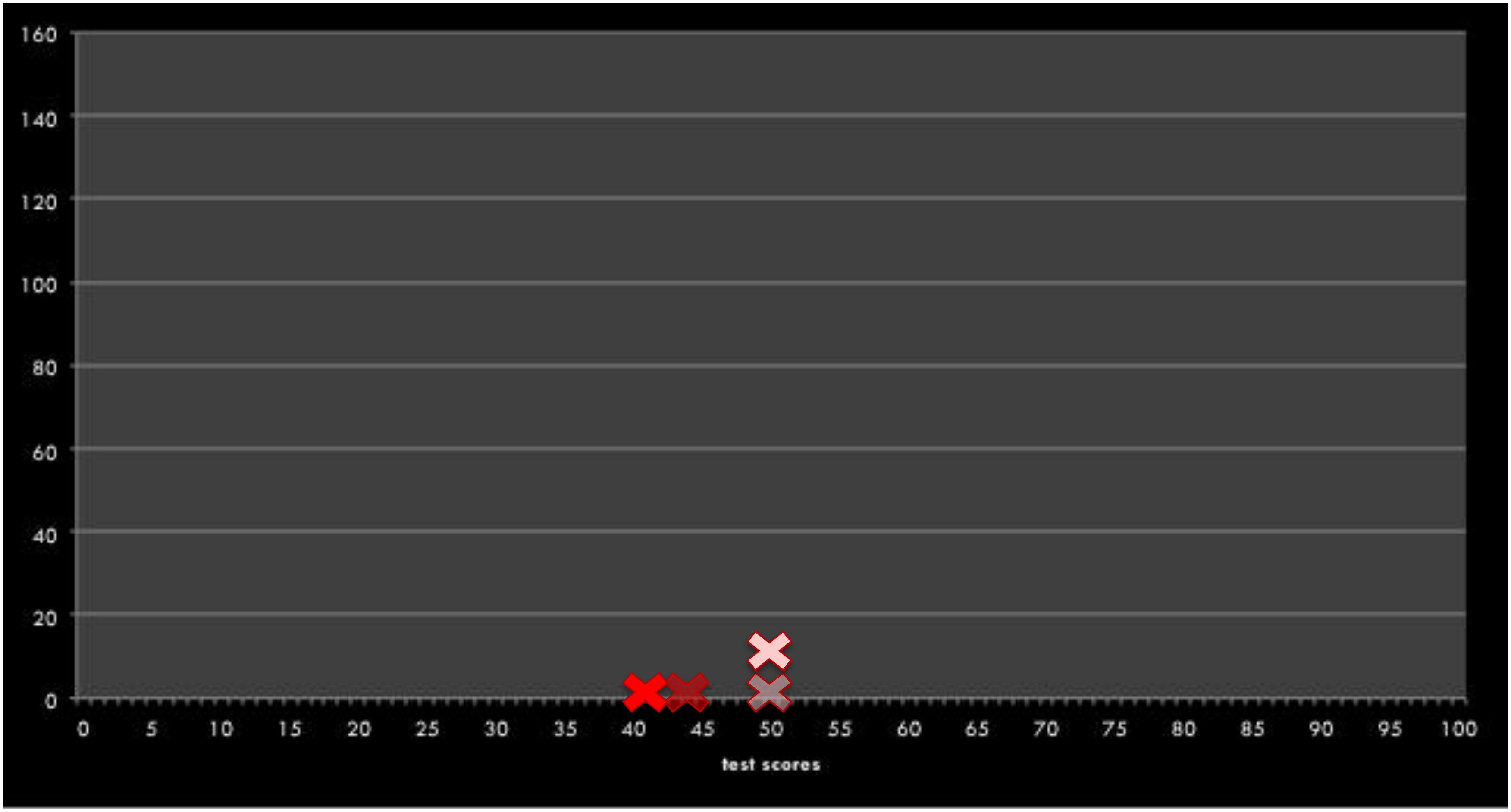
One experiment: 6 points



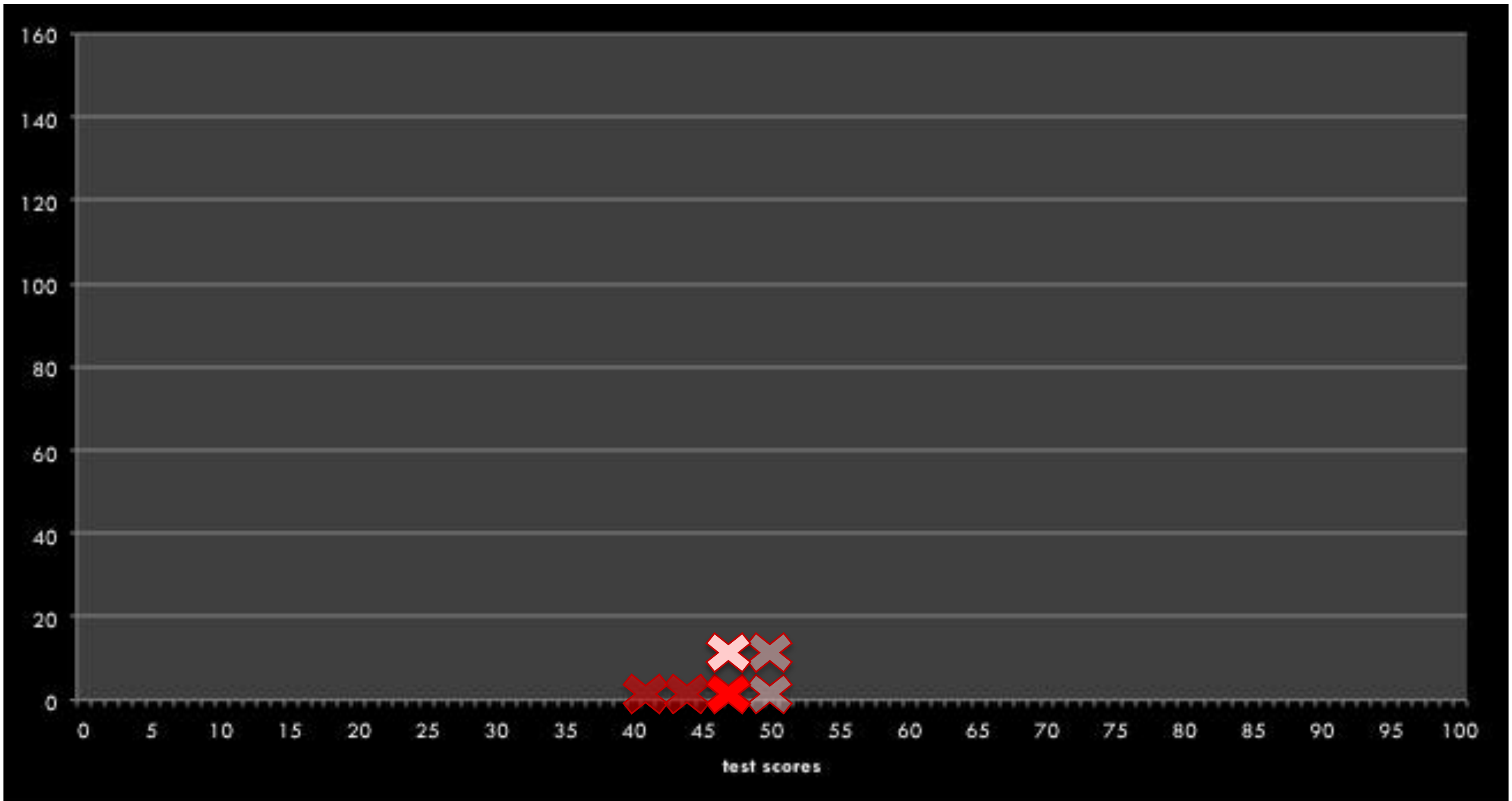
One experiment



Two experiments

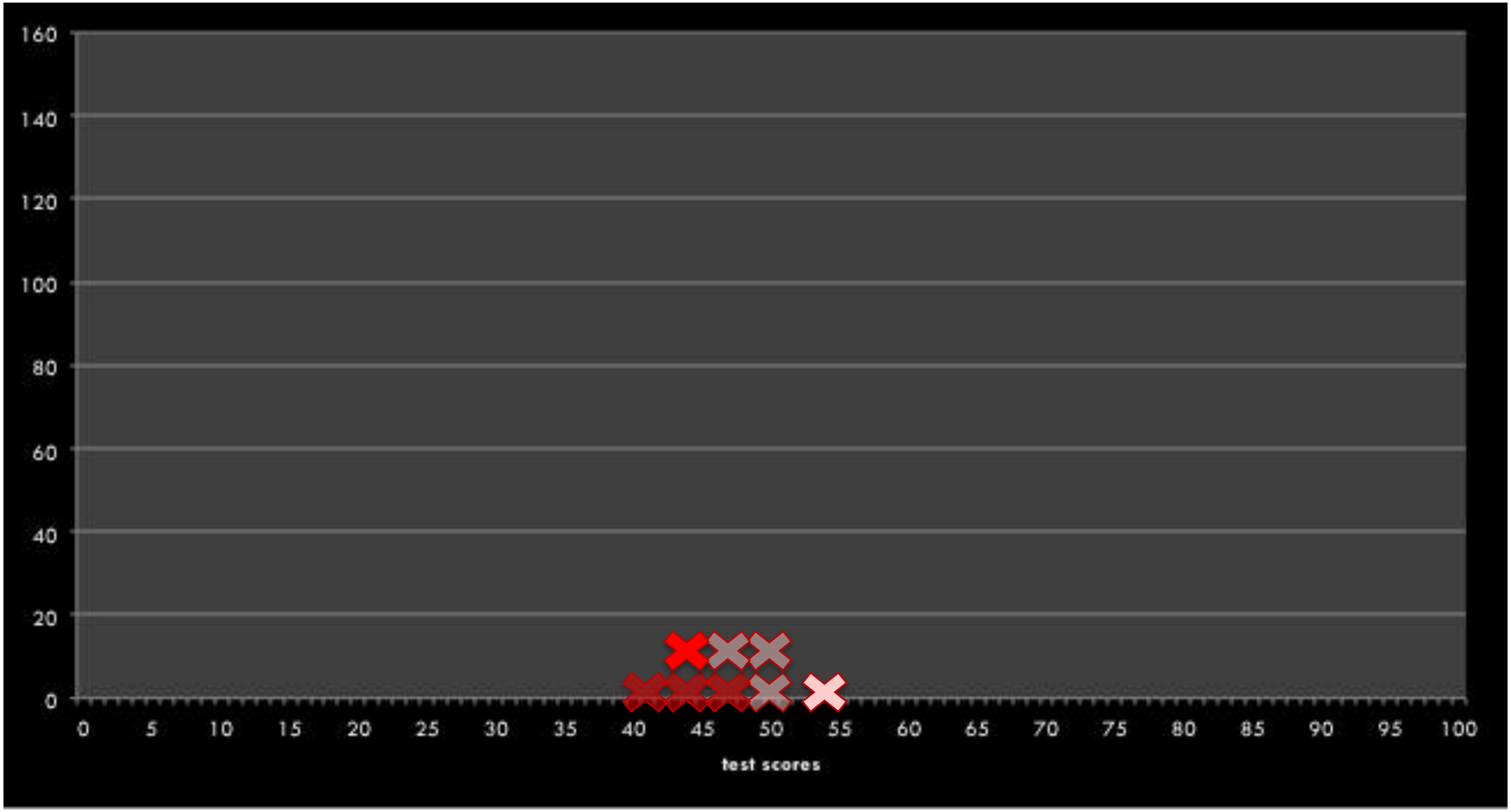


A few more…

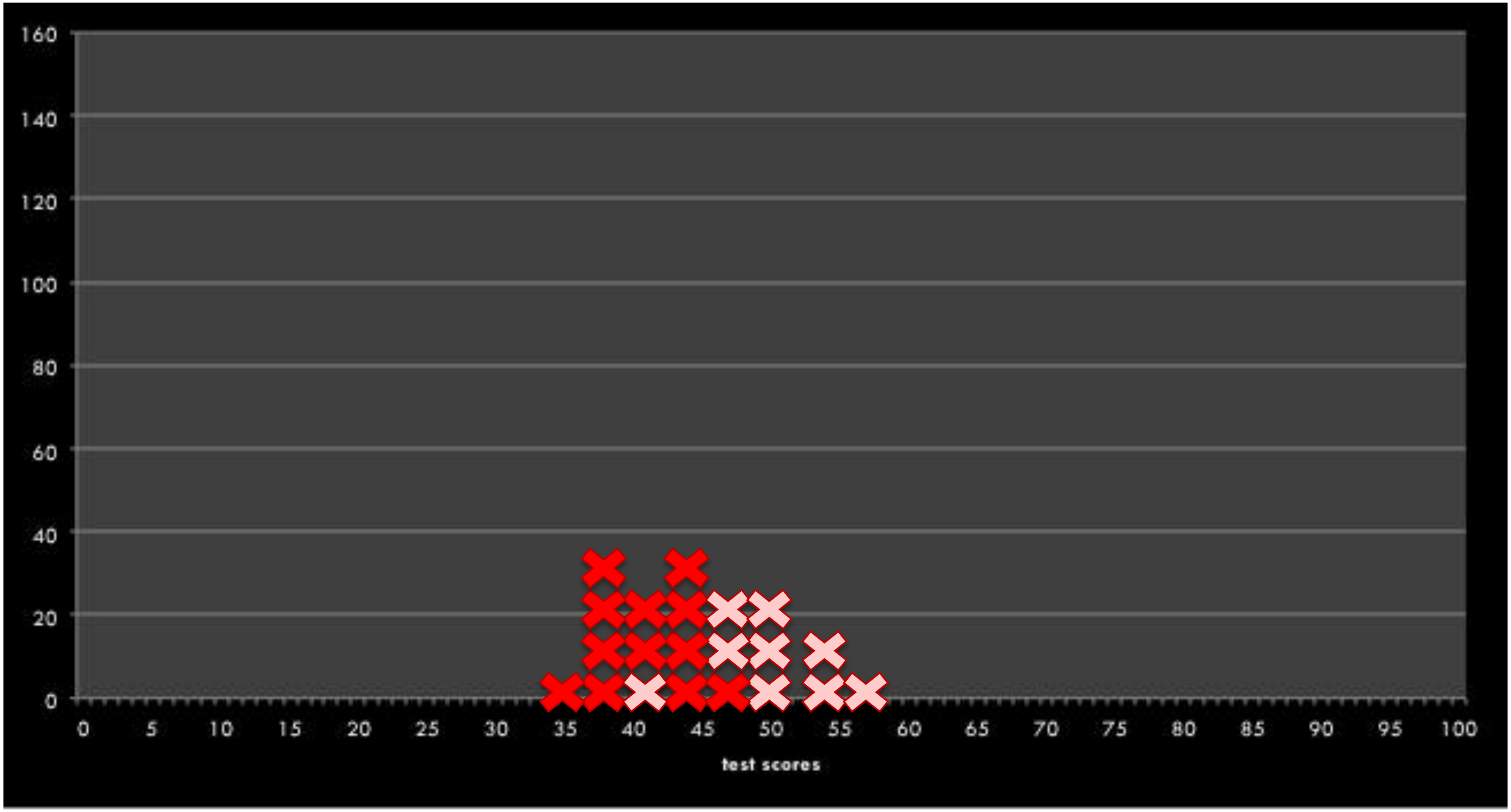


160
140
120
100
80
60
40
20
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
test scores

A few more…

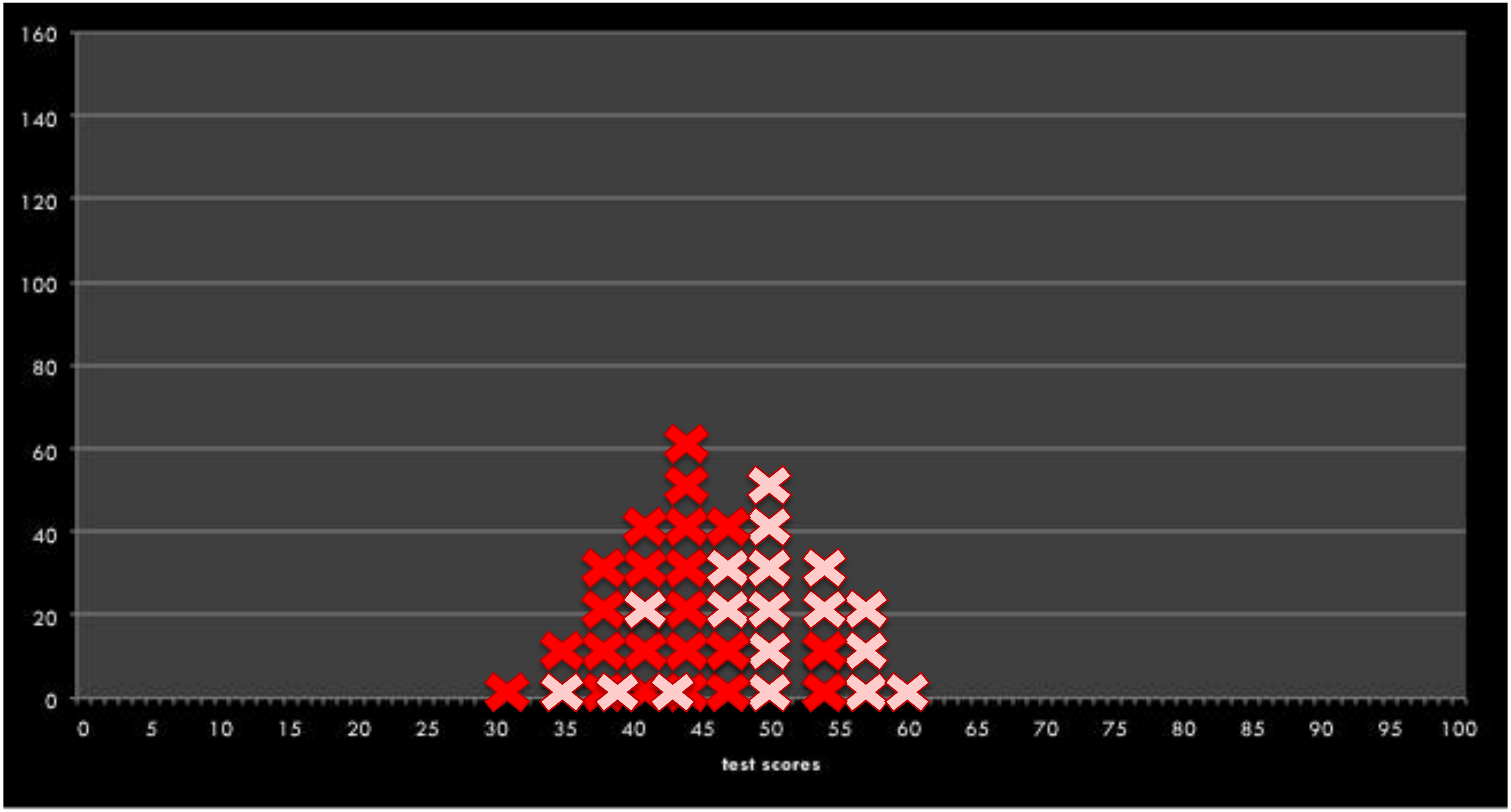

160
140
120
100
80
60
40
20
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
test scores

Many more…

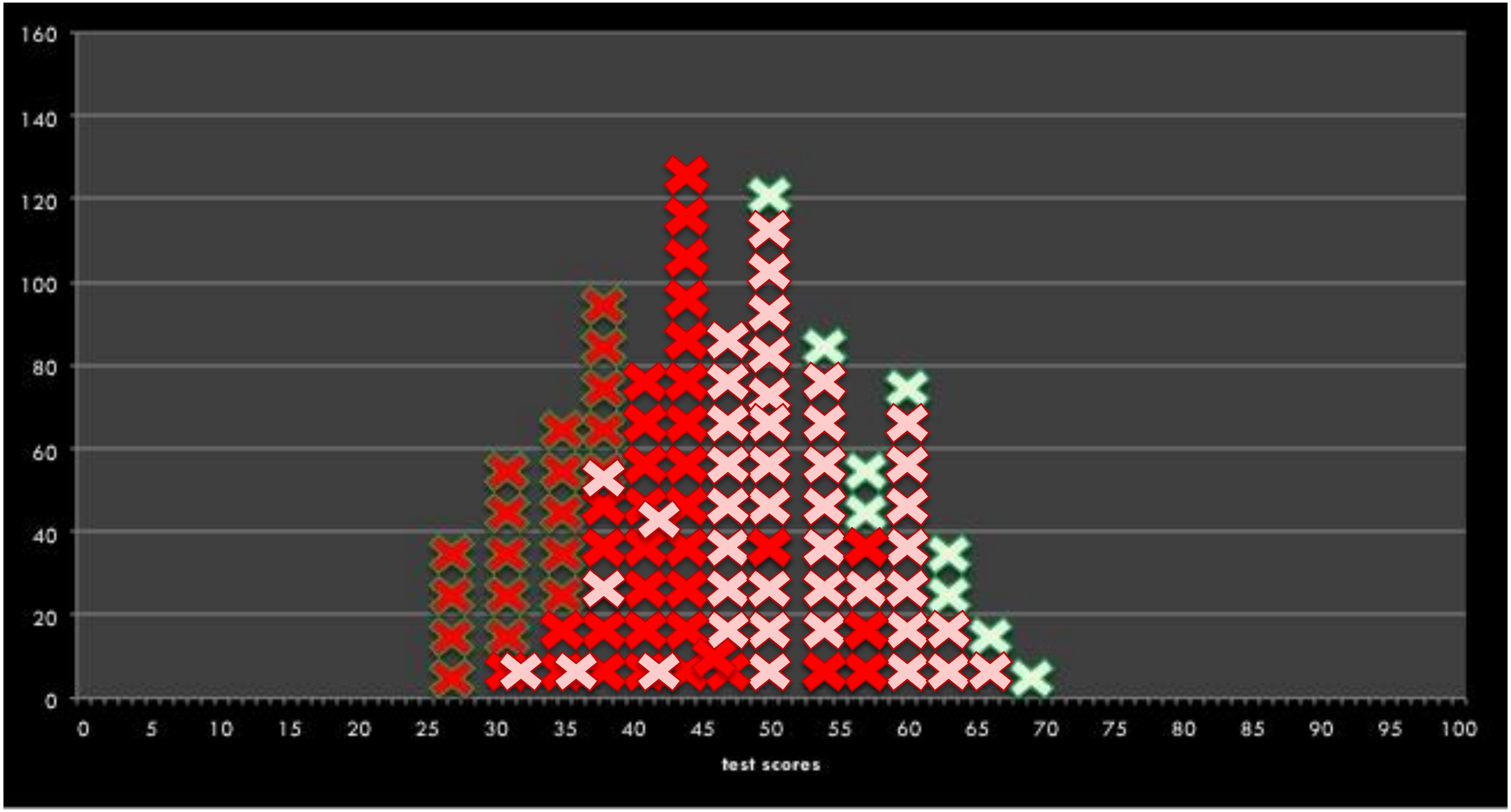


160
140
120
100
80
60
40
20
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
test scores

A whole lot more…

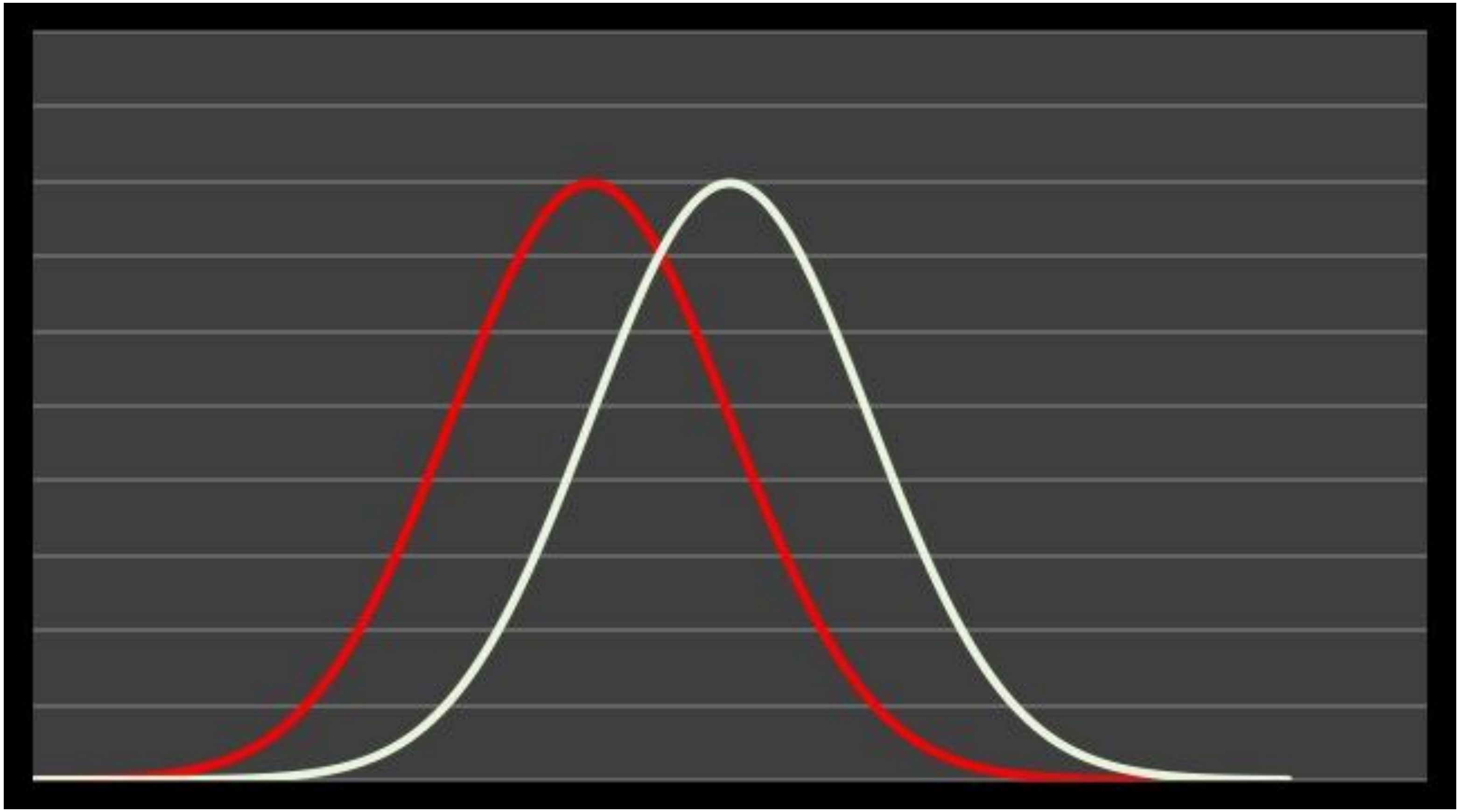


…



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

Running the experiment thousands of times…

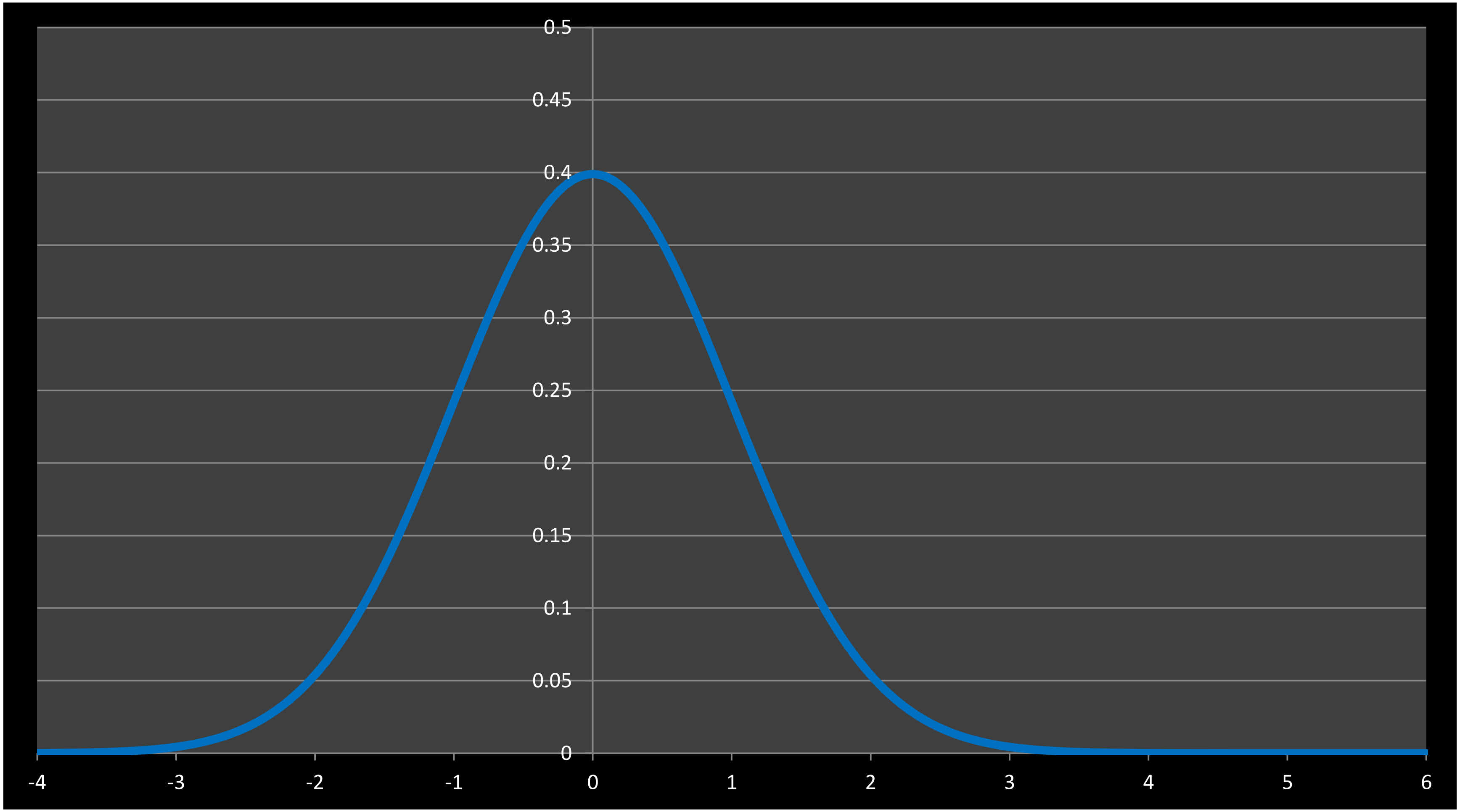


By the Central Limit Theorem, these are normally distributed

The assumption about your sample

The Central Limit Theorem and the Law of Large Numbers hold if the sample is **randomly sampled** from your population

Theoretical Sampling distribution



So let's look at hypothesis testing

- In criminal law, most institutions follow the rule: "innocent until proven guilty"
- In program evaluation, instead of "presumption of innocence," the rule is: "presumption of insignificance"
- The "Null hypothesis" (H_0) is that there was no (zero) impact of the program
- The burden of proof is on the evaluator to show a significant difference
  - Think about how this relates to the discussion of ethics on Sunday.

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

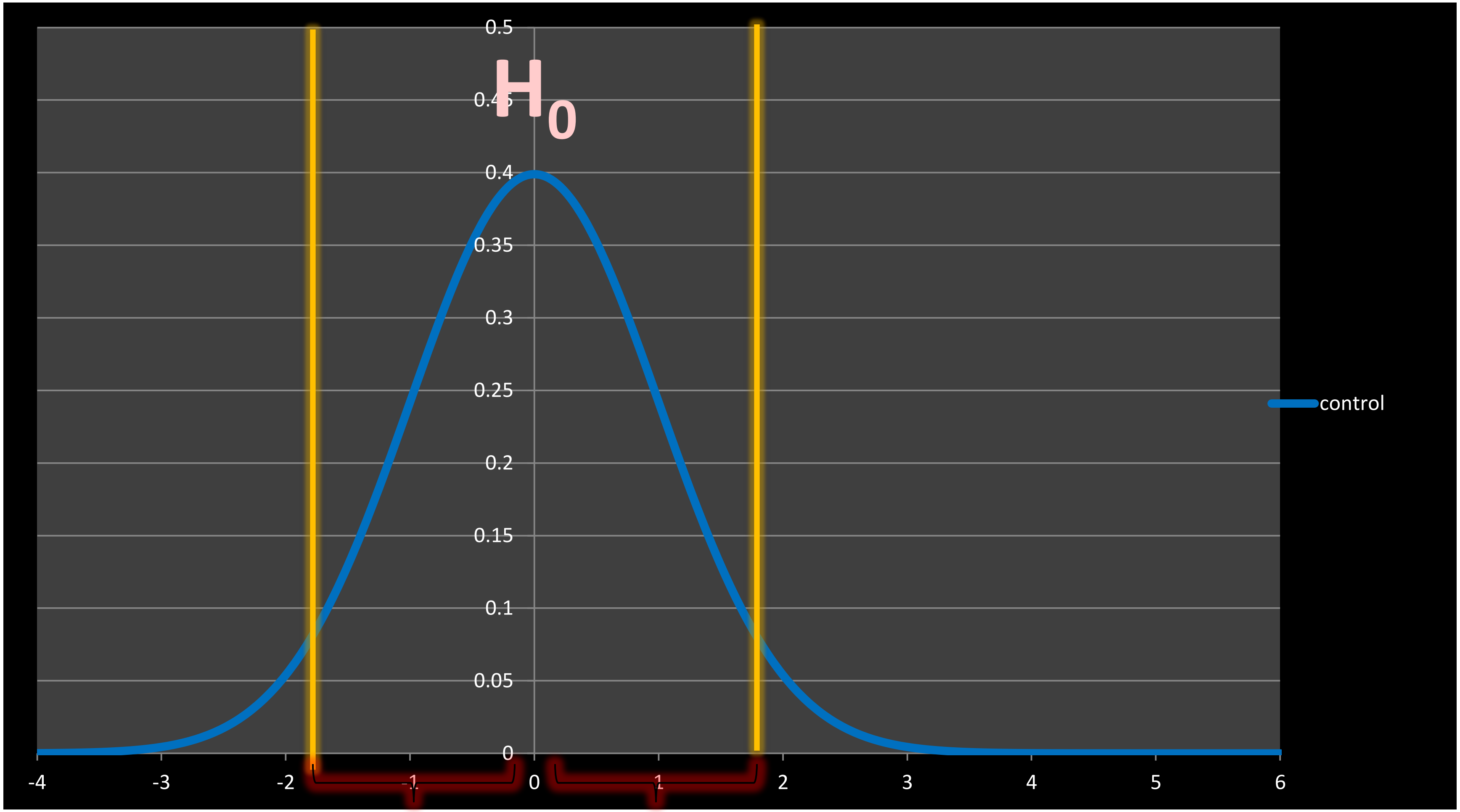
1. Determine the (size of the) sampling distribution around the null hypothesis H_0 by calculating the standard error

2. Choose the confidence interval, e.g. 95% (or significance level: α) (α=5%)

3. Identify the critical value (boundary of the confidence interval)

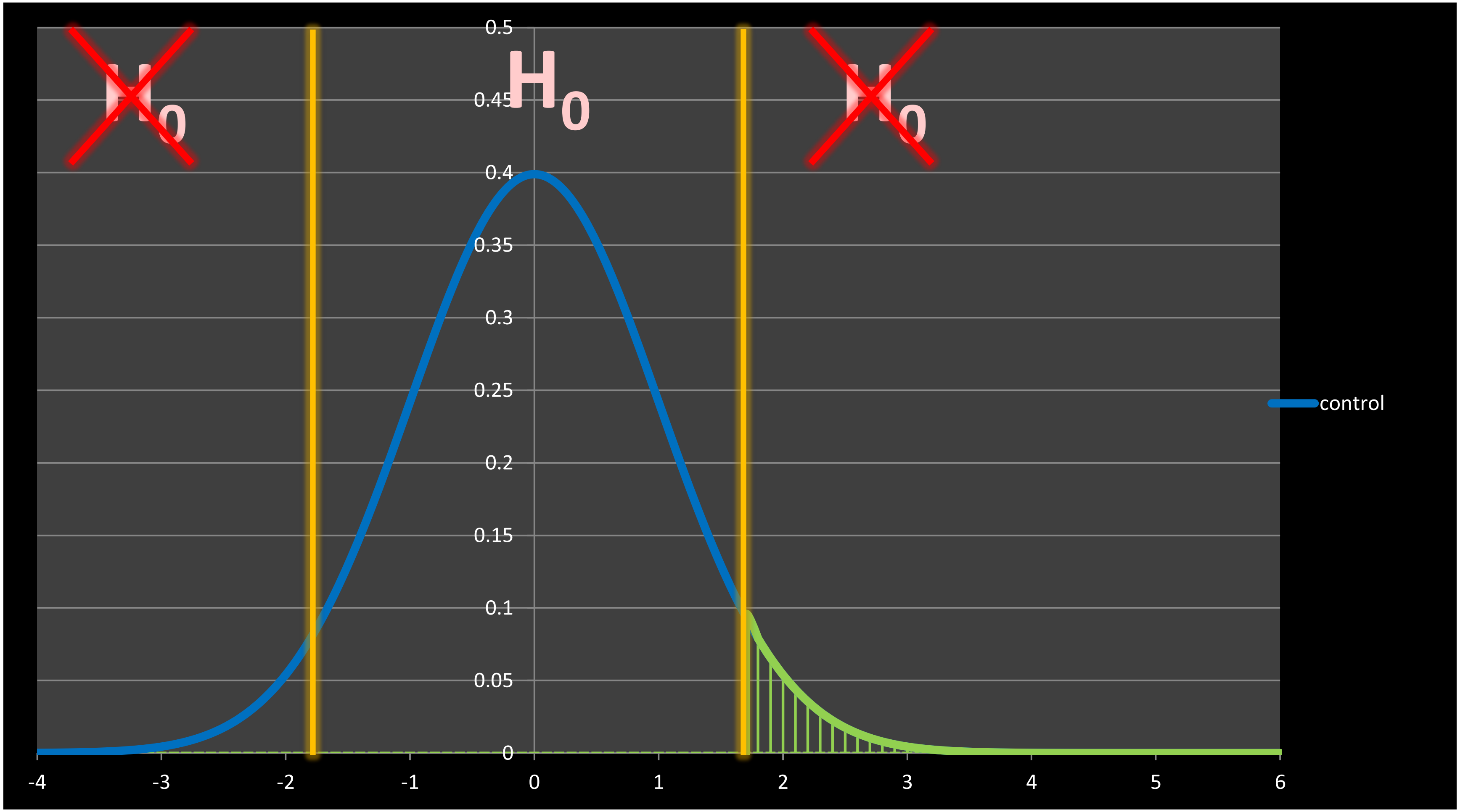
4. If our observation falls in the critical region we can reject the null hypothesis

Remember our 95% Confidence Interval?



1.96 SD **1.96 SD**

Impose significance level of 5%



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

What is Power?

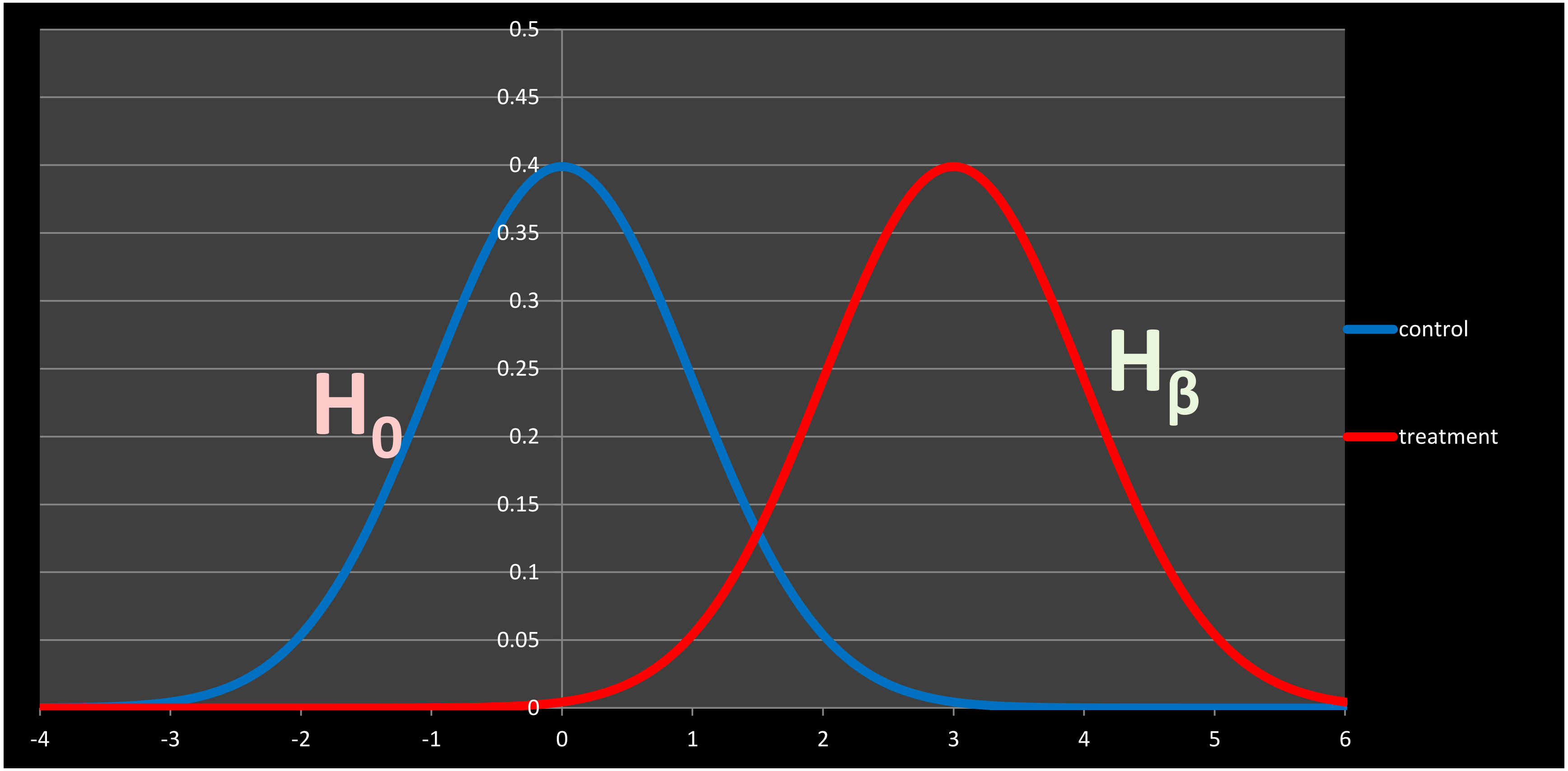
- **Type II Error:** Failing to reject the null hypothesis (concluding there is no difference), when indeed the null hypothesis is false.

- 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: 95% confidence

		YOU CONCLUDE	
		Effective	*No Effect*
THE TRUTH	*Effective*	☺	**Type II Error** (low power) ☹
	No Effect	**Type I Error** (5% of the time) ☹	☺

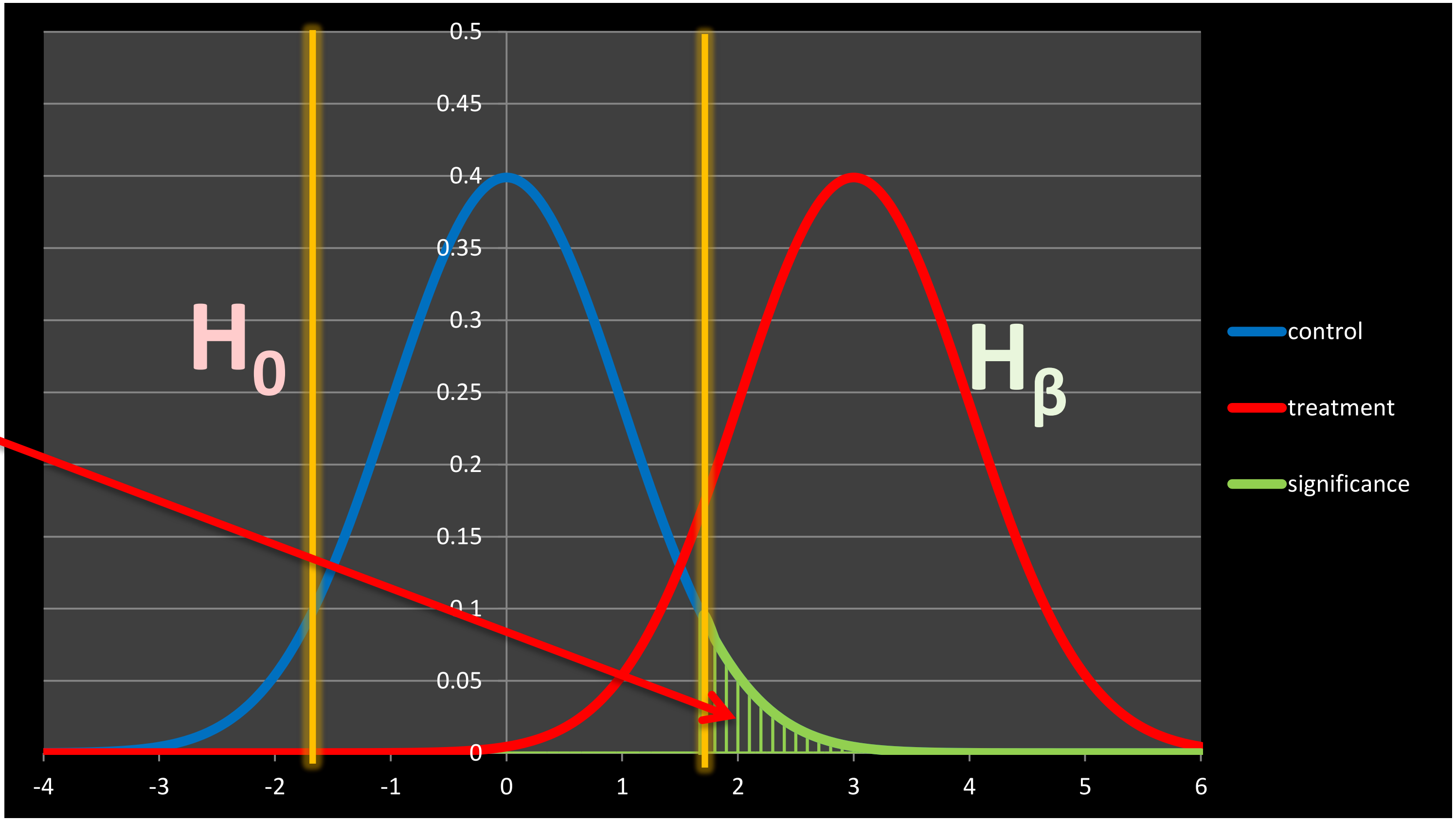
Before the experiment



Assume two effects: no effect and treatment effect β

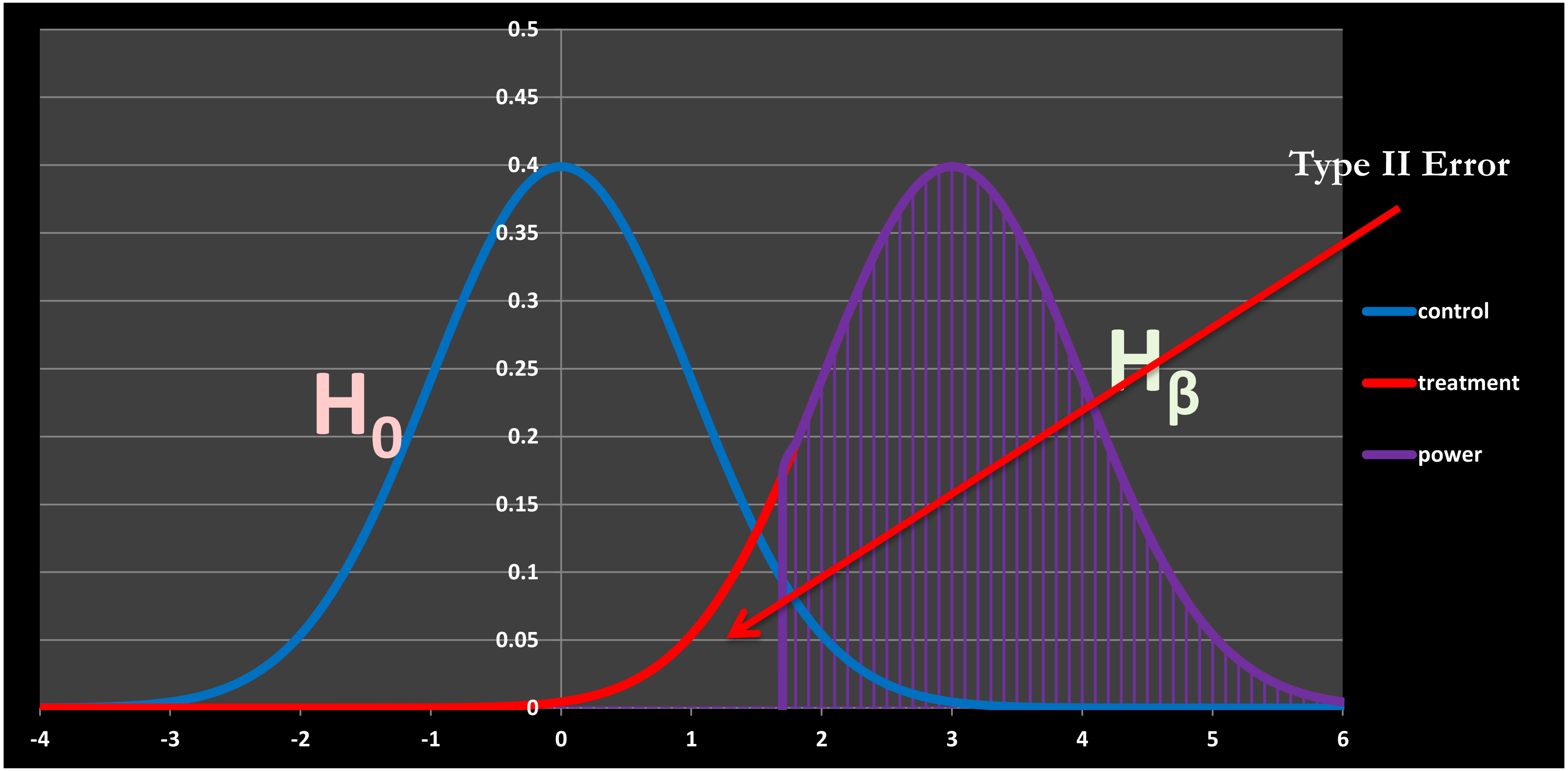
Impose significance level of 5%

Type I Error



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.

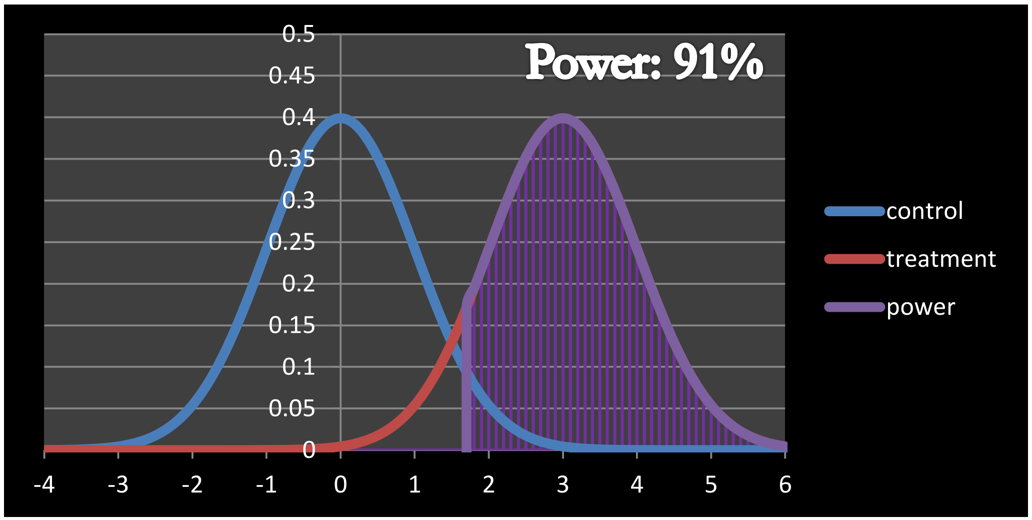
Power: main ingredients

1. Sample Size (N)
2. Effect Size (δ)
3. Variance (σ)
4. Proportion of sample in T vs. C
5. Clustering (ϱ)
6. Non-Compliance (akin to $\delta\downarrow$)

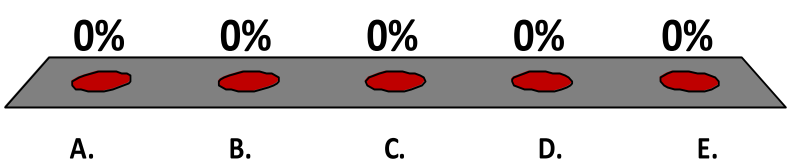
Power: main ingredients

1. **Sample Size (N)**
2. Effect Size (δ)
3. Variance (σ)
4. Proportion of sample in T vs. C
5. Clustering (ρ)
6. Non-Compliance (akin to $\delta\downarrow$)

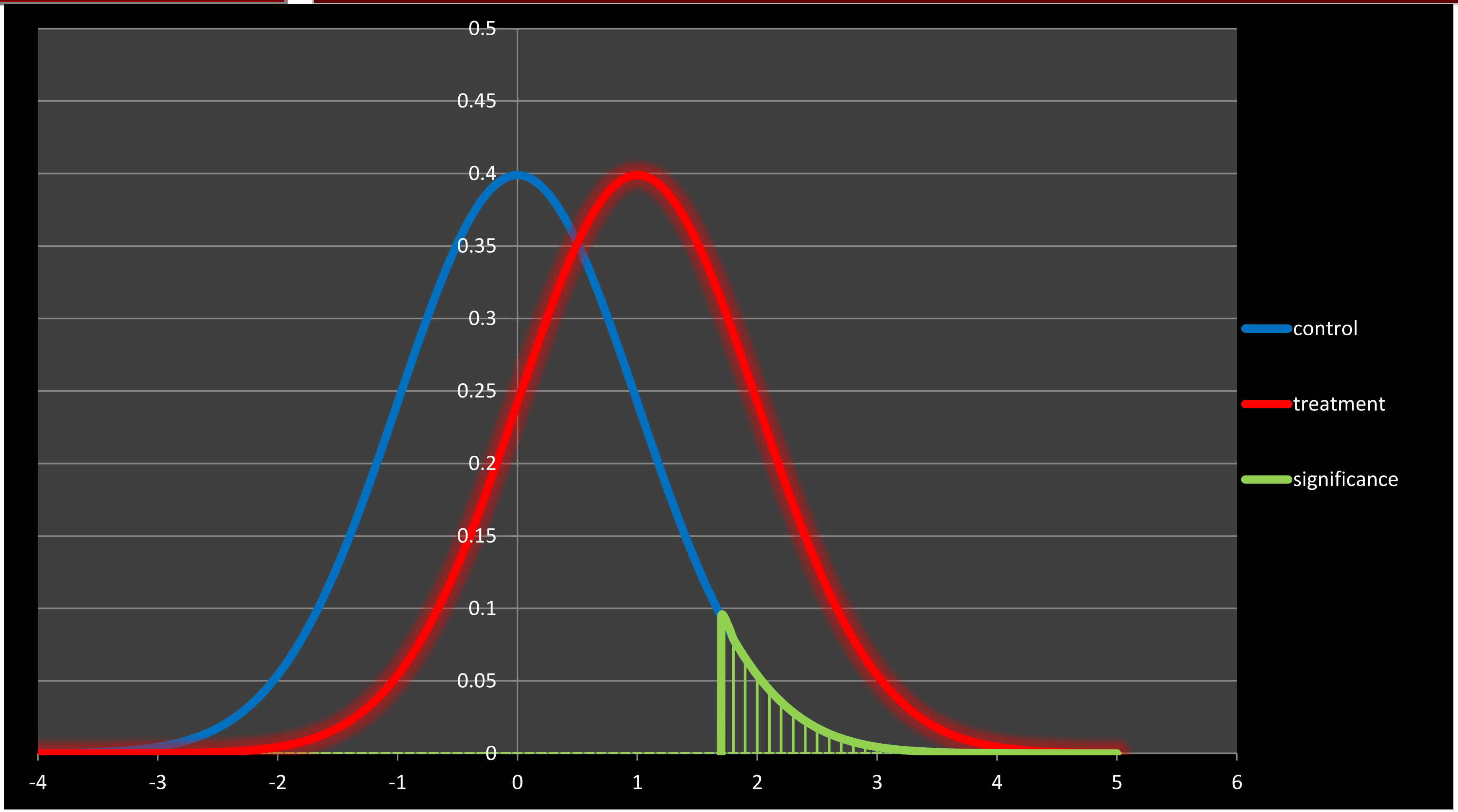
By increasing sample size you increase…



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

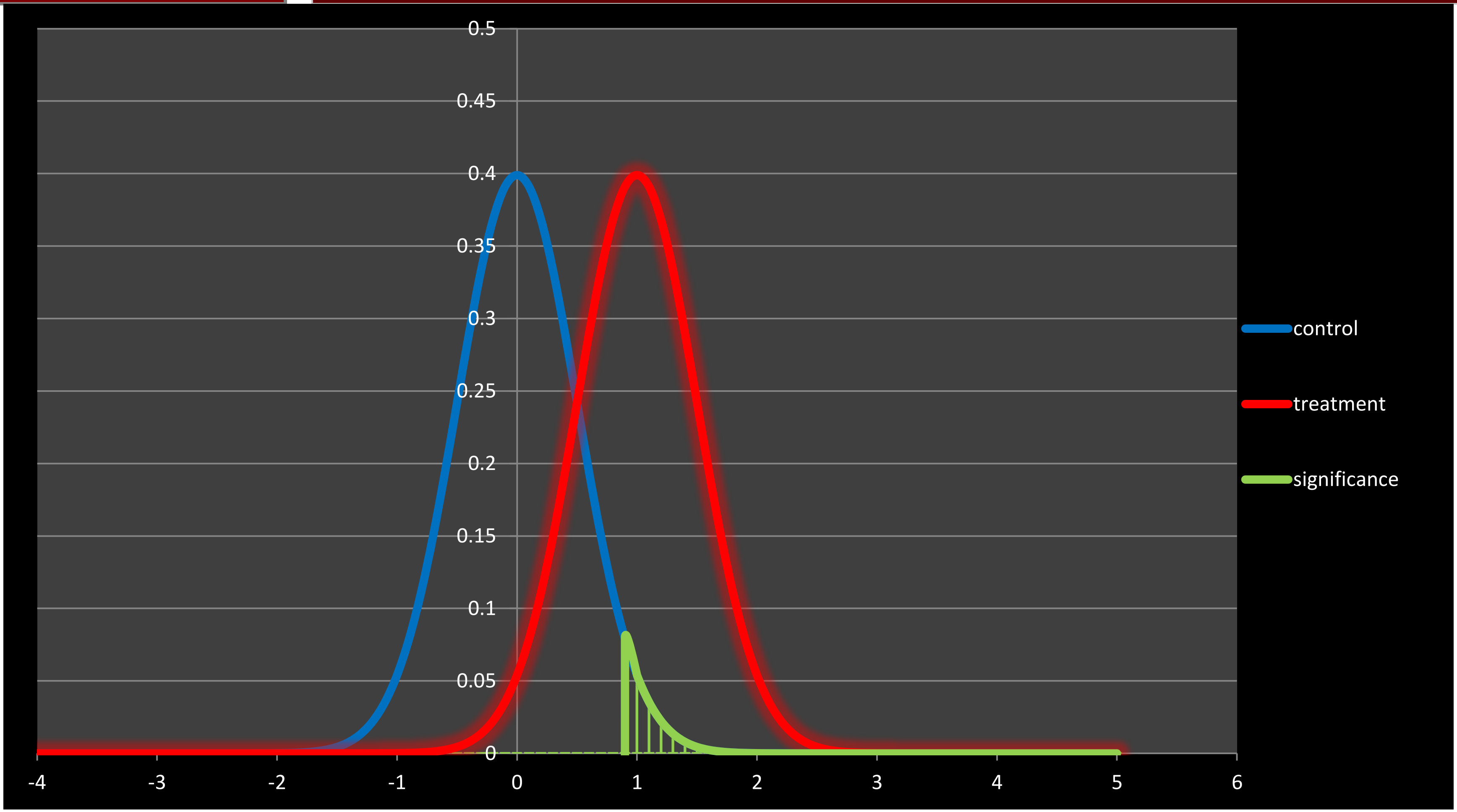


Power: Effect size = 1SE, Sample size = N

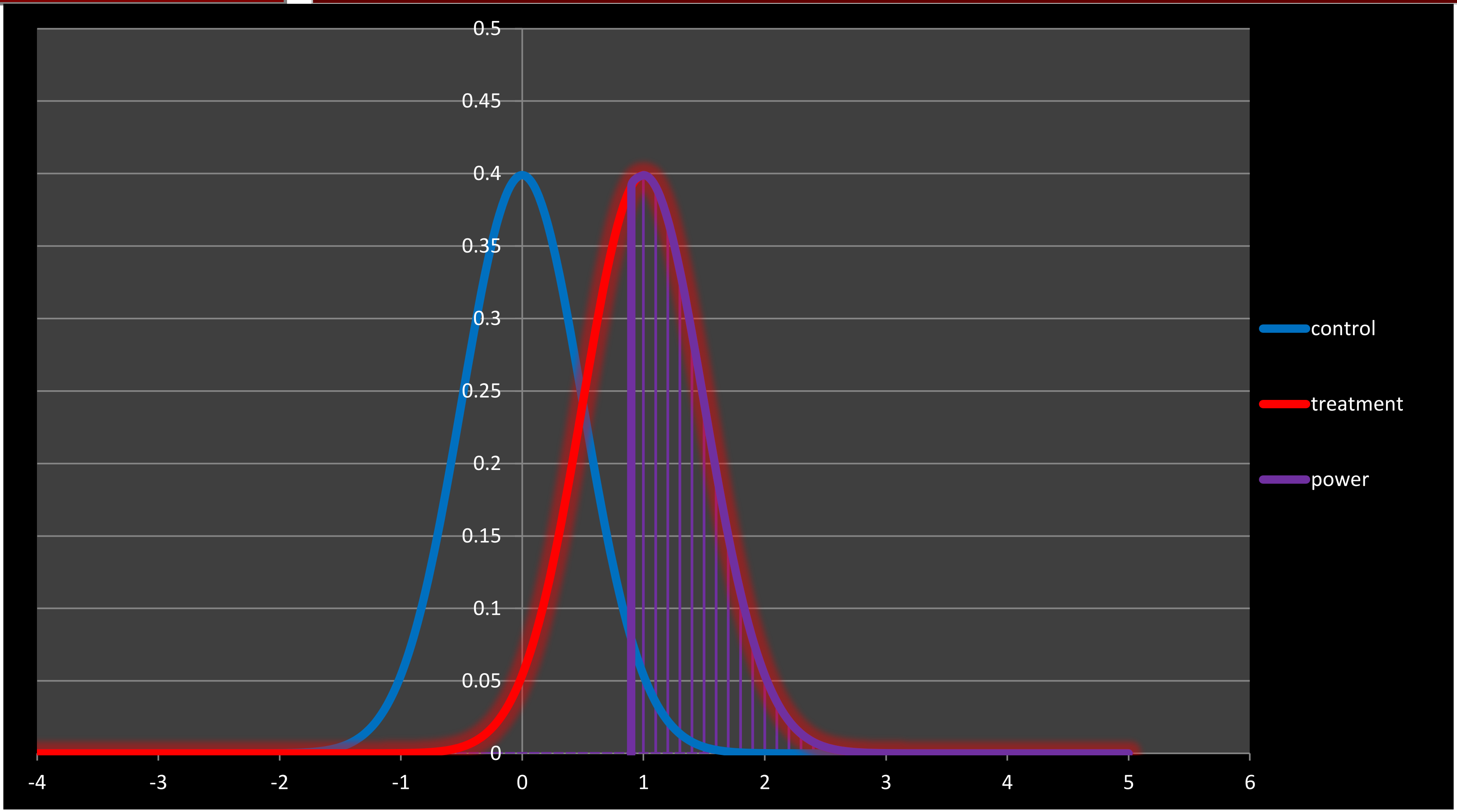


Remember, your sampling distribution becomes narrower as N↑

Power: Sample size = 4N

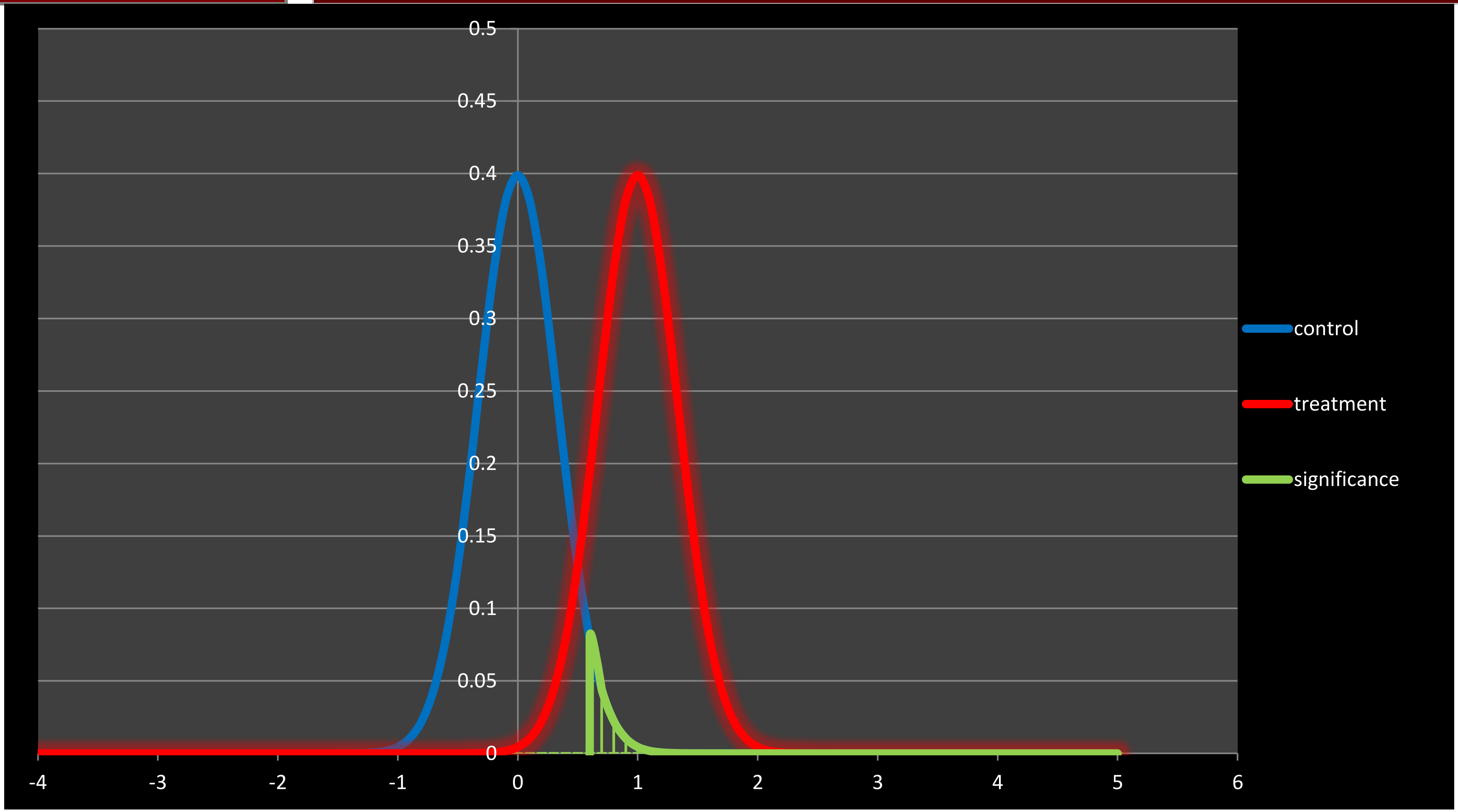


Power: 64%

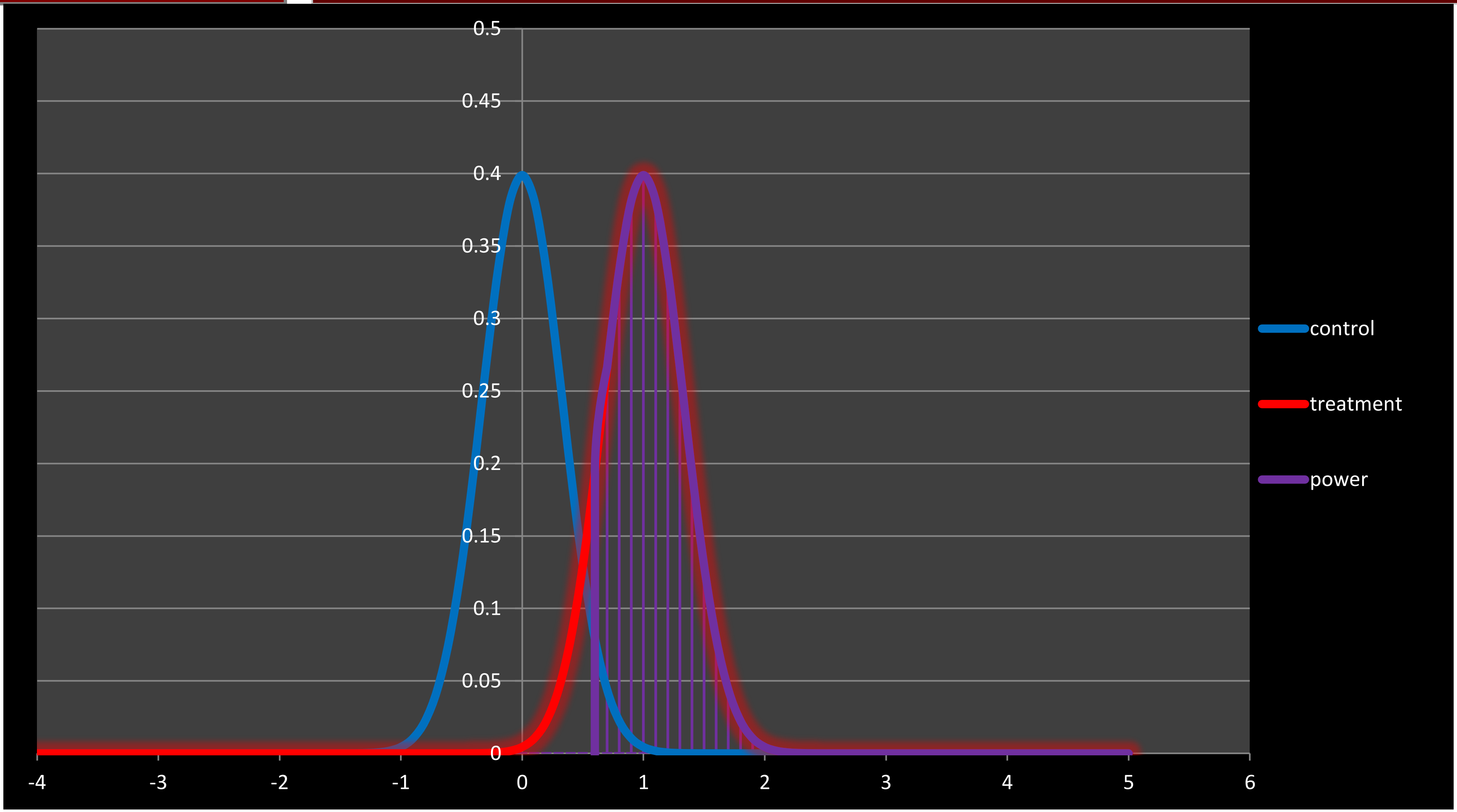


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6
control
treatment
power

Power: Sample size = 9N



Power: 91%

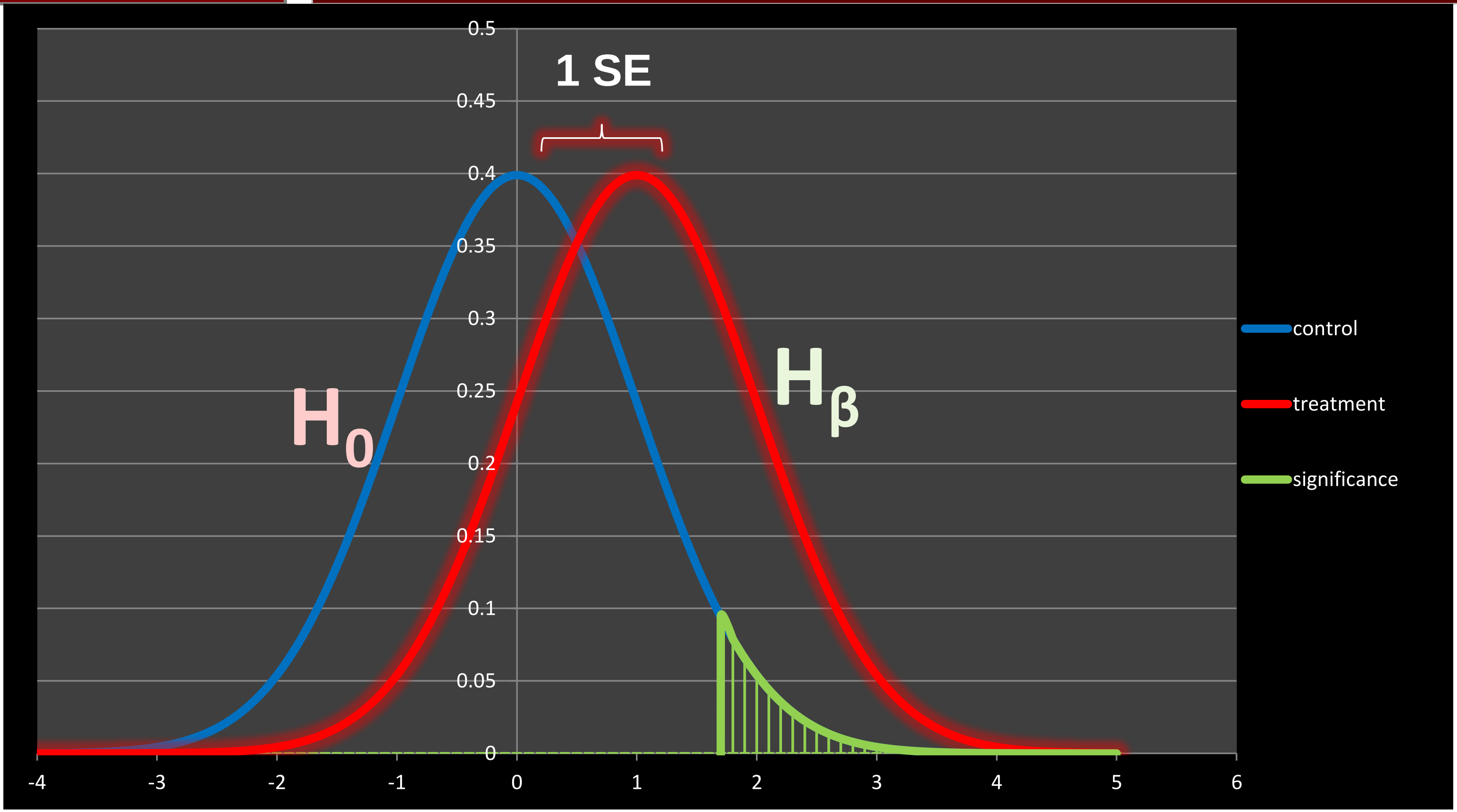


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6
control
treatment
power

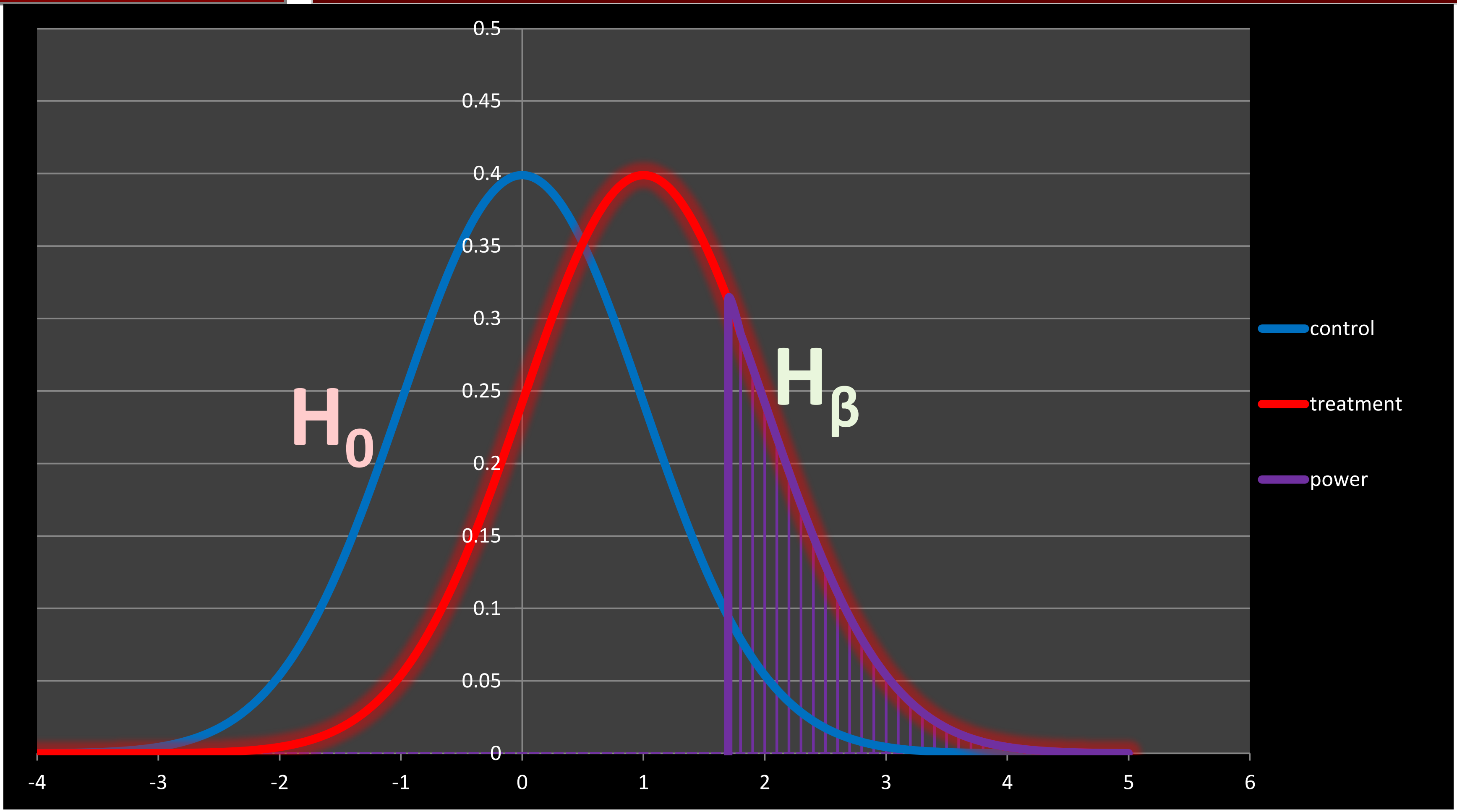
Power: main ingredients

1. Sample Size (N)
2. **Effect Size (δ)**
3. Variance (σ)
4. Proportion of sample in T vs. C
5. Clustering (ρ)
6. Non-Compliance (akin to δ↓)

Effect size = 1*SE

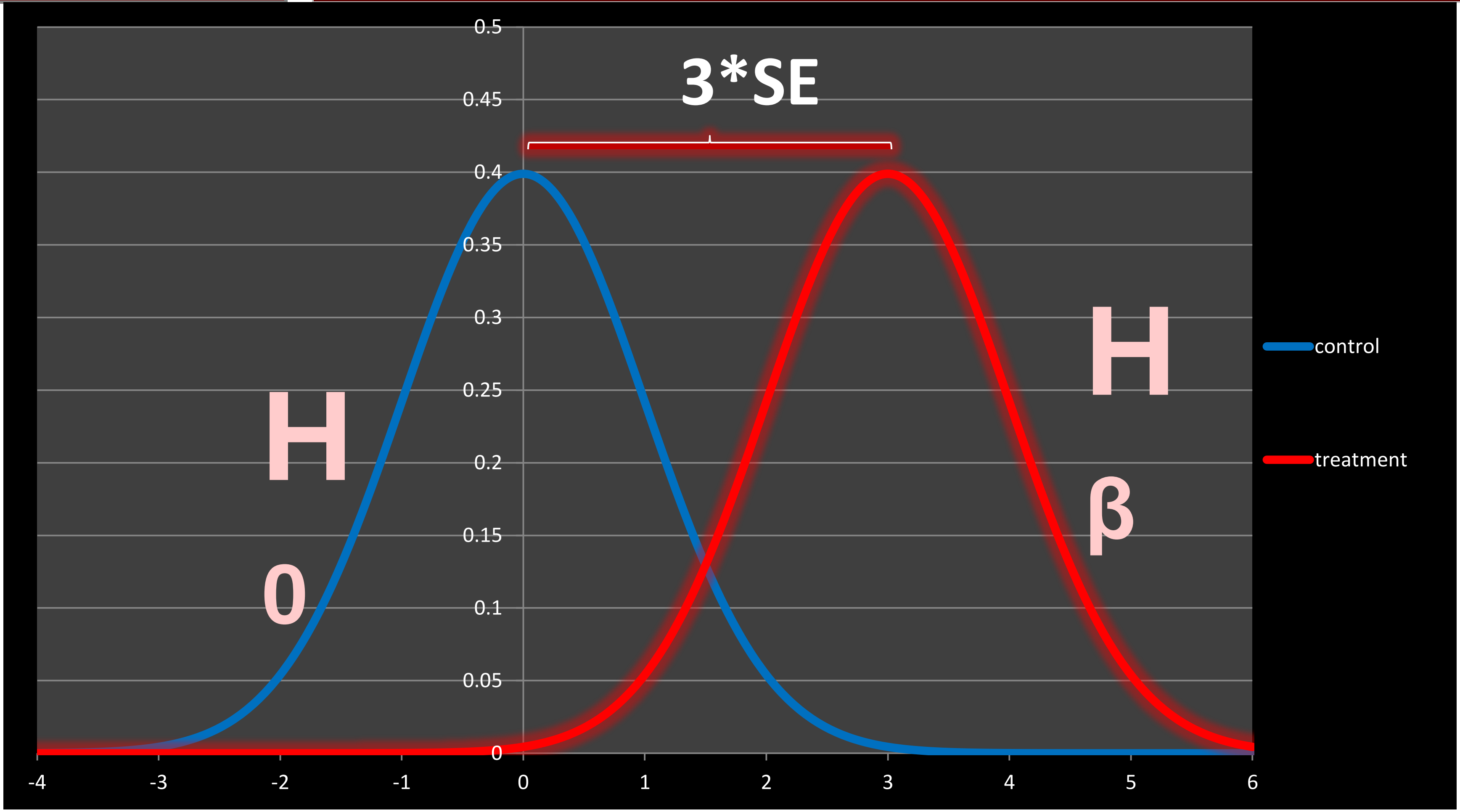


Effect size = 1*SE: Power = 26%



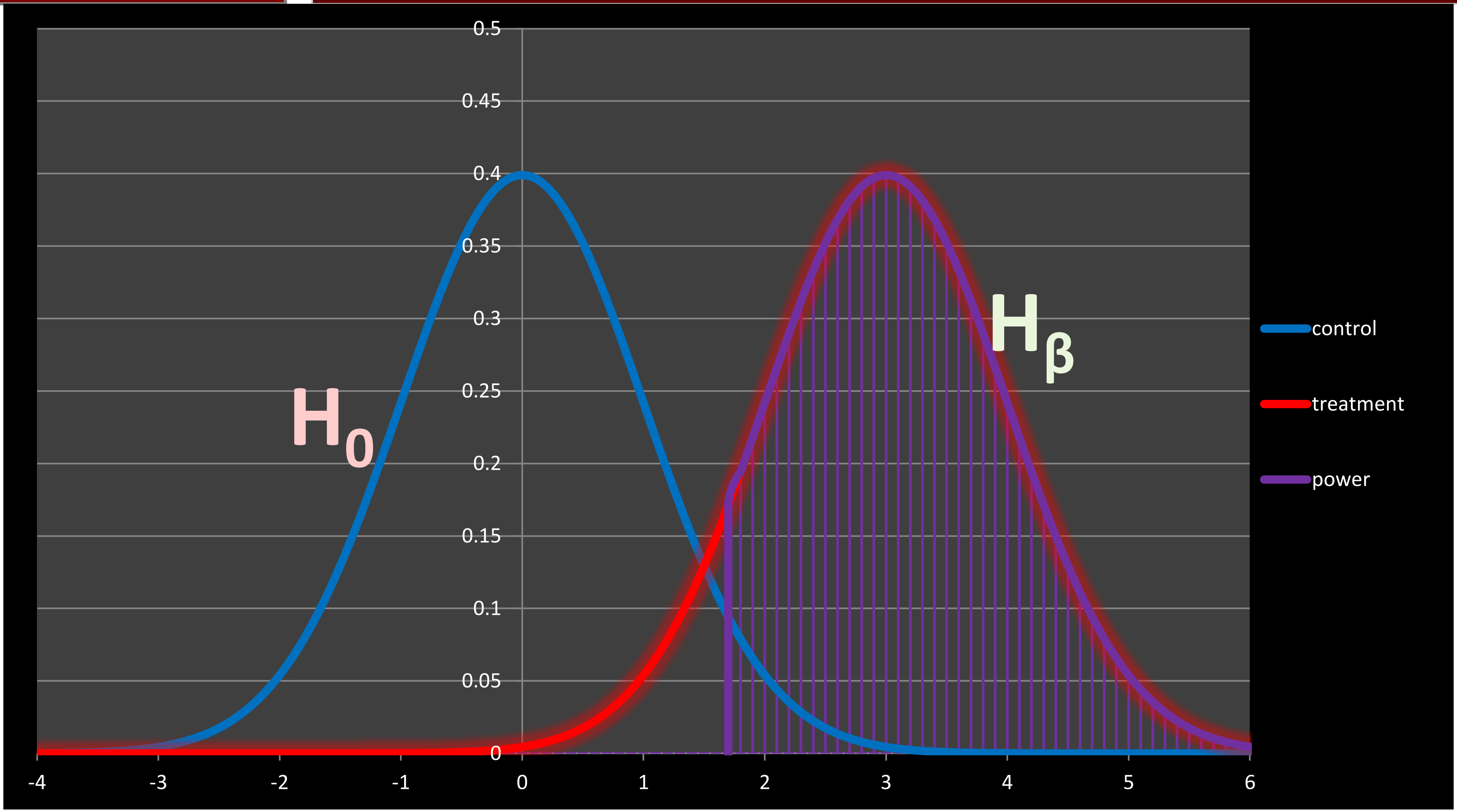
The Null Hypothesis would be rejected only 26% of the time

Effect size = 3*SE



Bigger hypothesized effect size→ distributions farther apart

Effect size = 3*SE: Power = 91%



Bigger Effect size means more power

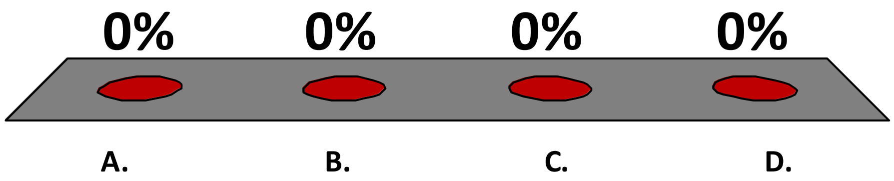
What effect size should you use when designing your experiment?

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

- What effect size should we pick while calculating the optimal sample size, assuming no other constraints?
- Ideally, we design our experiment to detect the **smallest effect size that is still interesting.**
  - Interesting, as long as the value of that answer is worth the cost of the evaluation.
- This is where "substantive significance" matters.

Power: main ingredients

1. Sample Size (N)
2. Effect Size (δ)
3. **Variance (σ)**
4. Proportion of sample in T vs. C
5. Clustering (ρ)
6. Non-Compliance (akin to δ↓)

Variance

- There is sometimes very little we can do to reduce the noise
- The underlying variance is what it is- just a characteristic of the population at hand!
- 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

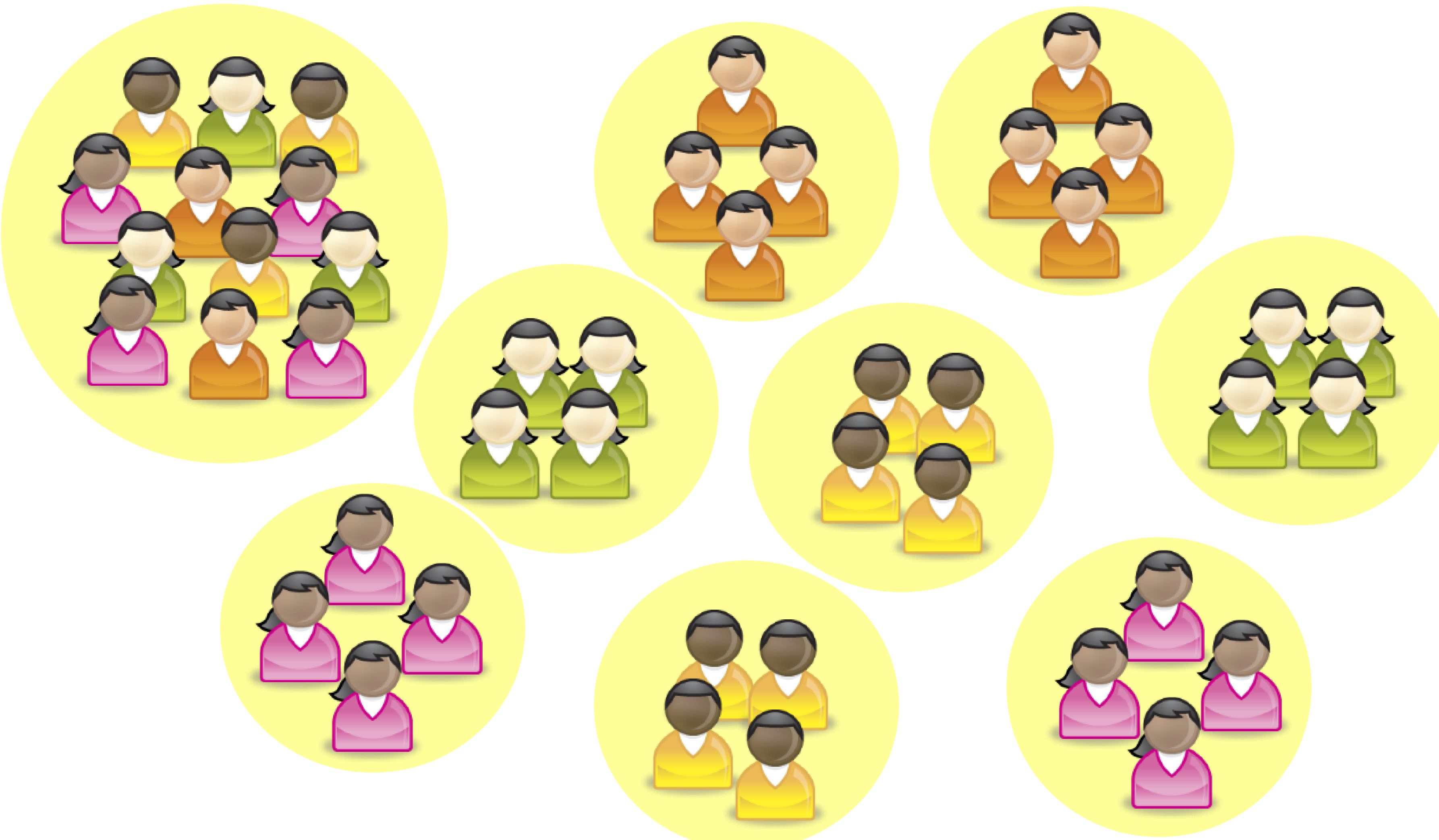
Power: main ingredients

1. Sample Size (N)
2. Effect Size (δ)
3. Variance (σ)
4. Proportion of sample in T vs. C
5. **Clustering (ϱ)**
6. Non-Compliance (akin to $\delta\downarrow$)

Clustered design: intuition

- You want to know how close the upcoming state elections will be
- Method 1: Randomly select 50 people from entire state (N=50)
- Method 2: Randomly select 5 families in the state, and ask ten members of each family their opinion (N=50)

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



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

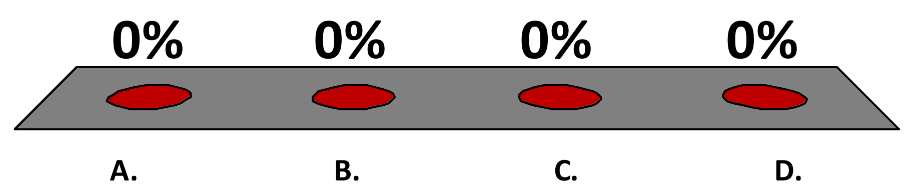
All uneducated people live in one village. People with only primary education live in another. College grads live in a third, etc. ICC (ρ) on education will be..

A. High

B. Low

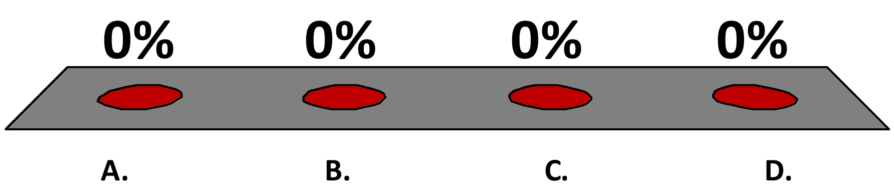
C. No effect on rho

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



BONUS SLIDES (TIME PERMITTING…)

Testing multiple treatments

Control Group

Balsakhi

↑ 0.15 SD ↓

↖ ↗ 0.05 SD

100

← 0.10 SD →

100

↑ 0.10 SD ↓

CAL program

Balsakhi + CAL

Power: main ingredients

1. Sample Size (N)
2. Effect Size (δ)
3. Variance (σ)
4. **Proportion of sample in T vs. C**
5. Clustering (ρ)
6. Non-Compliance (akin to δ↓)

Power!

Significance Level

Variance

Effect Size

Power

$$EffectSize = \left(t_{(1-\kappa)} + t_{\alpha}\right) * \sqrt{\frac{1}{P(1-P)}} * \sqrt{\frac{\sigma^{2}}{N}} \sqrt{1 + \rho(m-1)}$$

Proportion in Treatment

Sample Size

ICC

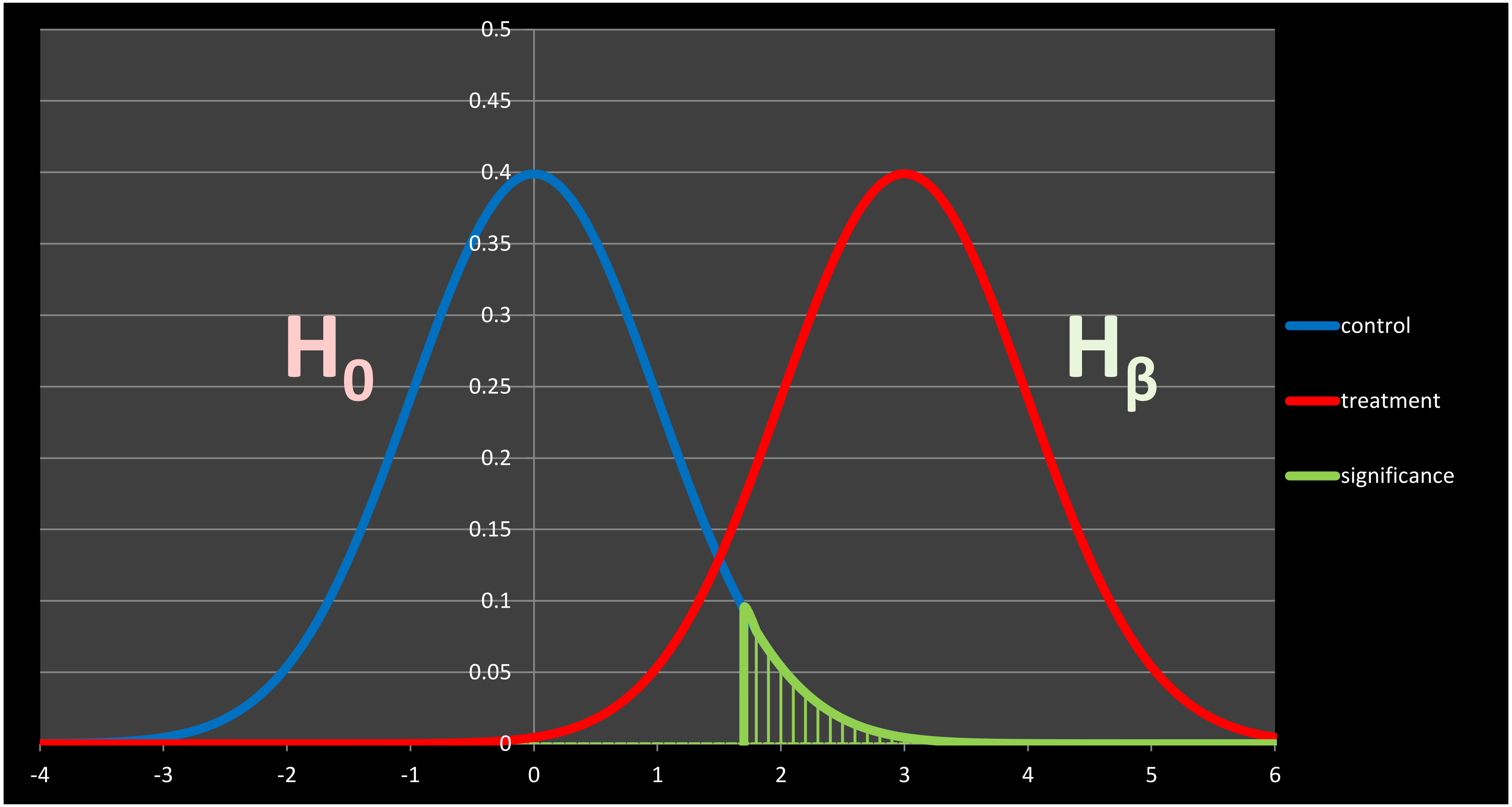
Average Cluster Size

Power!

$$EffectSize = \left(t_{(1-\kappa)} + t_{\alpha}\right) * \sqrt{\frac{1}{P(1-P)}} * \sqrt{\frac{\sigma^2}{N}} \sqrt{1+\rho(m-1)}$$

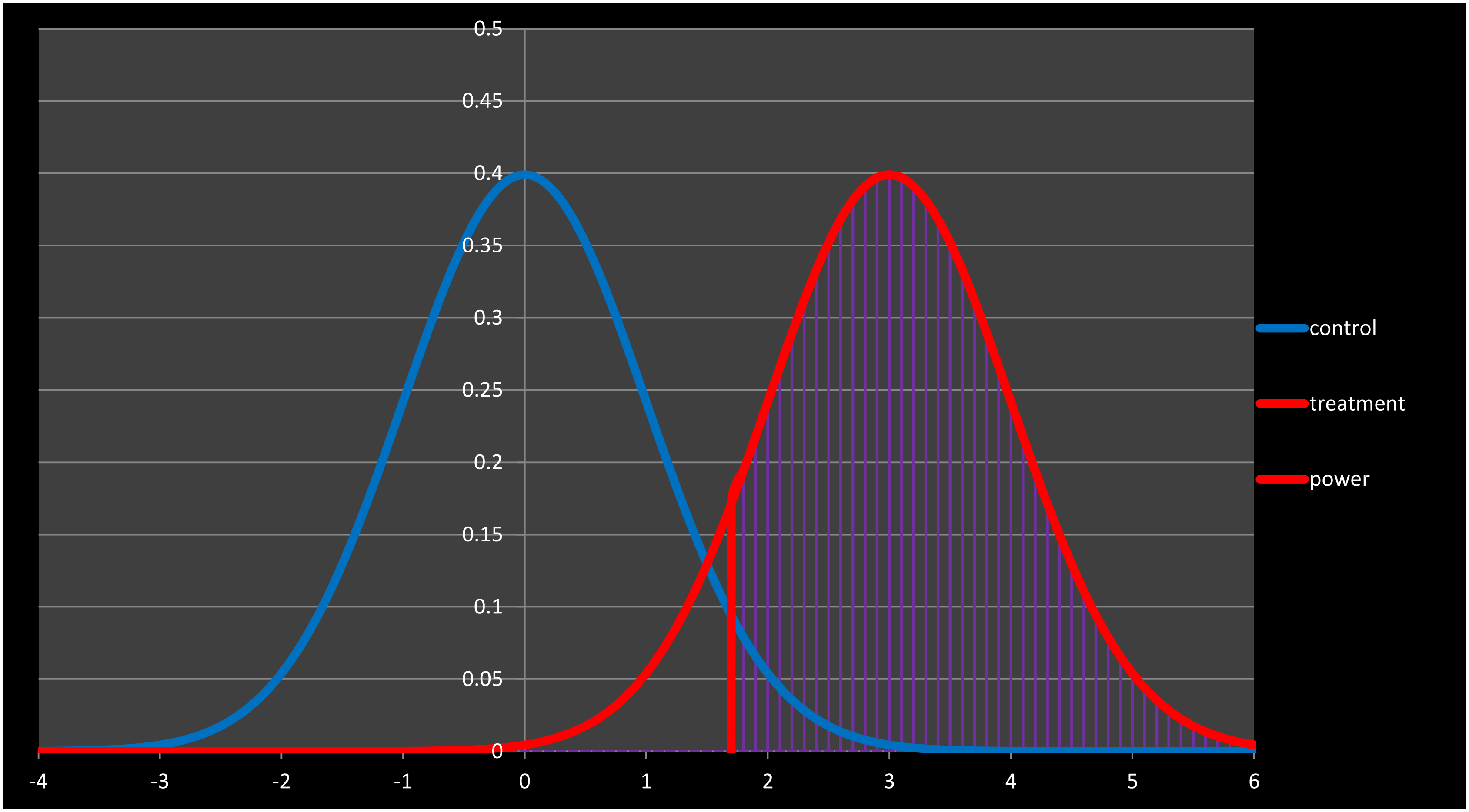
Proportion in Treatment

Sample split: 50% C, 50% T



Equal split gives distributions that are the same "fatness"

Power: 91%



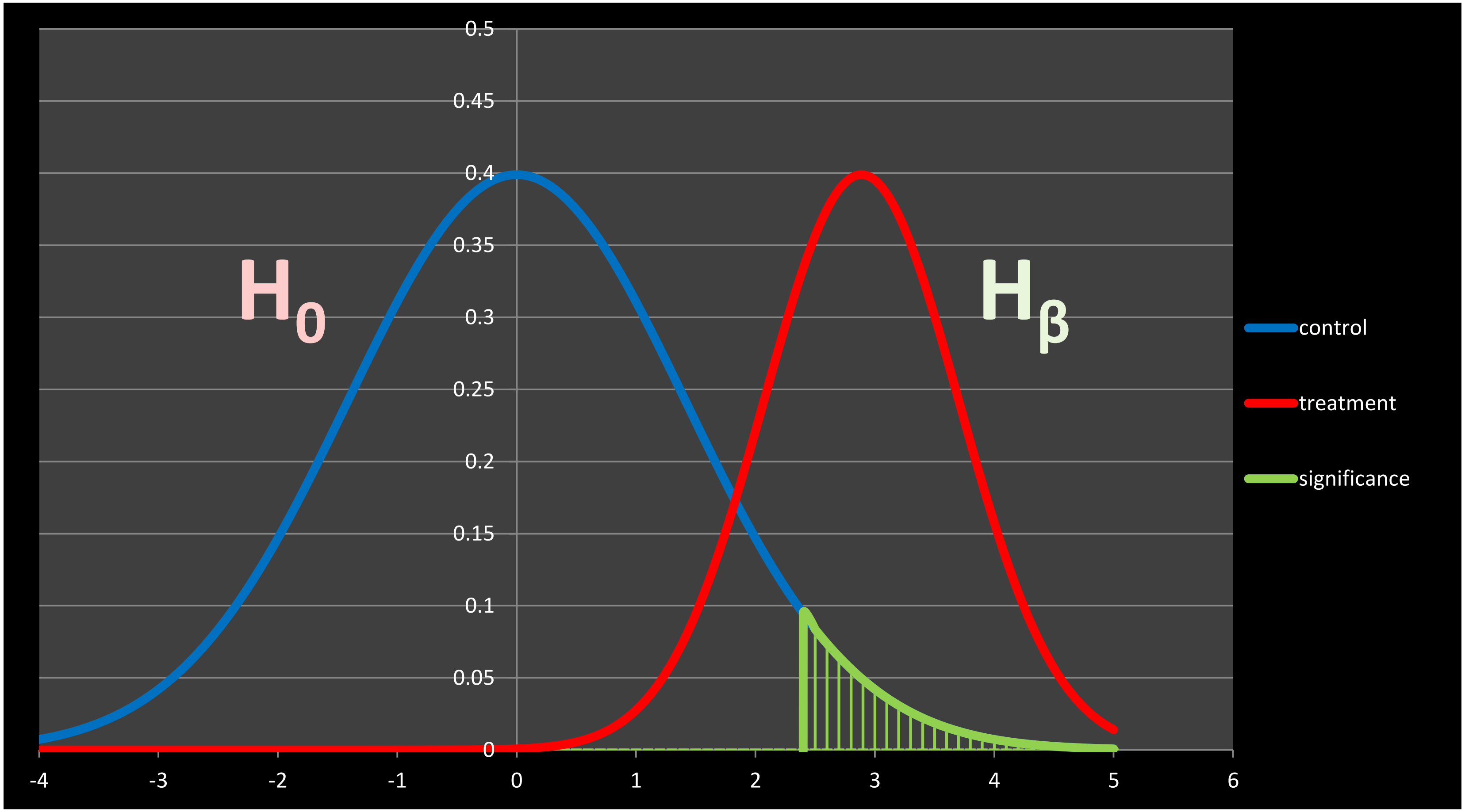
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6
control
treatment
power

If it's not 50-50 split?

- What happens to the relative fatness if the split is not 50-50?

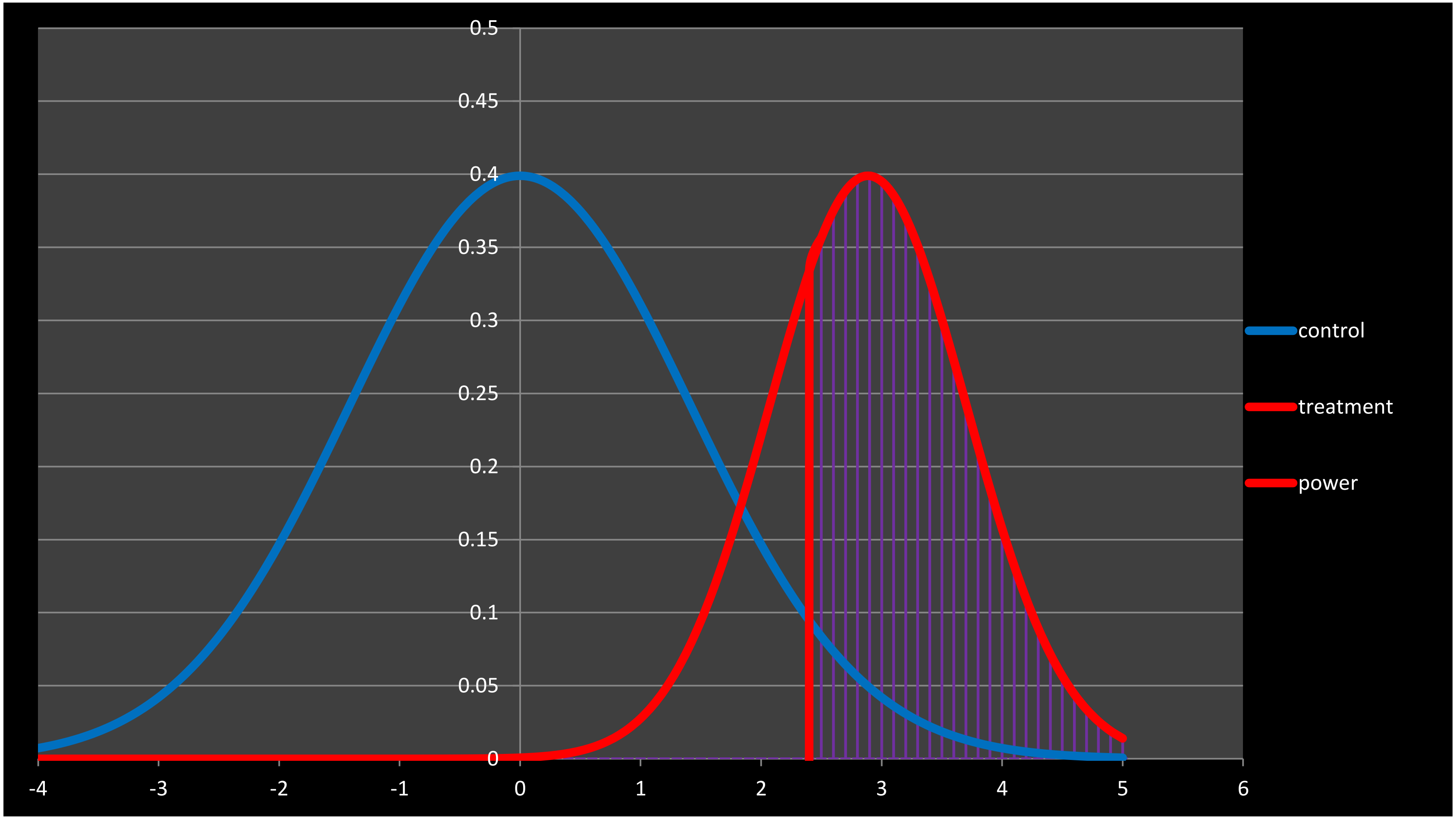
- Say 25-75?

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

END!