Summer Statistics Workshop

Session 1: Basic Concepts

Sam Johnson
Mark Sheskin

This workshop might be right for you if…

…you don't have much exposure to statistics.

…you need a refresher on basic concepts and tests.

…you want to learn the basics of different stats packages.

…you don't have much experience with real data.

…you want to work on your scientific writing.

…you want to be a responsible producer of knowledge.

This workshop might be NOT right for you if…

…you are already a stats wizard.

…you want to learn fancy statistical techniques.

…you want to learn stats packages in depth.

…you are a pro with real data (or you dislike reality).

…you don't want to do any writing.

…you aspire to be an unscrupulous scientist.

Topics

- Today: Basic Concepts
- June 21: Working with real data
- June 28: T-tests
- July 5: Analysis of variance (ANOVA)
- July 12: Regression
- July 19: Researcher degrees of freedom

Statistical tools

- Excel
    - Working with data
    - Useful tricks
    - T-tests
    - Making bar plots and scatterplots
- SPSS
    - Two-way ANOVA
- R
    - Simple and multiple linear regression
- G*Power

Today

- Types of variables
- Distributions
- Confidence intervals and standard error
- Hypothesis testing
- Statistical risk management

Mr Ollivander needs your help!

Research project on wand quality

Research Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Does the wand choose the wizard? Determinants of satisfaction in wand-wizard dyads.

Garrick K. Ollivander, Mark Sheskin, Samuel G. B. Johnson, and the Yale Summer Interns

Psychological Science
2016, Vol. 27(12) 1573–1587

DOI: 10.1177/0956797616666074
pss.sagepub.com
SAGE

What are some possible IVs?

- *Independent variable (IV)*: What we are manipulating (or taking as given/exogenous).

- Examples:
  - Wood type, wand core, wand length
  - Years of magic experience, gender, country
  - Affinity for dark arts, risk tolerance
  - Spell being tested

What are some possible DVs?

- *Dependent variable (DV)*: What we are measuring (or taking as influenced/endogenous).

- Examples:
  - Wizard's satisfaction, frequency of wand use
  - Wizard preferences across wands
  - Wand/wizard bonding scale
  - Success or failure (dichotomous) at various spells

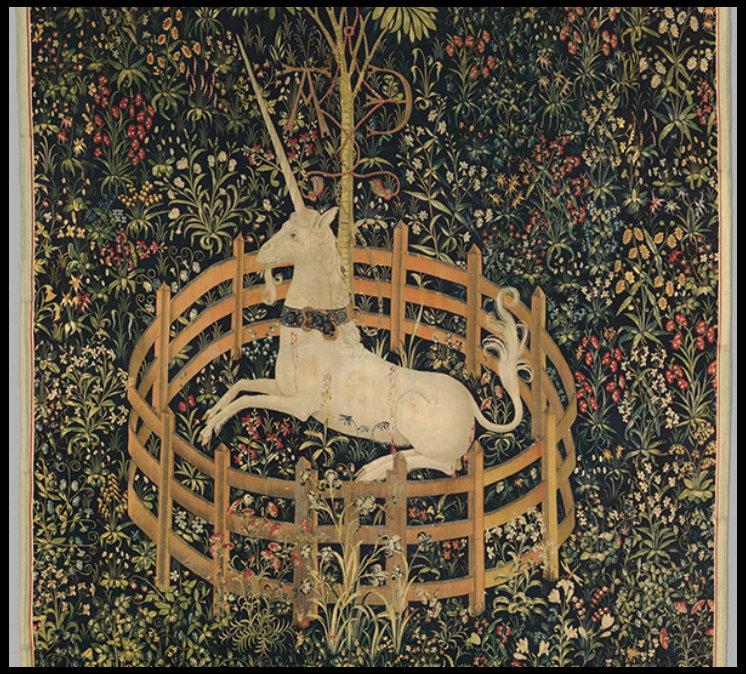
Continuous vs. discrete variables

- Continuous:
  - Can take a range of values
    - Ordinal (ranks)
    - Interval (temperature)
    - Ratio (distance)
- Discrete:
  - Takes on only a small number of values (not necessarily ordered)
  - If it's an IV, discrete variables are called "factors" and the different values are called "levels."

Which test?

Dependent Variable	Independent Variable: *Continuous*	Independent Variable: *Discrete*
Continuous	Correlation Linear regression	*t*-test ANOVA
Discrete	Logistic regression	Chi squared

What makes for the best wand?



Mr. Ollivander hypothesizes:

- Unicorns evolved from non-magical ancestors more recently than dragons
- Therefore, it stands to reason that dragons would have had more time to accumulate beneficial magical mutations
- So, a wand made from dragon heart should be more powerful than a wand made from unicorn horn

Wingardium leviosa!

- $T_{unicorn}$ = Feather levitation time with a unicorn wand
- T_{dragon} = Feather levitation time with a dragon wand

Levitation time with unicorn wand

8 seconds

9 seconds

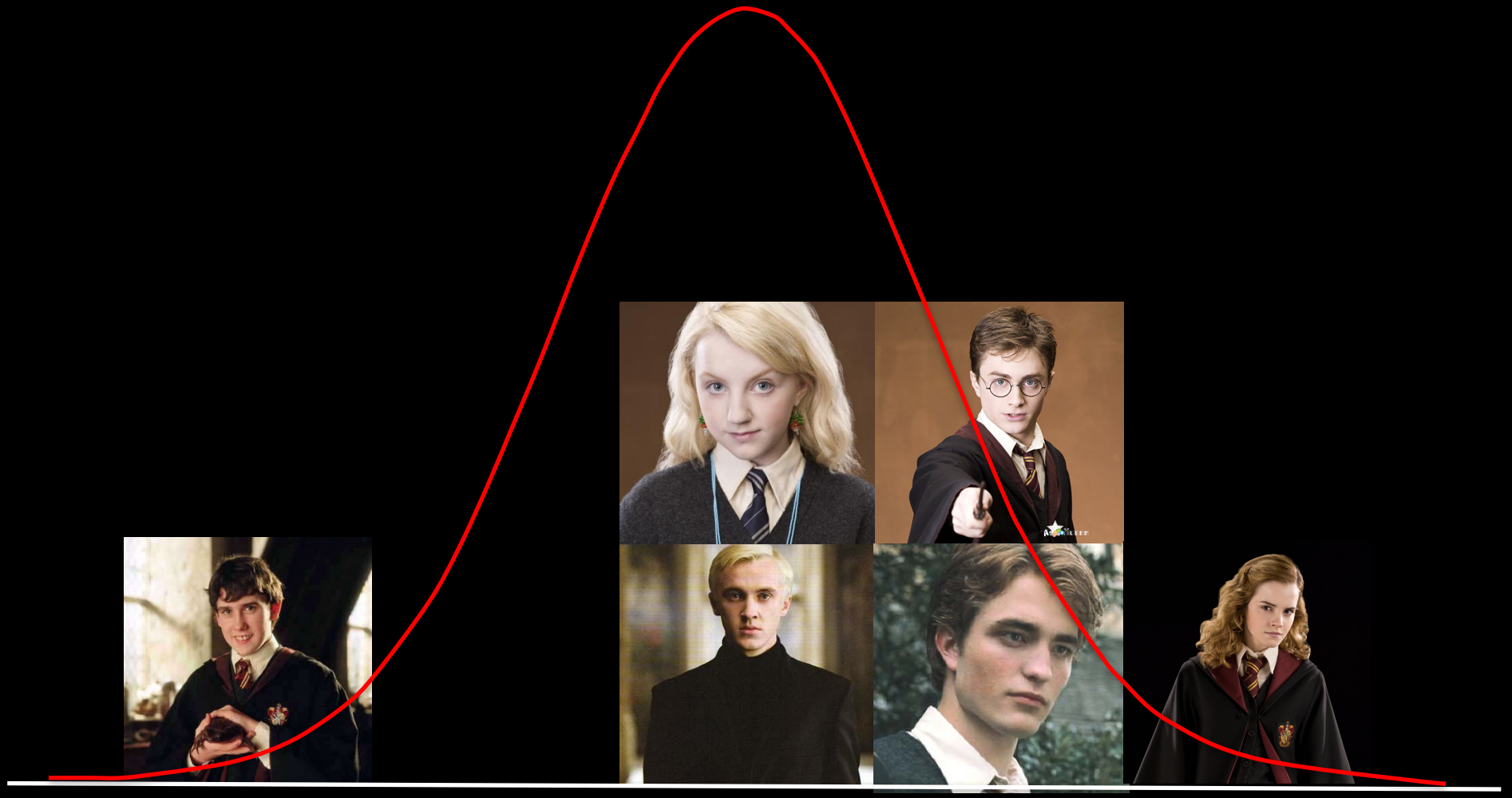
7 seconds

7 seconds

5 seconds

8 seconds

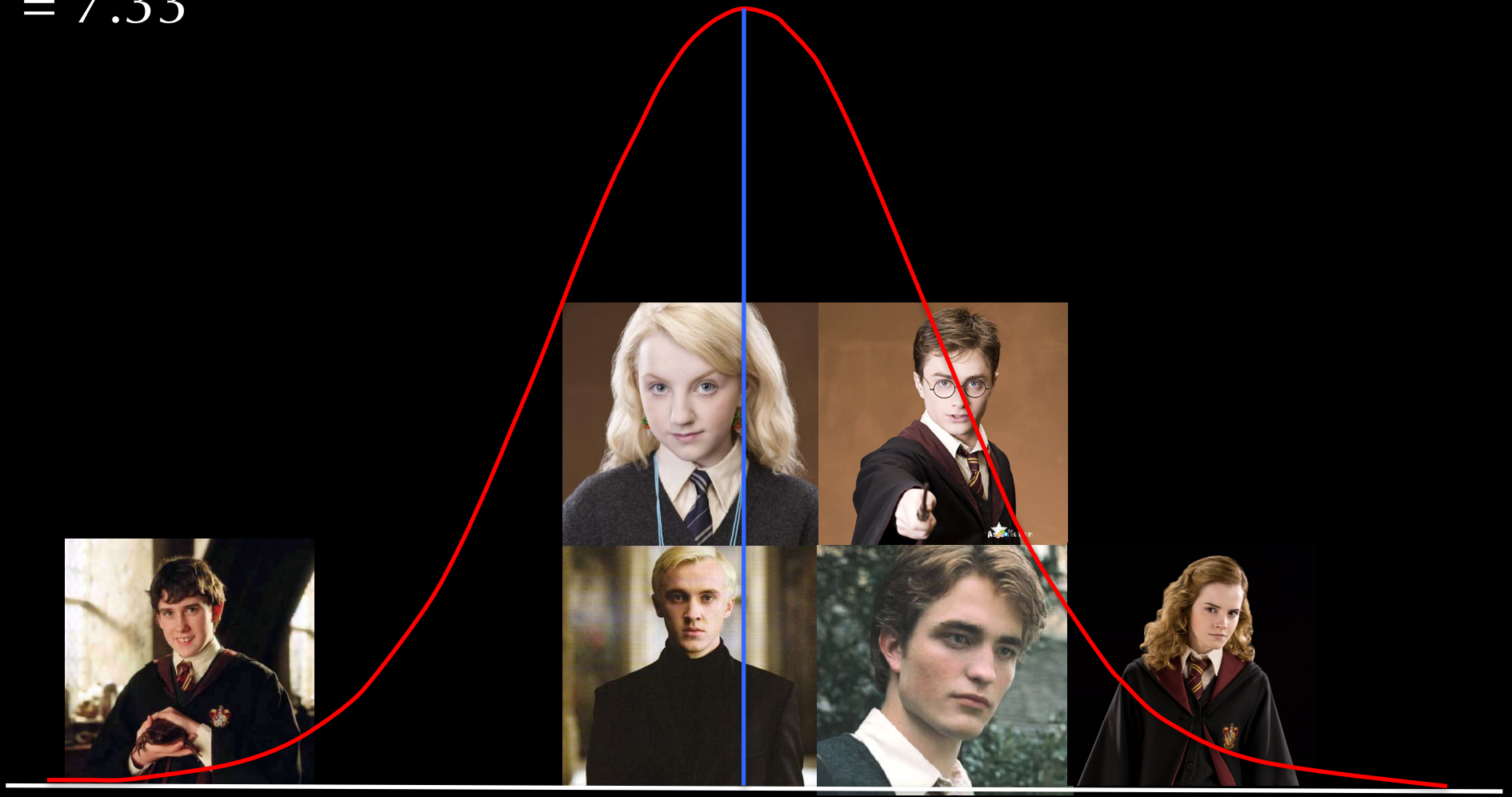
Distribution of a variable



Mean

- The average score
- Toward the middle of a distribution

$M = (5 + 7 + 7 + 8 + 8 + 9) / 6$
$= 7.33$



5

7

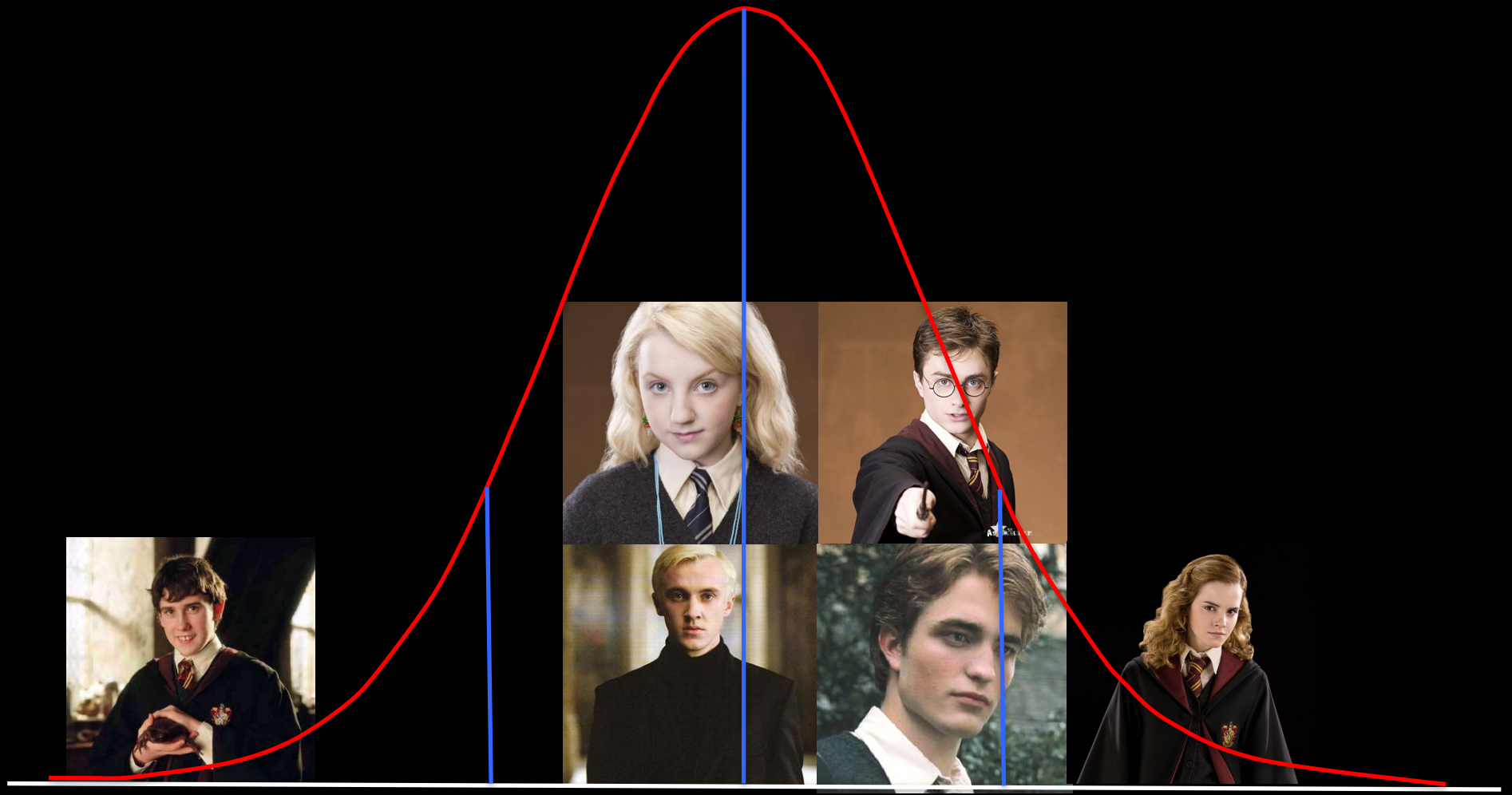
9

Levitation time (in seconds)

Standard Deviation

- Tells you how “spread out” the scores are
- Normally, most scores fall within 1 SD of the mean

SD = 1.37



5

7

9

Levitation time (in seconds)

Z Score

- How many SDs an individual is from the mean

- So, if z = -2 for Neville, then Neville is 2 SDs below the mean

$$z = \frac{(x - M)}{SD}$$

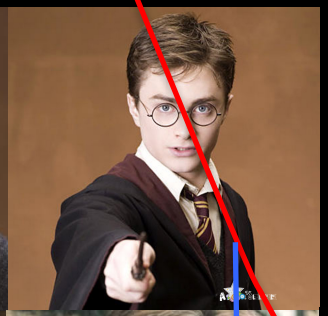
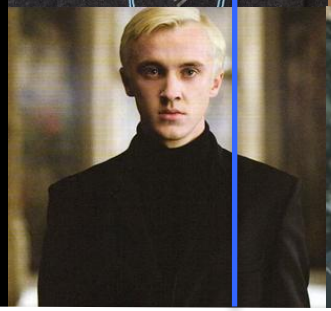
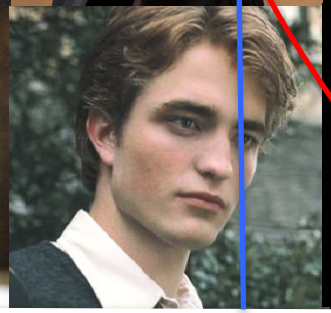
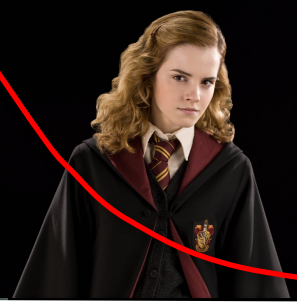
$M = 7.33$
$SD = 1.37$

score = 9 sec.

z = ???
= 1.67/1.37 = 1.25

$$z = \frac{(x - M)}{SD}$$



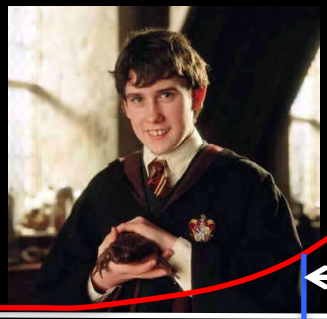





5 7 9

Levitation time (in seconds)

So, what fraction of people can levitate feathers longer than Neville?

$z = -2$



Experiments

- Usually, when we do experiments, we're not interested in individuals
- Instead, we're interested in the differences between different *groups* or *conditions*
- The way we look for these differences is by estimating what the *population* mean (μ) is for each condition

Within-subjects experiment

- IV is manipulated so that *each person* is in *every condition*

Between-subjects experiment

- IV is manipulated so that *different people* are in *different conditions*

Unicorn Group

8 seconds

7 seconds

5 seconds

Dragon Group

13 seconds

8 seconds

9 seconds

	Unicorn Condition	Dragon Condition	Difference
	8 seconds	10 seconds	+2 seconds
	7 seconds	6 seconds	-1 second
	5 seconds	6 seconds	+1 second

Randomness is your friend

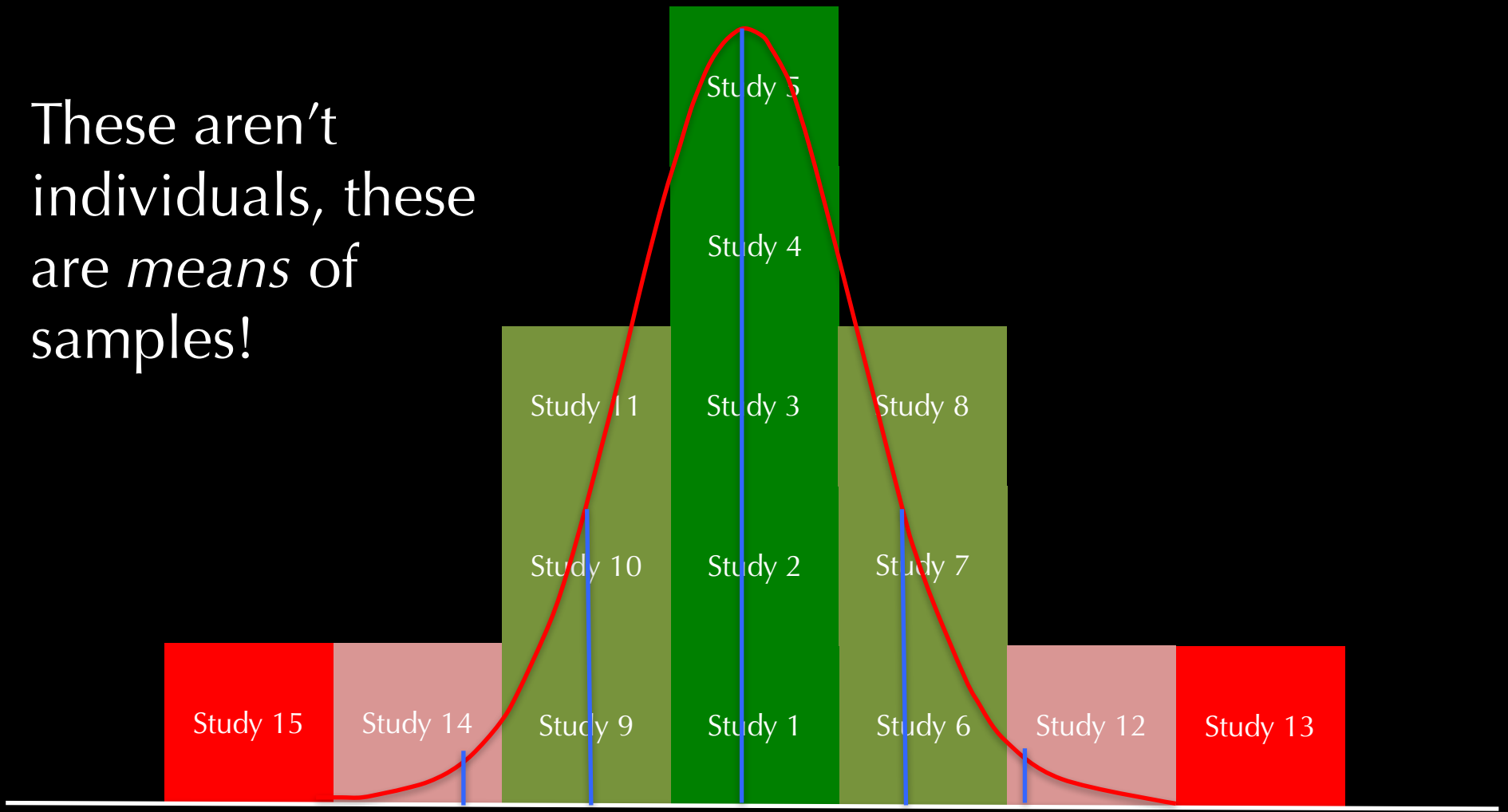
- What if some people happen to better at levitating feathers than others?
  - Random *assignment* to experimental condition: Allows *causal claims*.
- What if part of the population is different from other parts?
  - Random *sampling* from target population: Allows *generalization* from sample to population.

Populations and samples

- Goal is to estimate *population* mean (μ) in each condition.
- We almost never know the true population mean (if we did, we wouldn't need to do an experiment!)
- Instead, we use samples to *guess* what are likely values of the population mean.
- Figuring out what guesses are plausible and which are not plausible is the fundamental problem of statistics.

The sampling distribution

These aren't individuals, these are *means* of samples!



μ

Levitation time (in seconds)

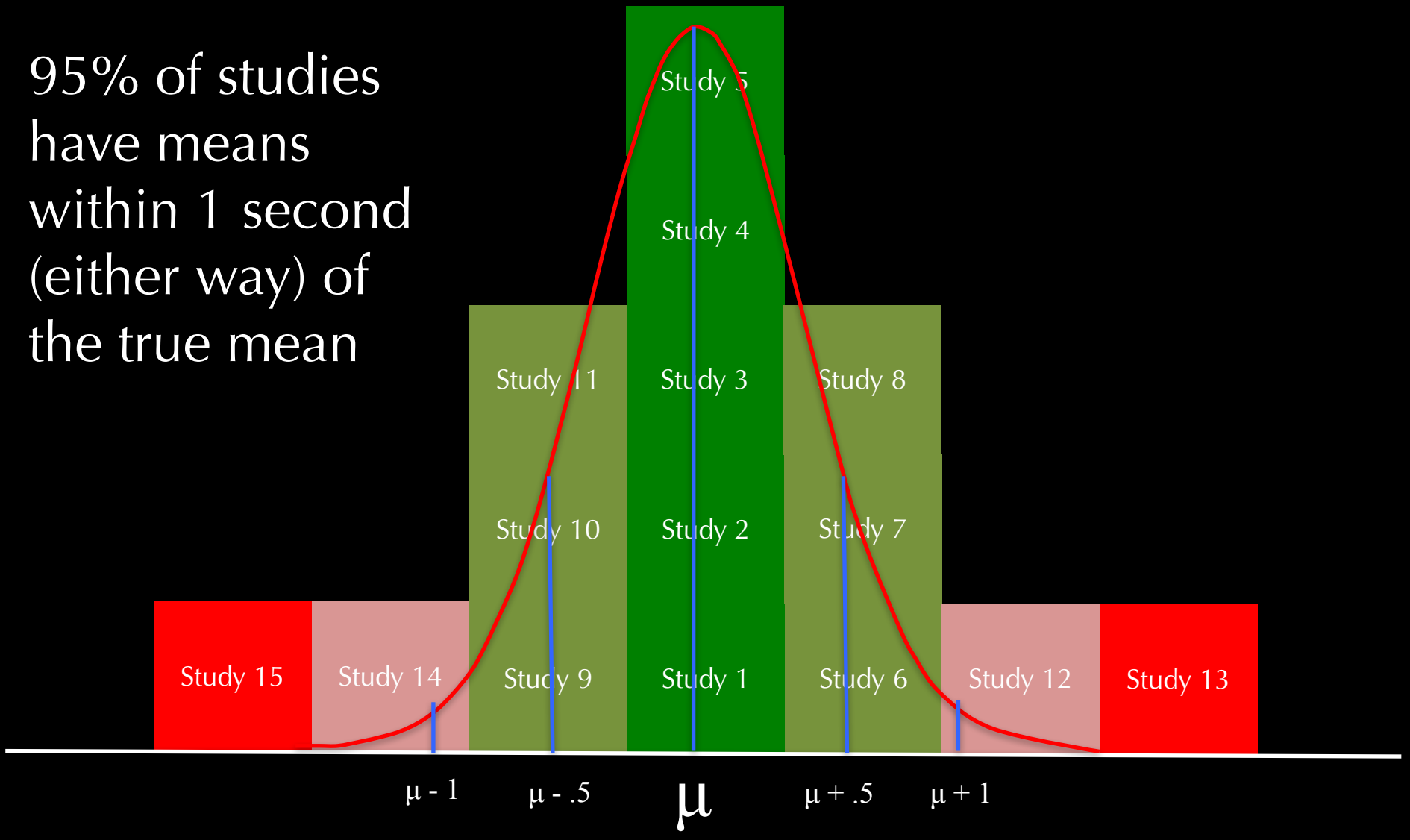
Standard error

- The standard deviation of the sampling distribution is called the *standard error* (SE)
- In other words, if we repeated our experiment over and over again, how much would the results vary?
- We can estimate SE from just *one* sample!

$$SE = SD / \sqrt{N}$$

- For our study, SE = 1.37/sqrt(6) = 0.5

95% of studies have means within 1 second (either way) of the true mean



Levitation time (in seconds)

Testing Ollivander's hypothesis

- Suppose Ollivander commissioned a survey of all the wizards in the land, and he knows that T_{dragon} is exactly 6 seconds on average in the population

- Now, Ollivander just wants to know whether $T_{unicorn}$ is different from 6 seconds

Null hypothesis significance testing

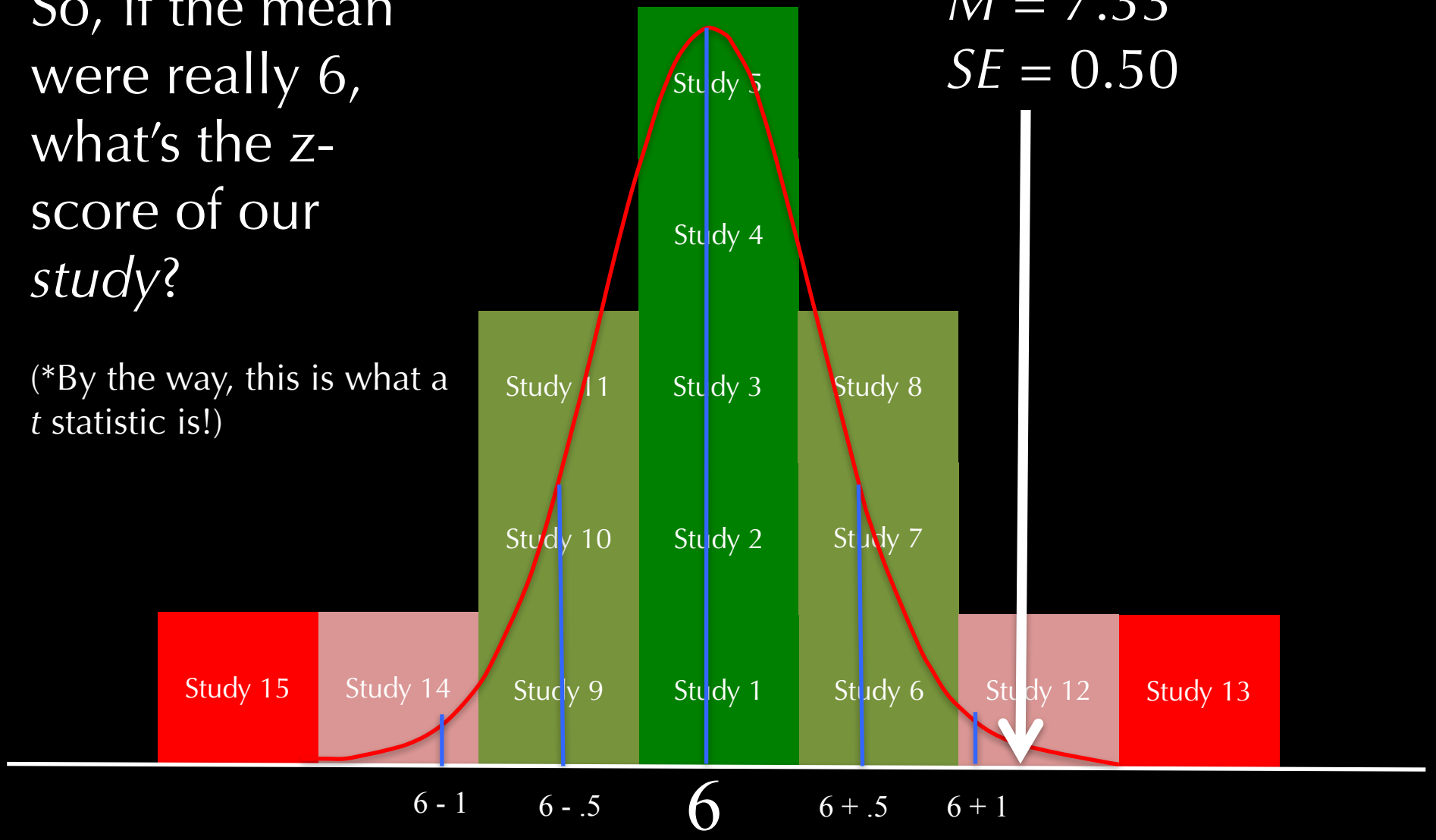
- Answers the question: If we were wrong (i.e., there is no difference), how likely would our results be?

- If not very likely, then there probably *is* a difference.

- So, how likely would our study results be if the true population mean were 6?

So, if the mean were really 6, what's the z-score of our *study*?

(*By the way, this is what a *t* statistic is!)



- So, 6 isn't a very likely value of the population mean for $T_{unicorn}$, since fewer than 5% of studies would have a sample mean of 7.33

- But what *are* the plausible values of the population mean for $T_{unicorn}$?

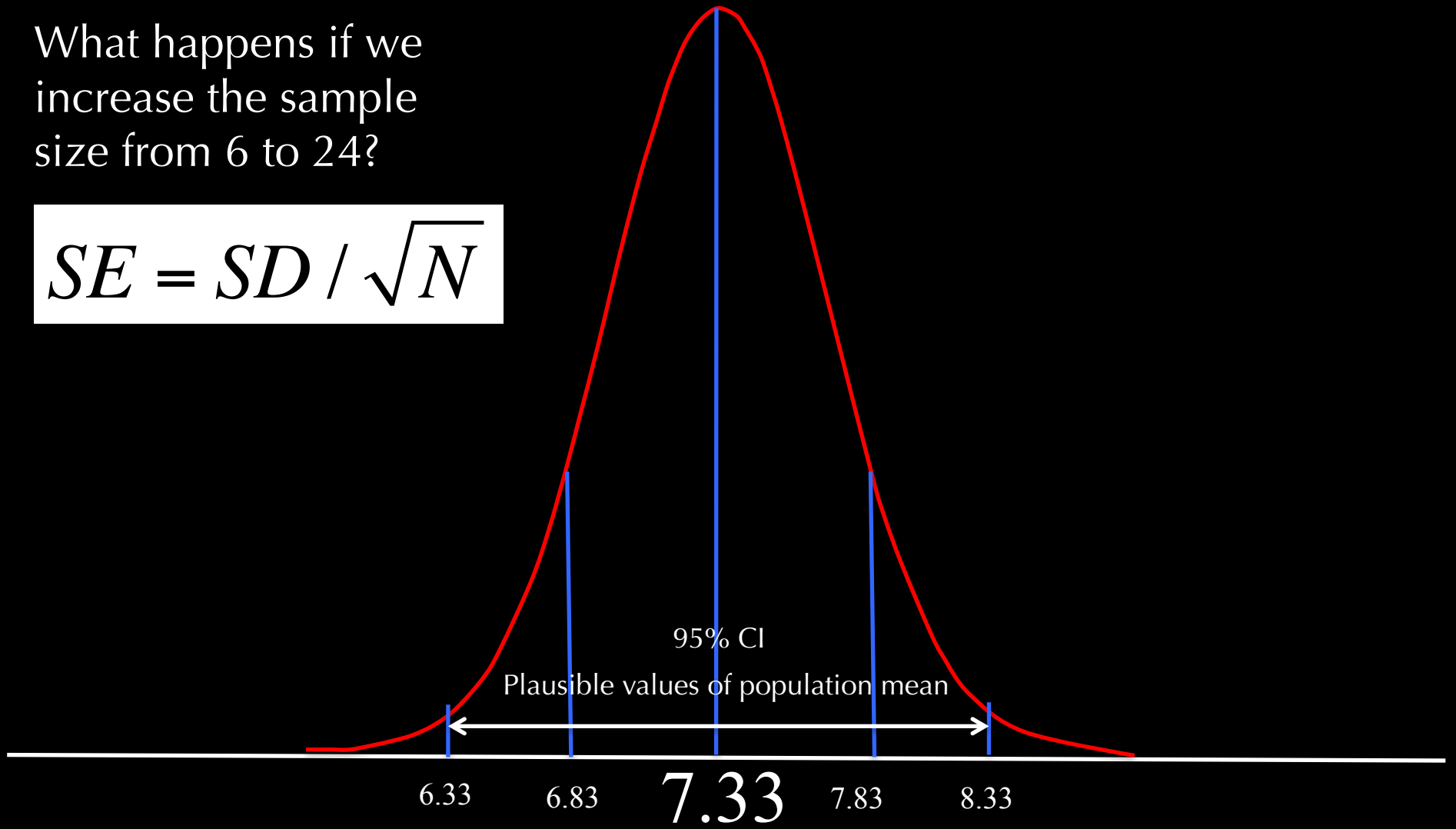
95% confidence intervals

- If we repeated our study many times and the population mean equaled our sample mean, then about 95% of the repetitions would fall within the 95% confidence interval (CI).
- Guesses for the population mean are considered plausible if they fall within the 95% CI.

- $M \pm 2$ SE

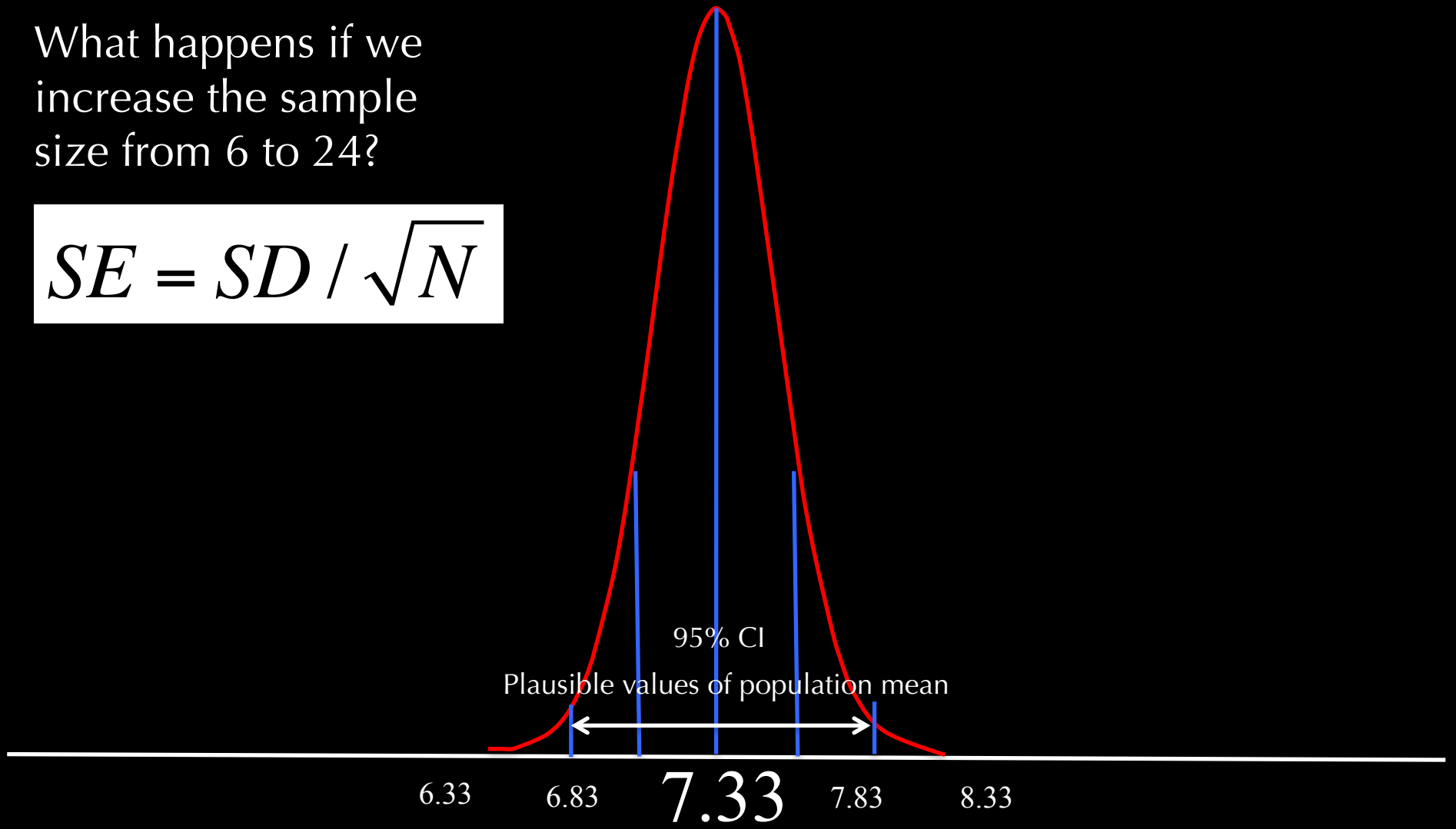
What happens if we increase the sample size from 6 to 24?

$$SE = SD / \sqrt{N}$$



What happens if we increase the sample size from 6 to 24?

$$SE = SD / \sqrt{N}$$



Testing Ollivander's hypothesis

- *Null hypothesis (H_0)*: True if there is no difference between groups

  $T_{unicorn} = T_{dragon}$

- *Alternative hypothesis*: True if there is a difference between groups

  $T_{unicorn} \neq T_{dragon}$

- So, if Ollivander is right, we should *reject* H_0

Science is risk-taking!
And statistics is risk management.

Types of Risk

Conclusion of test

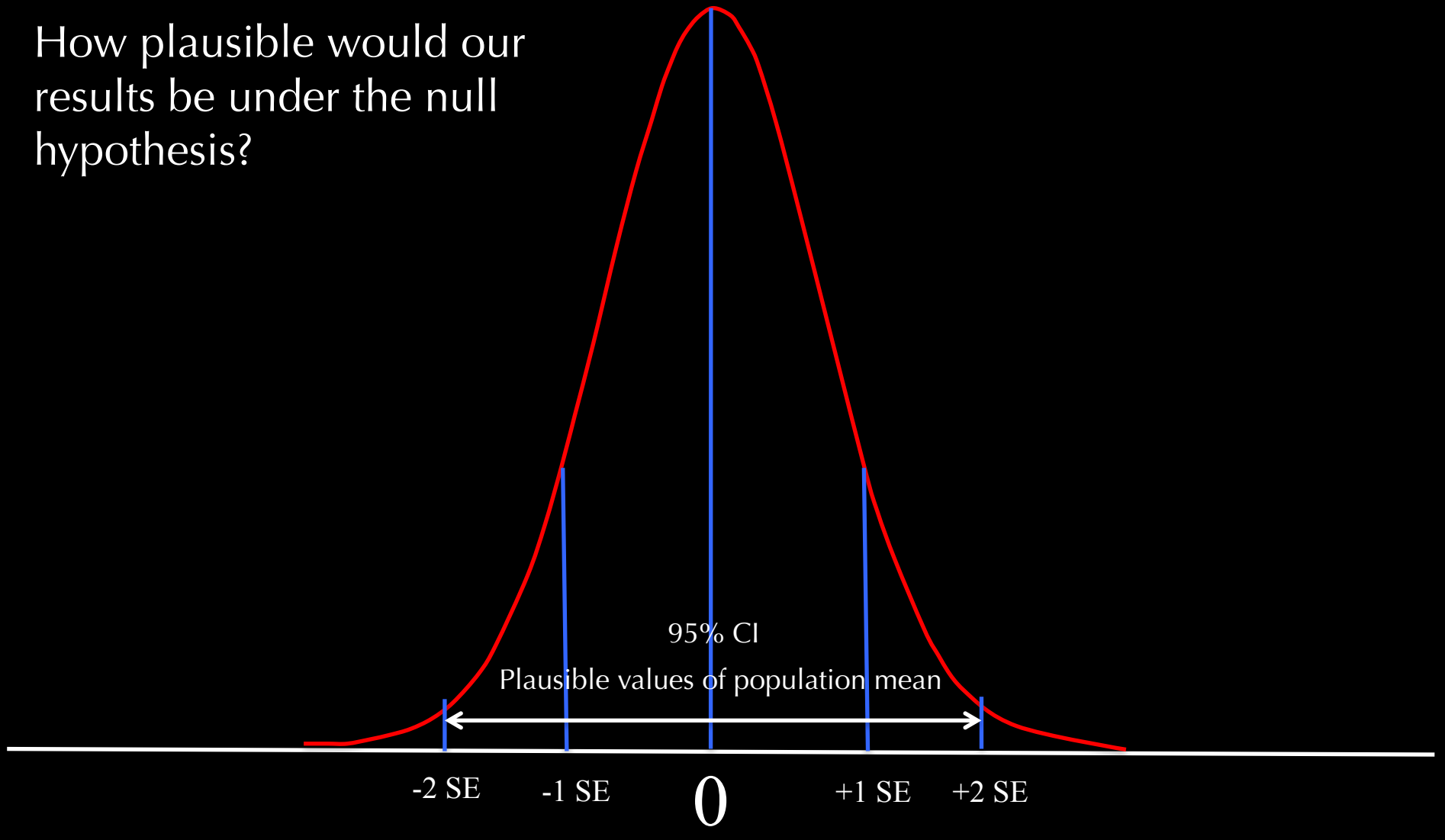
Reality	*"Yes, there's a difference" (Reject H_0)*	*"No, there's no difference" (Accept H_0)*
There really is a difference!	You discovered something!	TYPE II ERROR
Really, it makes no difference	TYPE I ERROR	You were wrong…but at least you didn't publish something false!

	Unicorn Condition	Dragon Condition	Difference
	8 seconds	10 seconds	+2 seconds
	7 seconds	6 seconds	-1 second
	5 seconds	6 seconds	+1 second

- Suppose we run this experiment with 100 subjects
- The mean difference is +1.0
  - That is, levitation times are 1 second more with the dragon wands
- Is this a *reliable* difference? *Significance level!*
- Is this a *big* difference? *Effect size!*

Imagine there were no difference

How plausible would our results be under the null hypothesis?

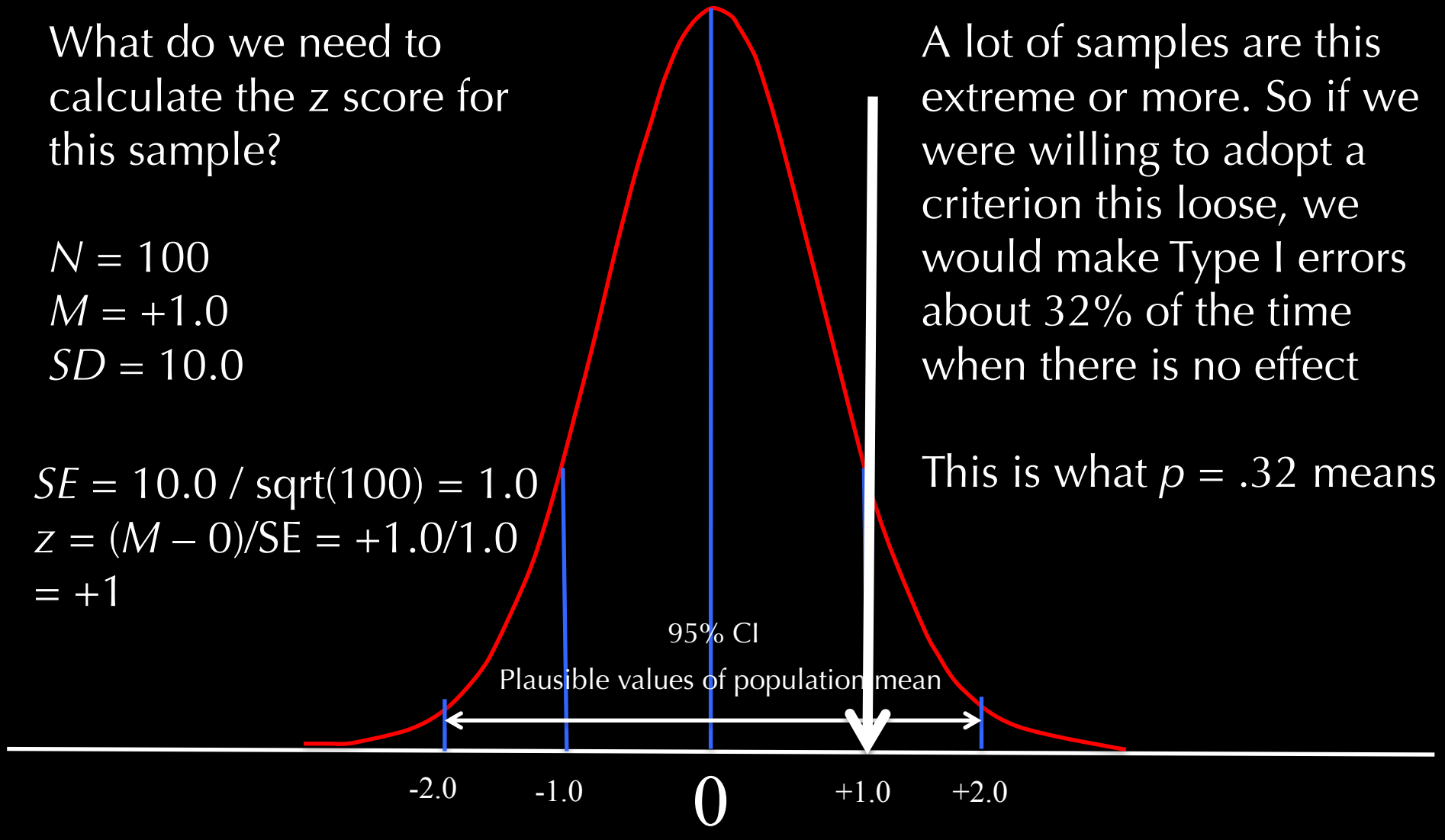


Difference in levitation time (sec.)

What do we need to calculate the z score for this sample?

$N = 100$
$M = +1.0$
$SD = 10.0$

$SE = 10.0 / \text{sqrt}(100) = 1.0$
$z = (M - 0)/\text{SE} = +1.0/1.0 = +1$



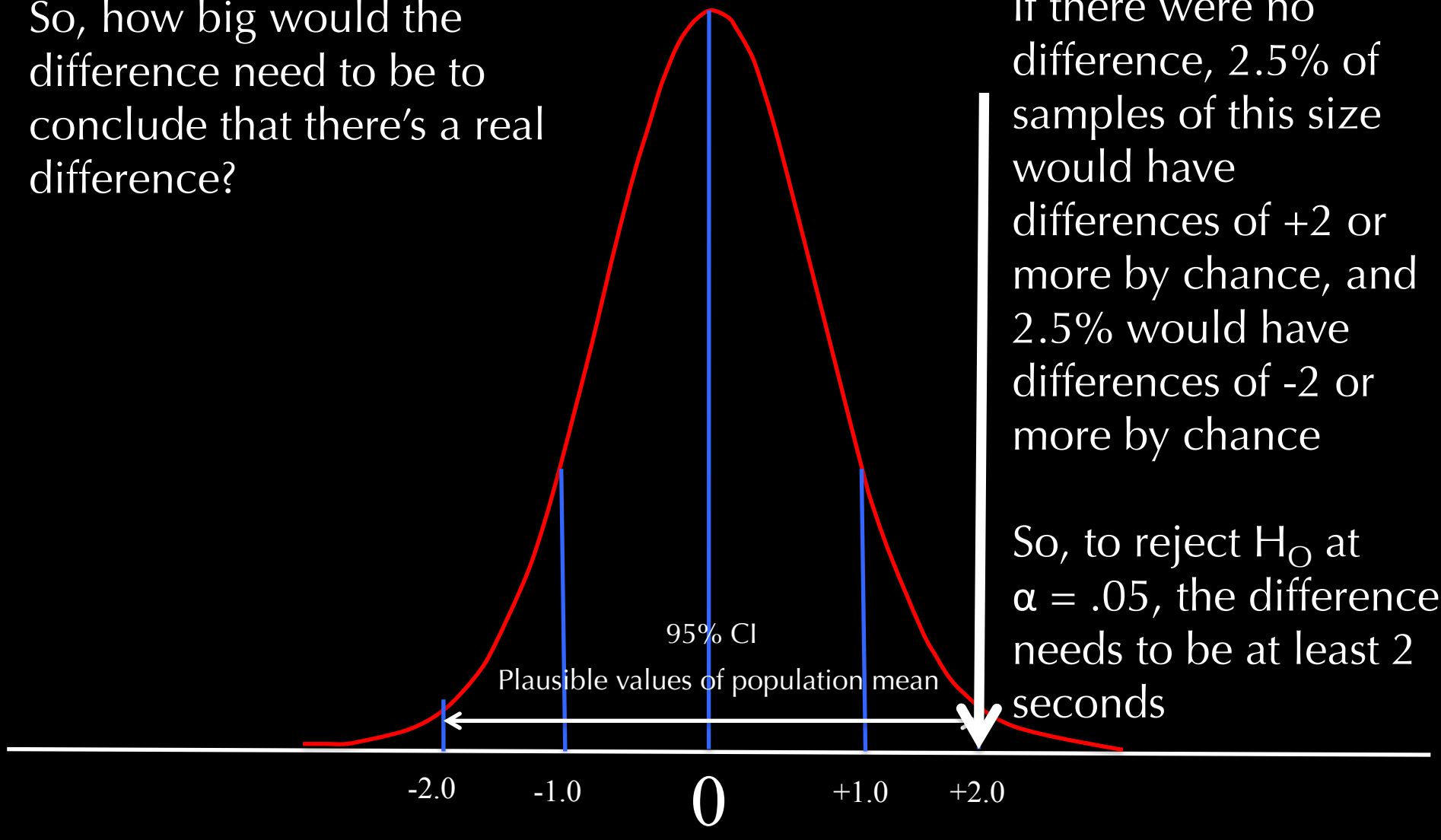
A lot of samples are this extreme or more. So if we were willing to adopt a criterion this loose, we would make Type I errors about 32% of the time when there is no effect

This is what $p = .32$ means

Controlling Type I risk

- The scientific community has decided that we are willing to tolerate Type I errors 5% of the time when there really is no effect
  - In other words, $\alpha = .05$
- We therefore adopt a rule to reject the null hypothesis when $p < .05$
- In our wand study, we do <u>not</u> reject the null hypothesis, and conclude that there is no evidence for Ollivander's hypothesis

So, how big would the difference need to be to conclude that there's a real difference?



If there were no difference, 2.5% of samples of this size would have differences of +2 or more by chance, and 2.5% would have differences of -2 or more by chance

So, to reject H_O at $\alpha = .05$, the difference needs to be at least 2 seconds

But could this be a Type II error?

- Calculating the likelihood that our procedure makes Type II errors is called *power analysis*
- Power = (1 – Type II error rate)
- A study must have high power to be informative (usually, 80% or 90% is considered adequate).

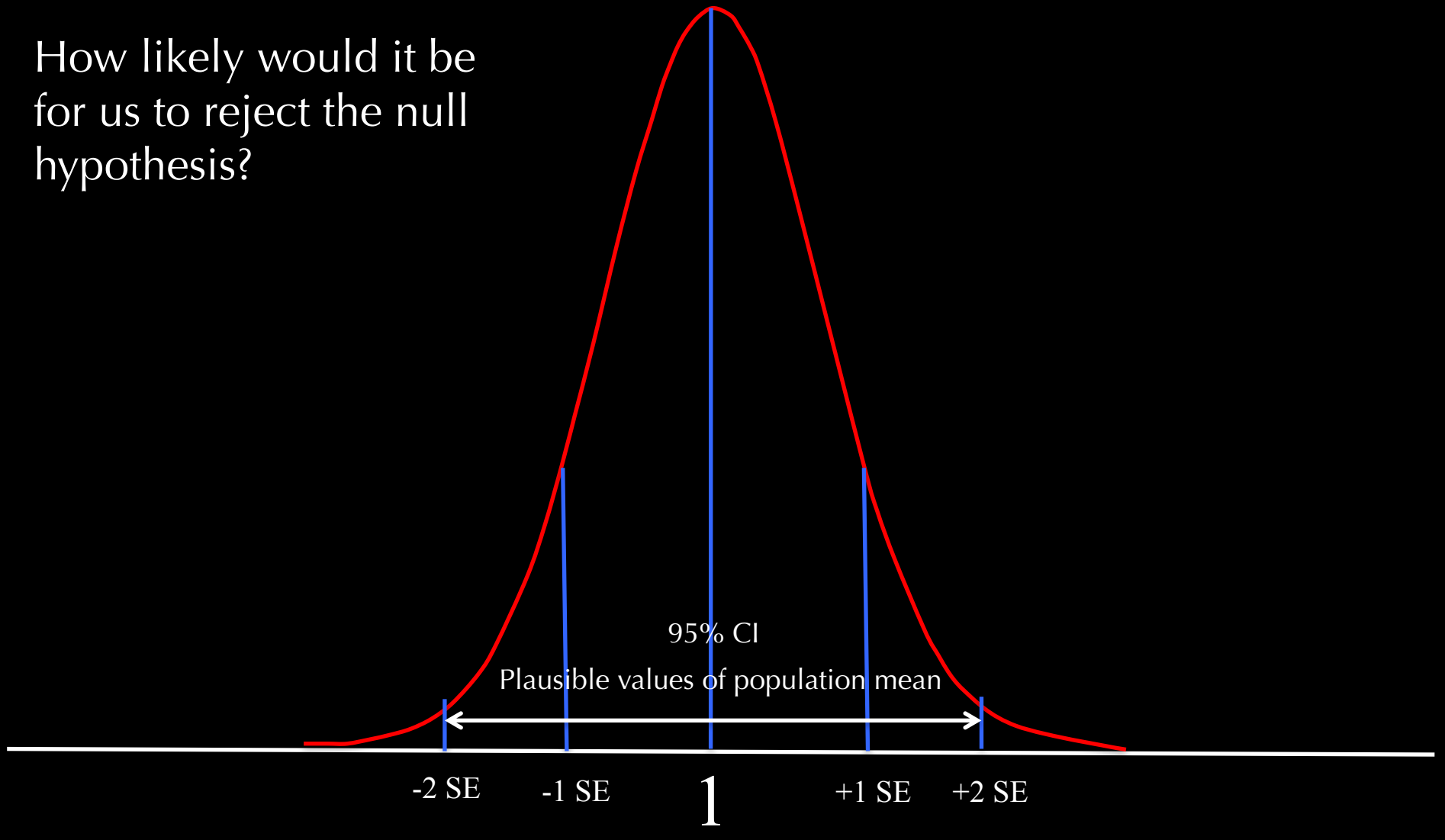
Effect size

- The size of the difference between conditions
  - In our example, the effect size would be +1 seconds
- Power is defined relative to a particular effect size
- The Type I error rate is the likelihood of declaring a difference significant when the effect size is 0
- Power is the likelihood of declaring a difference significant when the effect size is some number greater than 0

- Let's suppose there is a *real* difference between groups (i.e., between the *population* means), and it's 1 second
- What was the likelihood of observing a significant difference given our SD = 10.0 and N = 100?

Imagine there was a difference of 1 sec.

How likely would it be for us to reject the null hypothesis?



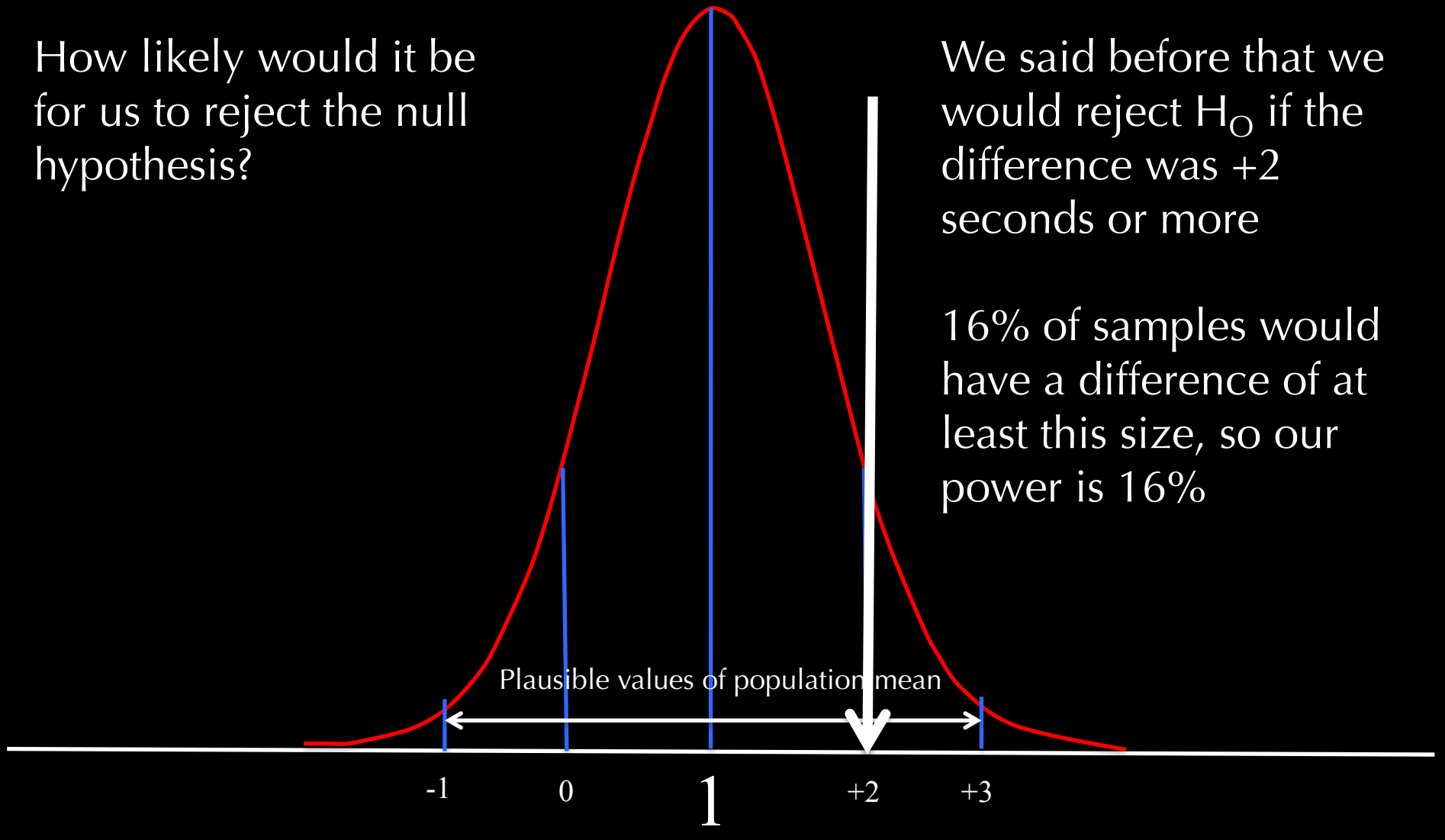
Difference in levitation time (sec.)

Imagine there was a difference of 1 sec.

How likely would it be for us to reject the null hypothesis?

We said before that we would reject H_O if the difference was +2 seconds or more

16% of samples would have a difference of at least this size, so our power is 16%



Tips for maintaining adequate power

$$SE = SD / \sqrt{N}$$

- Think of how big you think the effect might be, and pick a sample size that will be informative

- Minimize the variability as much as you can

Next time

- Using Excel to explore data
- Excel tips and tricks
- Basic graphing (bar plots, scatter plots)

Before next time

- Exercise 1
  - Exercises are optional but strongly encouraged
  - Thinking about designs for simple "factorial" experiments
  - Writing (very brief!) methods sections
  - If you want feedback, please email to both of us (sgbjohnson@gmail.com and msheskin@gmail.com) before the start of Session 2