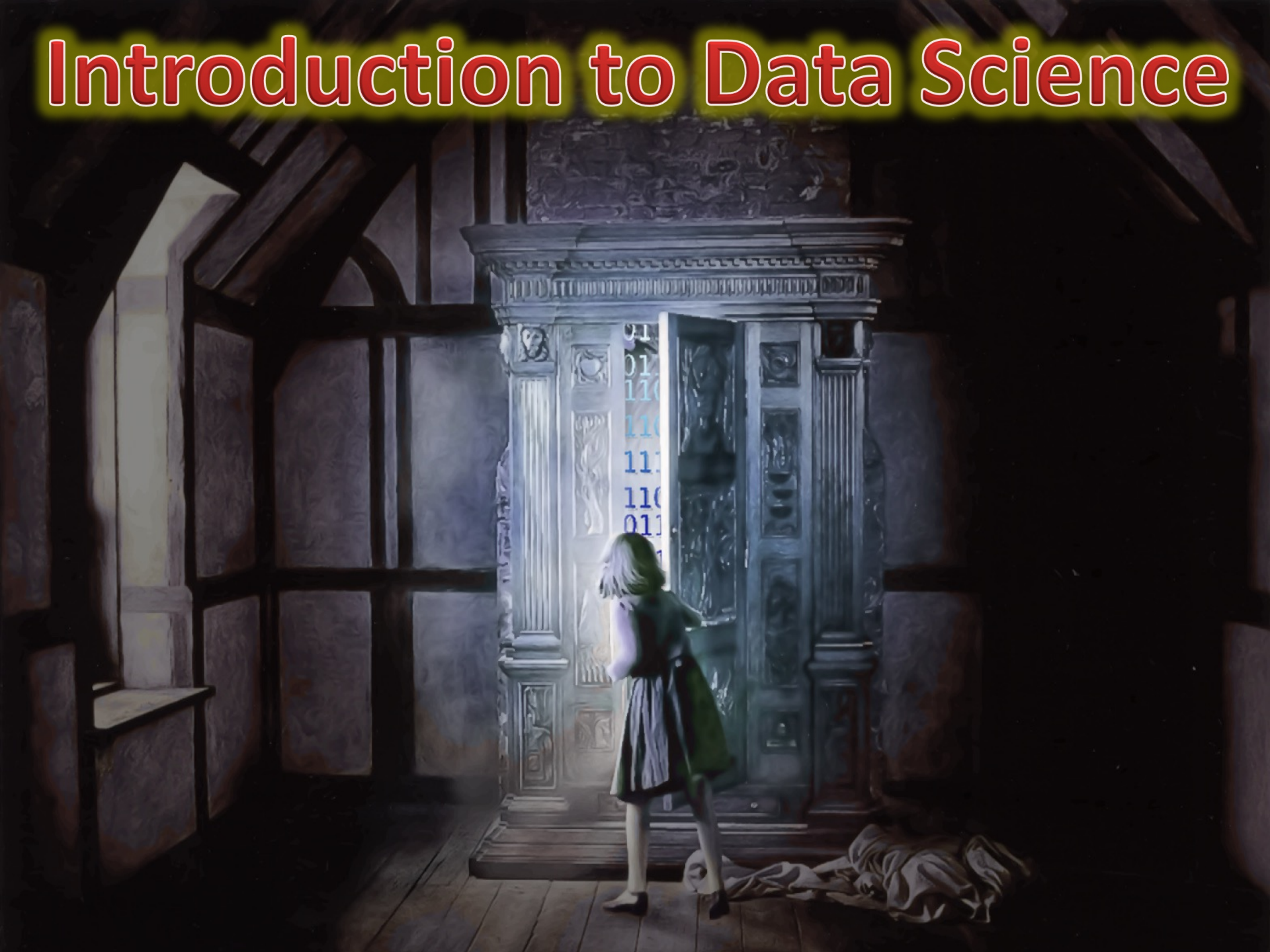
Smallest font

Please turn off and put away your cell phone

Calibration slide

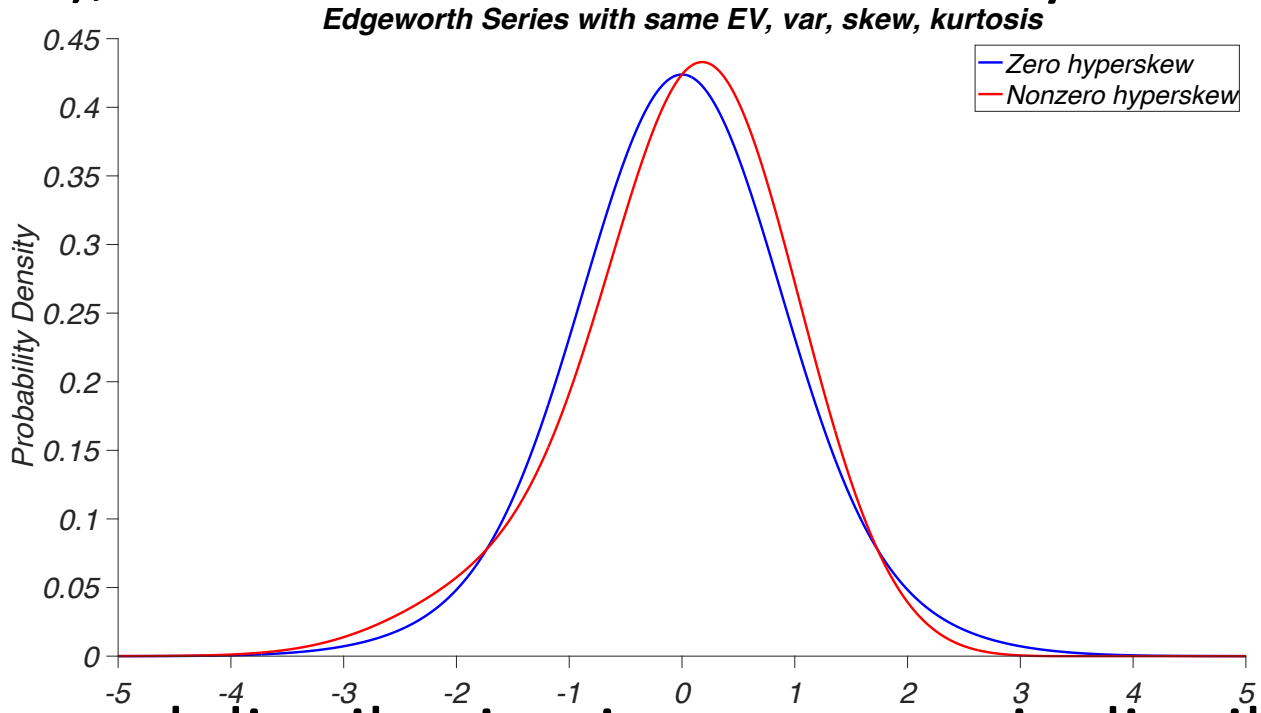
Smallest font

Introduction to Data Science



We ended with moments and the normal distribution

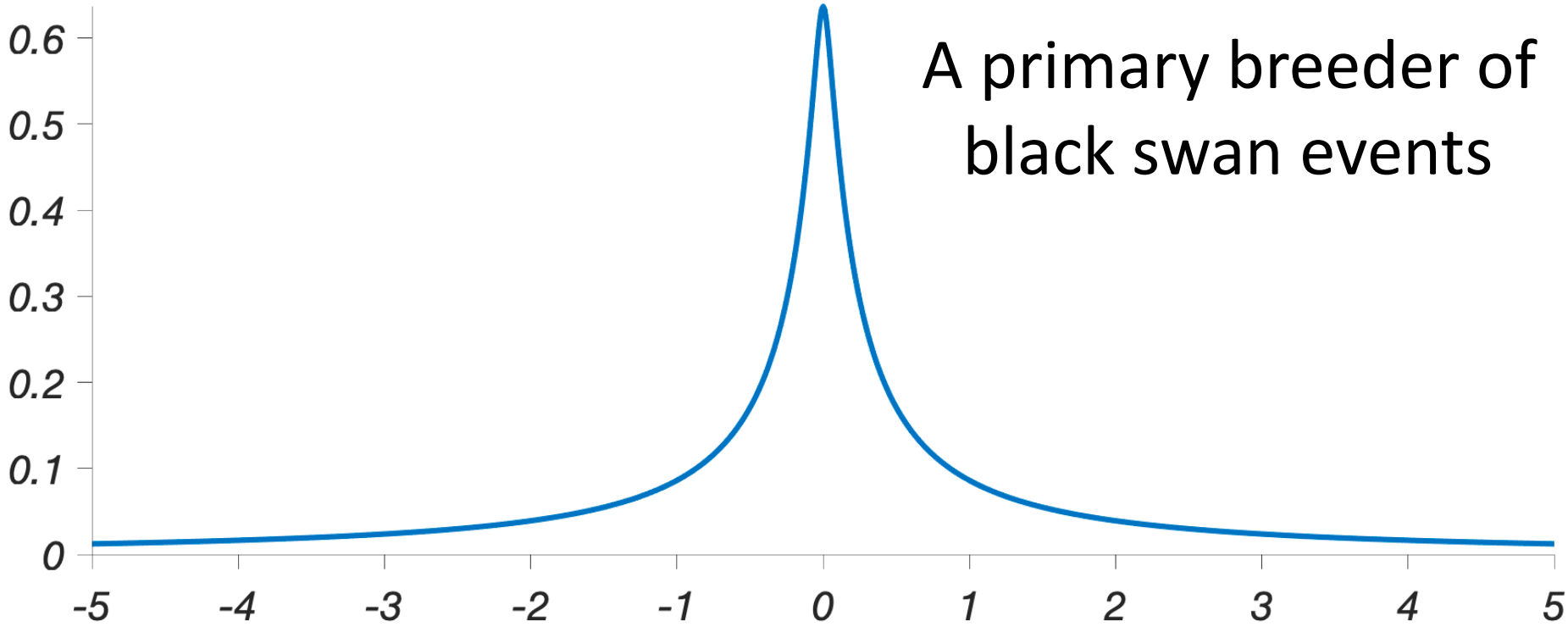
- As a reminder, moments characterize probability distributions (e.g. expected value, variance, skew, kurtosis), however sometimes not fully or uniquely:



- The normal distribution is a symmetric distribution that is fully characterized by the first two moments.
- It results from the combination of many independent events or factors and has “thin tails”, ie extreme valued outcomes are extremely unlikely.

The contrast: The Cauchy distribution

- Both expected value and variance are not defined (!)
- Higher raw moments are also undefined, as their expected value does not converge.
- Thus, given a Cauchy distribution, extreme events are unlikely, but not extremely unlikely ("fat tails").



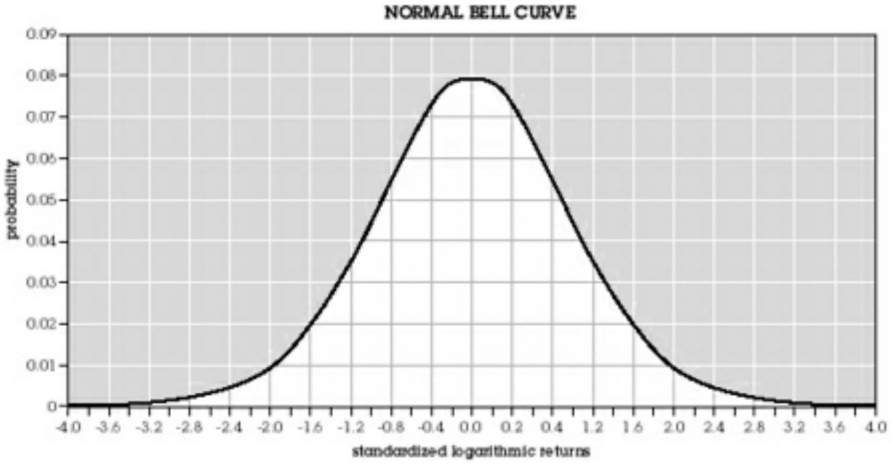
A primary breeder of black swan events

How Cauchy distributions come about

This is not just a theoretical concern: The case of Long Term Capital Management

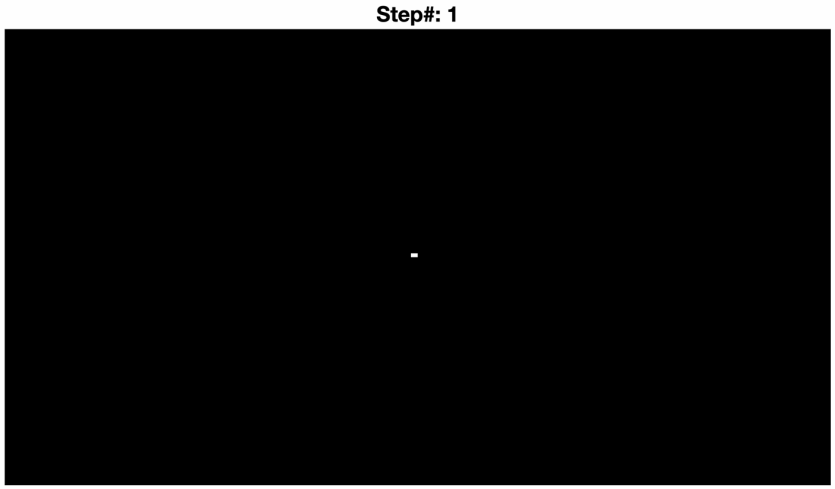
US 3Y | US 5Y The convergence trade

source: tradingeconomics.com



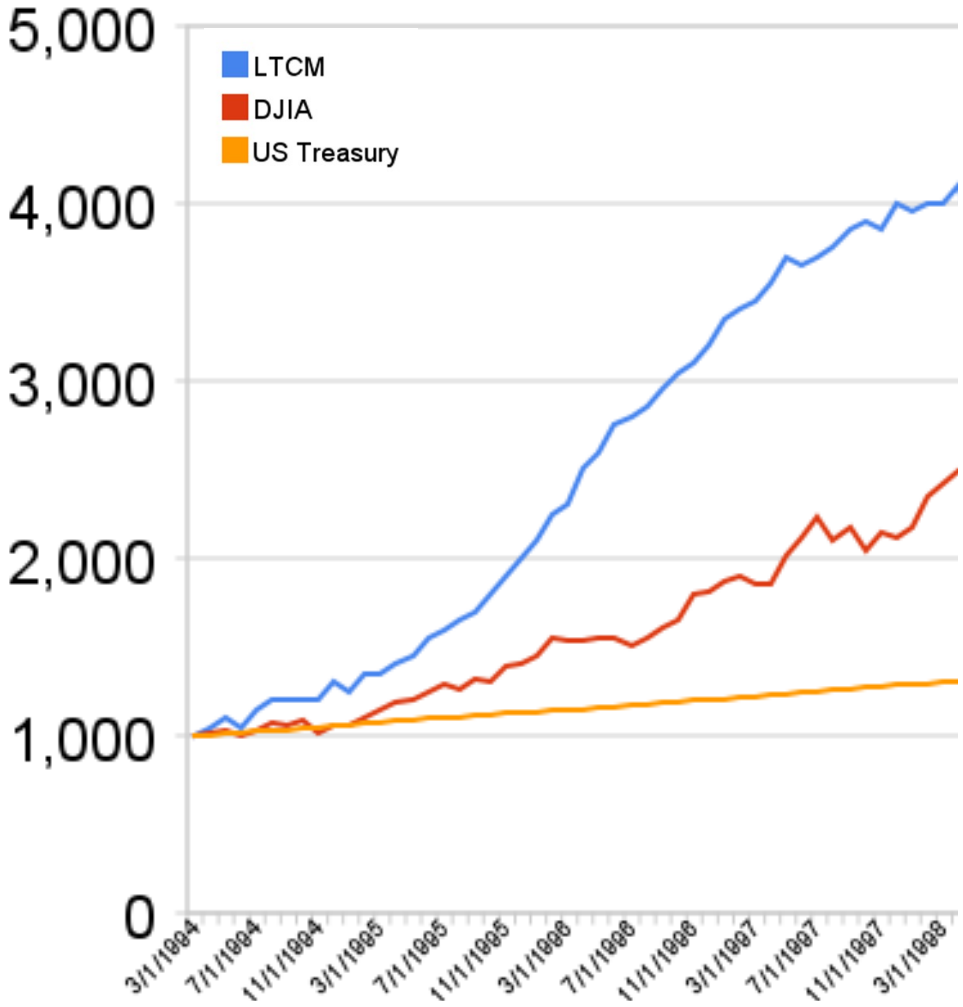
Merton Scholes
Nobel 1997

Brownian motion/
Diffusion modeling:



"We will assume ideal conditions in the market for the stock and for the option. The stock price follows a random walk in continuous time."

The rise and fall of LTCM

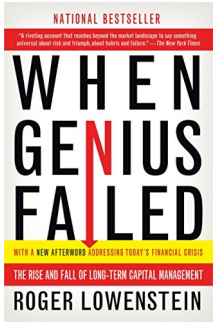


They modeled risk without modeling uncertainty

This led to overconfidence:

"Roughly, over a long period of time," the letter stated, "investors may experience a loss of 5% or more in about one month in five, and a loss of 10% or more in about one month in ten." Only *one year in fifty* should it lose at least 20 percent of its portfolio —and the Merton-Scholes encyclical did not entertain the possibility of losing more.

Then the Russian financial crisis of 1998 hit, and markets started to correlate, leading to a 44% loss in August and more again in September



confident—Long-Term traded on a greater scale, and it kept squeezing nickels long after others had quit. "We focused on smaller discrepancies than other people," one trader said. "We thought we could hedge further and leverage further."

and anticipated risks. The partners assumed that, all else being equal, the future would look like the past. Therefore, in they went. Moreover, its models were hardly a secret. "You could pick up

Hypothesis testing

Highly Simplified View of How the Data Team Works (NBA)

NBA App Team
Basketball Teams
Email Team
Marketing Team
Partnerships Team
Retention Team
Strategy

Request or Question

Data Project Manager

Data Scientist

Model+ Forecast+ Test

Dashboards+ Reports+ Analyses

Data Analyst

Pipelines+ Infra

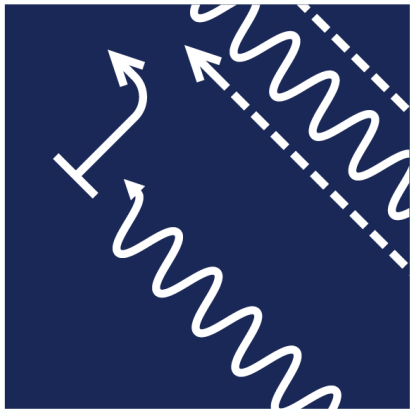
Data Product Manager

Build+Fix

Data Engineer

Blueprint

Data Architect



A Day in the Life of an NBA Data Scientist

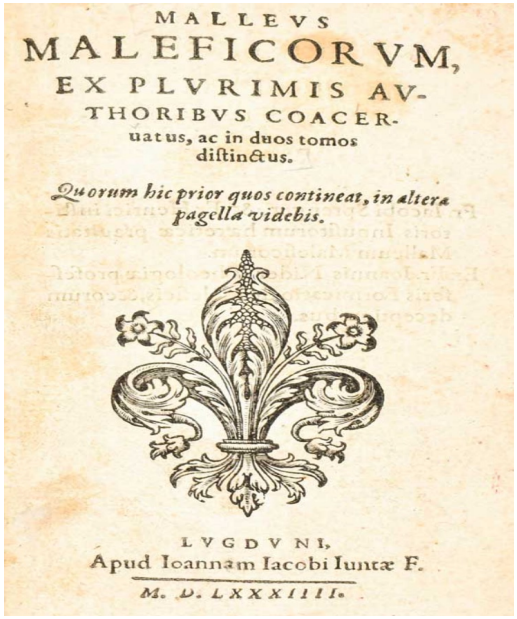
Hypothesis testing

- Hypothesis testing is the primary job of scientists.
- The hypothesis – and the quest to test it – guides the action of the (data) scientist.
- For instance, scientists make measurements to test specific hypotheses.
- These measurements yield data.
- From these data, we might be able to infer the causal processes that generated them.
- But why do this in the first place?
- It is an unusual thing to do. Most people, for almost all of human history have not done this.

The importance of hypothesis testing

Between 1580 and 1630, approximately 50,000 people were burned at the stake for witchcraft in Europe, mostly women

The proximal mechanism


MALLEVS
MALEFICORVM,
EX PLVRIMIS AV-
THORIBVS COACER-
uatus, ac in duos tomos
distinctus.

Quorum hic prior quos contineat, in altera pagella videbis.

LVGDVNI,
Apud Ioannam Iacobi Iuntæ F.
M. D. LXXXIIII.

Heinrich Kramer (1486)

+

The anatomy of a witch trial

Accusation → Looking for supporting evidence → Torture → Confession → Execution

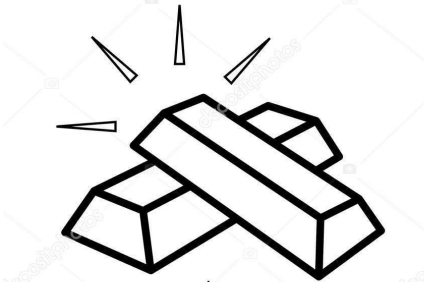
It is not just witchcraft

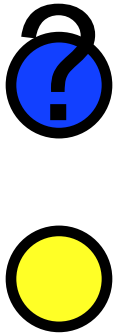
- Bloodletting has a 3,000 year history as a widely used treatment for *most* ailments.
- Theoretical basis: It works so universally because it rebalances the 4 basic humors (imbalances of which cause all disease, according to Hippocrates).
- It took until the late 1800s (!) for it to be shown to be ineffective at treating almost *any* ailment.
- This was shown by hypothesis testing - studies by Louis, Pasteur, Koch and Virchow could not substantiate a difference in outcomes between treatment and control groups.
- How could this go on for so long?

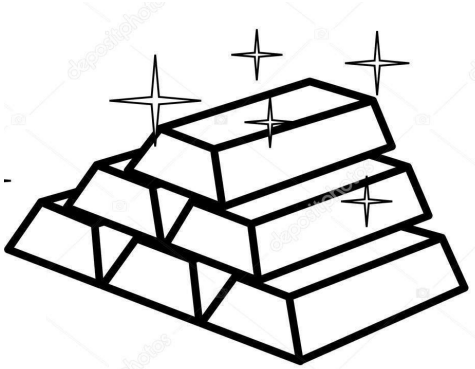
The underlying reason

- There is a human tendency to look for evidence in support of what they already believe.
- Psychologists call this “confirmation bias”.
- However, it might not be a bias, as it encourages rapid and consistent action.
- Such action confers a tremendous advantage.
- However, it is not conducive to eliciting truth under conditions of uncertainty and variability.
- The tendency persists because it is adaptive, not because it leads to truth.

In terms of adaptability, this is not a bias, as it confers first-mover advantage

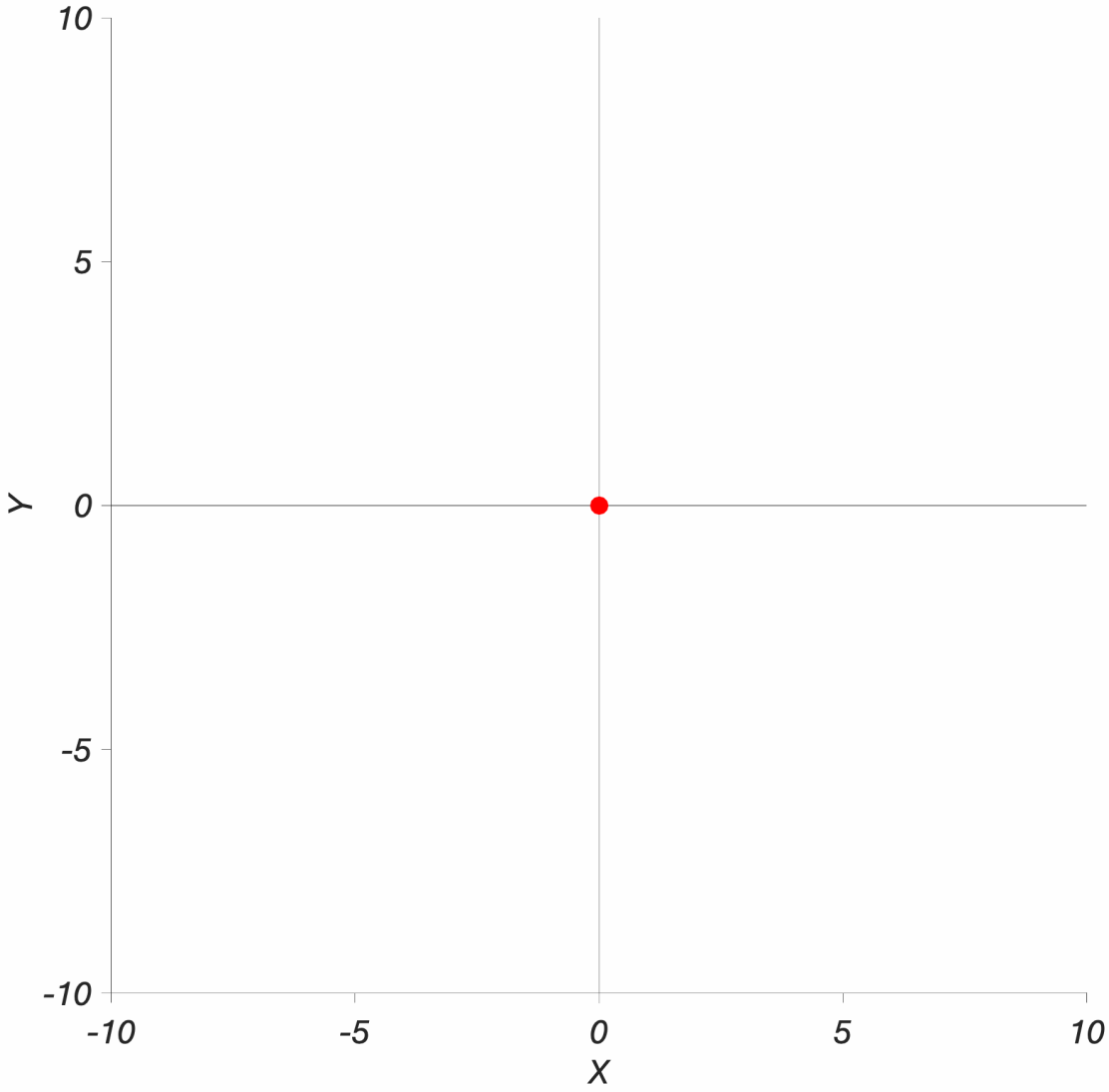






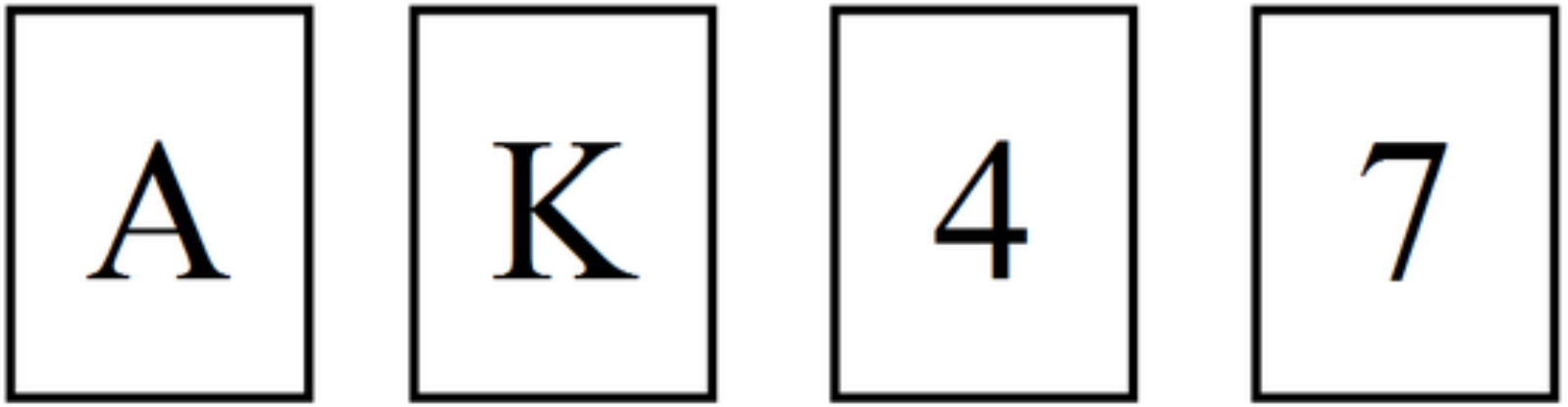
In addition, consistent action gets you much farther:

Distance from origin scales as n vs. sqrt(n) for random walk



Thus, thinking in terms of falsification does not come natural to most people:

Rule: If there is a vowel on one side, there is an even number on the other.

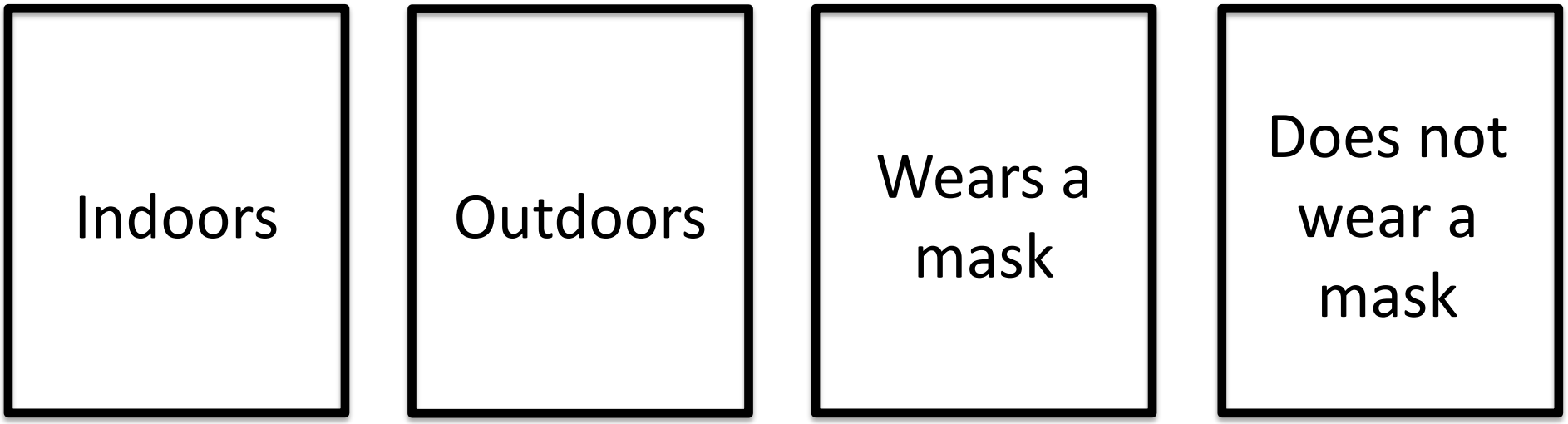


Which cards need to be turned over to test the rule?

Let’s try this again.

Rule:

If one is indoors, one has to wear a mask



Who has to be checked to see whether the rule holds?

We can only learn from instances that have the potential to falsify the rule

Science is a cultural adaptation with the purpose of counteracting the human tendency to seek out confirmation of beliefs that are already held

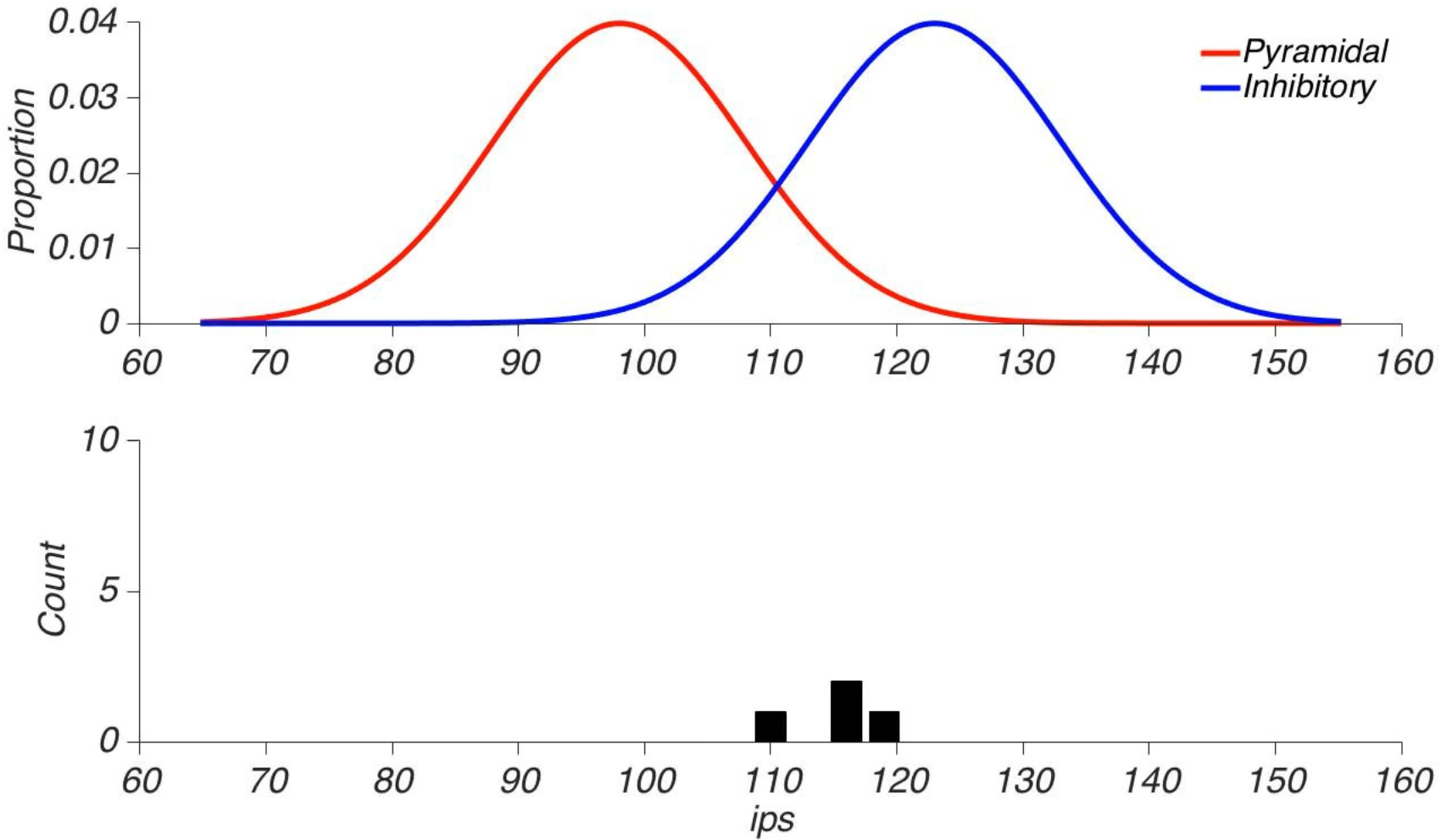
- As this is such an unusual thing to do (and easy to get wrong), we have developed a formal framework that standardizes this approach.
- This framework is called "**Null Hypothesis Significance Testing**" (NHST).
- It was developed in the first half of the 20th century and has since been adopted by most of science (and industry)
- It is extremely commonly used (90+% of papers).
- It is somewhat counter-intuitive and involves several critical concepts, so we have to build up to that.

Sampling

- We want to know whether something is true.
- For instance, the value of some parameter $\boldsymbol{\theta}$ in a **population**.
- The idea is that this parameter $\boldsymbol{\theta}$ is fixed, but unknown.
- It will always remain unknown.
- We need a method to estimate this parameter from a **sample**.
- A sample is the subset of the population that we (can) measure.
- We need to characterize this sample in terms of **statistics**, like the sample mean.
- For instance, the sample mean is an unbiased estimator of the population mean.
- If the sample size n increases, the sample mean converges to the population mean.
- Data -----> Sample statistics -----> Population parameters

Extract **Estimate**

Why do we need to characterize the sample by statistics (like the sample mean?) to make inferences about the population?



Example: You are an epidemiologist and want to know the prevalence of some disease in a small village?

How many do we have to sample to get a good estimate of the – unknown (to us) – overall proportion?

All of them? $\bar{X} = 0$, $n = 1$

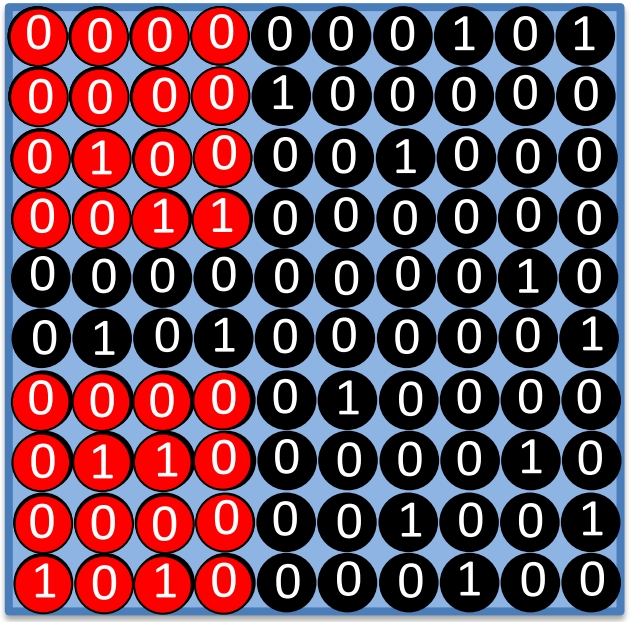
Just 1? $\bar{X} = 0$, $n = 4$

$\bar{X} = 0.11$, $n = 9$

$\bar{X} = 0.19$, $n = 16$

$\bar{X} = 0.25$, $n = 16$

Population parameter: $\theta = 0.2$



How large the sample size needs to be for a good estimate of the population parameter depends on how rare the event is

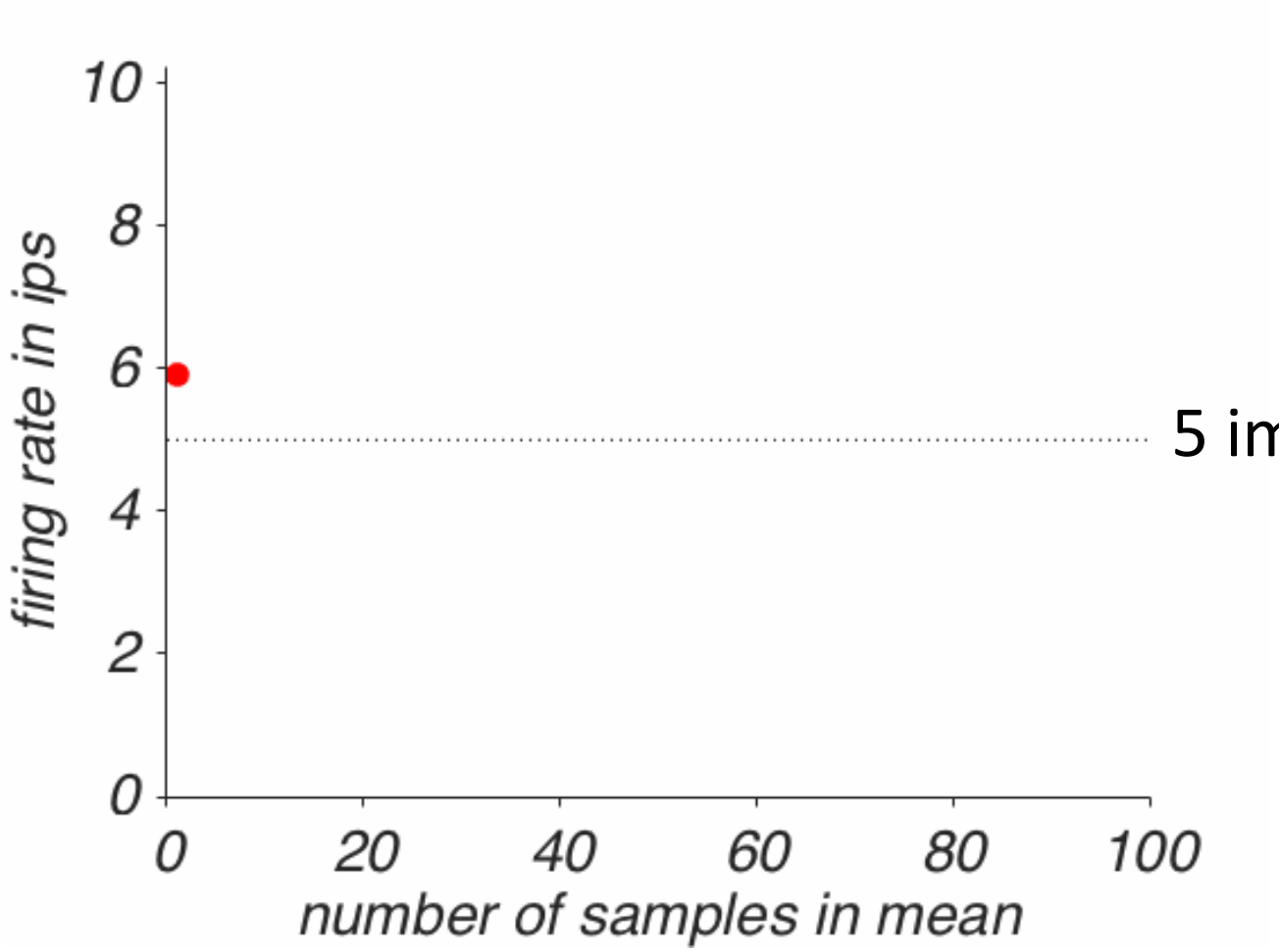
The (weak) law of large numbers

If the sampling from RVs is **iid** (independent and identical) and **representative** (each member of the population has an equal chance of being in the sample), then the sample mean will converge to the population mean, as sample size increases.

$$\lim_{n\to\infty} P(|\bar{x}_n - \mu| \geq \varepsilon) = 0$$

This often works well and serves as the cornerstone of the scientific process

What is the “true” firing rate of this neuron? Every measure we take will be ”contaminated” by random (ideally) variability:



We can go farther than that: **The central limit theorem**

- Briefly: If sampling randomly and independently, the sample means distribute normally as the sample size increases, regardless of how the underlying population is distributed.
- This is indeed absolutely central.
- If this was not true, inferential statistics (and (data) science as we know it) would not be possible.

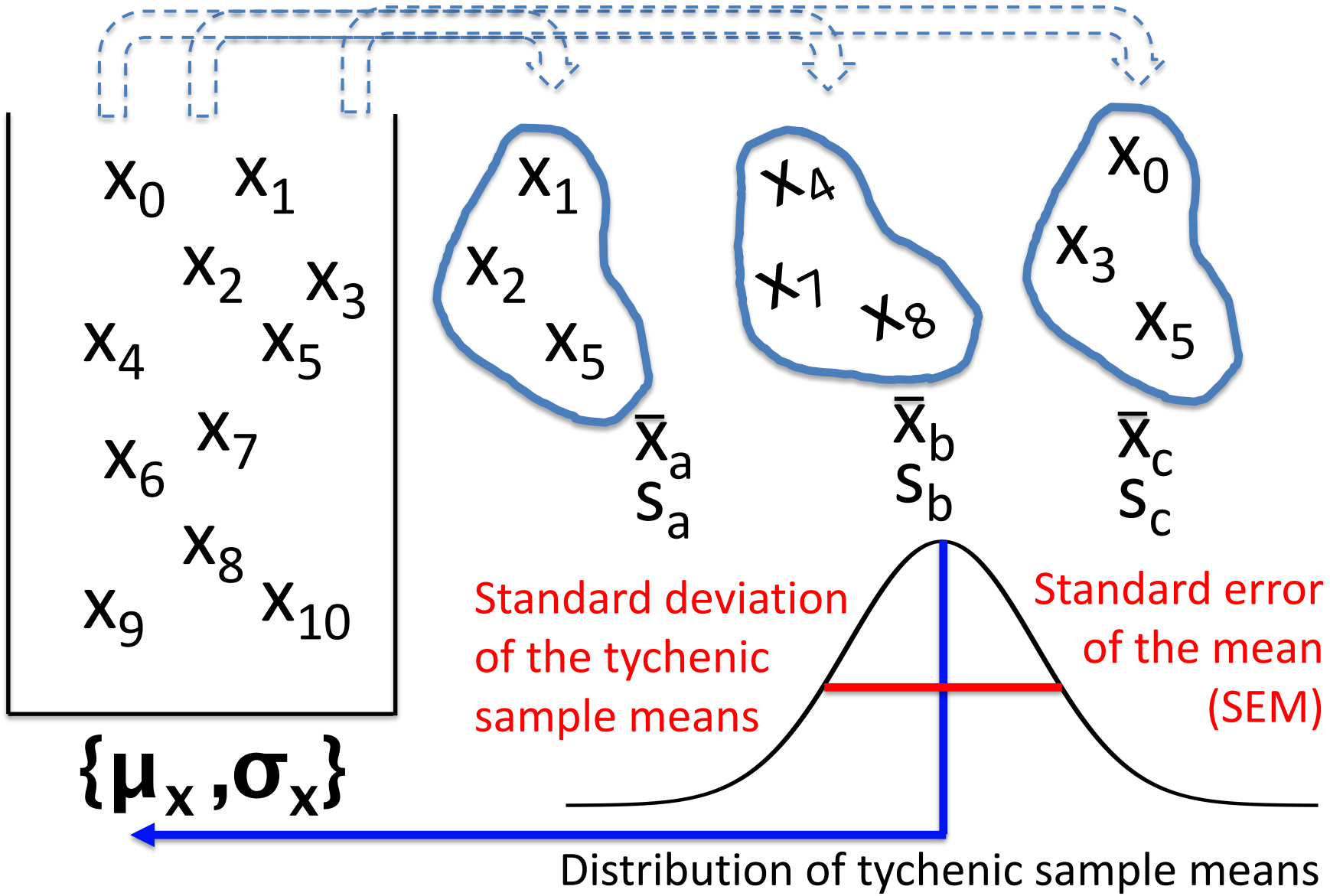
The likely **sampling error** relies on the central limit theorem (CLT)

We will use simulations to illustrate its properties

To illustrate the CLT and how it allows for a good estimate of the population mean as the sample size **n** gets large, we are going to draw repeatedly and randomly (random sampling) from the same population, **t** times, for each sample size **n** (the sample size in each sample). We take the sample mean $\bar{X}_t$ for each sample t_i and plot the distribution of these **tychenic** sample means.

Example: We want to estimate how many times per week the average person in a population watches a Netflix show

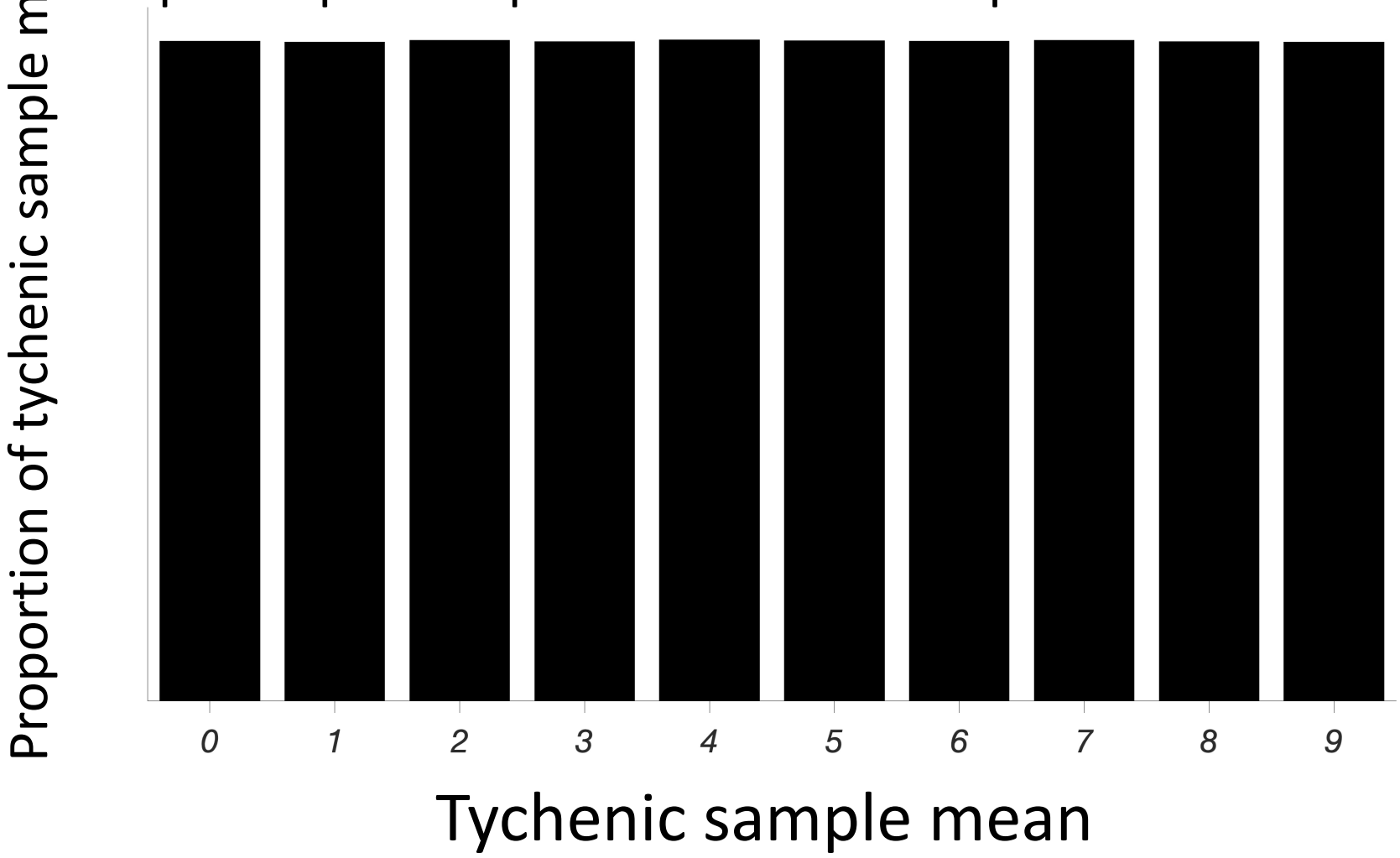
Some necessary conceptual clarifications



The tychenic sample means distribute normally, as n increases

n = Sample size, data points per sample.

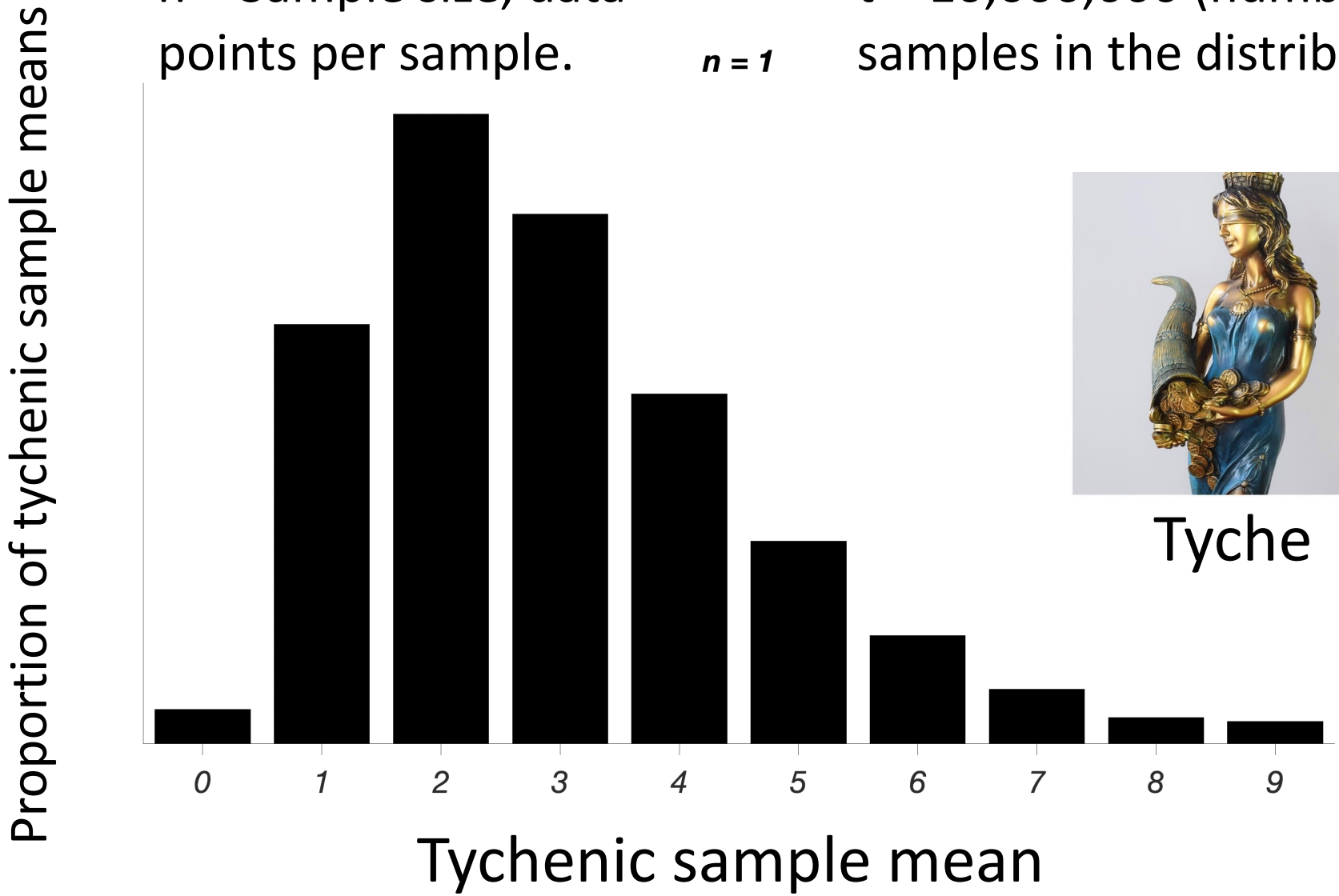
t = 10,000,000 (number of samples in the distribution)



Regardless of the shape of the underlying distribution in the population, e.g. gamma here.

n = Sample size, data points per sample.

t = 10,000,000 (number of samples in the distribution)



Tyche

What is the point of this?

- Here, we sampled 10,000 times from the underlying population (and its distribution), per sample size n.
- We could do this because this is a simulation.
- In reality, you will usually only be able to obtain a single sample (the dataset) and you will never know what the mean in the population (**ground truth**) really is.
- You want a reasonable chance that the sample mean of the one sample you have is a good (enough) estimate of the unknown population mean.
- For this to happen, we need to consider how scattered/variable these sample means are.
- This is captured by the standard deviation of the sample means (more commonly – referred to as **SEM**).

How does the standard deviation of the tychenic sample means develop as function of sample size?

- How large does our sample size n need to be, in order for the sample mean to be a close estimate of population mean?
- Enter the **standard error** (of the sample mean) or **SEM**, another term for the standard deviation of the sample means:

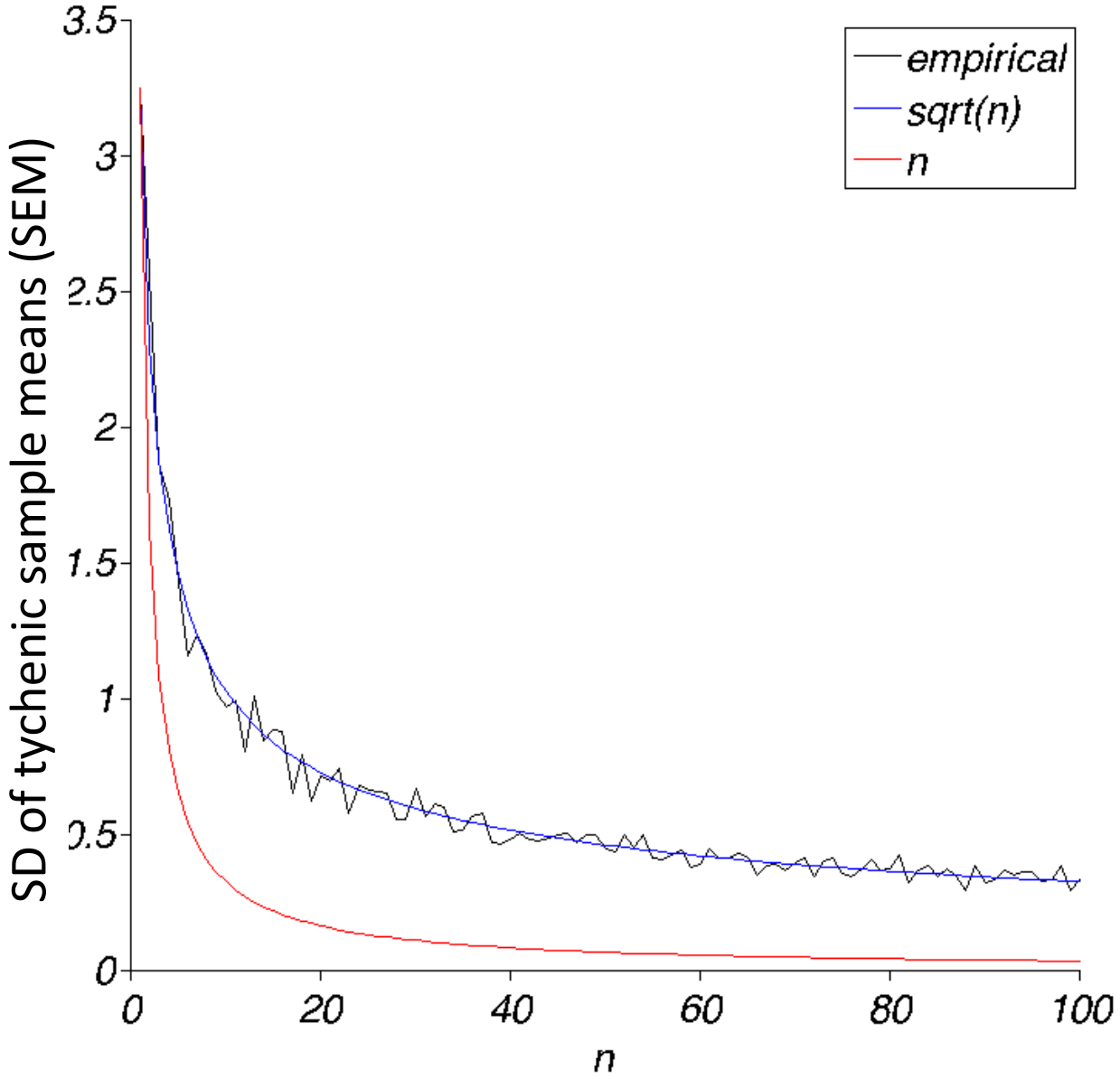
$$SEM = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Why take the square root of n?

- The variance of the sample mean scales inversely with n.
- The more samples go into the sample mean, the less the variance.
- But we are interested in standard deviation.
- So we take the square root on both sides.

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

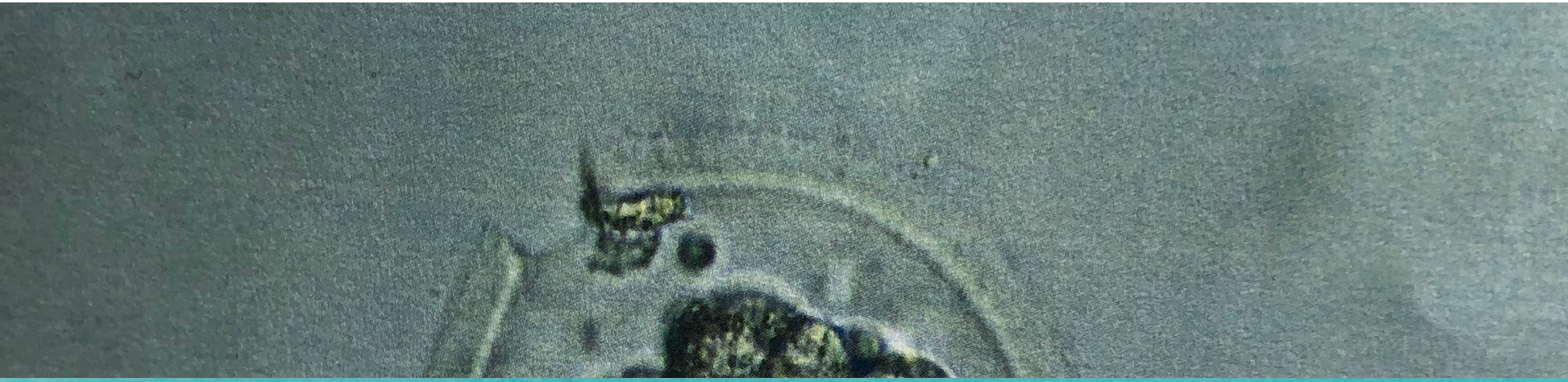
Visualizing how the SEM decreases as n increases



CLT caveats

- CLT reduces **sampling error** sufficiently, if the sample is sufficiently large.
- The mechanism behind this is that the larger the sample, the larger the chance for random error to cancel out, effectively amplifying the signal.
- However, this assumes – and applies *only* if we indeed sample randomly.
- If sampling is not random, there is potential for **sampling bias**.
- If sampling bias exists, the sampling error will be larger than suggested by CLT (and usually systematic, not random error), so the conclusions can be misleading.

Conclusions about embryo quality (and which embryo to transfer) depends on the representativeness of the sampling (biopsy)



ABNORMAL CELLS IN EMBRYOS MIGHT NOT PREVENT IVF SUCCESS

Study shows that chromosomal abnormalities in embryonic cells may be more common than previously thought and these conditions may lead to development of healthy babies during IVF

Doug Donovan / Jul 10

The way to control for these potential confounds and biases is by doing a scientific experiment.

What is a scientific experiment?

- Everyday language meaning of an experiment:

  "Try something and see what happens"

- **Mathematical experiment**: A procedure that has a well-defined set of possible (stochastic or deterministic) outcomes.
- **Scientific experiments** share components of these, but have a much stricter definition.

Scientific experiments

- An "**experiment**" has a specific meaning in science.
- It means that **units of analysis** (often participants/users) are ***randomly*** assigned to different experimental **conditions** ("**treatments"**) that you control.
- These treatments are called the **independent variable (IV)**.
- The advantage is that this random assignment – given a sufficiently large number of participants – effectively destroys (or "controls for") all other potential (pre)-existing systematic relationships in the data. Control emerges from randomization.
- This systematic **intervention** while controlling for all other factors by randomization allows to conclude causality by effectively eliminating all possible confounds, known or unknown.
- In other words, if one finds systematic differences in the measurements (**dependent variable, DV**) as a function of the IV, one can conclude that this intervention **caused** them.
- So whenever possible, one should do experiments if one is interested in causality.

What makes experiments special?

- Jointly:
- 1) **Systematically varying** some independent variable (IV), to create several experimental conditions.
- 2) **Randomization:** Randomly assigning units of analysis to these conditions – each to receive a particular level of the IV
- Together these constitute the **intervention**, not just observation
- *Then* observation**: Measurement** of a dependent variable (DV)
- Then math**: Compare** data from different conditions
- Then logic: Randomization ensures that only the IV systematically varies between conditions, so we *can* attribute **causality** (the effect of the IV on the DV), if there is a difference between the data from different conditions.

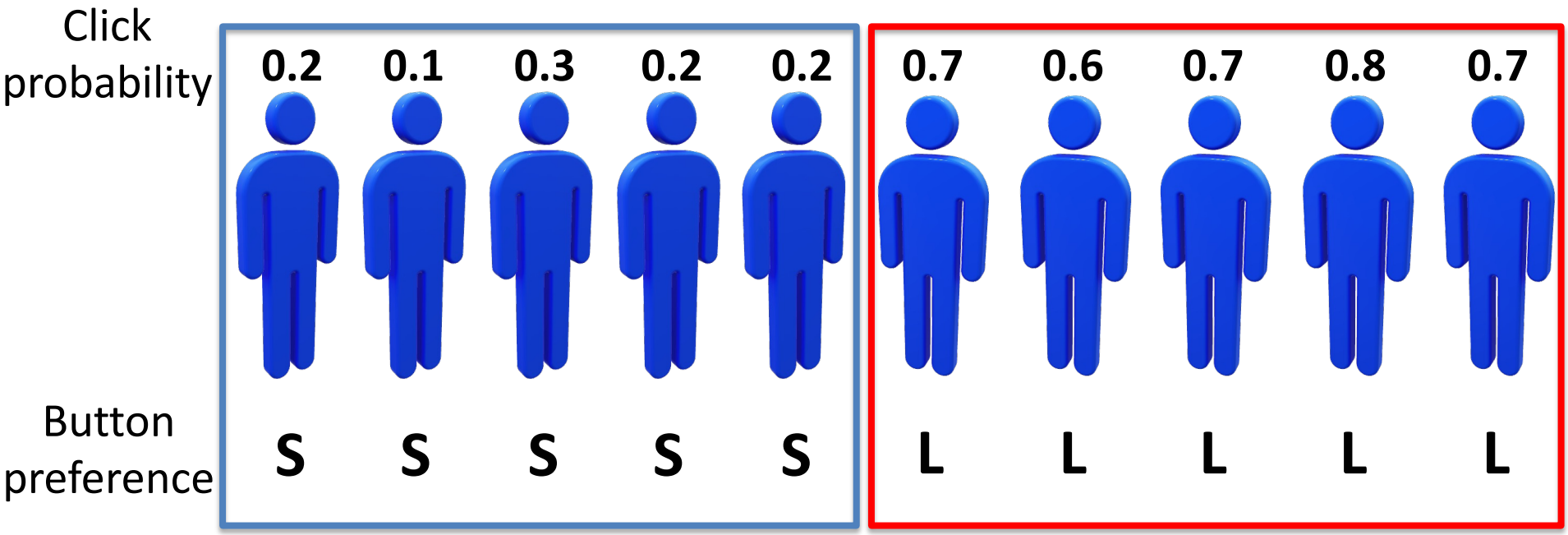
If done properly, this procedure provides **experimental control** of all possible confounds, even unknown ones, allowing strong causal inference

Let’s illustrate the logic of experiments with

A/B tests

- You want to know if making buttons bigger increases the probability of users clicking on them.
- How about this:
- We let users decide which site design they prefer (small or large buttons).
- Then, we measure the click rates of each group.
- We determine which design maximizes click rates.
- That’s the design we’ll adopt for everyone later.
- Yes? No?

What could have happened with this study design



Mean click rate: **0.2** **0.7**

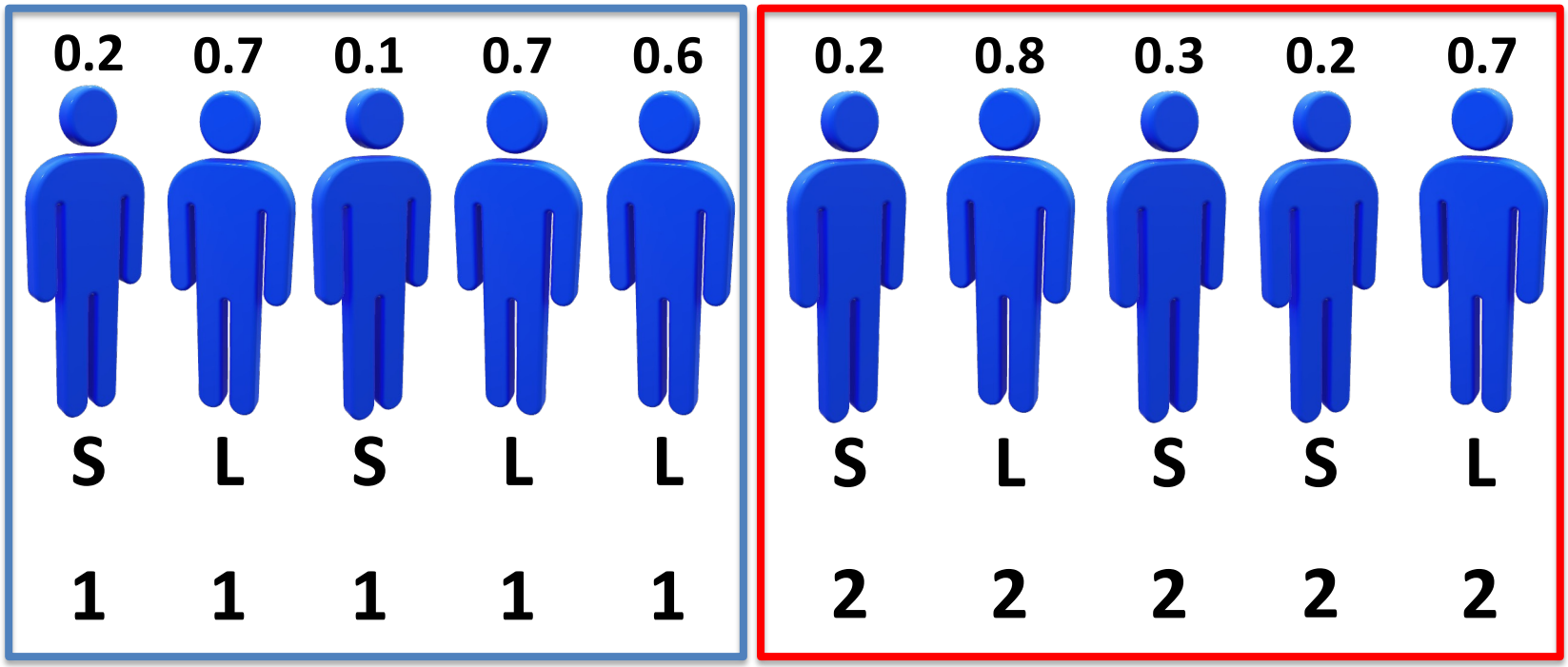
The data seems to strongly suggest that large buttons seem to be associated with higher click rates

Would you recommend to adopt large buttons for everyone?

Problem: There is a confound – for whatever reason, there is a correlation between button preference and click probability.

This would make larger buttons look more effective than they are

The same people and data, but now with an **experimental** design – the A/B test (no self-selection)



Mean click rate: **0.46** **0.44**

Assigning the same users randomly to conditions, regardless of their preference neutralized the confound and revealed that the effect of the large buttons on click rates was largely spurious

Whether this remaining difference is meaningful has to be interpreted in the framework of "**statistical significance**"

Before we can talk about **null hypothesis significance testing**, we first need to clarify what **statistical significance** is

- As seen on social media:
- “Last year, 19 people in the US were killed by lightning strikes, but in a country of 340 million, that is not statistically significant”
- “In 2019, there were 27 deaths associated with vaping, but given how many people smoke, that is not statistically significant”

An Association between Air Pollution and Mortality in Six U.S. Cities

Douglas W. Dockery, C. Arden Pope, Xiping Xu, John D. Spengler, James H. Ware, Martha E. Fay, Benjamin G. Ferris, Jr., and Frank E. Speizer

N Engl J Med 1993; 329:1753-1759 | December 9, 1993 | DOI: 10.1056/NEJM199312093292401

BACKGROUND

Recent studies have reported associations between particulate air pollution and daily mortality rates. Population-based, cross-sectional studies of metropolitan areas in the United States have also found associations between particulate air pollution and annual mortality rates, but these studies have been criticized, in part because they did not directly control for cigarette smoking and other health risks.

Full Text of Background...

METHODS

In this prospective cohort study, we estimated the effects of air pollution on mortality, while controlling for individual risk factors. Survival analysis, including Cox proportional-hazards regression modeling, was conducted with data from a 14-to-16-year mortality follow-up of 8111 adults in six U.S. cities.

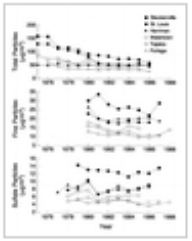
Full Text of Methods...

RESULTS

Mortality rates were most strongly associated with cigarette smoking. After adjusting for smoking and other risk factors, we observed statistically significant and robust associations between air pollution and mortality. The adjusted mortality-rate ratio for the most polluted of the cities as compared with the least polluted was 1.26 (95 percent confidence interval, 1.08 to 1.47). Air pollution was positively

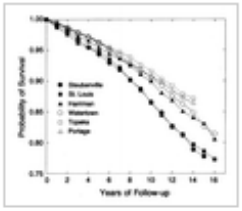
MEDIA IN THIS ARTICLE

FIGURE 1



Annual Average Concentrations of Total Particles, Fine Particles, and Sulfate Particles in the Six Cities.

FIGURE 2



Crude Probability of Survival in the Six Cities, According to Years of Follow-up.

ARTICLE ACTIVITY

2123 articles have cited this article ›

What does “statistically significant” mean?

A. The effect is large
B. The effect is important
C. The null hypothesis is false
D. The alternative hypothesis is true
E. The null hypothesis is true
F. The alternative hypothesis is false
G. The observed data is unlikely, assuming chance

Common misconceptions:

- "Statistically significant" is a synonym for
- Substantial
- Important
- Big/Large
- Real

- Statistical significance gives the
- probability that the null hypothesis is true
- probability that the null hypothesis is false
- probability that the alternative hypothesis is true
- probability that the alternative hypothesis is false

Statistical significance means:

- an observed **result** is **unlikely** to be due to **chance** alone.
- Put differently, it is not plausible that the data came from a **random number generator** (RNG)
- Formally:
- The probability of the **data**, **assuming** that the null hypothesis is **true** is less than the **chosen** significance level.
- It is a statement about the probability of data, *not* hypotheses.
- If something is statistically significant, you did *not* show that the null hypothesis is false.
- Just that it is sufficiently unlikely to be true, so we revisit our assumption that it was true - but it is a **choice**…
- At first glance, this seems surprising and backwards twice over:
- Naively, science should want: p(Experimental hypothesis | Data)
- What science actually computes: p(Data | Null hypothesis)
- The reason this is done is because once we assume the null hypothesis is true, we *can* compute the probability of the data.
- Inverting this requires Bayes' Theorem (and a prior).

Significance testing example

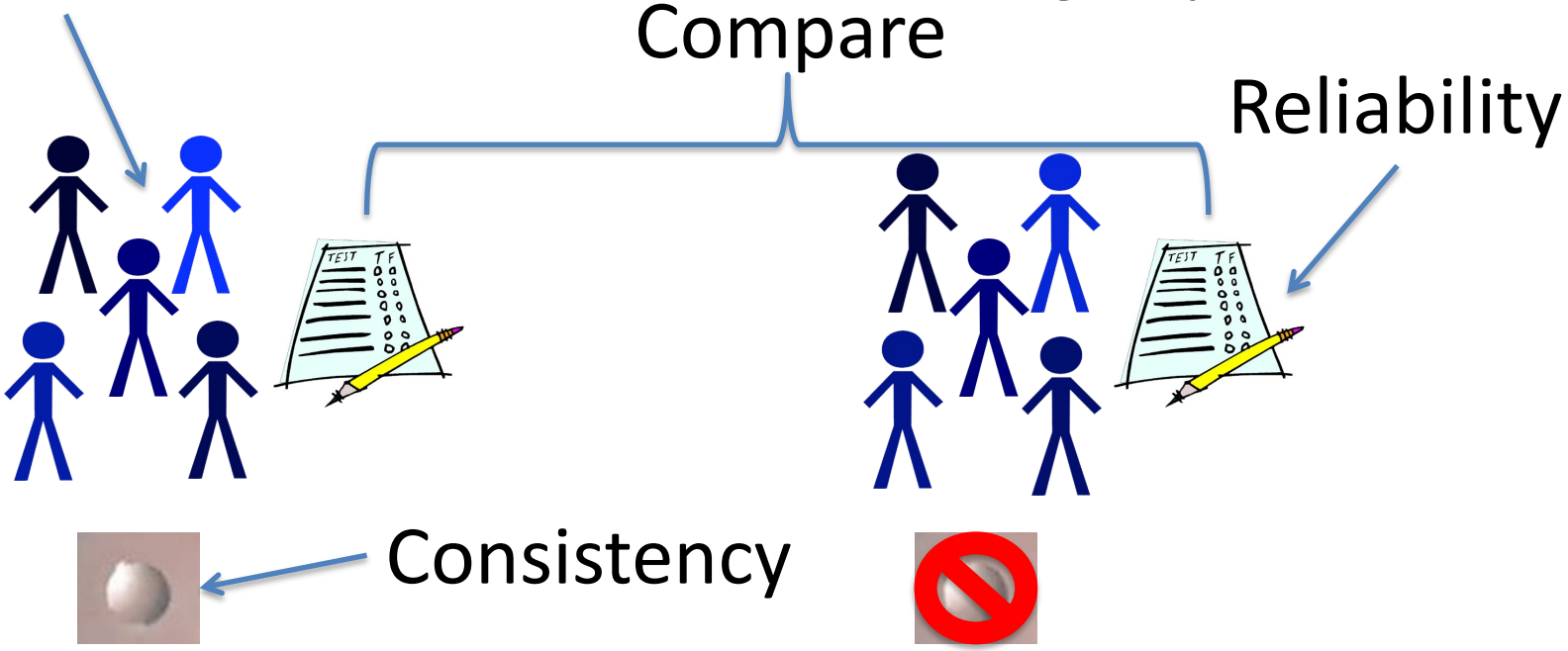
- You are a Data Scientist working for a major pharmaceutical company in New Jersey.
- Your chemists developed a new drug – NZT – that supposedly improves IQ.
- Question: Does NZT improve IQ?

Why are groups needed?

Natural variability

Compare

Reliability

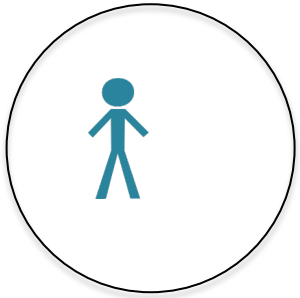


We want to know whether NZT increases IQ

- The issue is that in the real world, there are always sources of variability (this manifests as sampling error).
- Any observed result (in terms of difference between the two groups) - no matter how extreme - could be due to random chance alone (variability of uncertain origin that we do not control).
- So it could look like the drug is working even if it is not.
- The only question is how likely such an outcome – showing a difference – is, assuming chance.
- This likelihood can be assessed by the procedure of **null hypothesis significance testing**.

In the next several slides, we will illustrate this procedure by

- Sampling people from the population.
- We will then assess whether they are brighter (smarter) or dimmer (less smart) than average.
- We will represent people smarter than average in the sample with a bright blue pictogram: 
- We will represent people less smart than average in the sample with a dim blue pictogram: 
- To implement the null hypothesis significance testing framework, which is based on falsification, we assume that the drug does *not* work.
- In other words, all observed variability represents chance variability (due to sampling)



$p = 1/2^1$

$p = 0.5$

1) How likely is it to observe a given sample (assuming all variability is due to chance)?

2) How unlikely does the outcome have to be for you to abandon the assumption that this result is solely due to chance?

p = 0.25 **p = 0.125** **p = 0.0625** **p = 0.0313** **p = 0.0156**

$p = 1/2^2 = 1/4$

$p = 1/2^3 = 1/8$

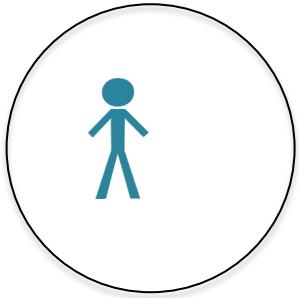
$p = 1/2^4 = 1/16$

$p = 1/2^5 = 1/32$

$p = 1/2^6 = 1/64$

Where would you draw the line?

In other words, what outcome would convince *you* that the drug *is* effective?



$p = 1/2^1$

$p = 0.5$

This is where science draws the line:
(as of today)

$p = 0.25$ $p = 0.125$ $p = 0.0625$ $p = 0.0313$ $p = 0.0156$





$p = 1/2^2 = 1/4$ $p = 1/2^3 = 1/8$ $p = 1/2^4 = 1/16$ $p = 1/2^5 = 1/32$ $p = 1/2^6 = 1/64$

0.05

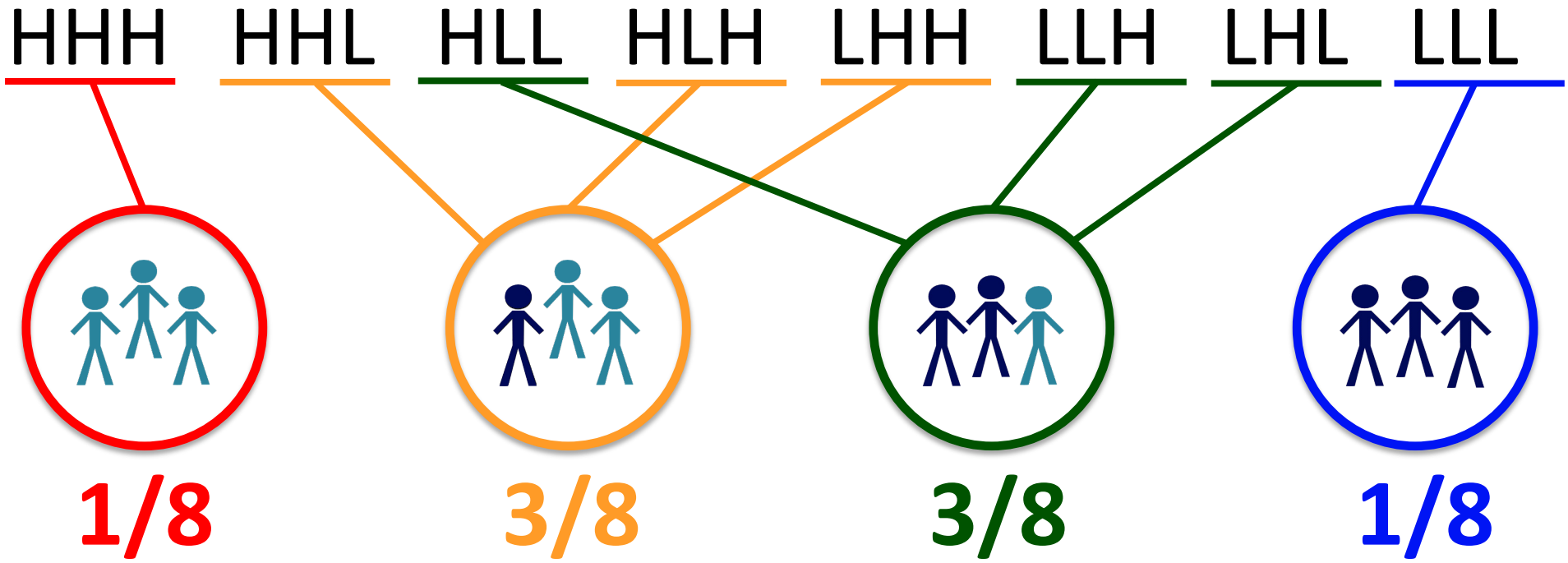
Where would you draw the line?

In other words, what outcome would convince *you* that the drug *is* effective?

How to calculate mixed outcomes

Example: Sampling 3 people from the population

Here are all 8 possible outcomes (given chance):



Probability of the outcome ”1 lower, 2 higher” than average (or more extreme) by chance:

$$p = 1/8 + 3/8 = 4/8 = 1/2 = 0.5$$

What does p<0.05 look like in this example?

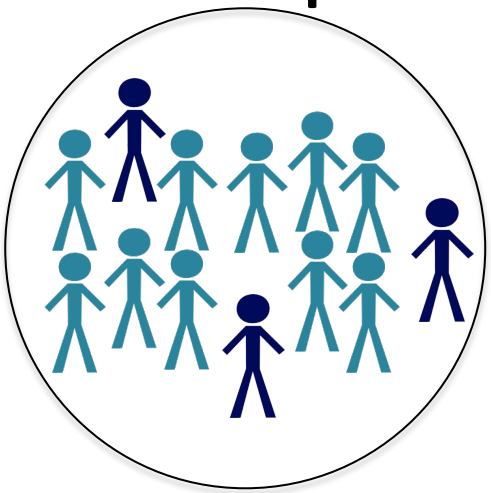
$\binom{13}{10}$ = 0.0461. Here is why:

There are 2^{13} = 8192 possible combinations

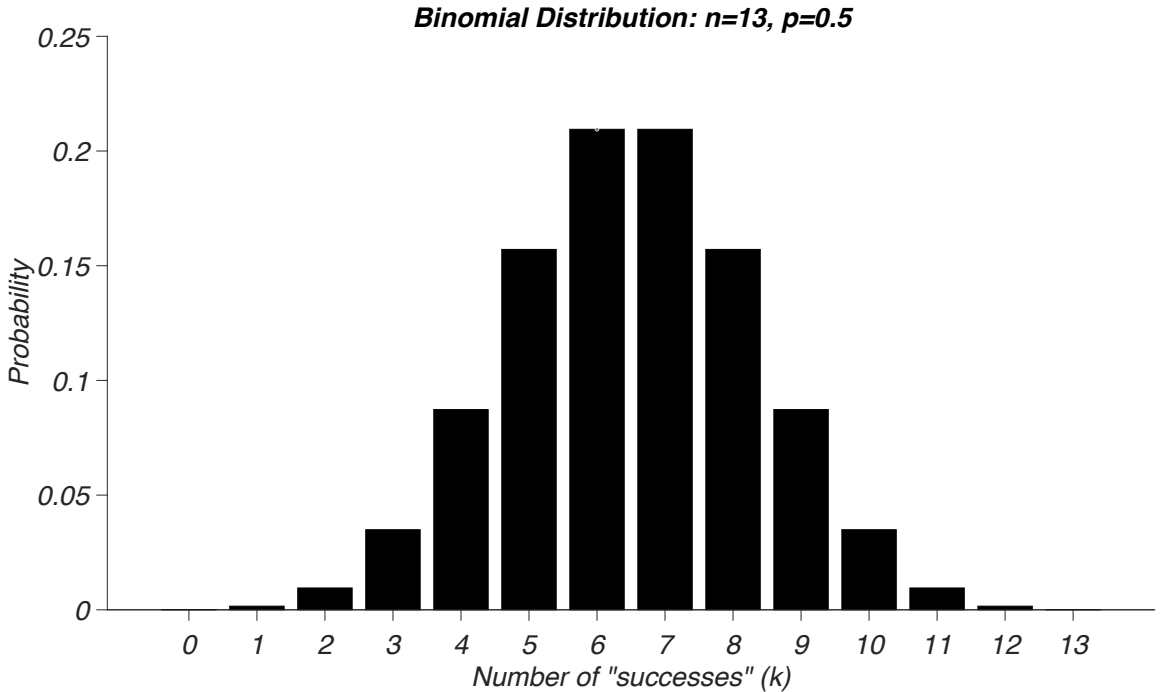
$\binom{13}{13}$ = 1. $\binom{13}{12}$ = 13. $\binom{13}{11}$ = 78. $\binom{13}{10}$ = 286.

1+13+78+286 = 378 combinations of interest

→10 *or more* = 378/8192 = 0.0461



More generally, the probability of any given outcome for discrete cases like this follows a binomial distribution:



Significance levels

- Are denoted by the Greek letter **α**.
- In principle, we can pick anything that we consider unlikely enough.
- In the scientific literature, the **consensus** is that a level of 0.05 or 1 in 20 is considered as unlikely enough.
- A level of 0.01 or 1 in 100 is considered “highly significant” or really unlikely.
- There have been proposals to move the consensus level of “too unlikely to be reasonably consistent with chance” from 0.05 to 0.005 (Benjamin et al., 2018).

Where does the standard significance level $\alpha = 0.05$ come from?

only once in 370 trials, while Table II. shows that to exceed the standard deviation sixfold would need nearly a thousand million trials. The value for which $P = \cdot 05$, or 1 in 20, is 1·96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion, we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice *if the data are insufficiently numerous to bring them* out, but no lowering of the standard of significance would meet this difficulty. Careful: Con-venience → Con-vention

From: Fisher RA (1925). *Statistical Methods for Research Workers*

Kn-ow(e)-l-edge

- More advanced experimental designs (beyond A/B tests)
- More advanced sampling techniques (e.g. multi-armed bandits)
- Causal inference
- Bayesian inference