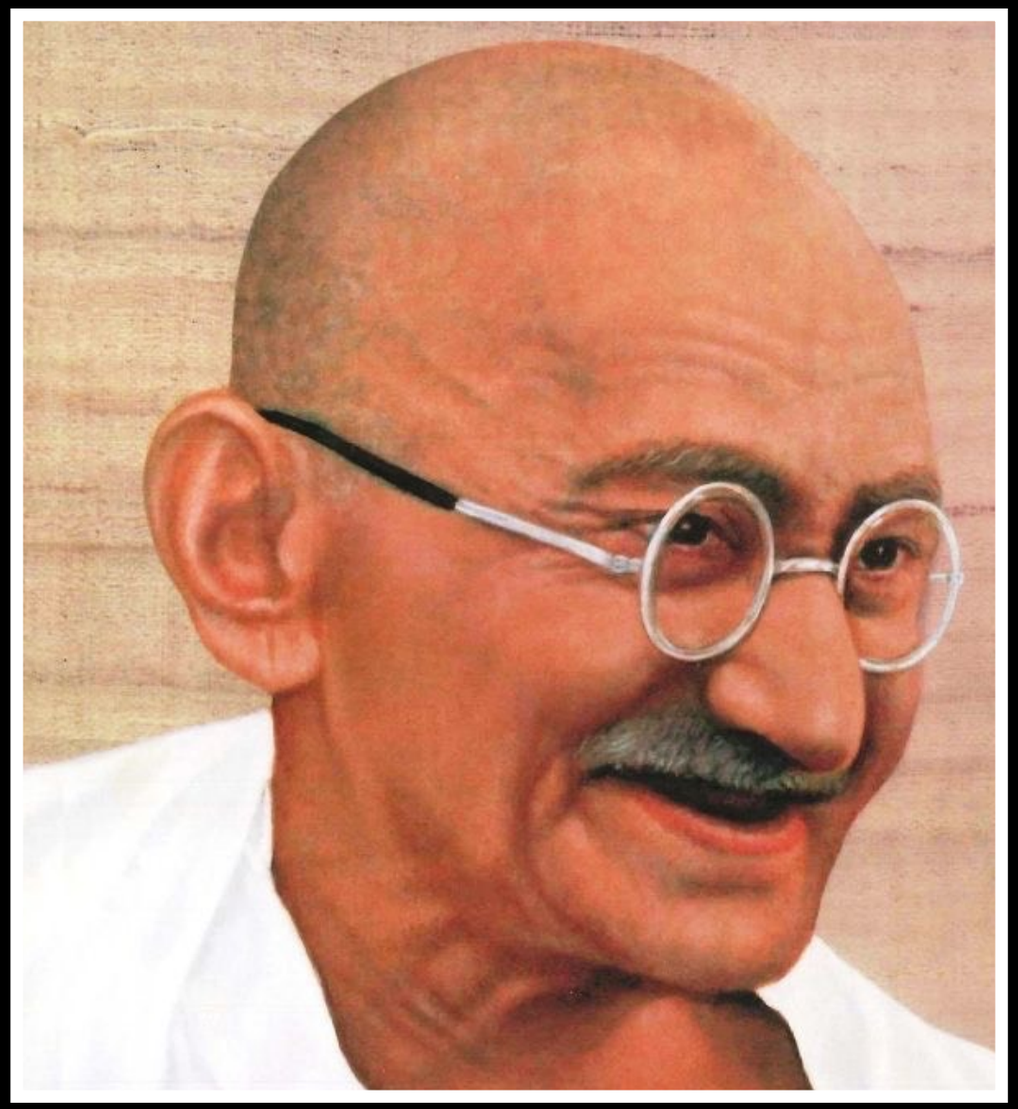
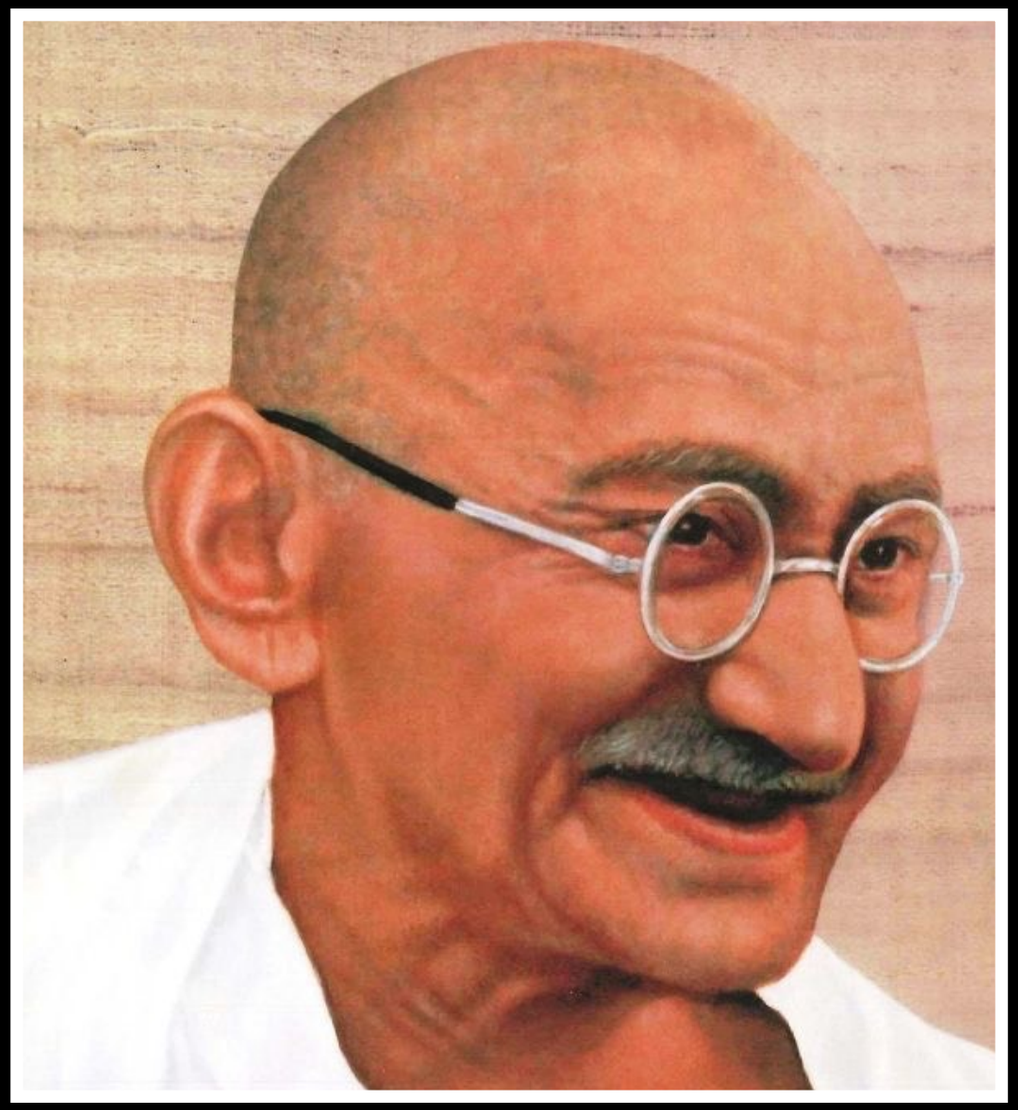


What is life worth without **Trials…**

- Mahatma Ghandi



What is life worth without **Trials...** and tribulations which are the salt of life.

- Mahatma Ghandi

Goal

Understand what <u>trials</u> are in the context of computing probabilities in high energy physics

Overview

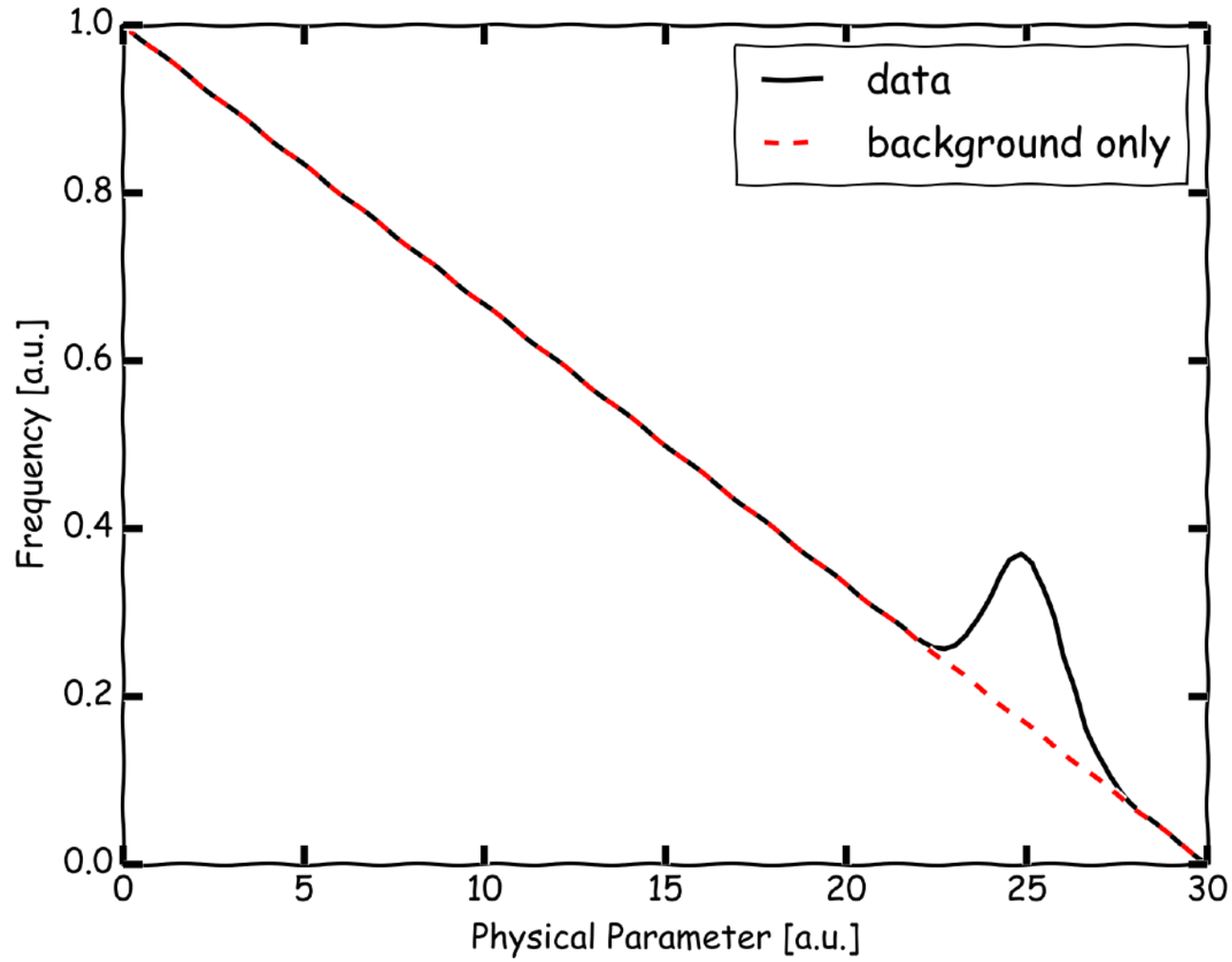
Recap of probability

Define trials, work through examples

Python Examples

Probability Recap

- When trying to discover new phenomena, we often talk about the **probability that an observation is a result of the background (null hypothesis)**



- Fancy people (statisticians) call this a **Type I error**

Probability Recap

- This Type I error probability can be measured in a number of ways. Incomplete list of methods:

brute force

$$\mathbf{P = \frac{\#\ bkg\ tests\ above\ threshold}{total\ bkg\ tests}}$$

analytic description

$$\mathbf{P(N, mu) = \frac{mu^{-N}\ exp(-mu)}{N!}}$$

Example: Poisson Dist.

likelihood formulation

$$\mathbf{TS = -2 \times \log(L_{bkg} / L_{sig+bkg})}$$

TS —> P

Probability Recap

- **Want to ensure the probability of Type I errors is small** to be confident we are observing real signal

- **P = 2.87 x 10^{-7}** (1-sided, 5σ) is the canonical threshold for confident detection of real signal

- **Important Caveat:** Often the formulation of **P** only accounts for a single analysis of the data.

  However, we frequently examine data more than once!

Overview

Recap of probability

Define trials, work through examples

Python Examples

Concept of Trials

- Each time you obtain the answer of an analysis from real data is called an **experimental trial***

- It is very common to perform multiple trials when working with a specific analysis technique:

Example:

A point source likelihood returns the probability that **one** location is consistent with background exp.

If we apply the analysis to a list of 30 source candidate locations from other measurements of known objects there are <u>30 trials</u>

*my definition, direct hate mail to me

Examples Continued

Example: Repeating an analysis with new cuts

Count number of photons/neutrinos with E > 1 TeV, E > 10 TeV in 1 year of data (2 trials)

Example: Repeating with different signal hypotheses

Look for diffuse Galactic plane neutrino emission using both a Fermi pi0 map and a KRA model (2 trials)

Examples

Example: Repeating existing analysis with new data

Independent Case:

Search for a new source/signal in 1 year of data,
Repeat the search in the next year of data (2 trials)

Correlated Case:

Search for a new source/signal in 1 year of data,
Repeat with 2 years of data, including the first year
(2 trials)

Does everyone have a reasonable sense of what a trial is?

Why Trials Matter

- The more times you examine data, the more opportunities you have to find a rare background fluctuation that can be mistaken for signal

Toy Example: Finding the Queen of Hearts

What's the probability for finding the Queen of Hearts if you pull a random card from a deck of cards?

$$\mathbf{P = 1/52}$$

What's the probability if you have two tries with replacement?

$$\mathbf{P = 1/52 \times 51/52 + 51/52 \times 1/52 + (1/52)^2 \sim 2 \times 1/52}$$

Why Trials Matter

- The more times you examine data, the more opportunities you have to find a rare background fluctuation that can be mistaken for signal

Toy Example: Finding the Queen of Hearts

Wh

key concept:

We roughly doubled our odds of finding something rare by looking twice!

Need to account for trials or else you can fool yourself into thinking signal is there just by looking at data more times

Terminology

- Probability of a single trial is called pre-trial prob.

$$\mathbf{P_{pre} = 1/52}$$

- Probability after 2 trials is called post-trial prob.

$$\mathbf{P_{post} = 1/52 \times 51/52 + 51/52 \times 1/52 + (1/52)^2}$$

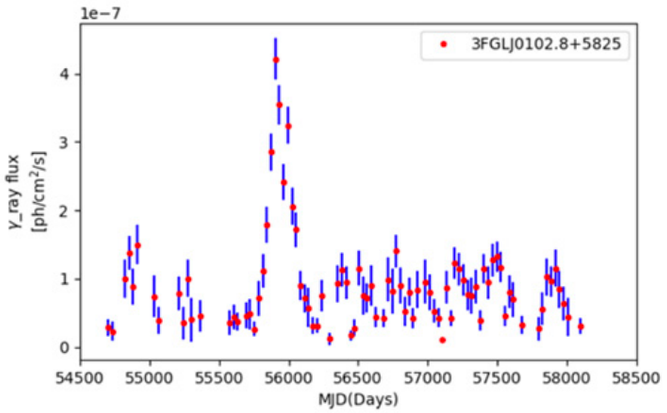
- For N independent trials, P_{post} can be written as

$$\mathbf{P_{post} = 1 - (1 - P_{pre})^N} \qquad \textbf{Šidák Correction}$$

- The Šidák correction can be approximated in the regime where both P_{pre} and P_{post} are small

$$\mathbf{P_{post} \sim N \times P_{pre}} \qquad \textbf{Bonferroni Correction}$$

Real World Example #1



Example:

You're looking for gamma-ray variability at five locations. What p-value(s) should you report?

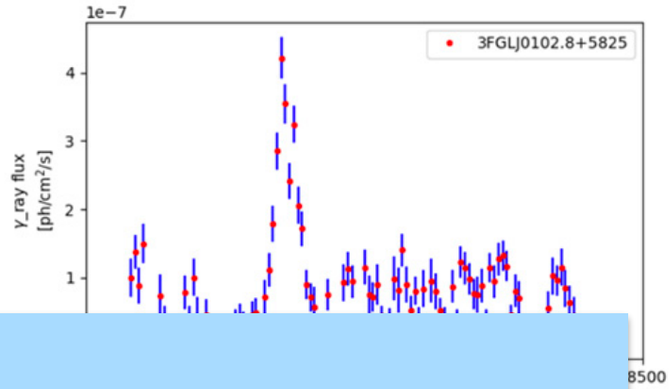
Source Name	P
Crab	0.2
Mrk 421	0.8
Mrk 501	0.5
J1908+06	0.003
Geminga	0.01

J1908+06 has the best pre-trial probability of 0.003 (2.7σ).

This corresponds to a post-trial prob of $1 - (1 - 0.003)^5 = 0.015$ (2.2σ) given the 5 trials taken to look at 5 independent locations.

What about the post-trial probability for other sources in the list? **NO!**

Real World Example #1



Example:

key concept:

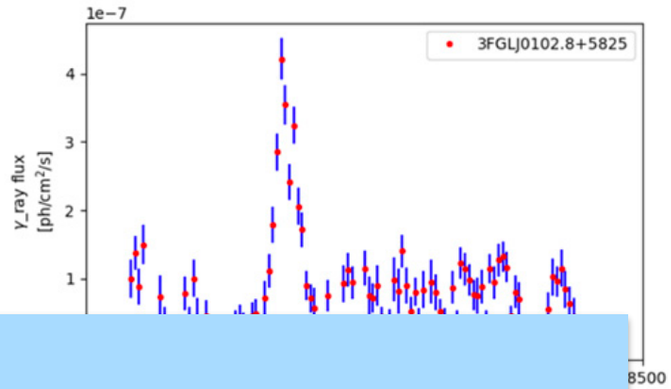
Trials are discussed in the context of a **first discovery** where you have not seen any source/phenomena yet.

Expectation is you will see, at most, 1 significant detection which will be the most significant answer out of all trials.

Entire framework is dedicated to computing probability of the most significant result in the list - not other values.

What about the post trial probability for other sources in the list? **NO!**

Real World Example #1



Example:

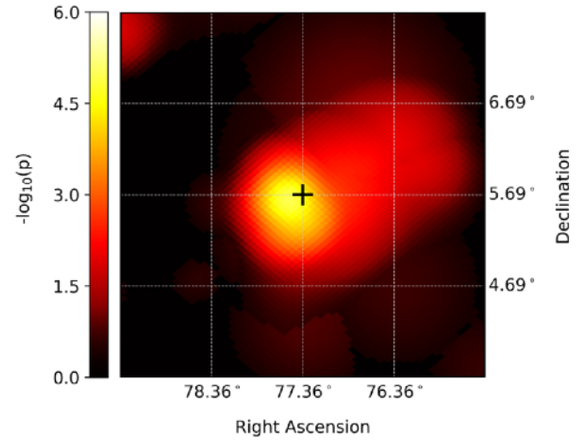
corollary:

If you have a list where you expect multiple detections it's usually better to quote the number of expected false positives (Type I errors).

For example, "We found 5 detections with an expectation of 0.5 being due to background."

What about the post trial probability for other sources in the list? **NO!**

Real World Example #2



Example:

You're looking for neutrino point sources with two lists. What p-value(s) should you report?

Source Name	P
Crab	0.2
Mrk 421	0.8
Mrk 501	0.5
J1908+06	0.003
Geminga	0.01

list 1

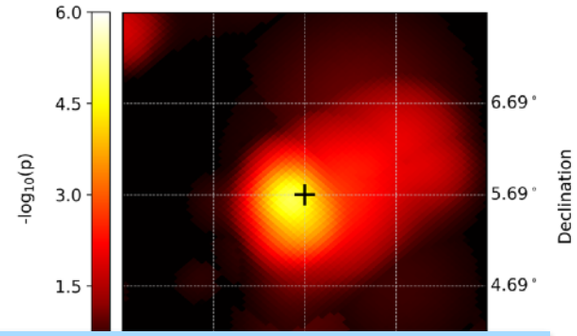
Source Name	P
PKS 1234+00	0.002
PKS 1235+01	0.3
PKS 1345+89	0.1

list 2

List 1 has a post-trial prob of $1 - (1 - 0.003)^5 = 0.015$ for J1908+06
List 2 has a post-trial prob of $1 - (1 - 0.002)^3 = 0.006$ for PKS 1234+00

Global post-trial prob from both lists is $1 - (1 - 0.006)^2 = 0.012$

Real World Example #2



Example:

key concept:

When there are multiple levels of selection, trials are scoped such that trials from unselected branches do not contribute to the final answer

5 trials

3 trials

x2 trials = 6 trials

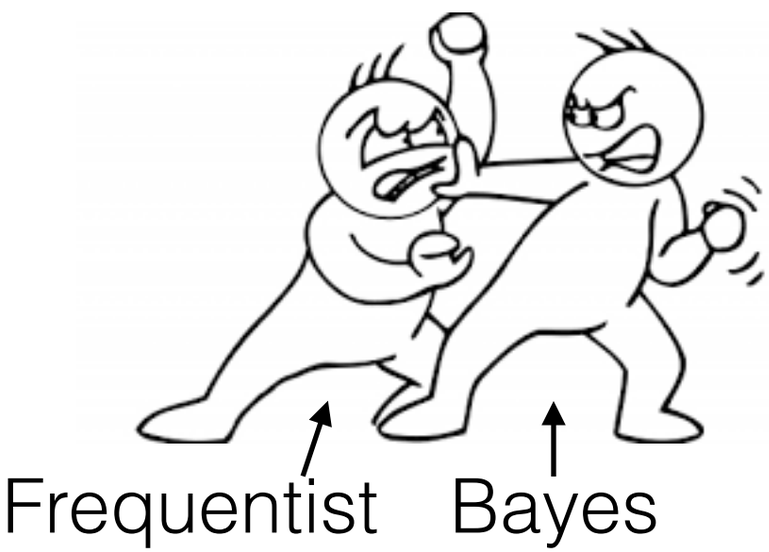
selection 1 **selection 2**

Global post-trial prob from both lists is 1 - (1 - 0.006)^2 = 0.012

Brain Rest!

Frequentist vs Bayesian

- So far the real-world examples have been straightforward lists of independent trials determined prior to looking at data



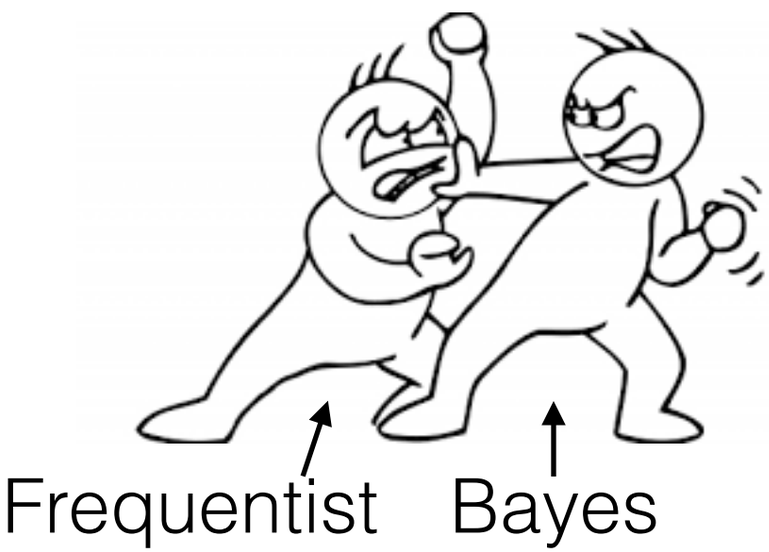
- **Less clear example:** LIGO issues alerts for 10 different GW detections. You plan to analyze all 10 alerts but examine data for the first alert and find a 3σ detection. How many trials are there?

Frequentist: I looked once, there is one trial. I haven't looked at the other 9 yet.

Bayesian: Your ideas are bad and you should feel bad.

Frequentist vs Bayesian

- So far the real-world examples have been straightforward lists of independent trials determined prior to looking at data



- **Less clear example:** LIGO issues alerts for 10 different GW detections. You plan to analyze all 10 alerts but examine data for the first alert and find a 3σ detection. How many trials are there?

Frequentist: I looked once, there is one trial. I haven't looked at the other 9 yet.

Bayesian: ~~Your ideas are bad and you should feel bad.~~ There are 10 trials.

Frequentist vs Bayesian

- So far the real-world examples have been straightforward lists

key concept:

There are cases when it can be difficult to determine the number of trials involved.

This is why the concept of **blindness** is often used. The idea is to **first think** about all the trials you plan to perform and **then analyze** the data.

Blindness also reduces the number of trials taken for the final result as it's less likely to "tune-as-you-go"

Frequentist vs Bayesian

- So far the real-world examples have been straightforward lists

key concept:

The **5σ threshold is engineered to be robust** against $O(10)$ errors in the reported number of trials

10 x 2.87e-7 —> 4.5σ

pro tip:

If you're in a heated debate over a difference of <10 trials: Stop, relax, go do something fun instead.

Brain Rest!

What about correlations?

- Šidák correction is exact for the case of N independent trials. **Does not work for correlations.**

- However, we can still apply the form of the Šidák equation with one notable change:

$$\mathbf{P_{post} = 1 - (1 - P_{pre})^{Neff}}$$

- Neff represents an **effective number of trials** which must be bounded by (1 < Neff < N) where N is the total number of correlated trials

What about correlations?

- Šidák correction is exact for the case of N independent trials. **Does not work for correlations.**

- However, we can still apply the form of the Šidák equation with one notable change:

$$P_{post} = 1 - (1 - P_{pre})^{Neff}$$

- Neff represents an **effective number of trials** which must be bounded by $1 \leq Neff \leq N$, where N is the total number of correlated trials

Physicists at Work

Example

Consider a poisson counting experiment where you measure the number of events recorded between:

(1) 0 - 10 seconds
(2) 5 - 15 seconds (50% overlap to window 1)

Goal is to look for the largest number of events to find signal on top of background

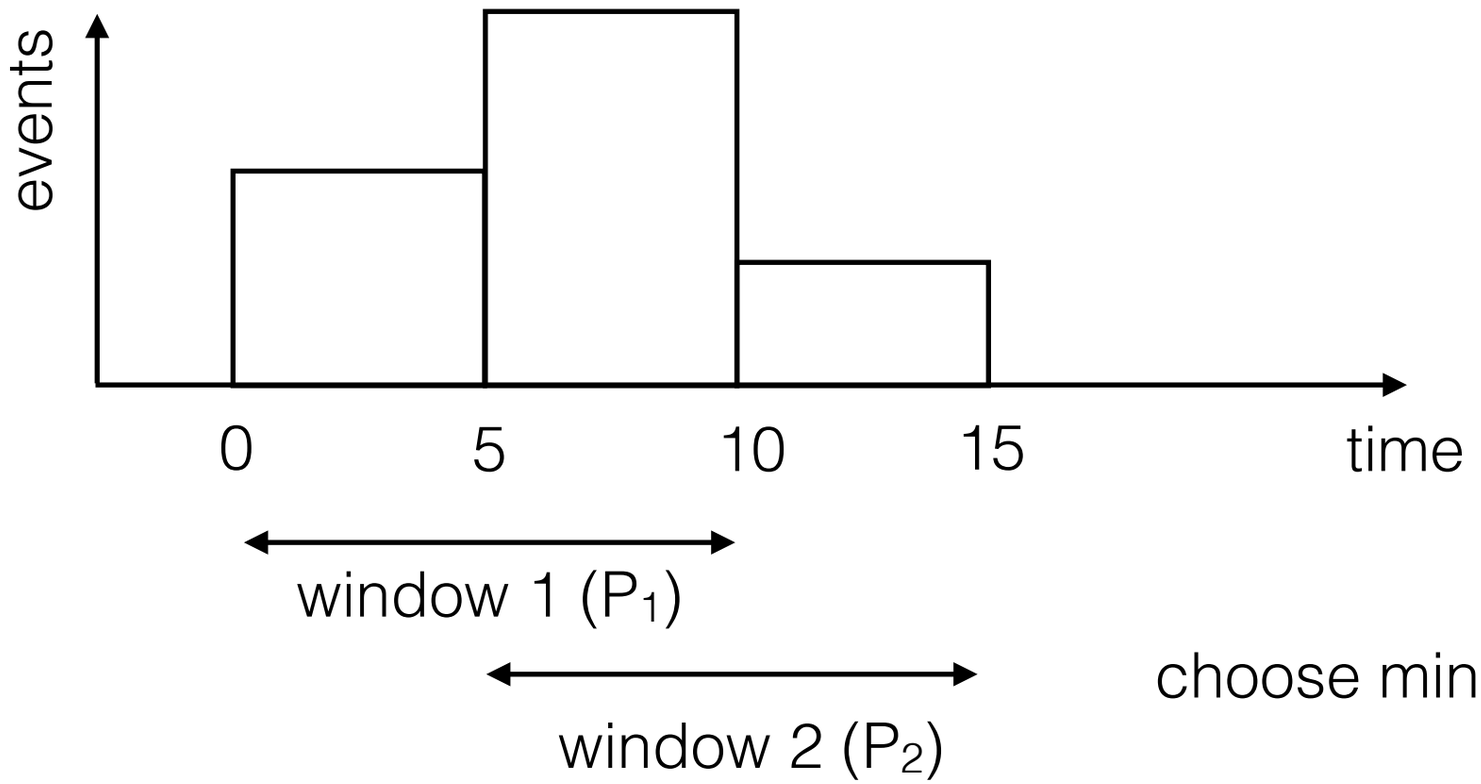
The background expectation is 1 event/sec

Select the window with the largest number of events as the answer. What's the number of trials?

Example

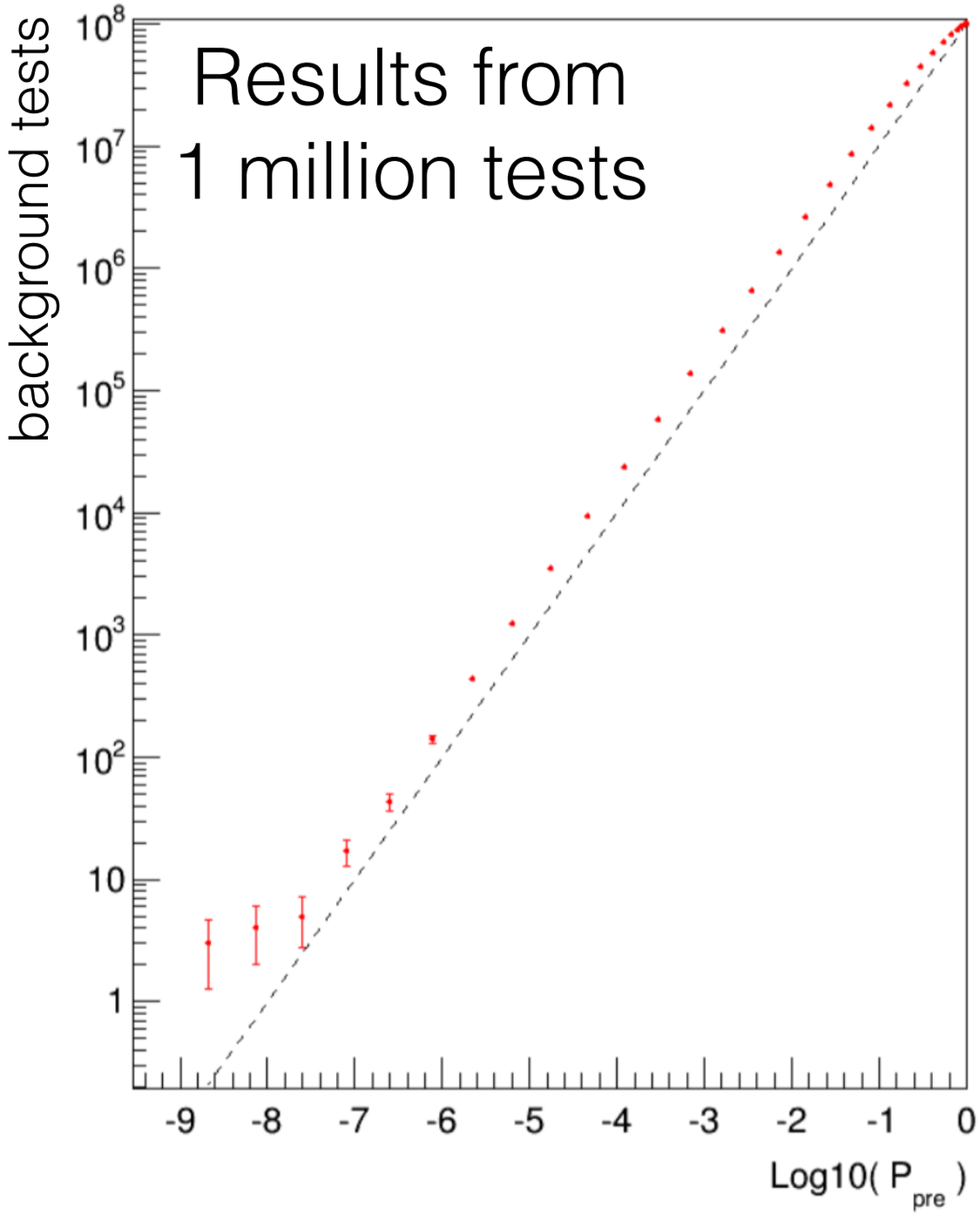
Given this simple setup, we can run a simulation to determine P_{post} and P_{pre}.

Use random number generator to randomly distribute background events in three bins, each with a mean of 5



choose min(P_1, P_2) as answer

Example

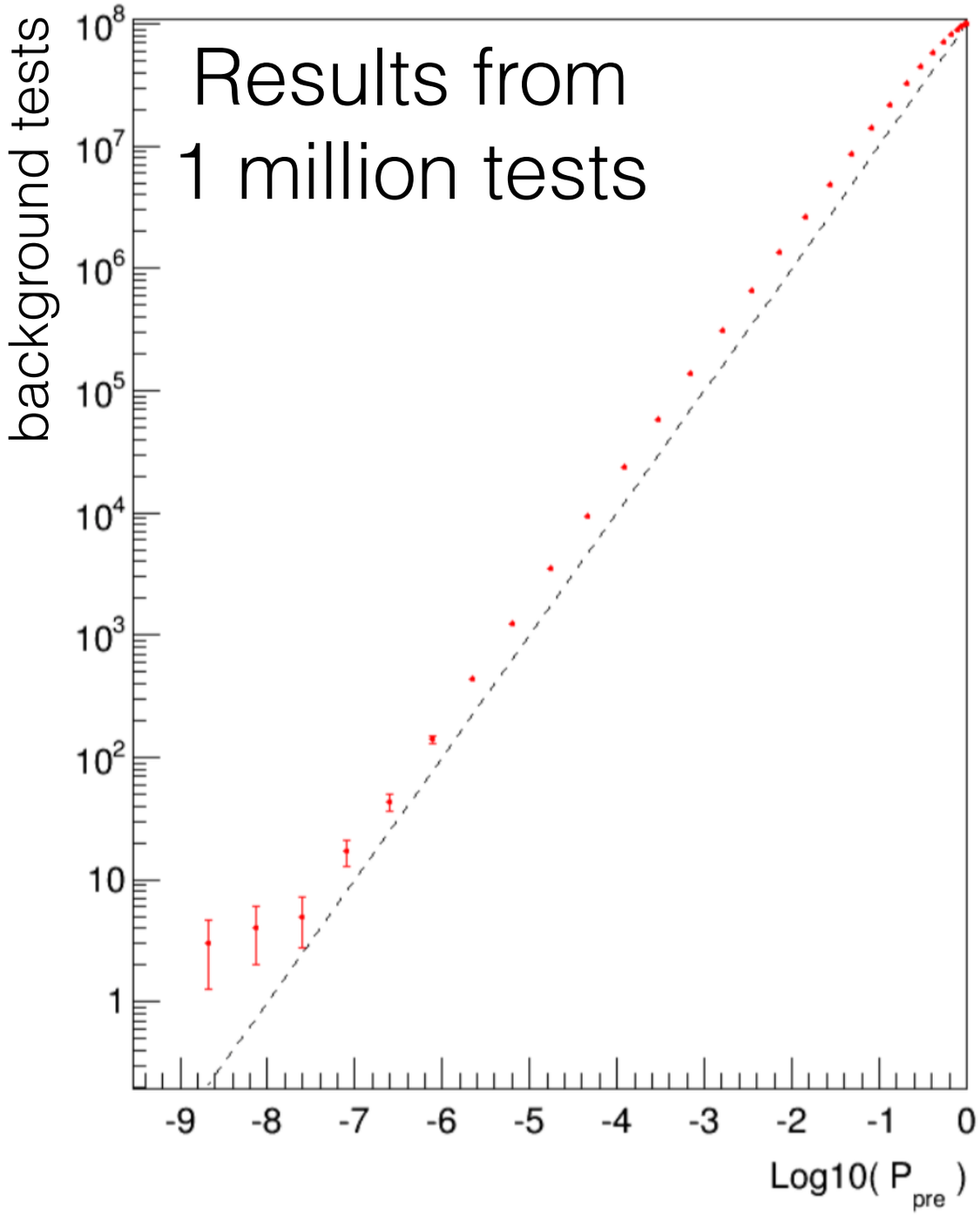


μ = 10 sec x 1 event/sec

$$P_{pre}(\geq N) = \sum_{i=N}^{\infty} \frac{10^{-i}\exp(-10)}{i!}$$

Cumulative Poisson

Example



μ = 10 sec x 1 event/sec

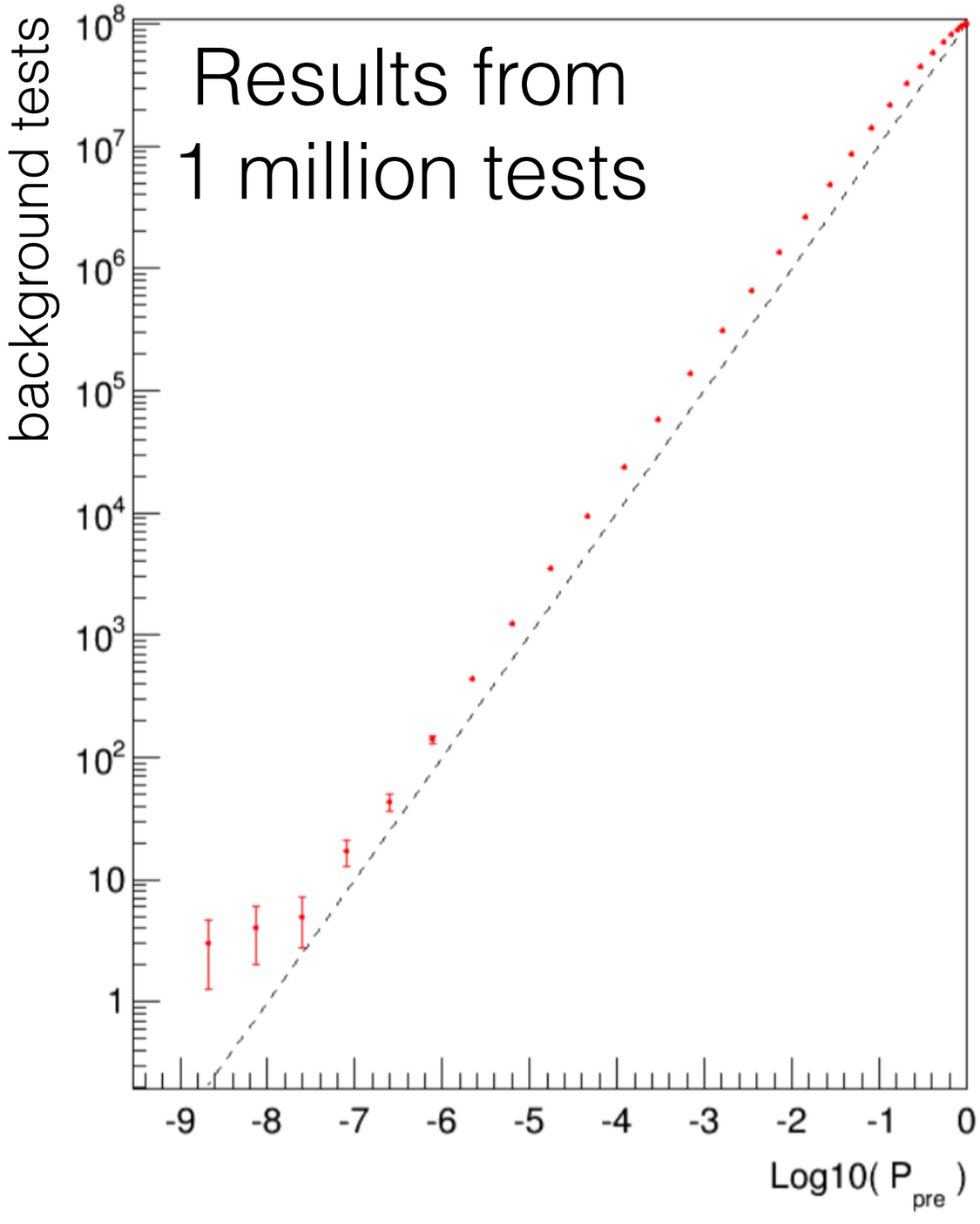
$$P_{pre}(\geq N) = \sum_{i=N}^{\infty} \frac{10^{-i}\exp(-10)}{i!}$$

Cumulative Poisson

$$P_{post}(P_{pre}) = \frac{\#\text{ tests} \geq P_{pre}}{\text{total tests}}$$

Brute Force

Example



μ = 10 sec x 1 event/sec

$$P_{pre}(\geq N) = \sum_{i=N}^{\infty} \frac{10^{-i}\exp(-10)}{i!}$$

Cumulative Poisson

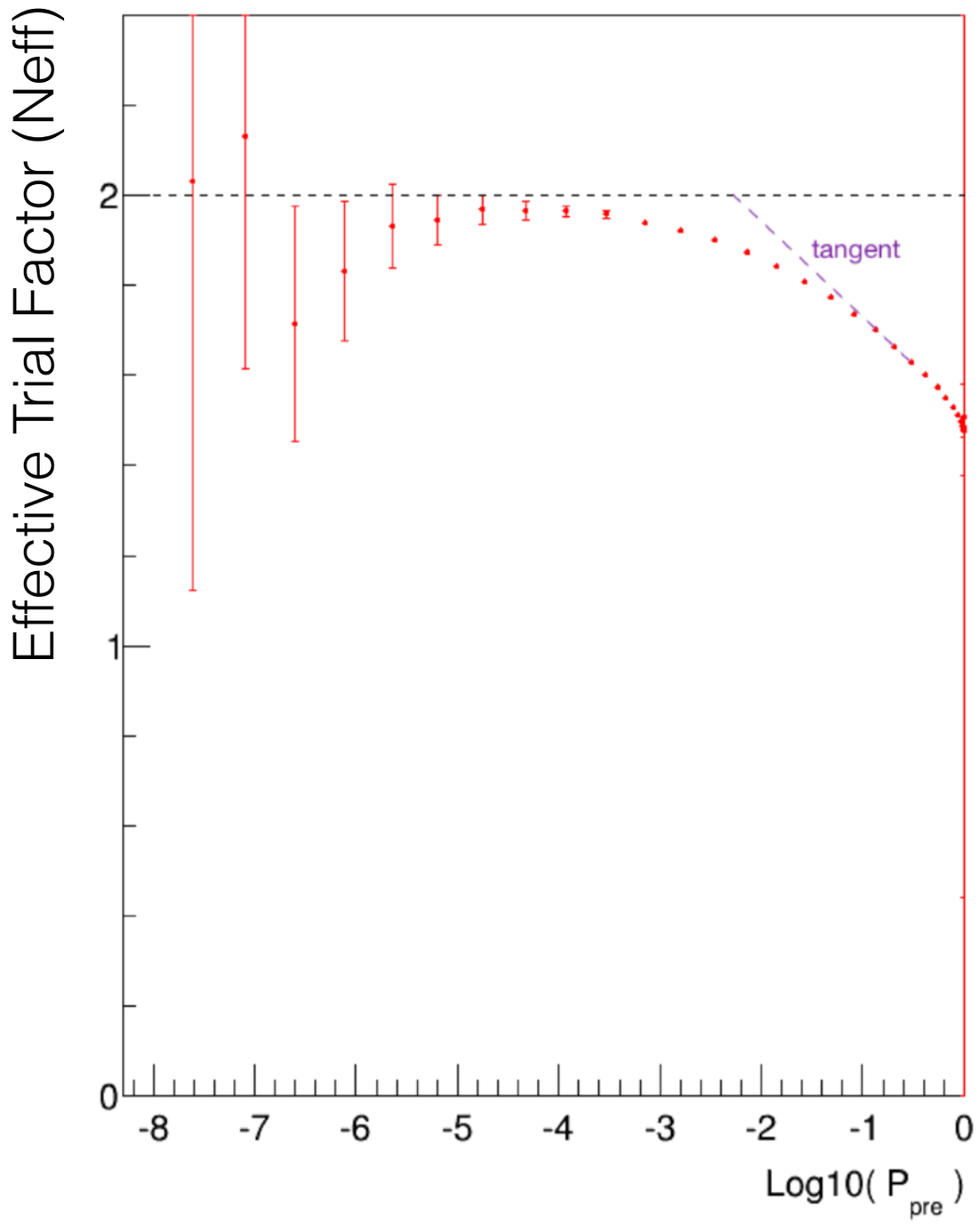
$$P_{post}(P_{pre}) = \frac{\#\text{ tests} \geq P_{pre}}{\text{total tests}}$$

Brute Force

$$Neff = \frac{\log(1 - P_{post})}{\log(1 - P_{pre})}$$

Because I can….

Example



μ = 10 sec x 1 event/sec

$$P_{pre}(\geq N) = \sum_{i=N}^{\infty} \frac{10^{-i}\exp(-10)}{i!}$$

Cumulative Poisson

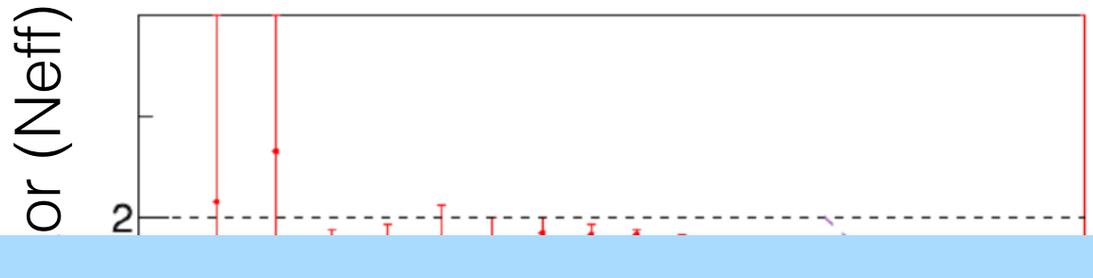
$$P_{post}(P_{pre}) = \frac{\#\text{ tests} \geq P_{pre}}{\text{total tests}}$$

Brute Force

$$Neff = \frac{\log(1 - P_{post})}{\log(1 - P_{pre})}$$

Because I can….

Example



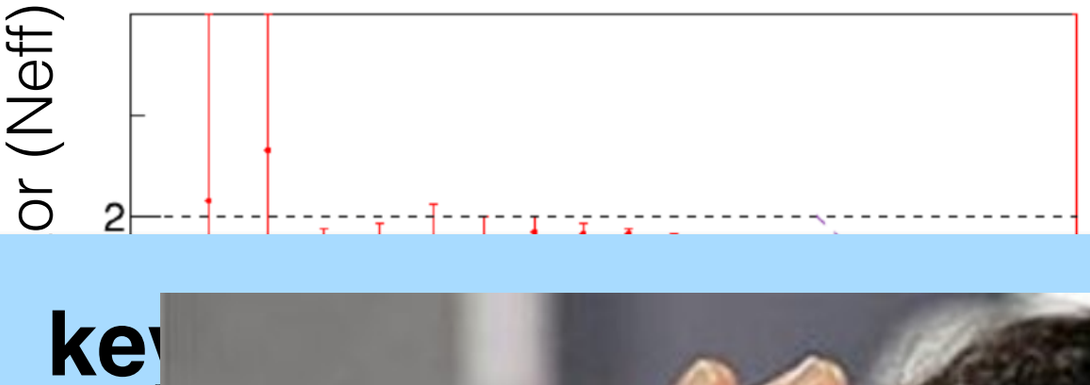
$$P_{pre}(\geq N) = \sum_{i=N}^{\infty} \frac{10^{-i}\exp(-10)}{i!}$$

key concept:

Rare background fluctuations have a larger number of effective trials in the case of correlated trials

—> Neff is a function of Ppre
—> 1 < Neff < N where N is total # of correlated trials

Example

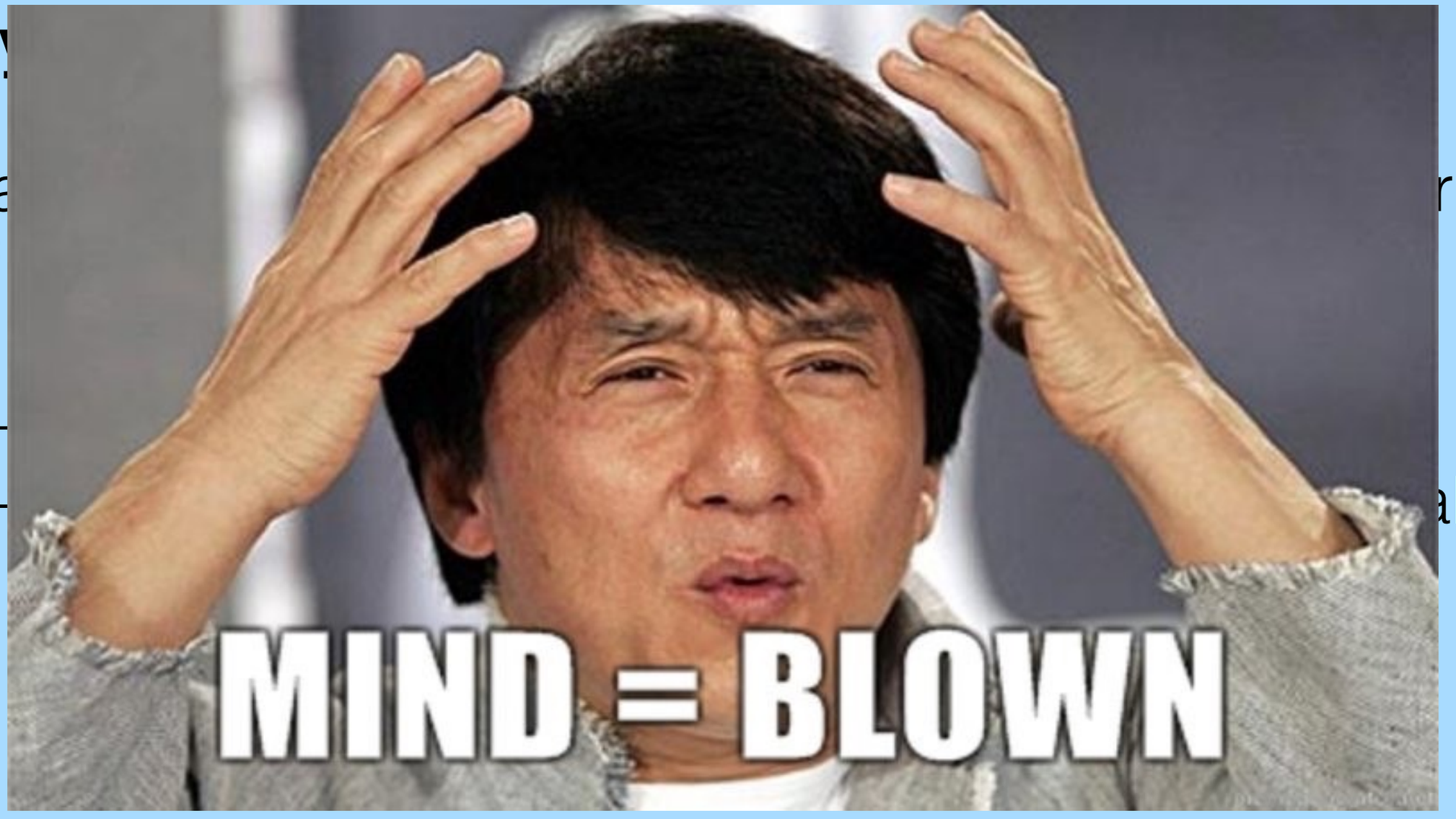


$$P_{pre}(\geq N) = \sum_{i=N}^{\infty} \frac{10^{-i}\exp(-10)}{i!}$$

key

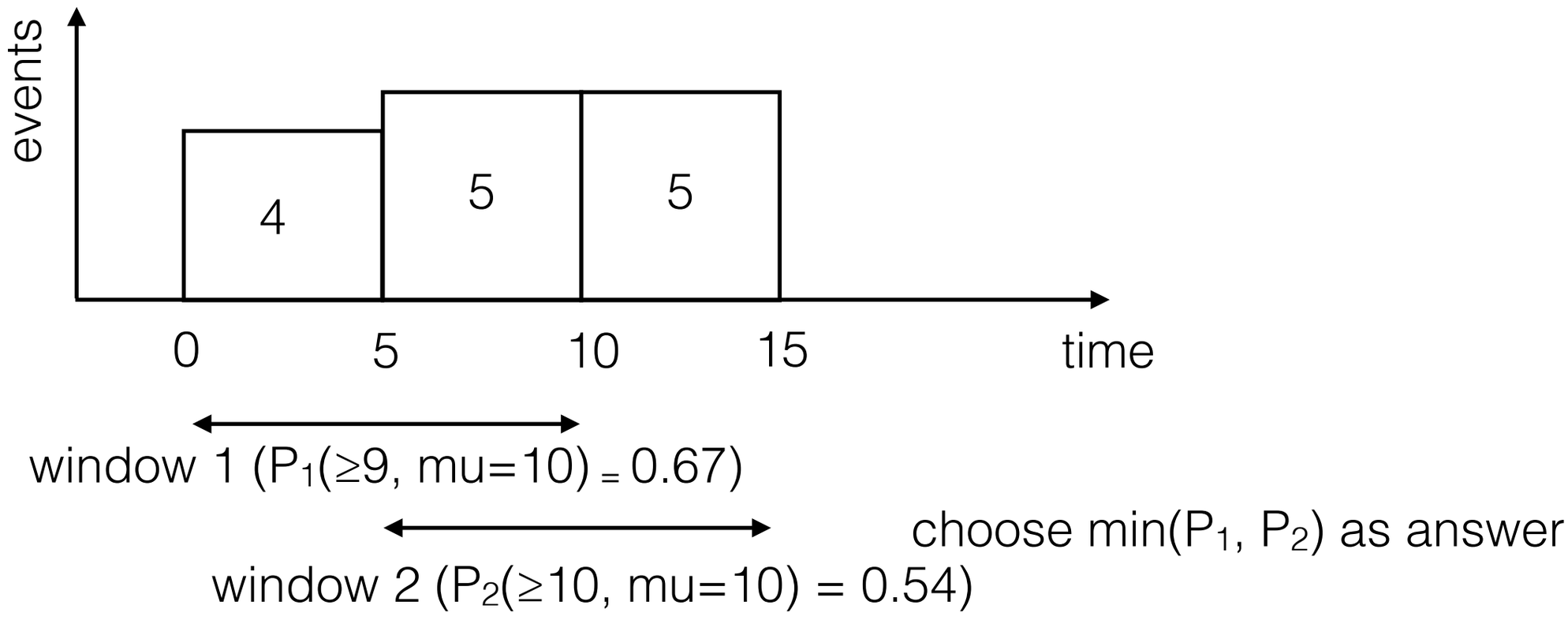
Ra r
of

als



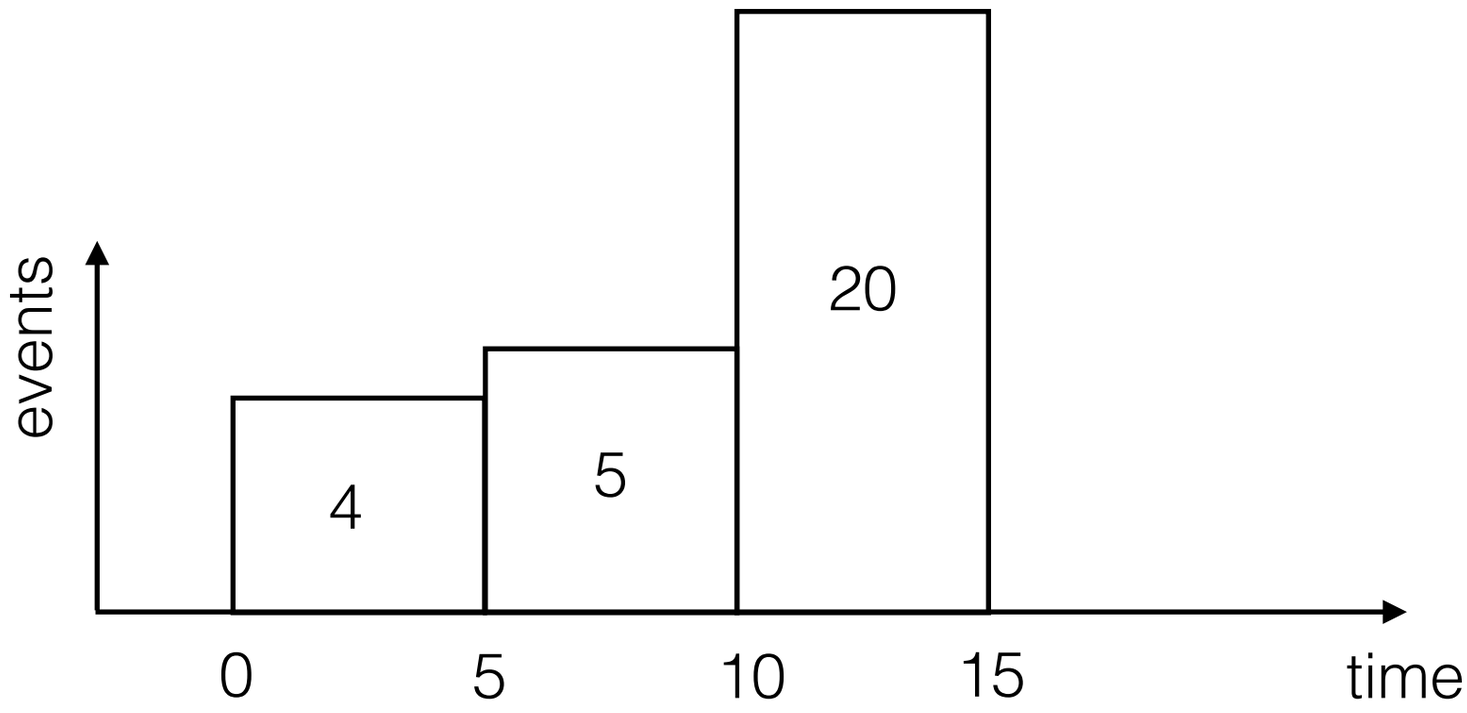
What's going on?

- Correlations mean your trials **share information**.
- If you observe a common background fluctuation in one trial (P ~ 1), you pretty much know a second trial sharing most of the events/info will have P ~ 1



What's going on?

- If you observe a rare background fluctuation in one trial, the p-value of a second trial with overlapping events can be quite different



window 1 ($P_1(\geq 9, mu=10) = 0.67$)

choose min(P_1, P_2) as answer

window 2 ($P_2(\geq 25, mu=10) = 4.7 \times 10^{-5}$)

Summary

- An **experimental trial** is when you look at the result of an analysis on data

- A simple rule of thumb in the case of independent trials is $\mathbf{P_{post} \sim N \times P_{pre}}$ where N is number of trials

- $\mathbf{P_{post} \sim N \times P_{pre}}$ **is an upper limit for correlated trials**

- **5σ criterion** is engineered to be robust against under-reported trials

- **Blindness** procedure is good practice for defining trials beforehand, minimizing total trials taken