Nonparametric Statistics

C. Alex Simpkins Jr., Ph.D
UC San Diego, RDPRobotics LLC

Department of Cognitive Science
rdprobotics@gmail.com
csimpkinsjr@ucsd.edu

Lectures :http://casimpkinsjr.radiantdolphinpress.com/pages/cogs108_ss1_23/index.html

Plan for today

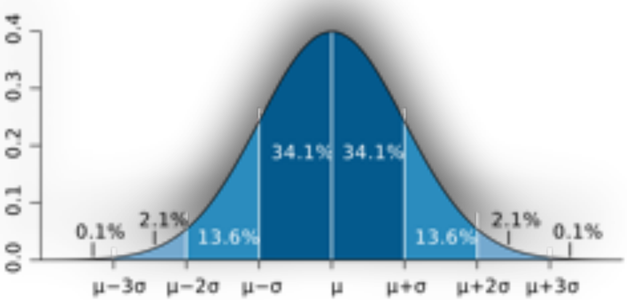
- Announcements
- Project CP2: EDA discussion
- Non-parametric statistics
- Geospatial analysis

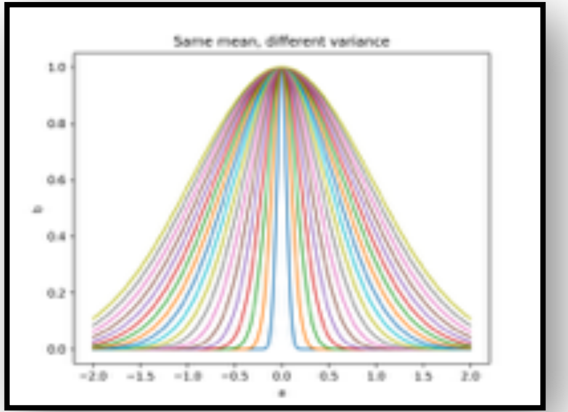
Announcements

- Checkpoint 2: EDA Friday
  - https://github.com/drsimpkins-teaching/COGS108/blob/main/project/EDACheckpoint_groupXXX.ipynb
- Working on feedback for CP1: Data, proposals are being released
- Upcoming deadlines Friday (may be extended to Monday to provide time)
  - D5, D6, A3, Q3

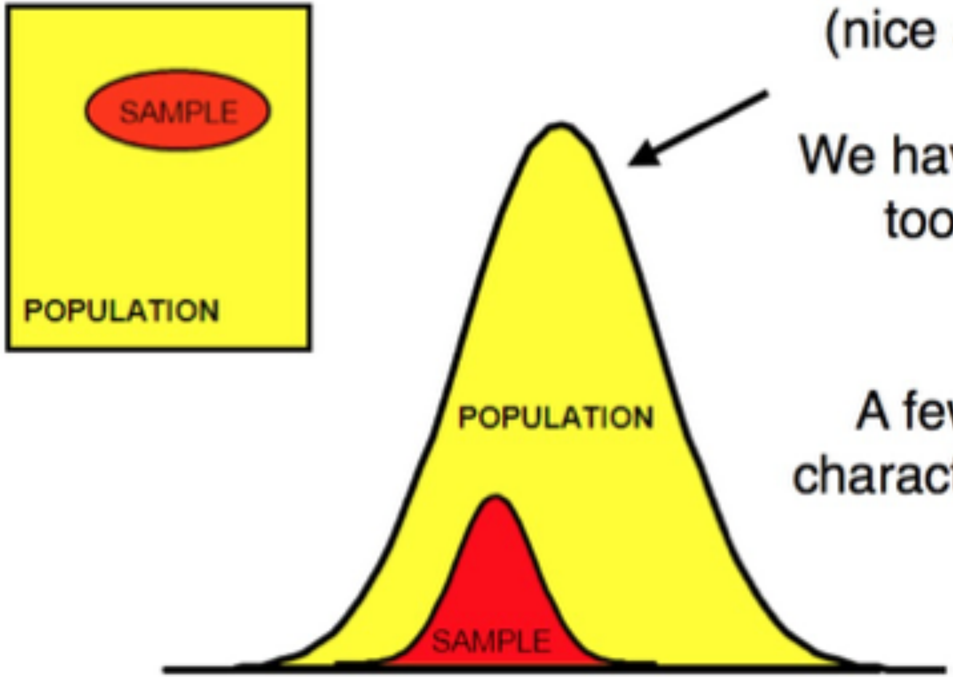
Parametric statistics

- Uses ***parameters*** such as the **mean, standard deviation, variance, correlation, covariance**, etc.
- Using observational data we can **estimate parameters to describe the distribution**
    - These are the characteristics of the system generating the data
- Data is assumed to come from a **normal distribution** and has a population mean, population standard deviation, etc.
- We can use the data with the **assumption** that our **sample distribution is representative of the population distribution**





Non-parametric Statistics: The Why

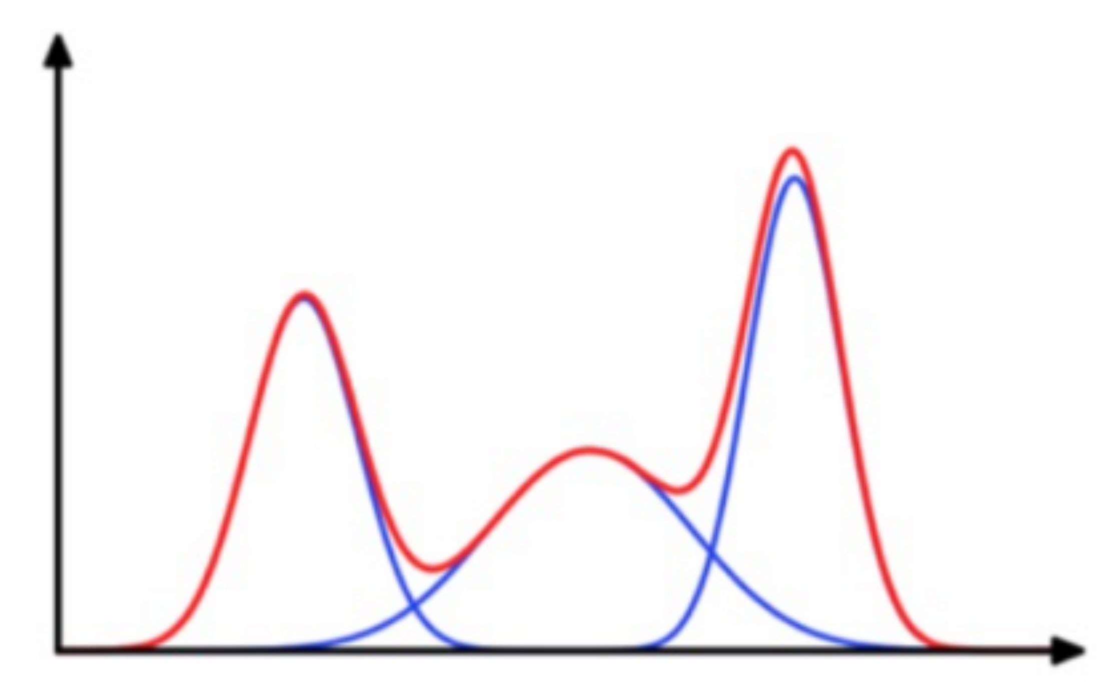


Normal distribution
(nice and friendly)

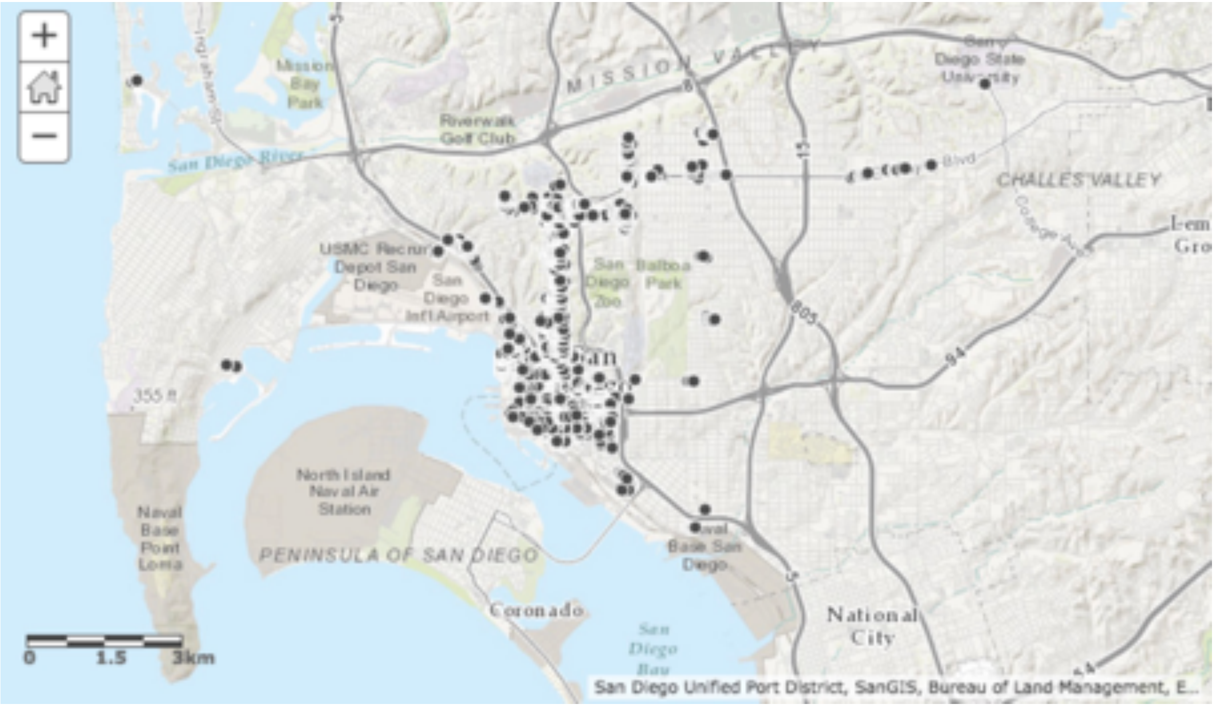
We have good math tools for this.

A few parameters **fully** characterize the distribution.

Non-parametric Statistics:
What if your distribution looks like this?



Non-parametric Statistics:
...or like this?



Parameters (like mean and variance) cannot fully and accurately capture this distribution!

Hence, we require **non-parametric statistics**.

When to turn to non-parametric statistics...

- When underlying distributions are non-normal, skewed, or cannot be parameterized simply.

- When you have ranked (ordinal) data, *e.g.*, preferences.

Like	Like Somewhat	Neutral	Dislike Somewhat	Dislike
1	2	3	4	5

- When you need to build an empirical “null” distribution.

Why is ordinal data for nonparametric stats?

- Ordinal data is classified into categories within a variable that have a natural rank order. However, the distances between the categories are uneven or unknown.

- For example, the variable “frequency of physical exercise” can be categorized into the following:
  - ***1. Never | 2. Rarely |3. Sometimes | 4. Often | 5. Always***

- There is a clear order to these categories, but we cannot say that the difference between “never” and “rarely” is exactly the same as that between “sometimes” and “often”. Therefore, this scale is ordinal.

Non-parametric Statistics: distribution-free

- ***Myth***: Non-parametric statistics does not use parameters.
- ***Fact***: Non-parametric statistics does not make ***assumptions about*****/parametrize** the underlying distribution generating the data.
- **“Distribution-Free” statistics**
    - Meaning, it does not assume data-generating process (like heights) result in, *e.g.*, normally-distributed data
    - An alternative to t-tests, ANOVA, etc that only work if data satisfies some criteria/assumptions

Ordinality

Which of the following variables contains ordinal data?

A
Favorite Pet (ie.e dog, cat, fish, horse, etc.)

B
Distance traveled by car each day (miles)

C
Survey responses (scale from Dislike to Like)

D
Human height (in inches)

E
Human hair color (i.e. black, brown, red, blonde, etc.)

Resampling in statistics

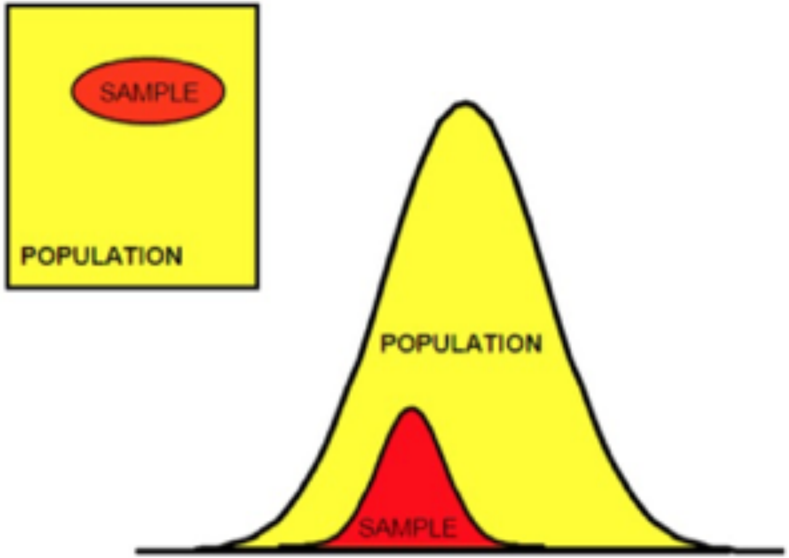
- Creating new samples based on one observed sample
- Many methods of doing this

Resampling statistics: The What

- Bootstrap (Monte Carlo)
- Rank Statistics (Mann Whitney U)
- Kolmogorov-Smirnoff Test
- Non-parametric prediction models

1) Bootstrapping (resampling)

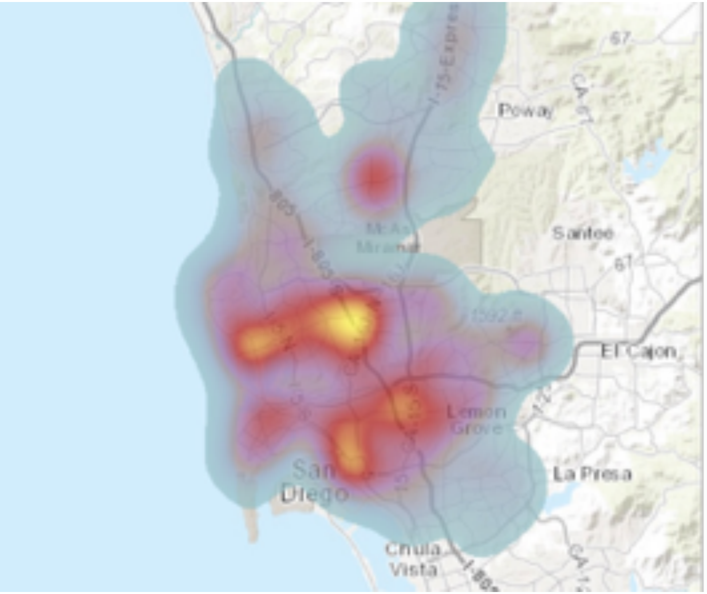
- How can we build a more realistic “null distribution” for the sample estimate without knowing the population it’s drawn from?

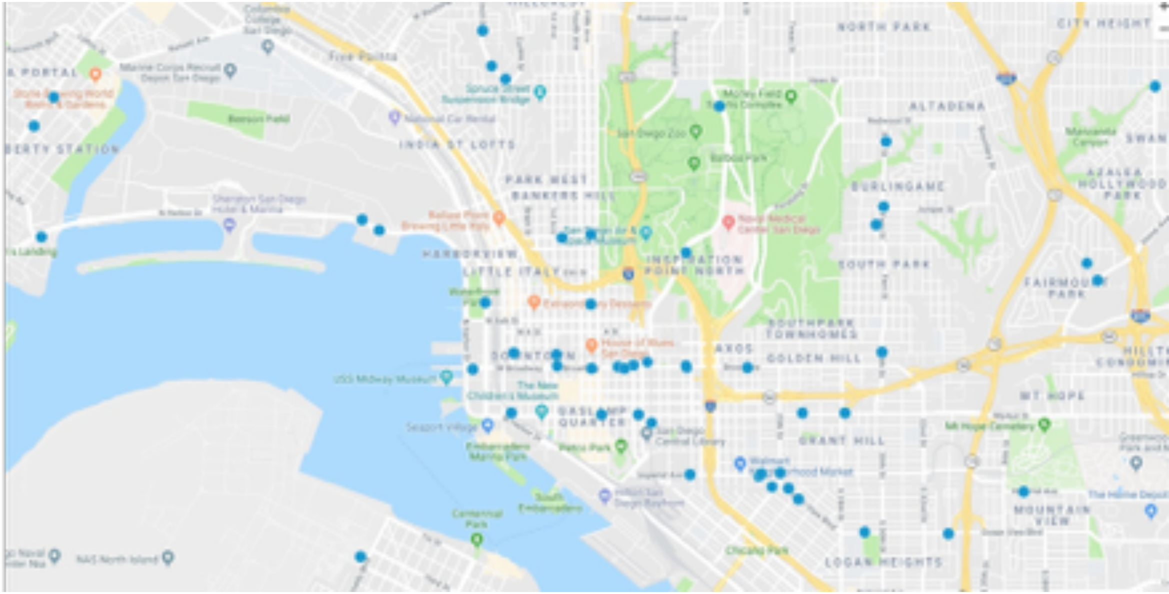


Bootstrapping (resampling)

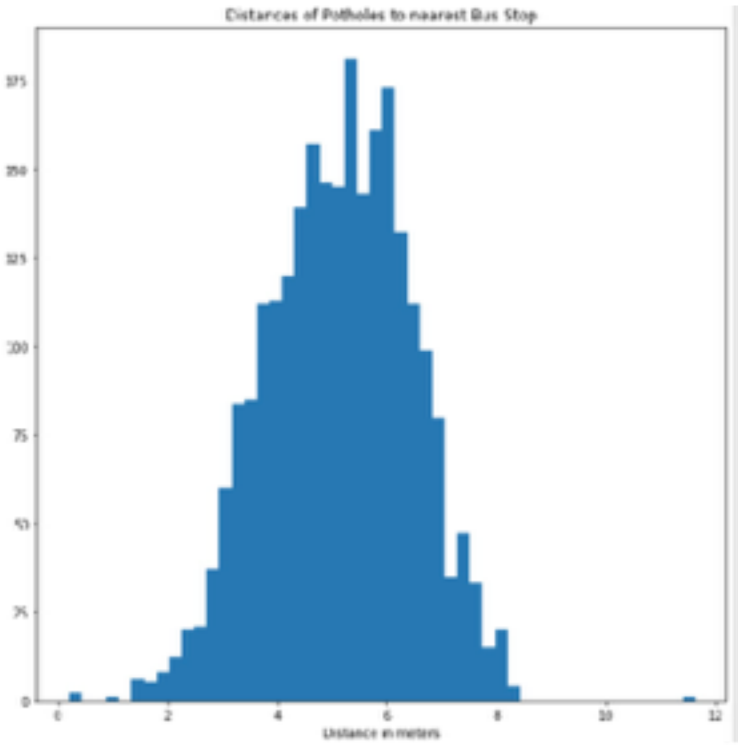
Example Question:

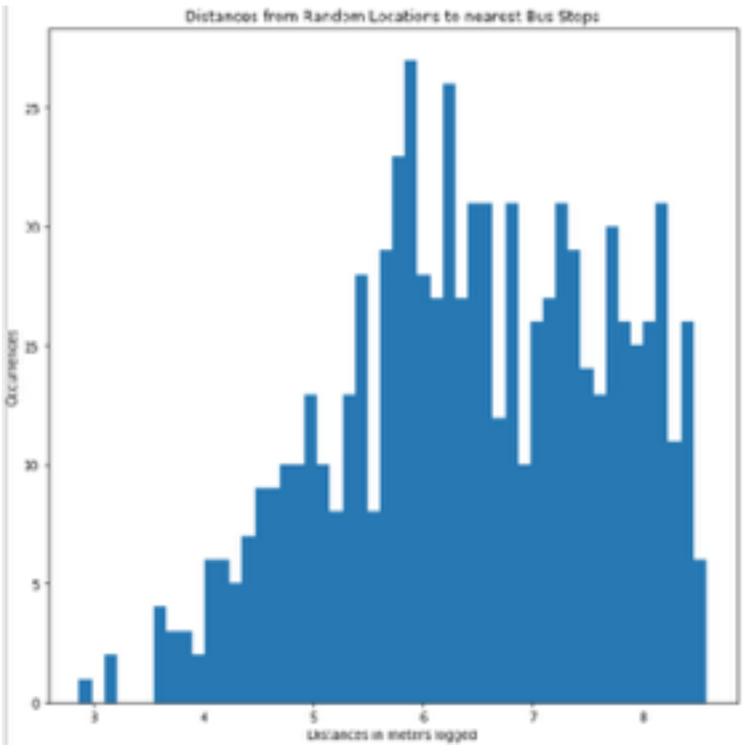
- Are San Diego’s pot holes closer to bus stops than not?





Bootstrapping (resampling)





Bootstrapping - has pluses and minuses

- Advantages - simplicity
  - can be applied to complex designs
  - can be asymptotically more accurate in confidence intervals than standard CI obtained with assumptions of normal distribution
  - Avoids costs of repeating experiment to obtain more data
- Disadvantages
  - Naive bootstrapping will not always yield asymptotically valid estimates
  - No finite sample guarantees
  - Can be time/processor intensive

2) Rank Statistics

We rank things in the real world *all the time!*

- International rankings (economics, happiness, government performance)
- Sports (teams, players, leagues)
- Search Engines
- Academic Journals’ prestige
- Reviews online (1-4 stars)

Rank Statistics

Data are transformed from their quantitative value to their rank.

quantitative data

1, 4.5, 6.6, 9.2

ordinal data

1, 2, 3, 4

Ordinal data - categorical, where the variables have a natural order

Particularly helpful when data have a ranking but no clear numerical interpretation (i.e. movie reviews)

Rank Time

What would the rank of the following list be?

[77, 49, 23, 10, 89]

A
[1, 2, 3, 4, 5]

B
[5, 4, 3, 2, 1]

C
[4, 3, 2, 1, 5]

D
[2, 1, 1, 1, 2]

E
I’m lost.

Wilcoxon rank-sum test (Mann Whitney U test)

- Determine whether two independent samples were selected from the same populations, having the same distribution
- Similar to t-test (but does not require normal distributions) & tests <u>median</u>

Assumptions:

- Observations in each group are independent of one another
- Responses are ordinal

H_o: distributions of both populations are equal

H_a: distributions are *not* equal

Python implementation

- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ranksums.html
- “The Wilcoxon rank-sum test tests the null hypothesis that two sets of measurements are drawn from the same distribution. ” [docs above]

Mann-Whitney U: question example

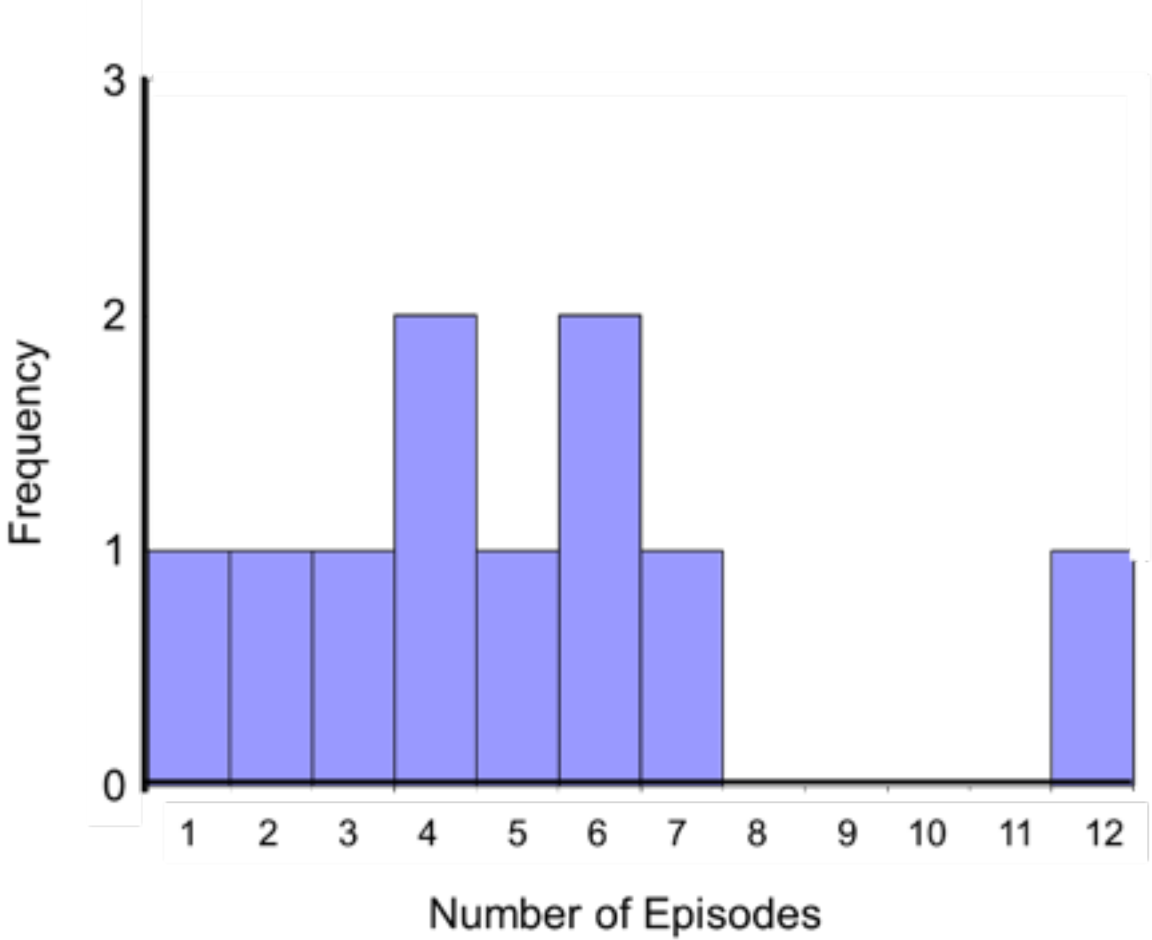
In a clinical trial, is there a difference in the number of episodes of shortness of breath between placebo and treatment?

Step 1: Participants record number of episodes they have.

Step 2: Episodes from both groups are combined, sorted, and ranked

Step 3: Resort the ranks into separate samples (placebo vs. treatment)

Step 4: Carry out statistical test



http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric4.html

		Total Sample (Ordered Smallest to Largest)	Ranks
Placebo	New Drug		
7	3		
5	6		
6	4		
4	2		
12	1		

Sum of ranks:

Placebo = 37

New Drug = 18

Mann-Whitney *U*: calculating the *U* statistic

$$U_A = n_a n_b + \frac{n_a(n_a+1)}{2} - T_A$$

The max possible value of TA

The observed sum of ranks for sample A

Ho: low and high scores are approximately evenly distributed in the two groups

Ha: low and high scores are NOT evenly distributed in the two groups (U <= 2)

n_a = number of elements in group A

n_b = number of elements in group B

$U_{Placebo} = 3$

$U_{treatment} = 22$

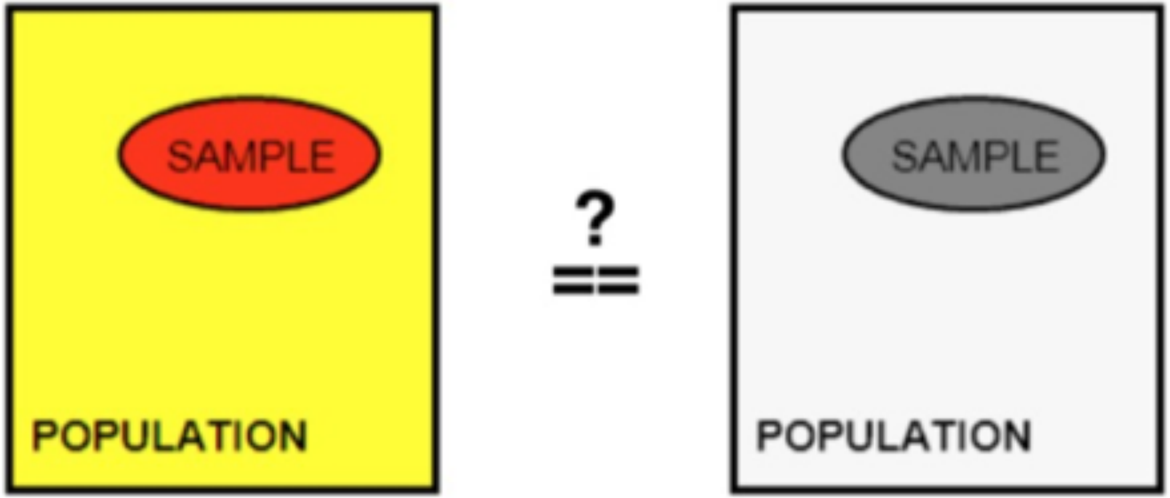
$0 < U < n_1 * n_2$

Complete separation → no separation

We reject the null if U is small.

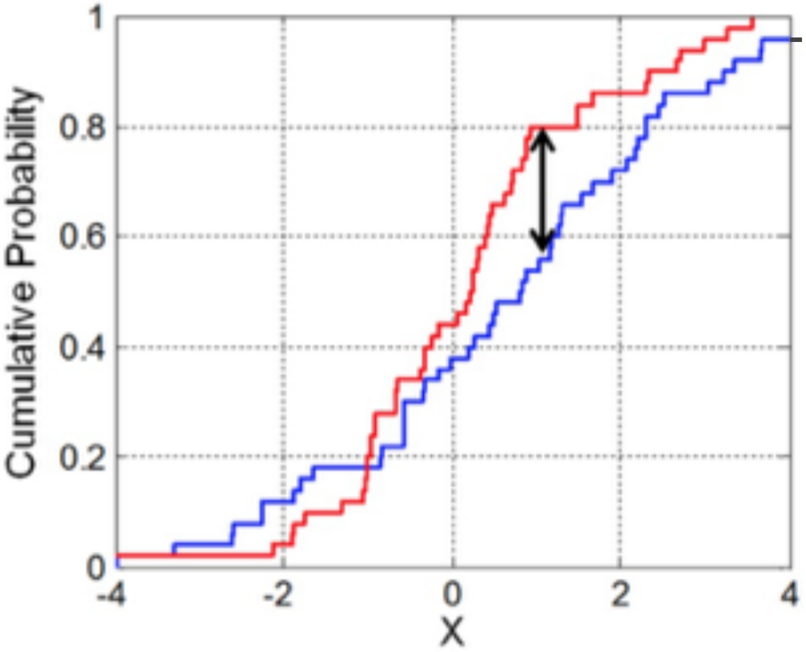
3) Kolmogorov-Smirnov (KS) test

- Given (limited) samples from two populations, how do we quantify whether they come from the same distribution?



Kolmogorov-Smirnov (KS) test

Comparing cumulative distributions empirically



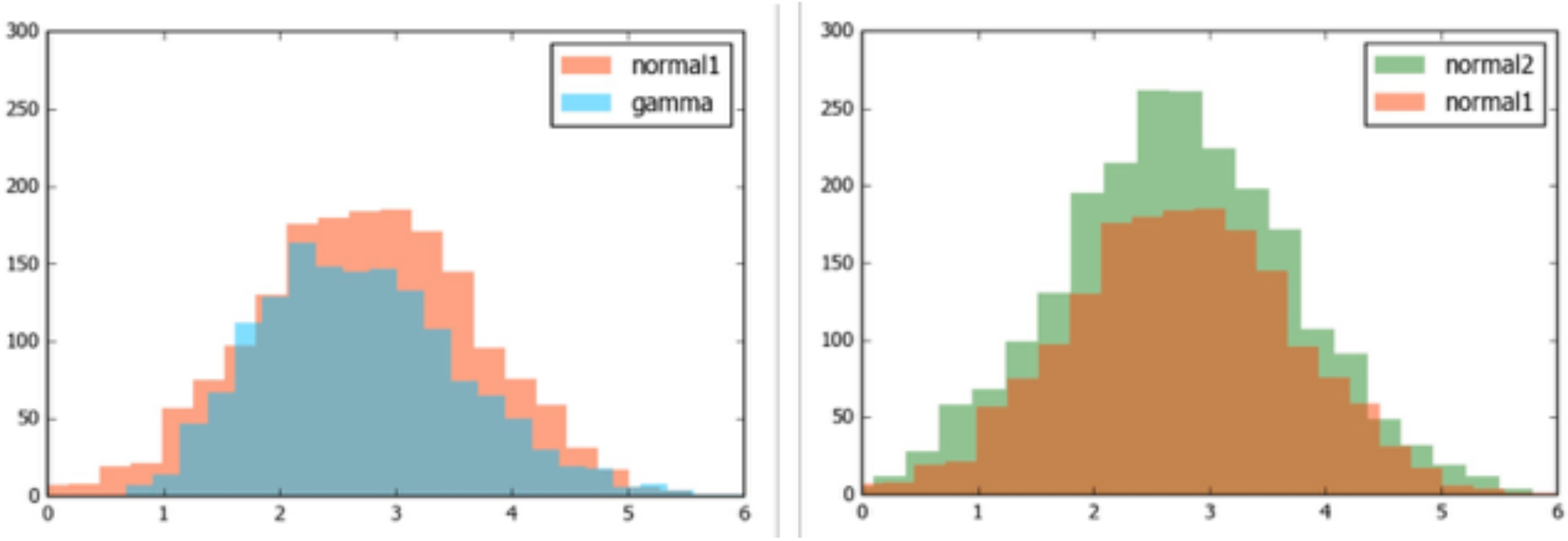
Find the maximum difference between the CDFs.

Tests:

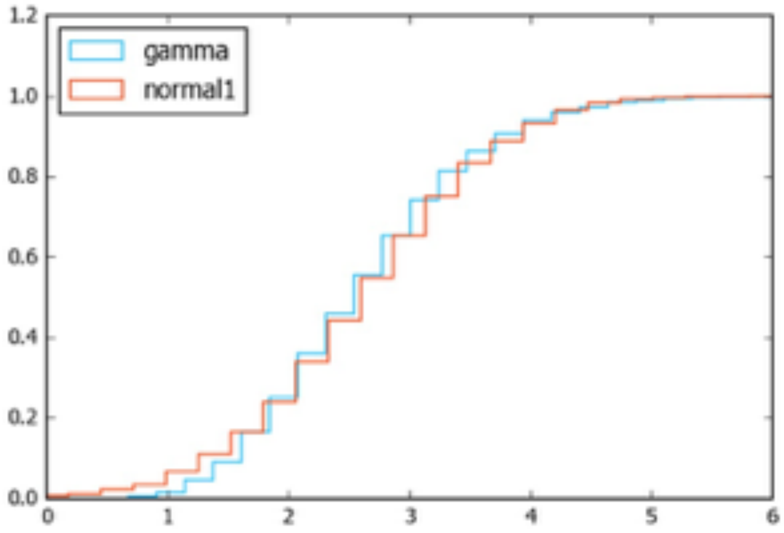
- whether a sample is drawn from a given distribution
- Whether two samples are drawn from the same distribution

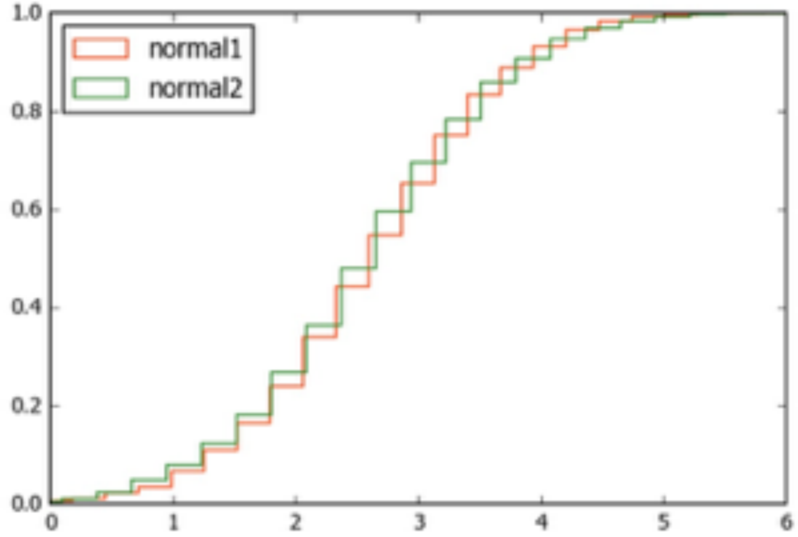
Kolmogorov-Smirnov (KS) test

- Given (limited) samples from two populations, how do we quantify whether they come from the same distribution?



Kolmogorov-Smirnov (KS) test






```
gamma vs. normal1: p = 0.0106803628411
normal1 vs. normal2: p = 0.55073598243
```

4) Non-parametric prediction models

- When you have lots of data and no prior knowledge
- When you're not focused/worried about choosing the right features
- Goal: fit training data while being able to generalize to unseen data
- Examples:
  - KNN (K-Nearest Neighbors)
  - Decision Trees (CART)
  - Support Vector Machines (SVM)

Why do we even teach/use parametric statistics anyway?

Parametric approaches:

- Lots of data follow expected patterns
- Require less data
- More sensitive
- Quicker to run/train/predict
- More resistant to overfitting

How does the brain really work? Why?

- Brain doesn’t operate according to average response
    - Differences in perception, conscious and unconscious processes, real-world embodied, embedded, situated issues (active perception), encoding, decision making
- Strong evidence for high dimensionality of encoding especially in pre-frontal cortex
    - Rigotti, M. et al (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585–590
- ***Must*** understand neural dynamics within a single trial
- Consider (Stokes and Spaak 2016)

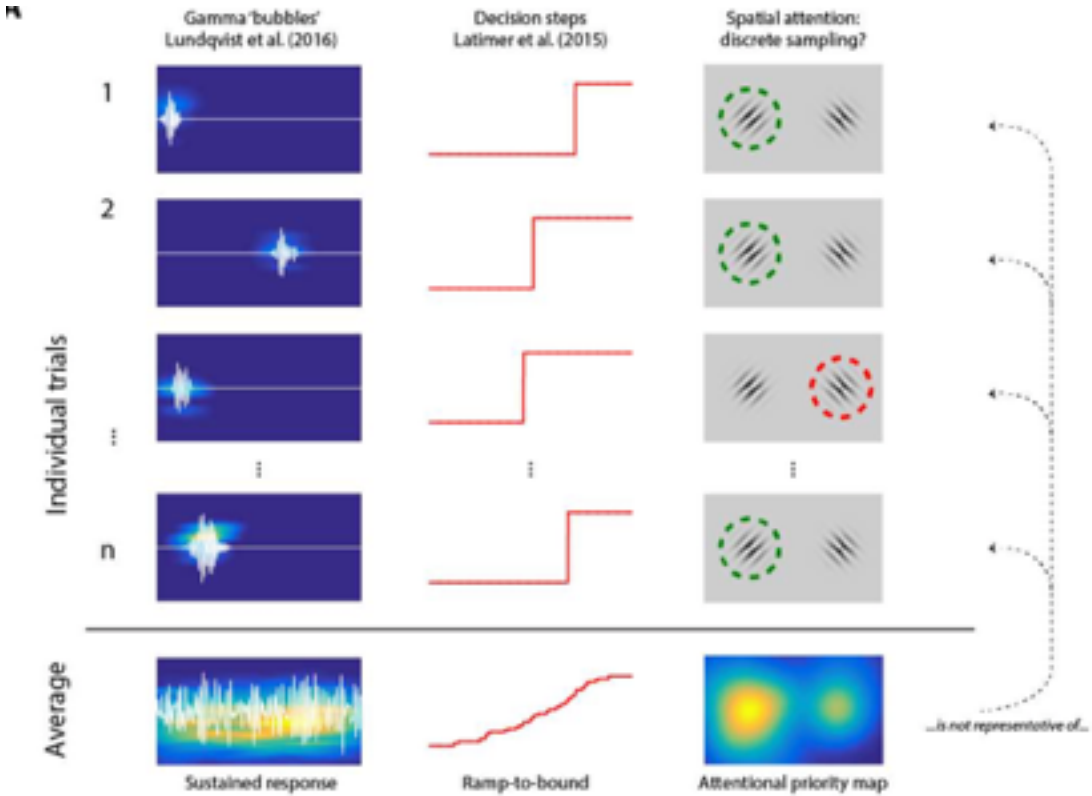
What they tested

- Lunqvist et al. developed novel method to characterize trial-wise dynamics in working memory tasks for primates
- Do we see sustained activity, as previously concluded as recorded from LFP in primate PFC at the single trial?
- Or is it different dynamically at single trial level from ‘average response?’

How did they test it?

- Developed a novel metric they refer to as ‘burstiness’ to quantify temporal gamma activity for single trials before averaging
- By using 2nd order average of this metric, found persistent activity consists of bursts of activity - not an unbroken chain of firing
- Memories stored in hidden neural states

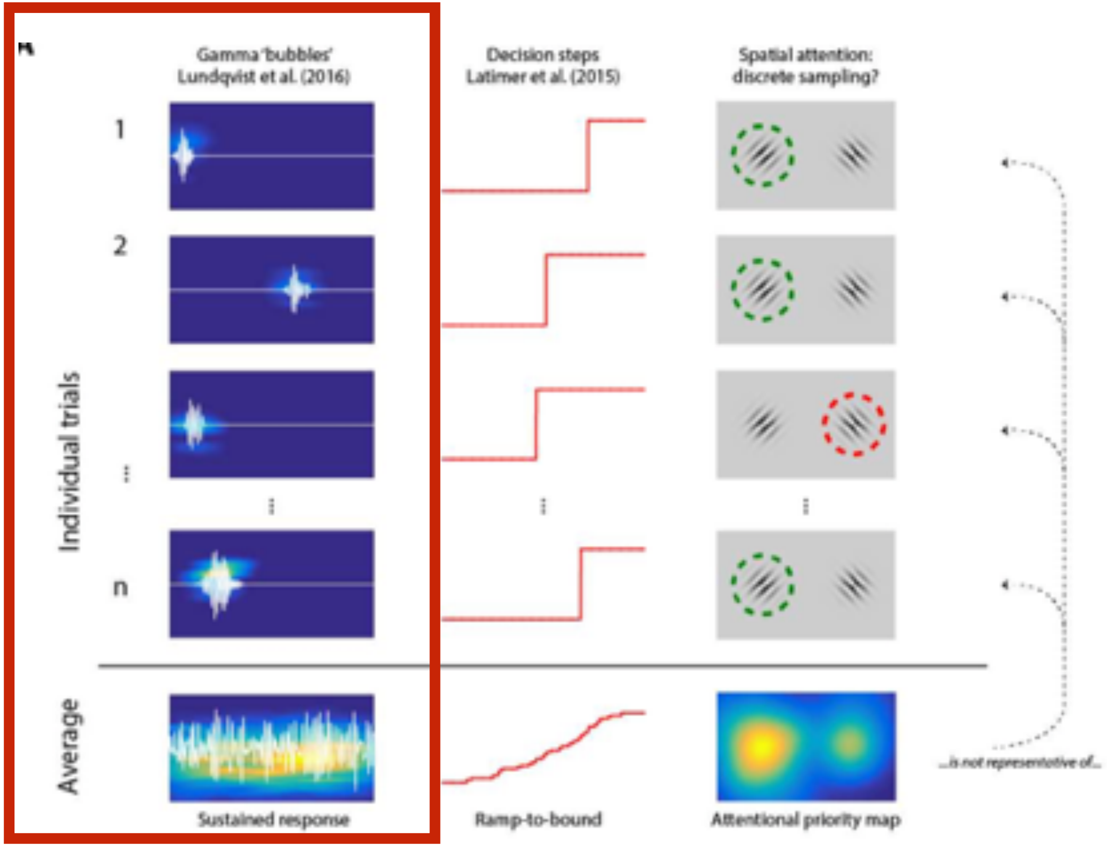
Combining by computing mean doesn't necessarily create a good representation



Combining by computing mean doesn't necessarily create a good representation

- Left: (time-frequency representations of power, with traces superimposed) in the prefrontal cortex during individual trials of working memory maintenance activity

- The average shows a familiar sustained gamma response, but qualitatively misrepresents the single trial dynamics.

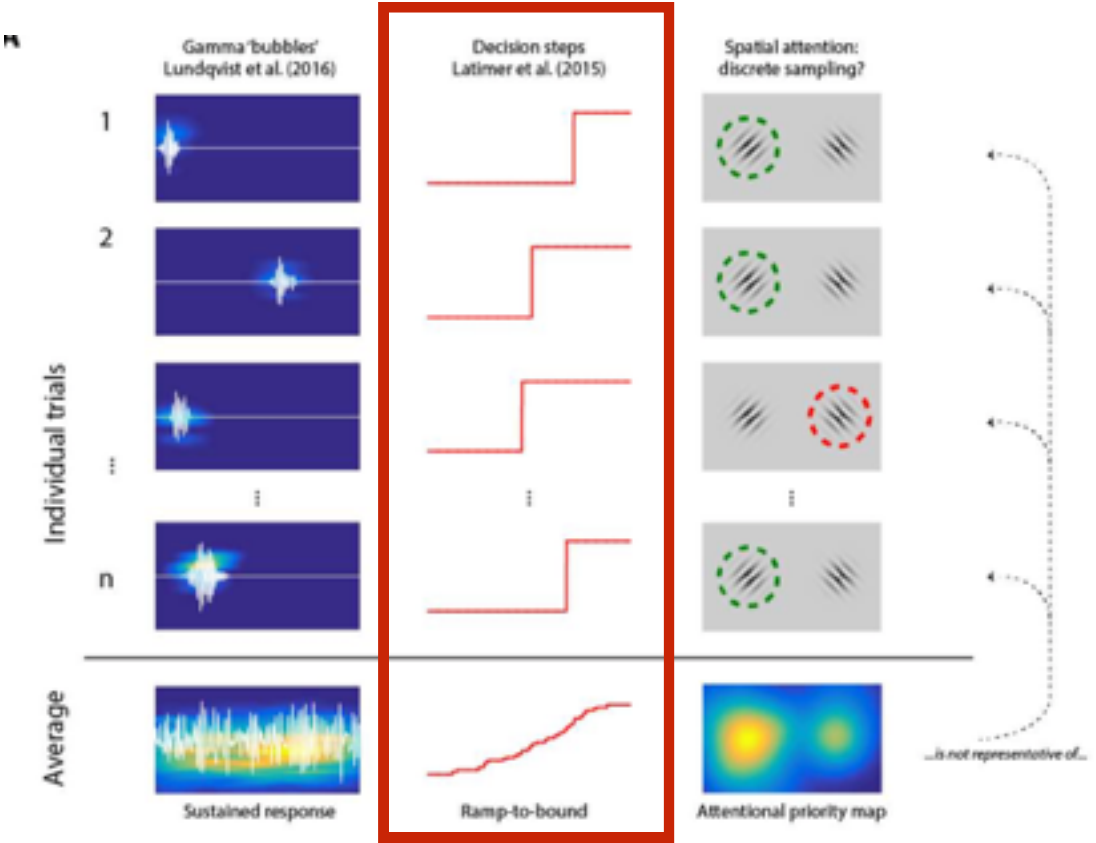


Combining by computing mean doesn't necessarily create a good representation

- Middle: neurons display discrete steps reflecting the time of sensory decisions

- The average response shows a classic ramp-to-bound process for the decision.

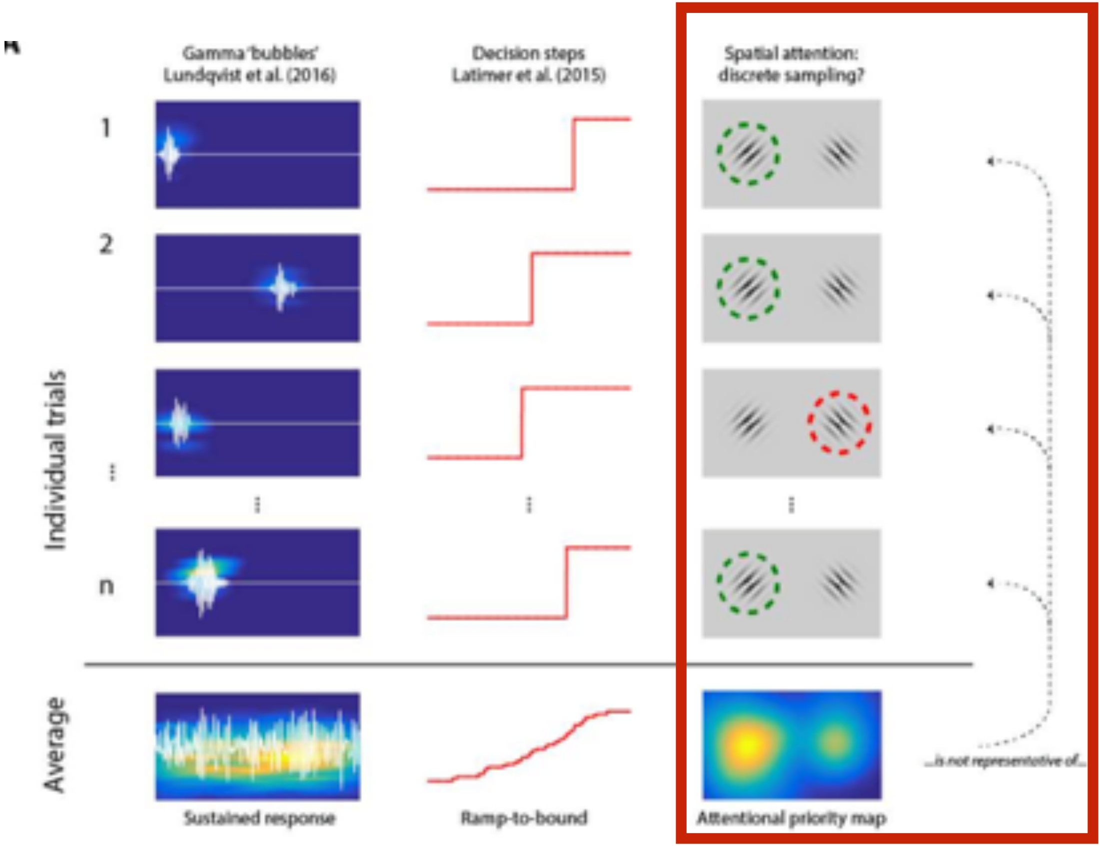
- Again not representing what's happening



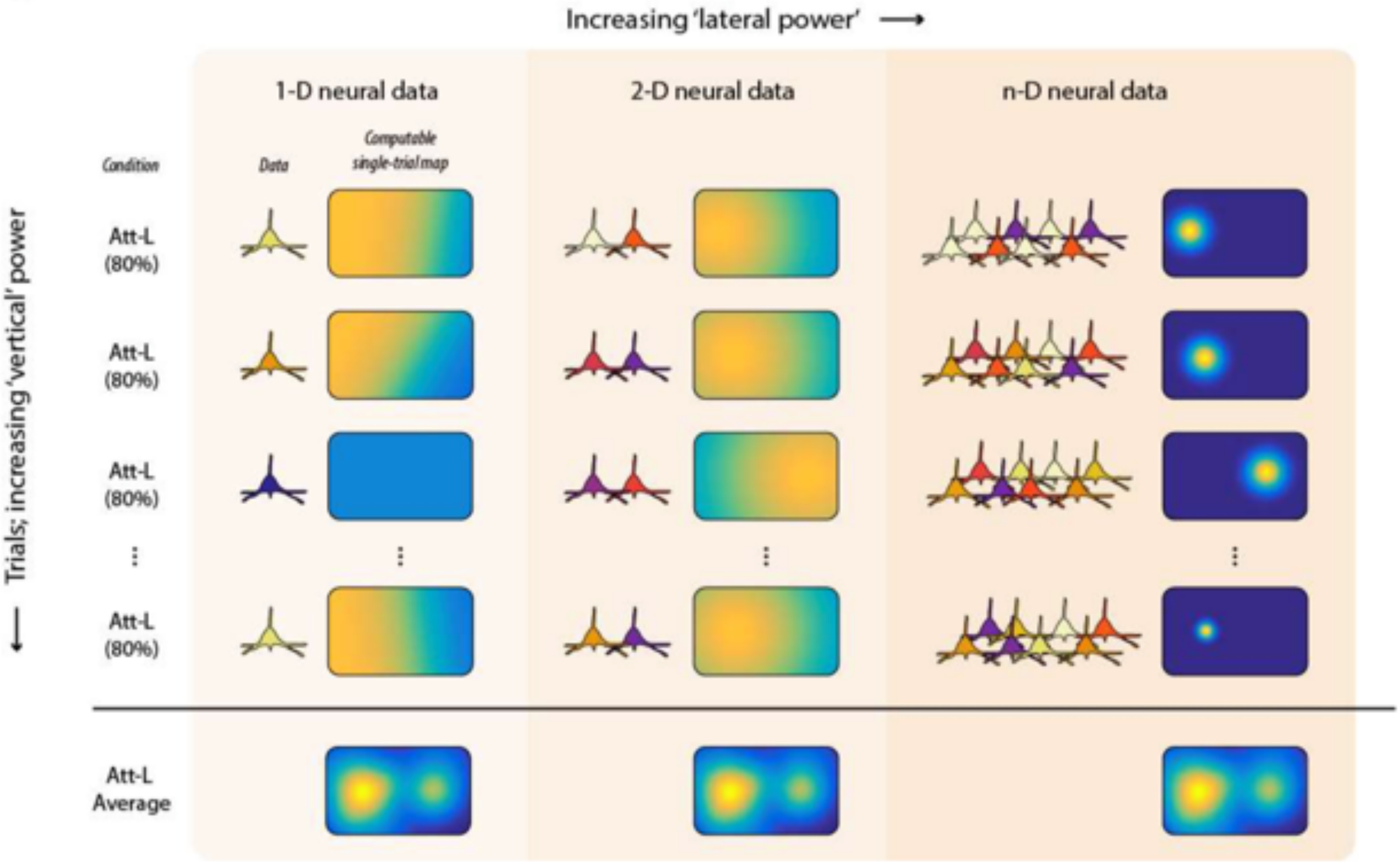
Combining by computing mean doesn't necessarily create a good representation

- Right: spatial attention ***might*** be distributed in a continuous fashion throughout the visual field (as in the average, bottom),

- but such an average profile ***could also be caused by*** individual trials sampling discretely from visual space (80% of trials on the left, 20% on the right).



B



Increasing 'lateral power'
1-D neural data
2-D neural data
n-D neural data
Condition
Data
Computable single-trial map
Att-L (80%)
Att-L (80%)
Att-L (80%)
Att-L (80%)
Att-L Average
Trials; increasing 'vertical' power

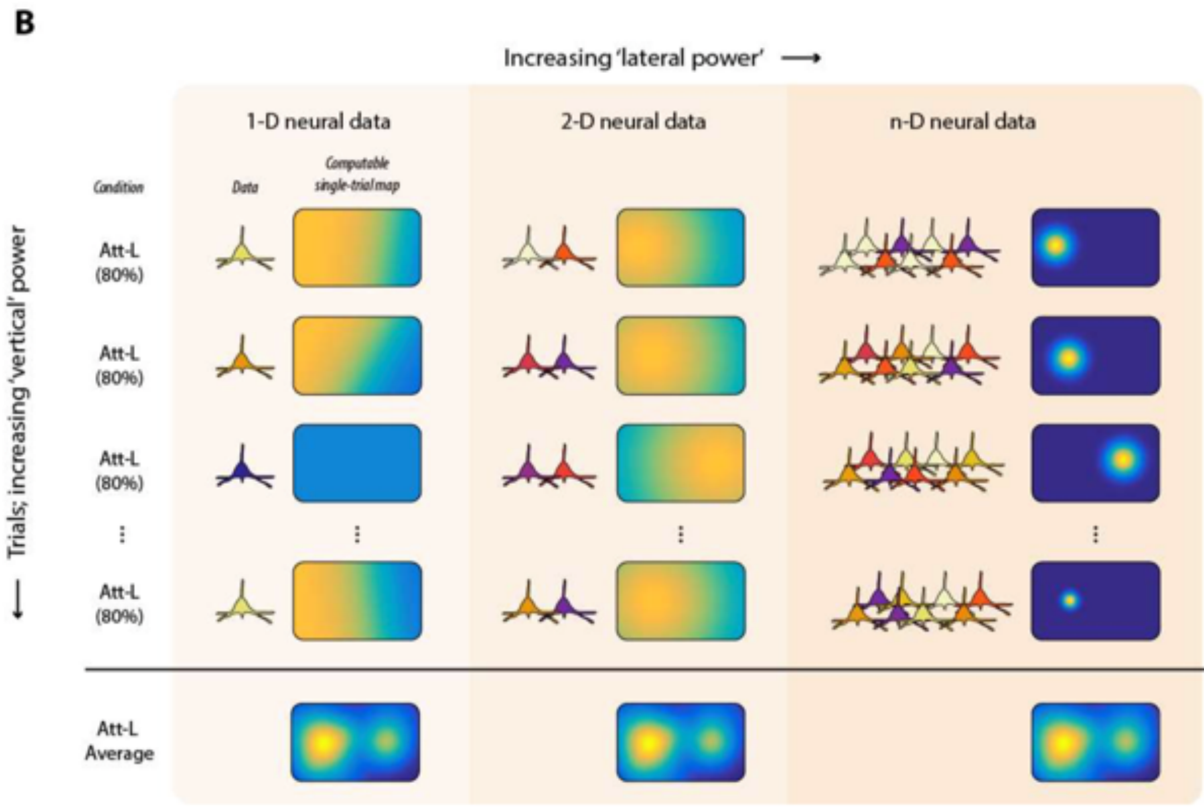
Combining by computing mean doesn't necessarily create a good representation

- Traditionally **statistical power** = more observations (i.e., trials) to average data

    - "Vertical power"

- **Lateral power:** adding more measurement density (spatial dimension).

    Larger lateral power-> Necessary for characterizing neural dynamics in single trial (we'll come back to this)



Motivation for single trial analysis

- We can thus miss important variability patterns by collapsing to the mean
    - Consider a large classroom with 500 students all talking while waiting for lecture.
    - If we record their conversations even for the same class each day all quarter then average the recording, will we recover what the individuals said?
    - No it doesn’t ‘average’ to the conversation as each day has subtle differences
- Solution -**Single trial analysis** studies variability across trials