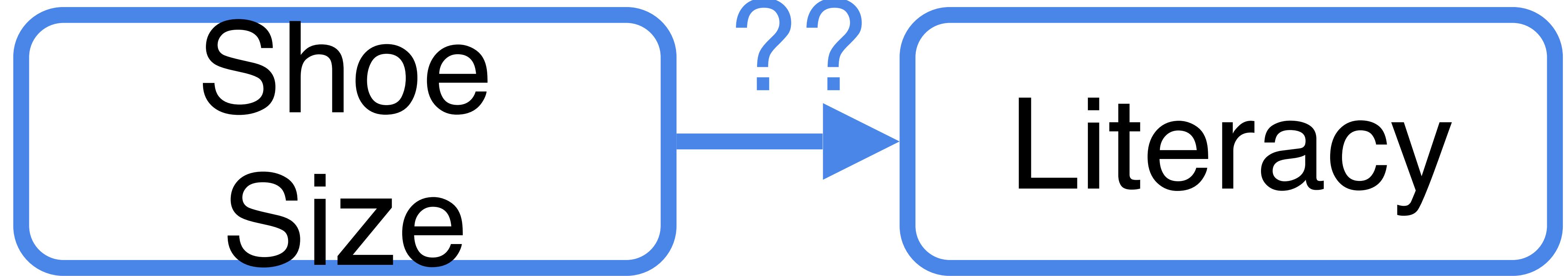


Confounding





Small shoes
Not literate
Child

Big shoes
Literate
Adult

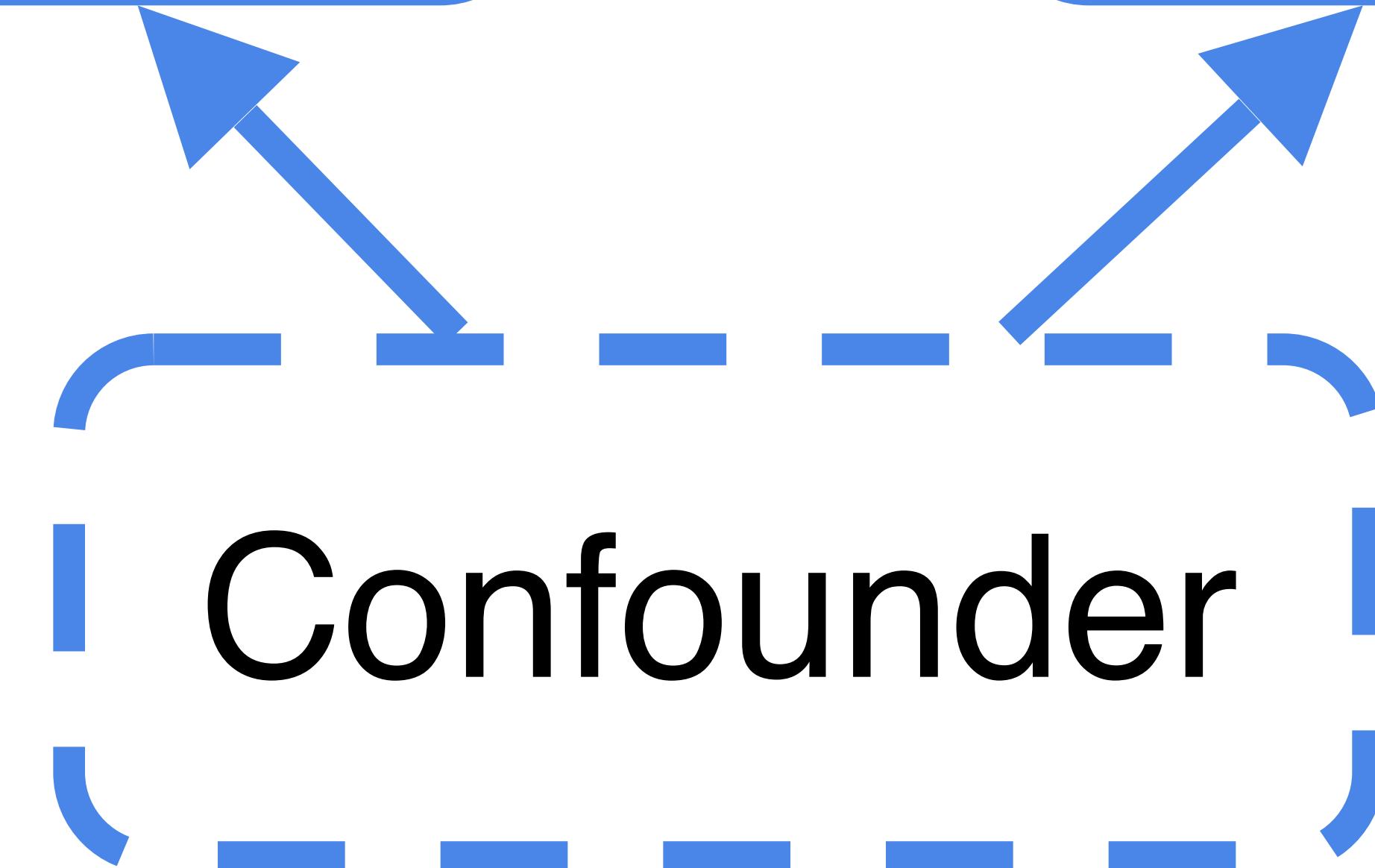
**Shoe
Size**

Literacy

Age

Variable1

Variable2



Confounding

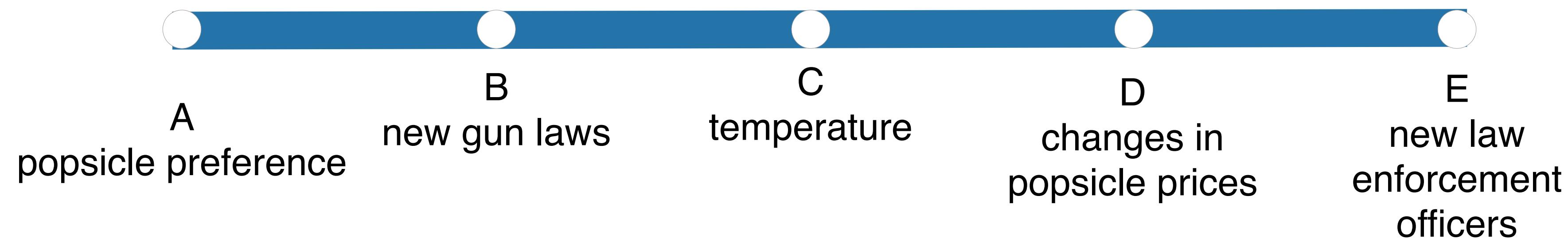


popsicles

crime rate



Your analysis sees an increase in crime rate whenever popsicle sales increase. What could confound this analysis?



You can plan ahead to avoid confounding and/or include confounders in your models to account for their role on the outcome variable.

Ignoring confounders will lead you to draw incorrect conclusions

Stratification changes results

Sample: 400 patients with index vertebral fractures

...looks like vertebroplasty was *way* worse for patients!

Vertebroplasty	Conservative care	Relative risk (95% confidence interval)
30/200 (15%)	15/200 (7.5%)	2.0 (1.1–3.6)

subsequent fractures

But wait...at time of initial fracture...

	Vertebroplasty N = 200	Conservative care N = 200
Age, y, mean \pm SD	78.2 ± 4.1	79.0 ± 5.2
Weight, kg, mean \pm SD	54.4 ± 2.3	53.9 ± 2.1
Smoking status, No. (%)	110 (55)	16 (8)

Age and weight are similar between groups. **Smoking Status** differs vastly.

So...let's stratify those results real quick

Smoke			No smoke		
Vertebroplasty	Conservative	RR (95% confidence interval)	Vertebroplasty	Conservative	RR (95% confidence interval)
23/110 (21%)	3/16 (19%)	1.1 (0.4, 3.3)	7/90 (8%)	12/184(7%)	1.2 (0.5, 2.9)

Risk of re-fracture is now similar
within group

Nonparametric statistics

Jason G. Fleischer, Ph.D.

Asst. Teaching Professor

Department of Cognitive Science, UC San Diego

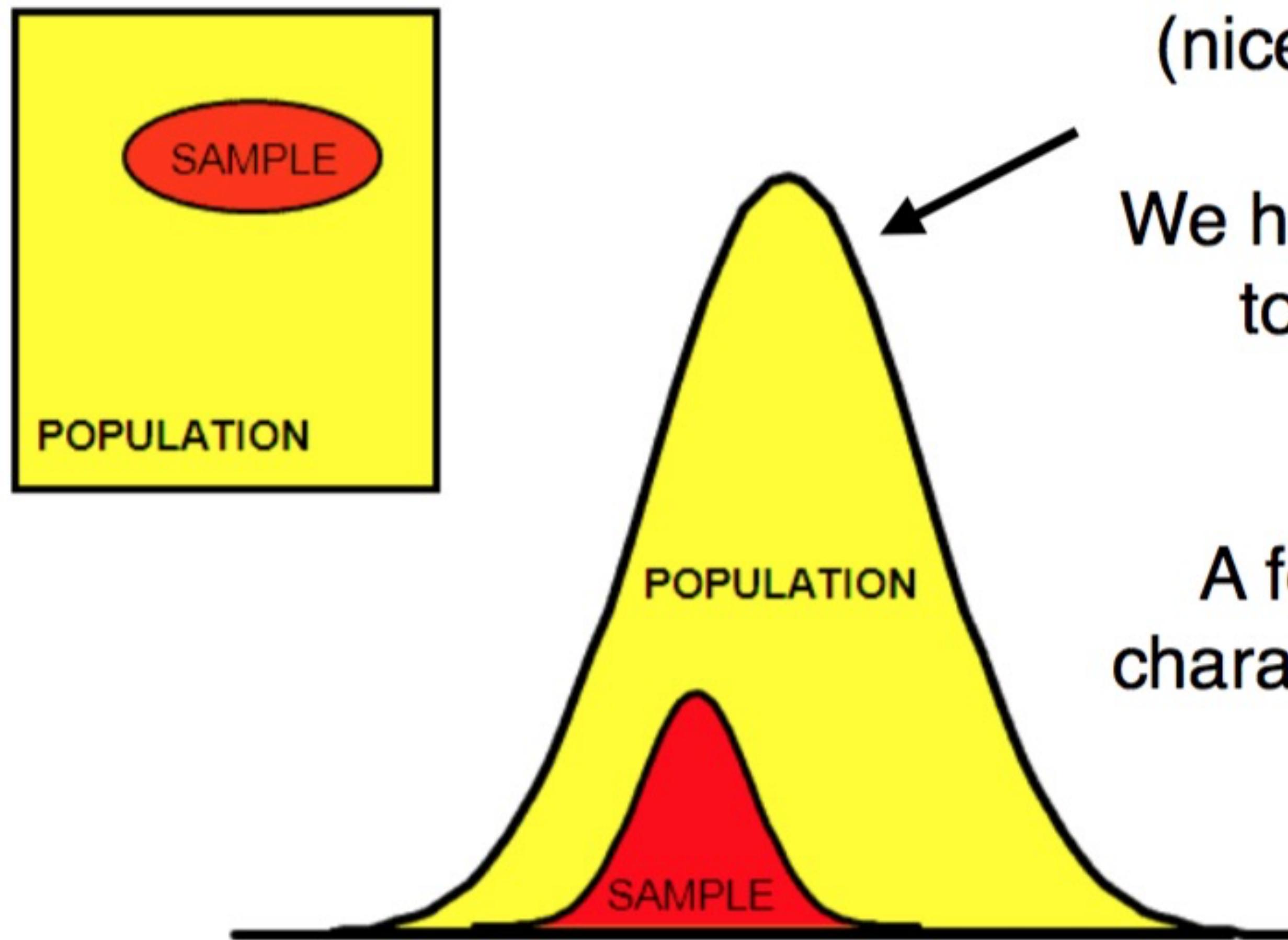
jfleischer@ucsd.edu



@jasongfleischer

<https://jgfleischer.com>

Non-parametric Statistics

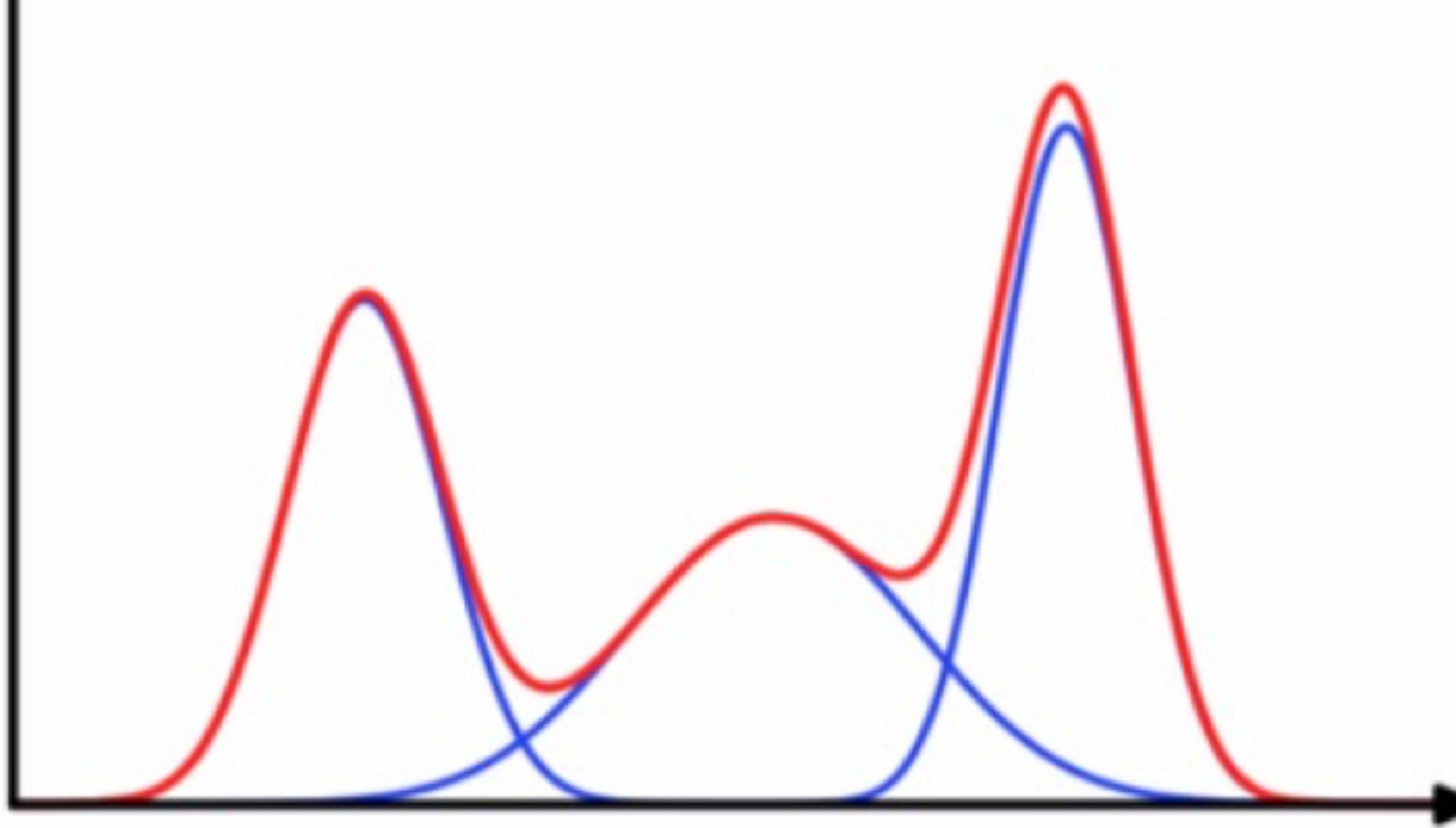


Normal distribution
(nice and friendly)

We have good math tools for this.

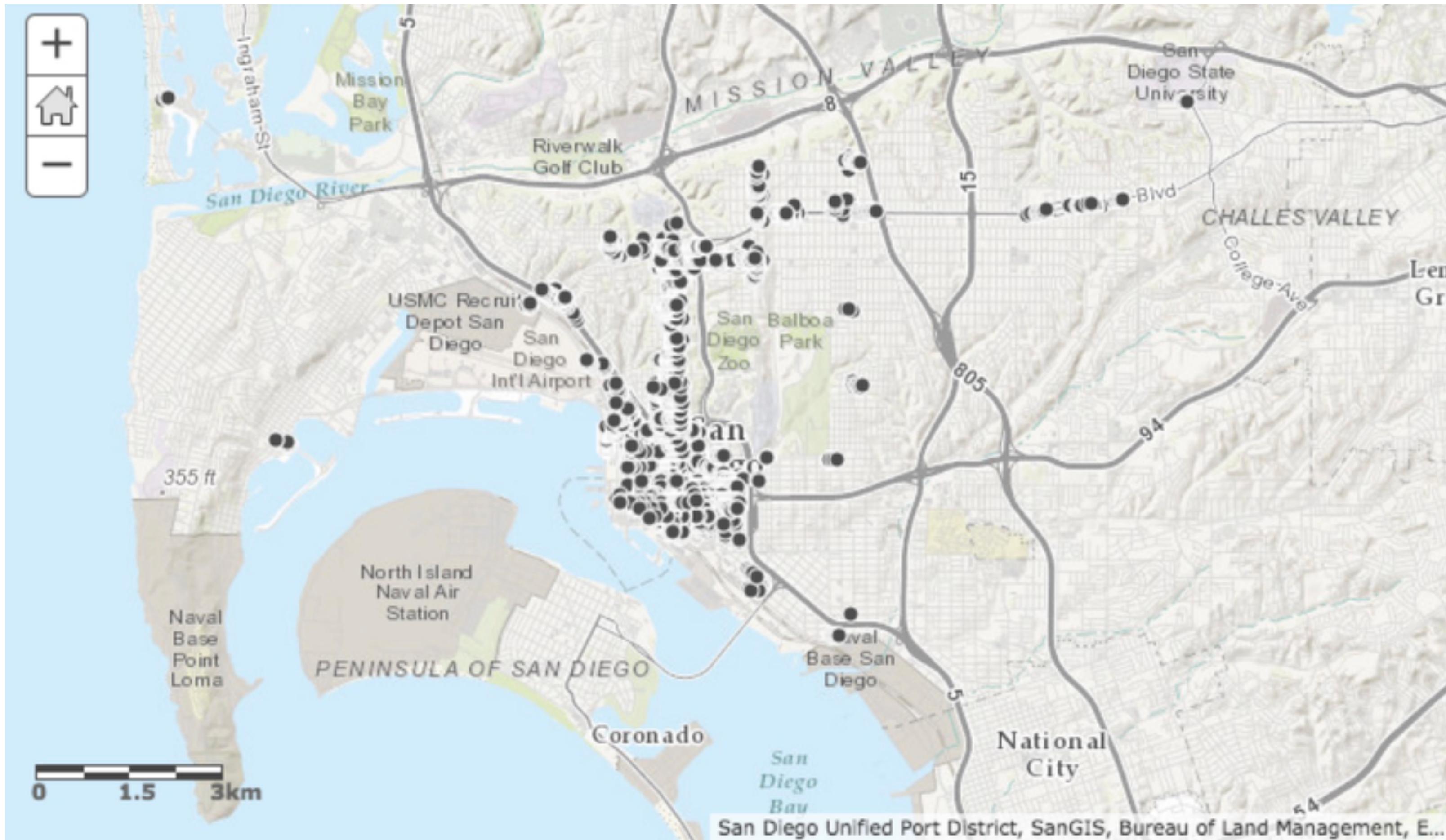
A few parameters **fully** characterize the distribution.

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



adapted from Brad Voytek

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



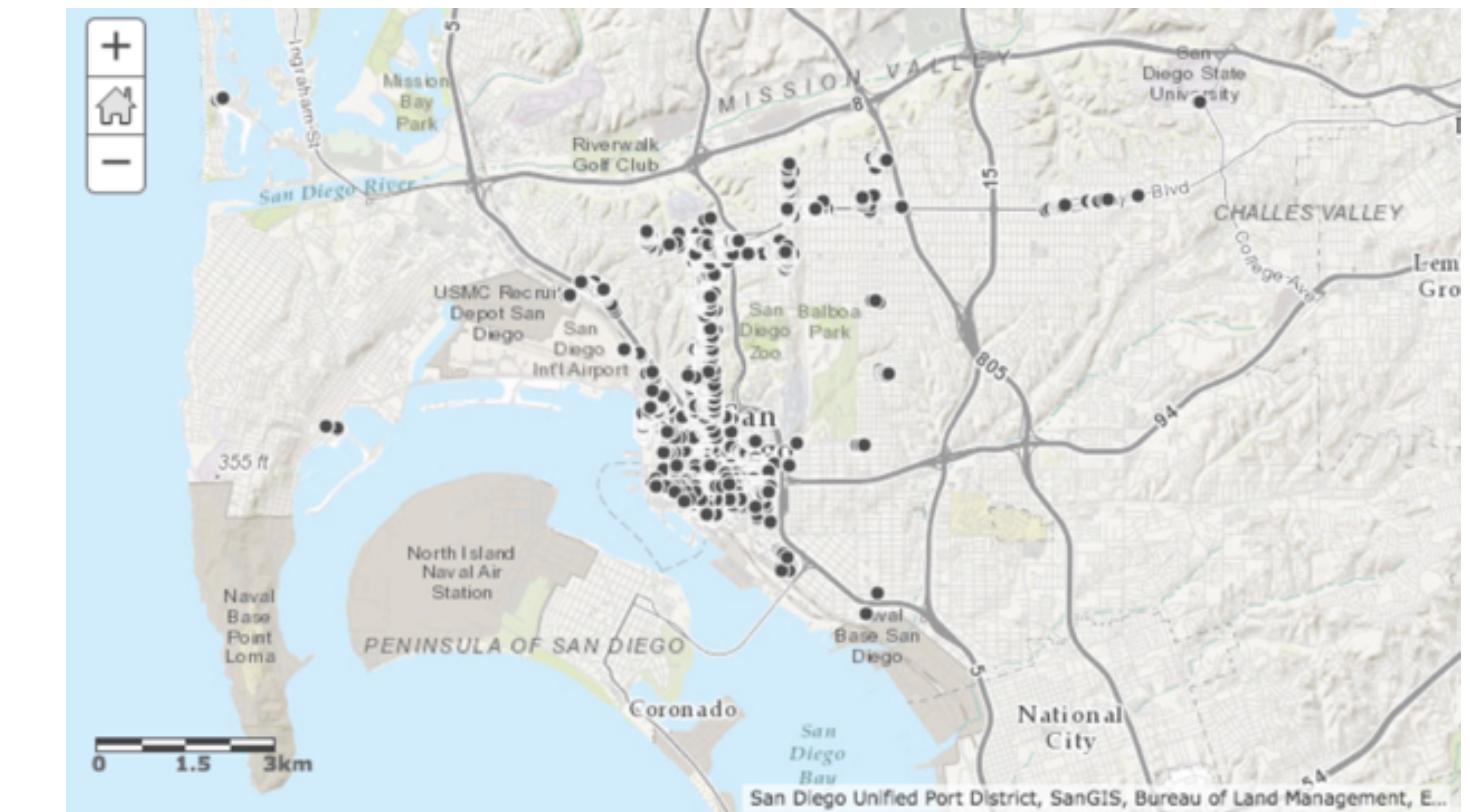
Parameters (like mean and variance) cannot fully and accurately capture this distribution!

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

adapted from Brad Voytek

When to turn to non-parametric statistics...

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



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

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

Resampling statistics: The What

Empirical null distribution (Monte Carlo)

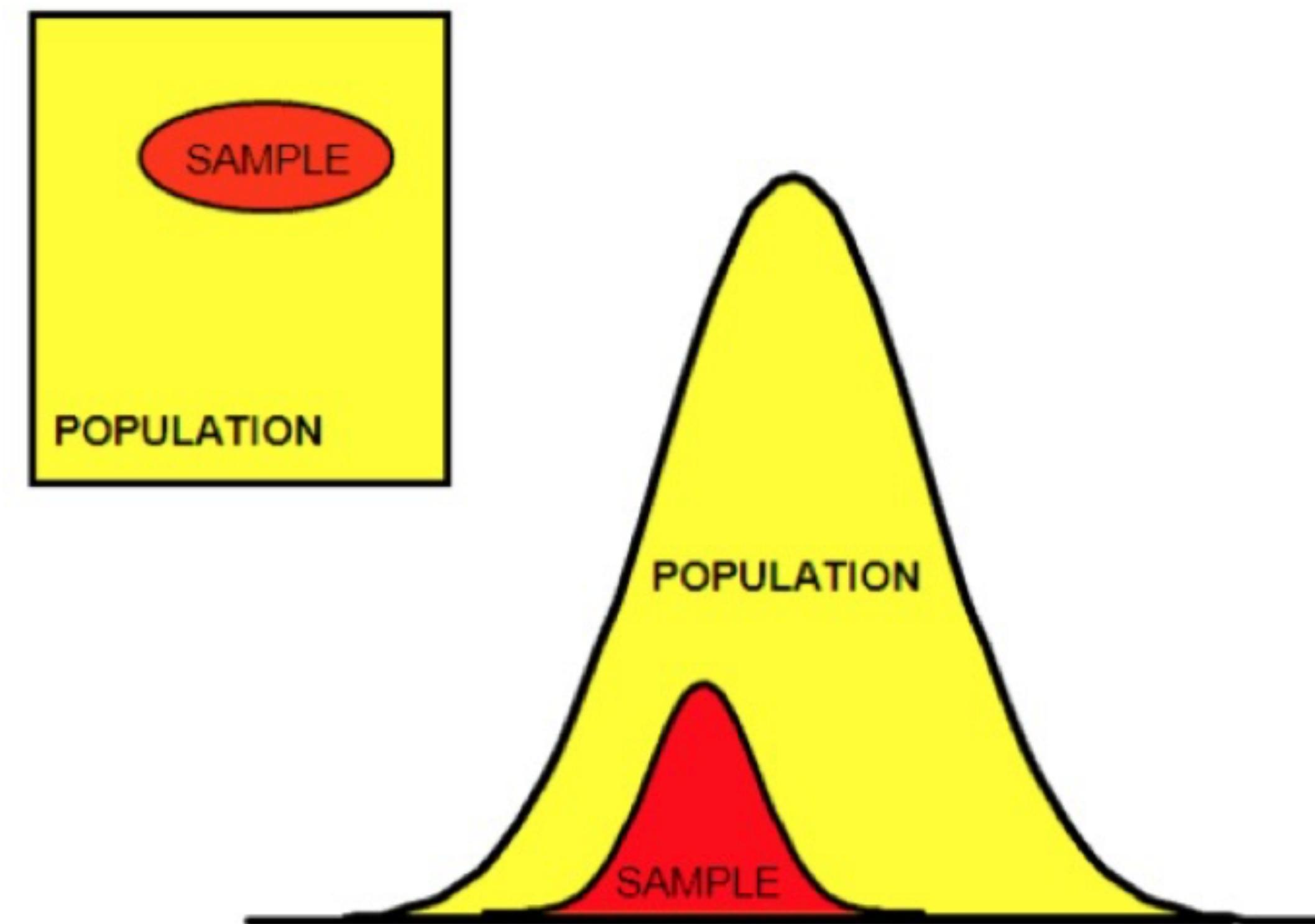
Rank Statistics (Mann Whitney U)

Kolmogorov-Smirnoff Test

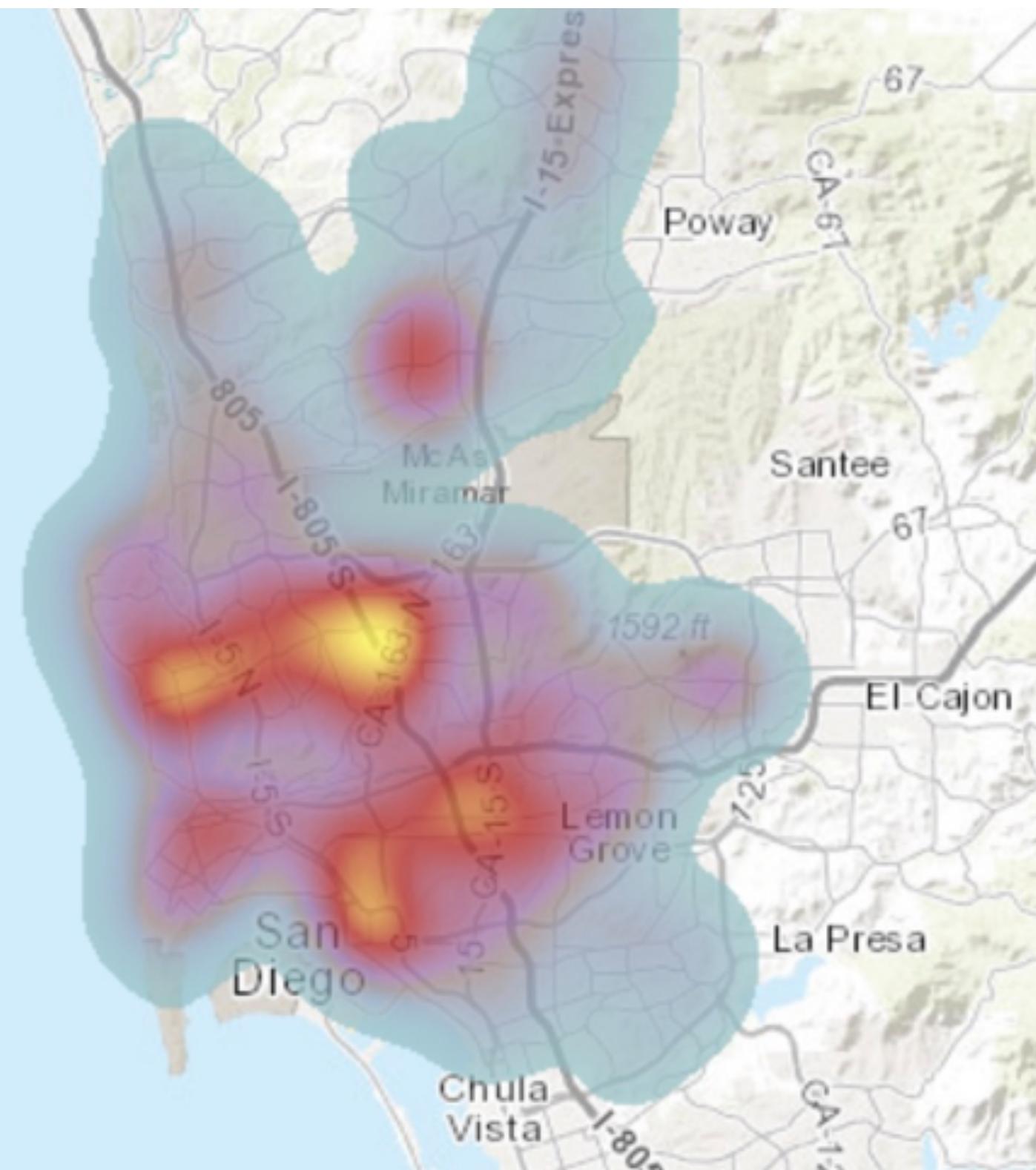
Non-parametric prediction models

1) Bootstrapping (resampling)

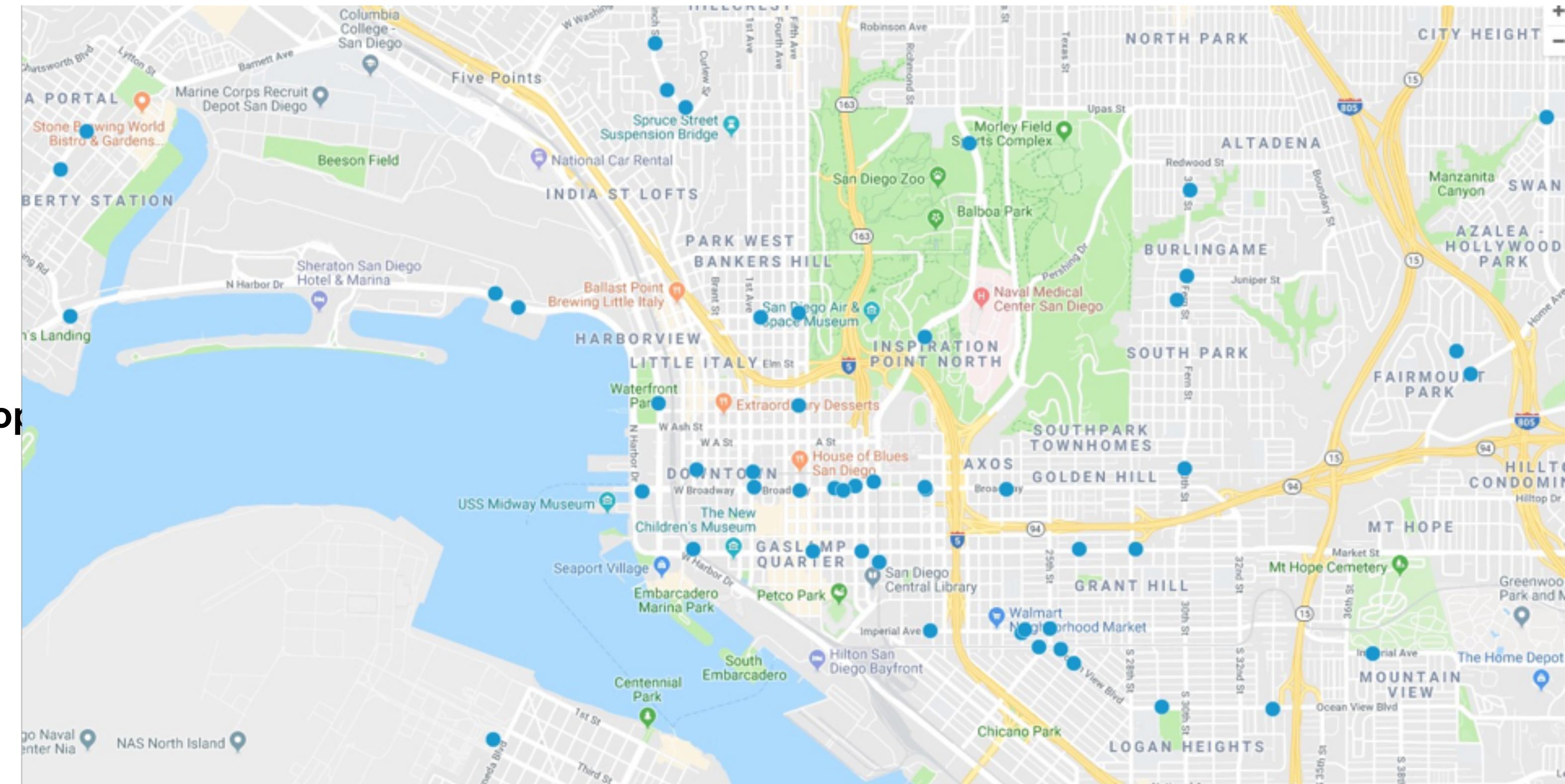
How can we build a more realistic “null distribution” for the sample estimate without knowing the population it’s drawn from?



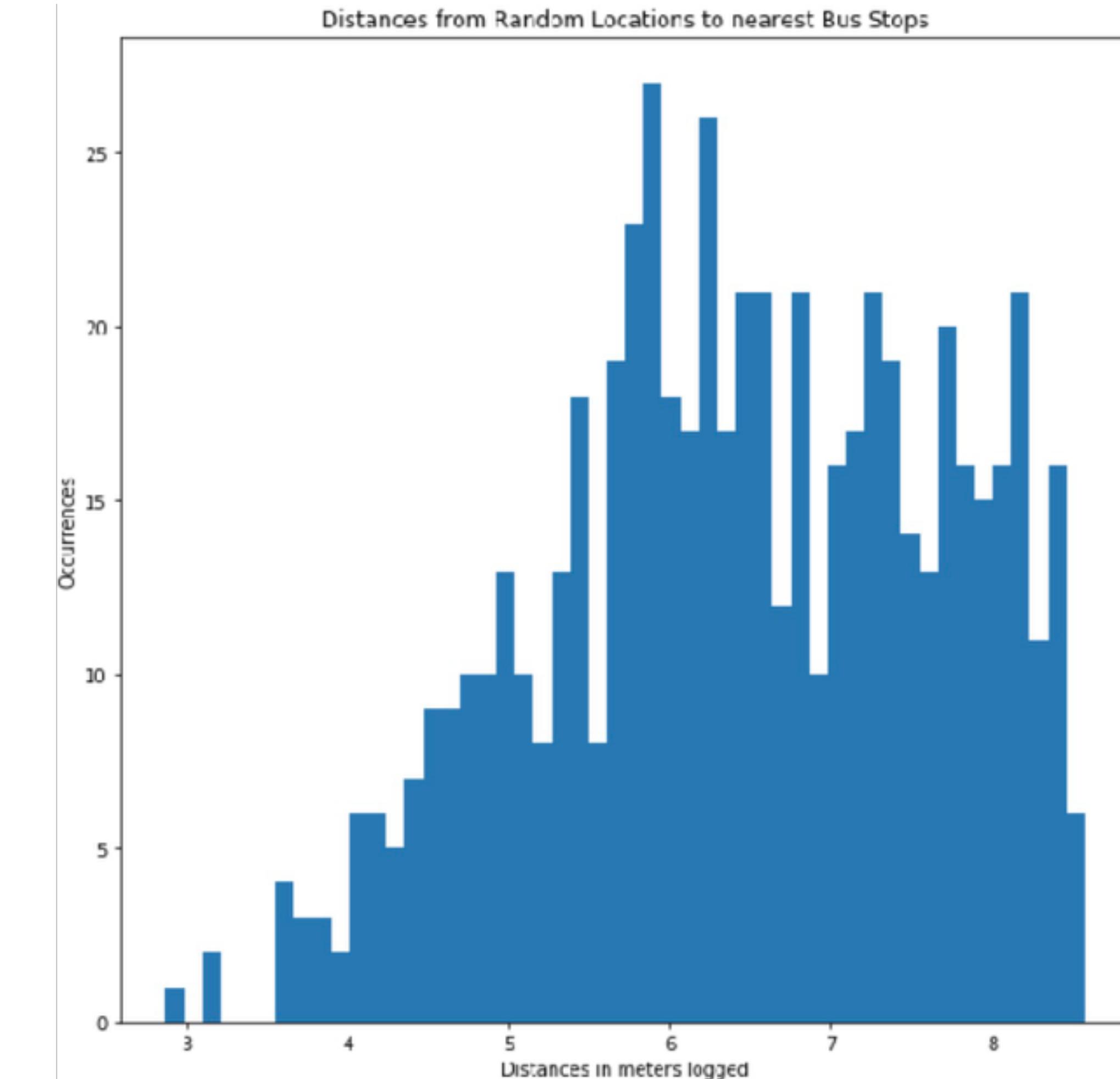
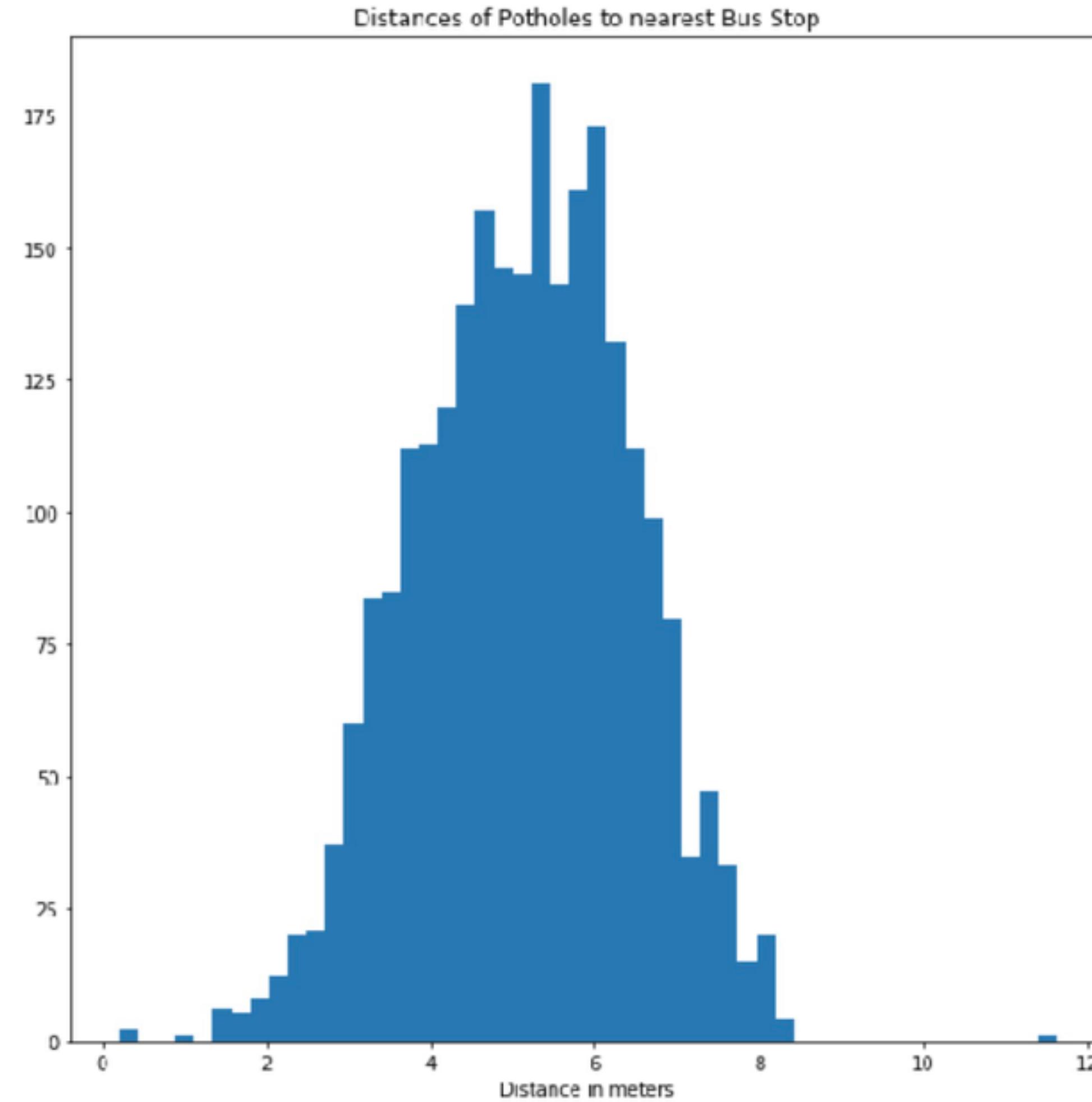
Bootstrapping (resampling)



stop



Bootstrapping (resampling)



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

quantitative data

1, 4.5, 6.6, 9.2

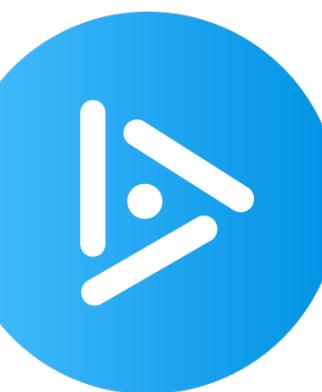
ordinal data

→ 1, 2, 3, 4

Data are transformed from their quantitative value to their rank.

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)



Ordinality

Which of the following variables contains ordinal data?

<https://forms.gle/UUbbQd9nyz8tgzVh6>



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 median

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

Mann-Whitney U: question example

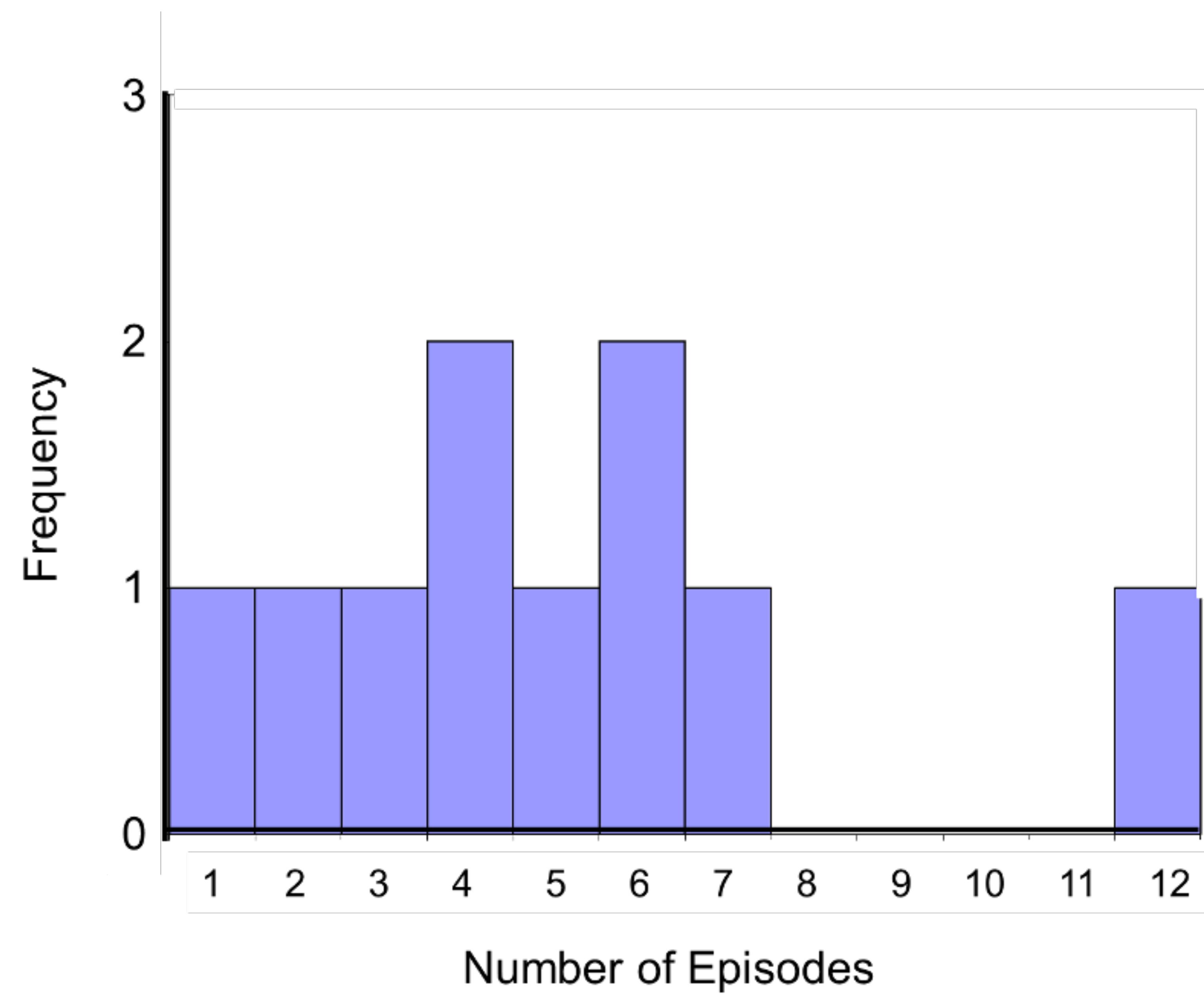
In a clinical trial, is there a difference in the number of episodes of shortness of breath between placebo and treatment?

Step 1: Participants record number of episodes they have.

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

Step 2: Resort the ranks into separate samples (placebo vs. treatment)

Step 3: Carry out statistical test



		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

H_0 : low and high scores are approximately evenly distributed in the two groups

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

H_a : low and high scores are NOT evenly distributed in the two groups ($U \leq 2$)

The max possible value of T_A

The observed sum of ranks for sample A

n_a = number of elements in group A

$U_{\text{Placebo}} = 3$

0 < U < $n_1 * n_2$

n_b = number of elements in group B

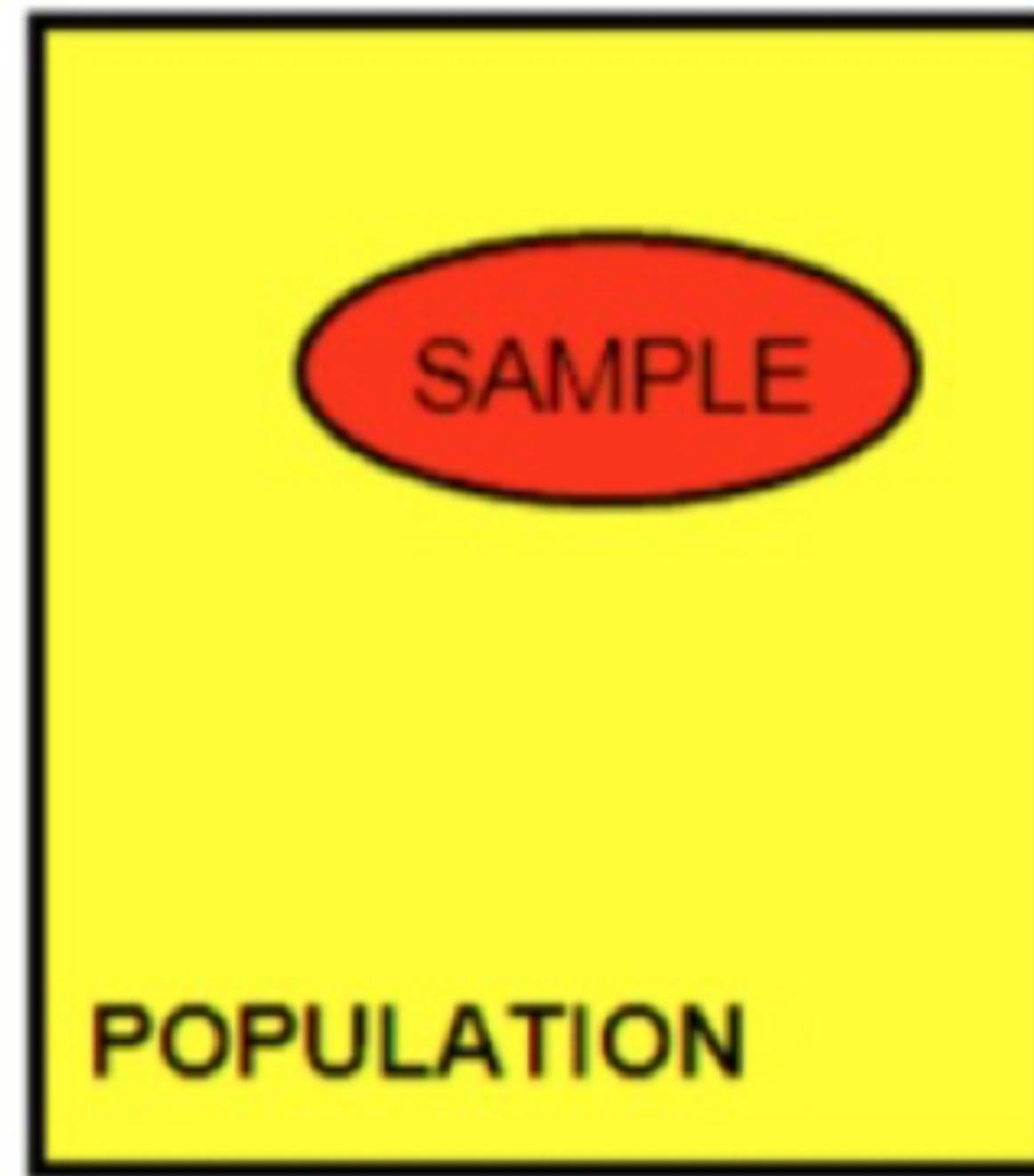
$U_{\text{treatment}} = 22$

Complete separation \rightarrow no separation

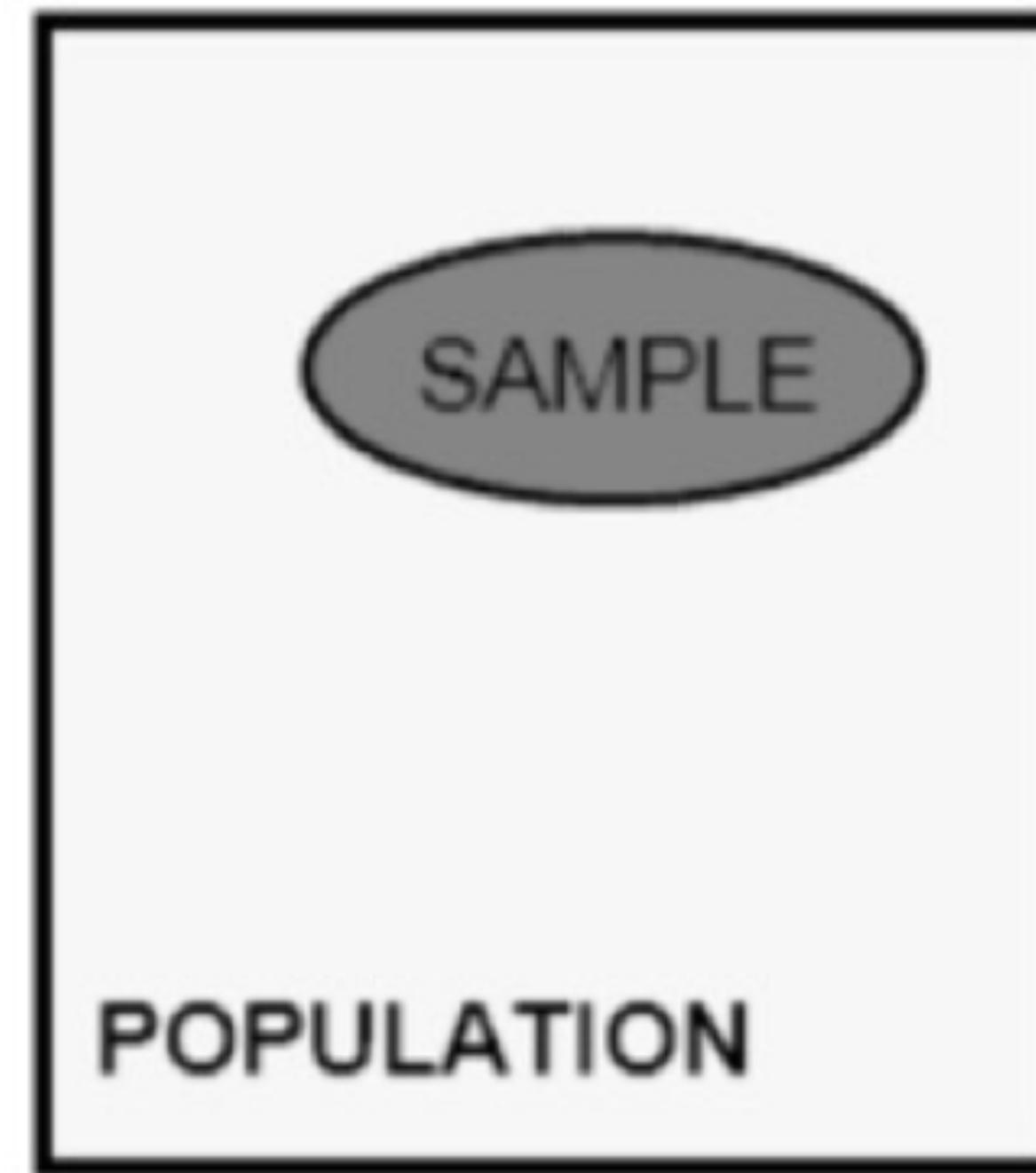
We reject the null if U is small.

3) Kolmogorov-Smirnov (KS) test

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



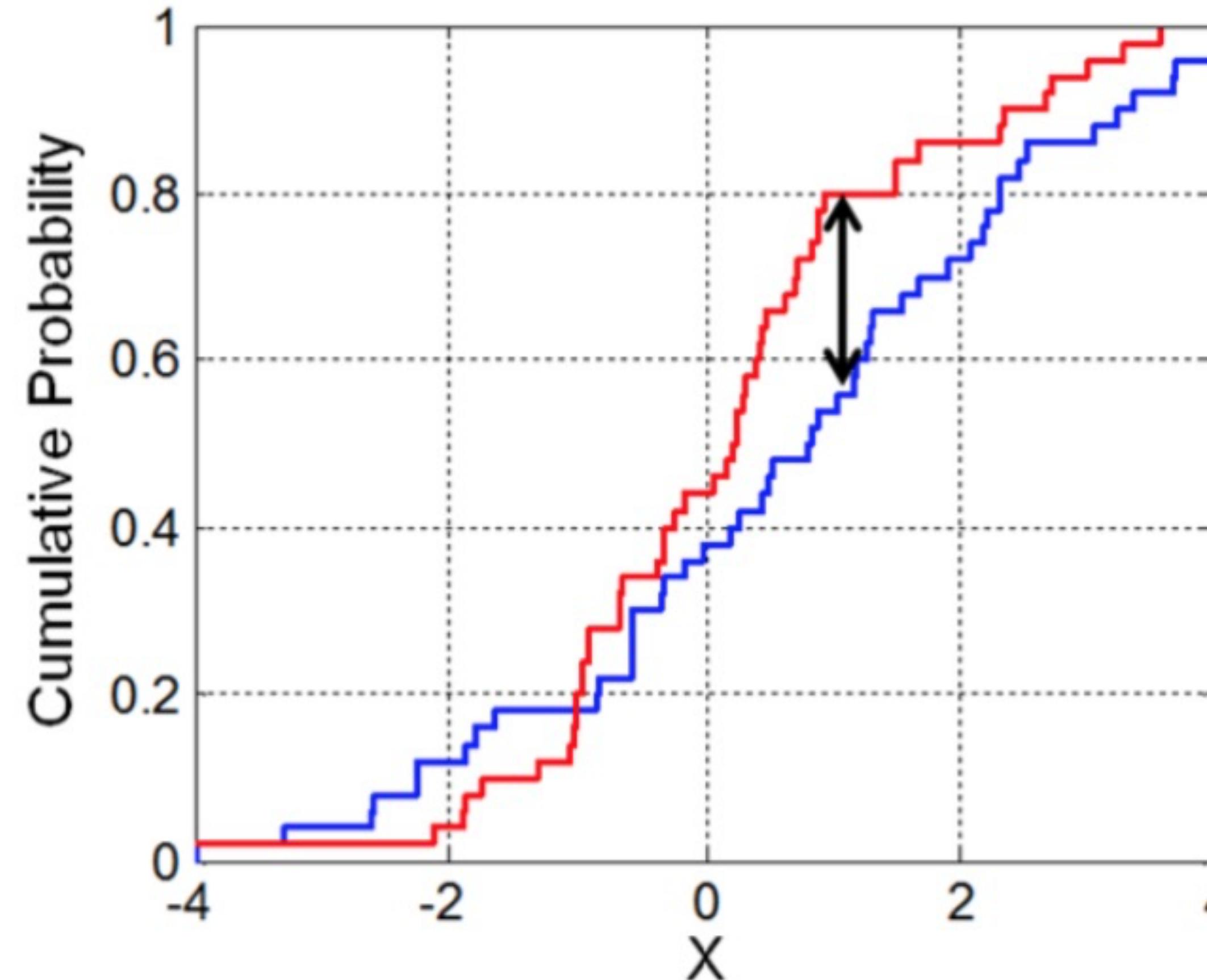
?
==



Kolmogorov-Smirnov (KS) test

Comparing cumulative distributions empirically

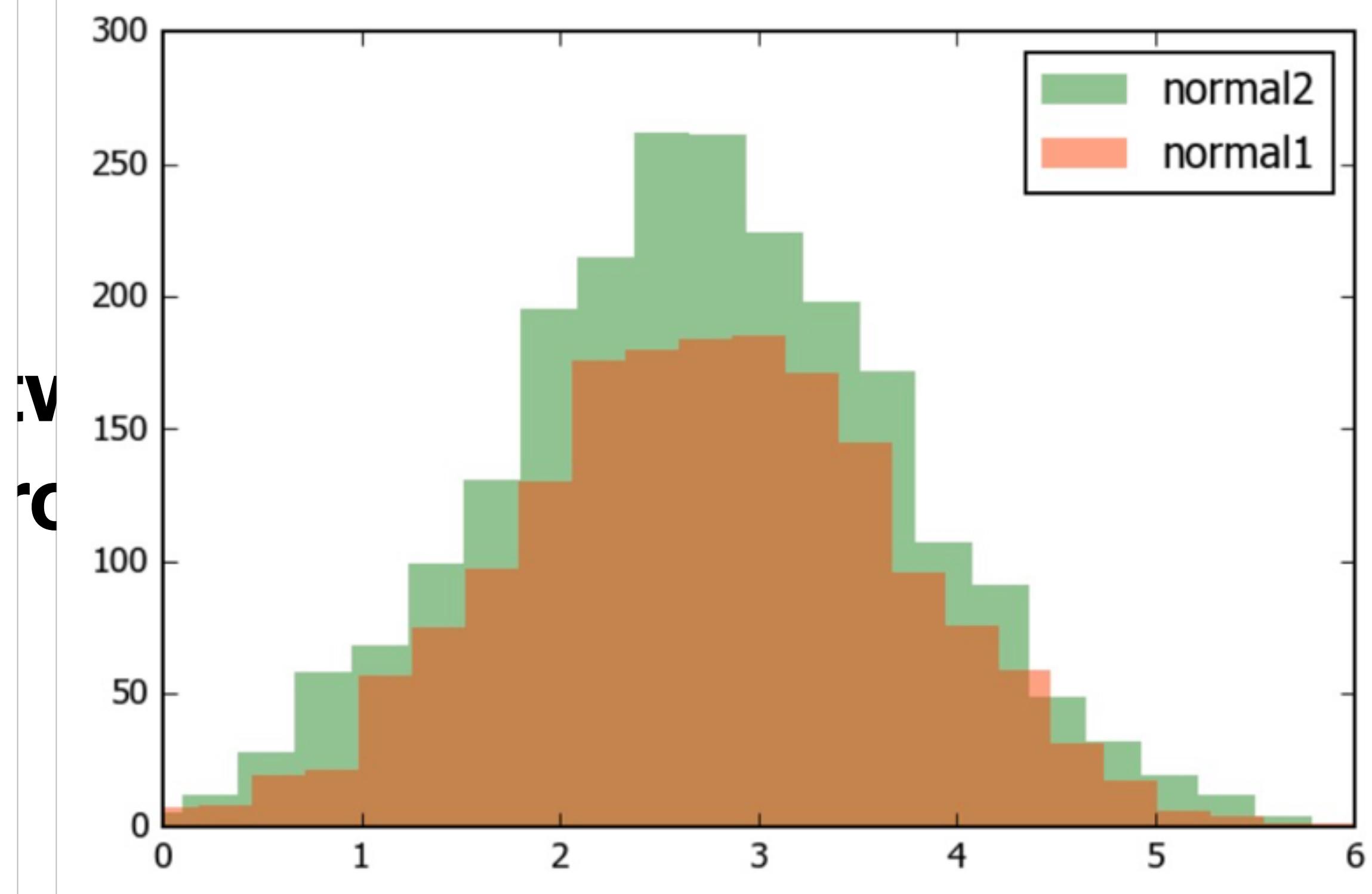
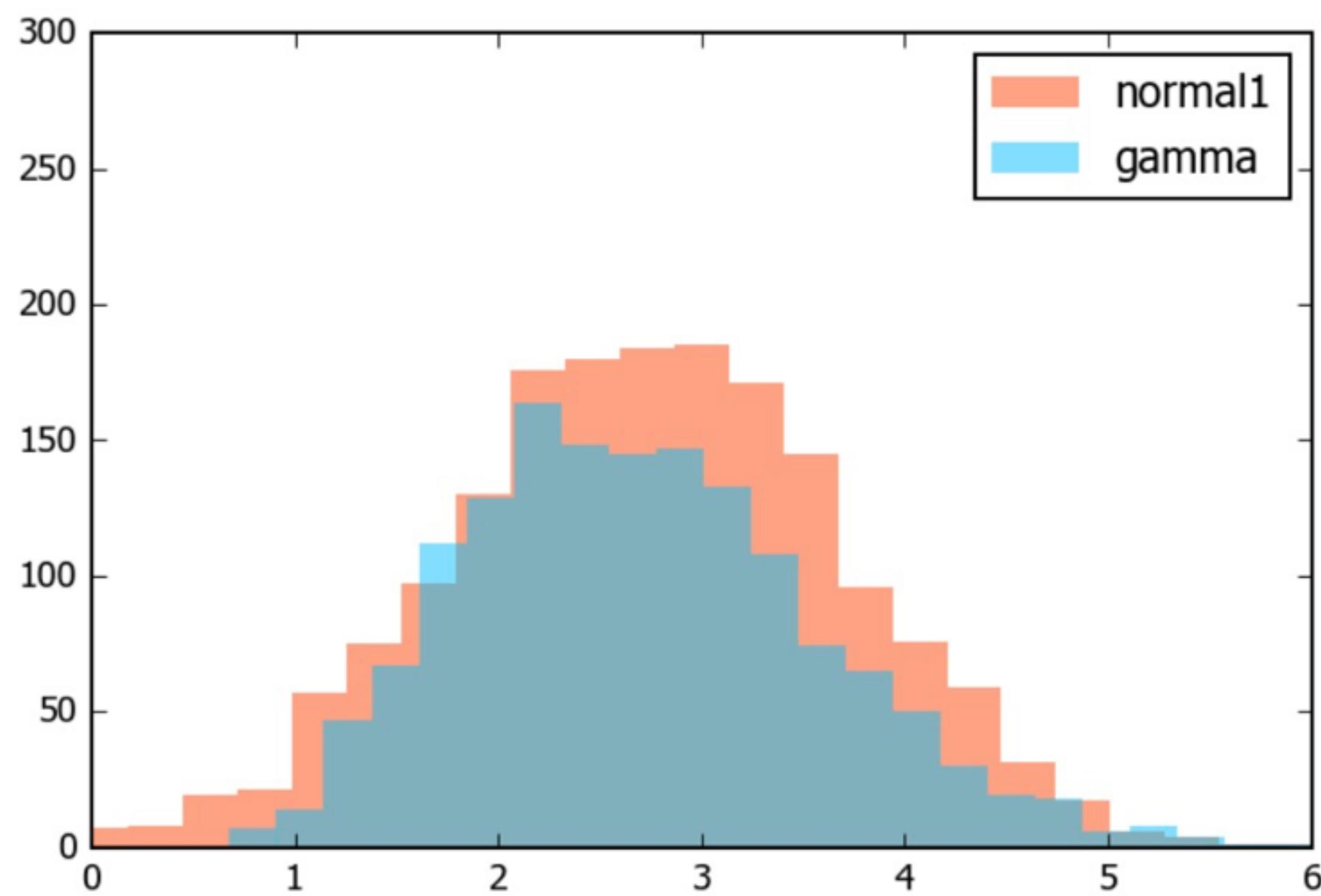
Find the maximum difference between the CDFs.



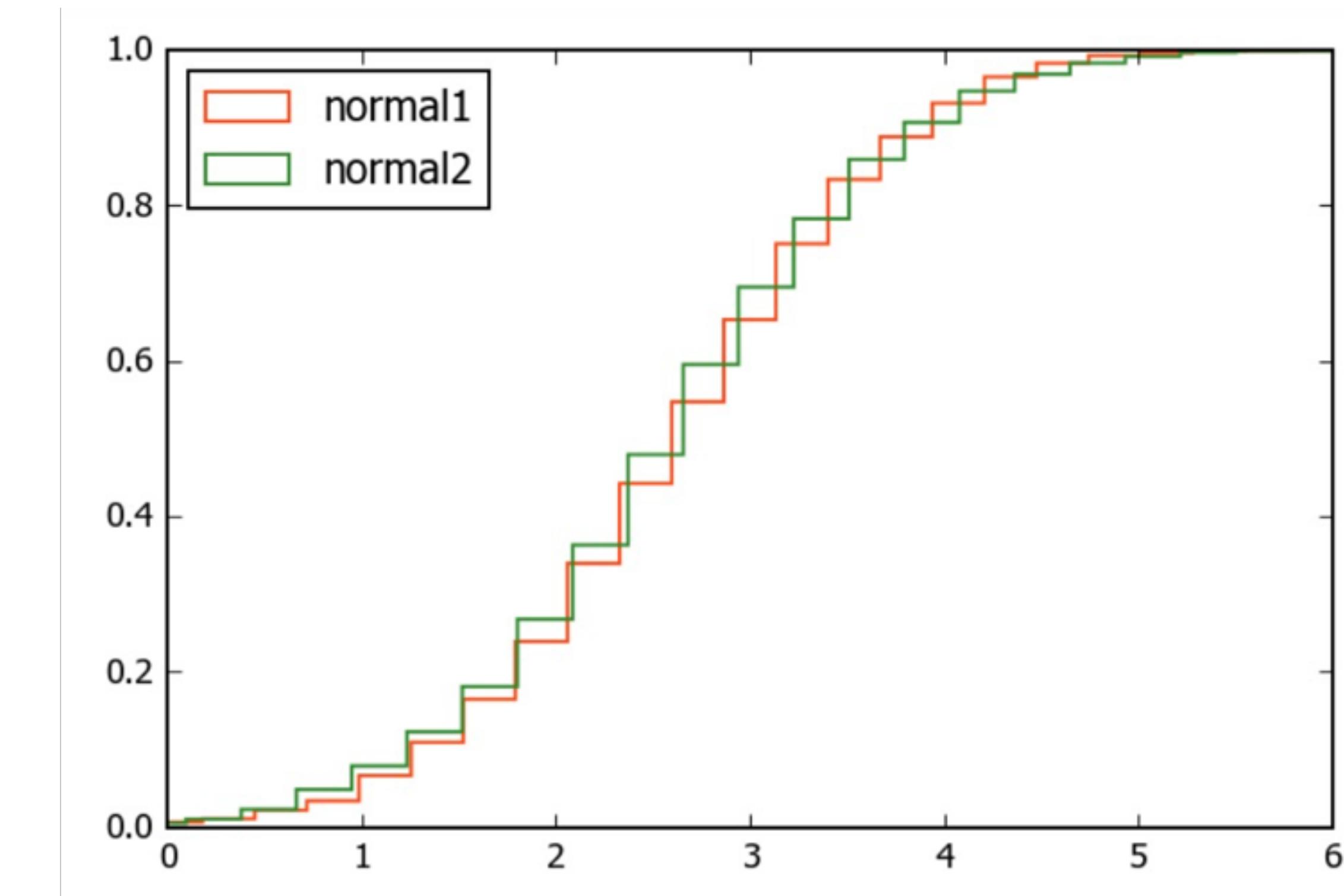
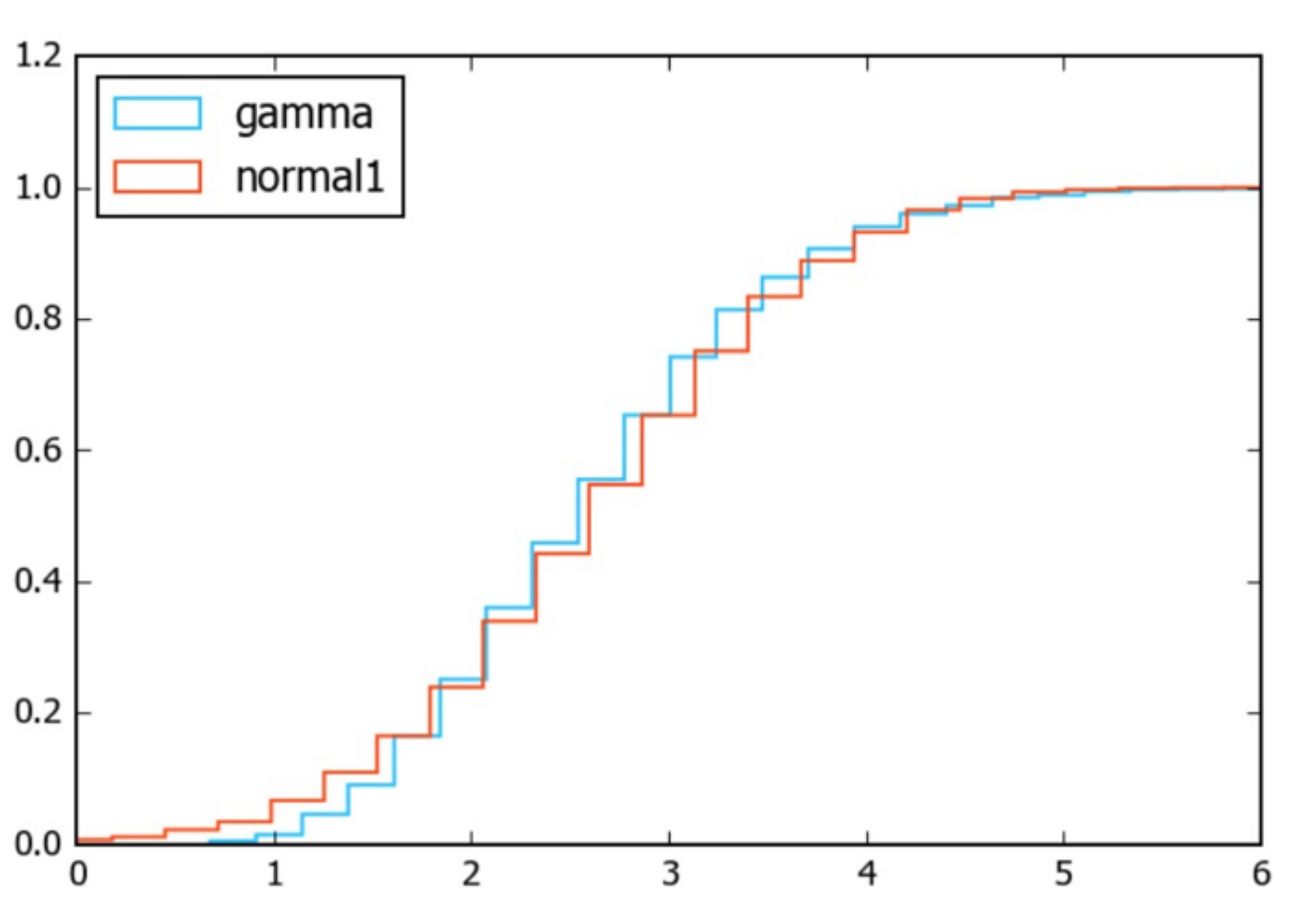
Tests:

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

Kolmogorov-Smirnov (KS) test



Kolmogorov-Smirnov (KS) test



gamma vs. normal1: $p = 0.0106803628411$

normal1 vs. normal2: $p = 0.550735998243$

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 and Decision Tree Regressions (CART)

Quantile regression

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