

Reminders

Upcoming due dates

Wed Oct 29th Project Proposal

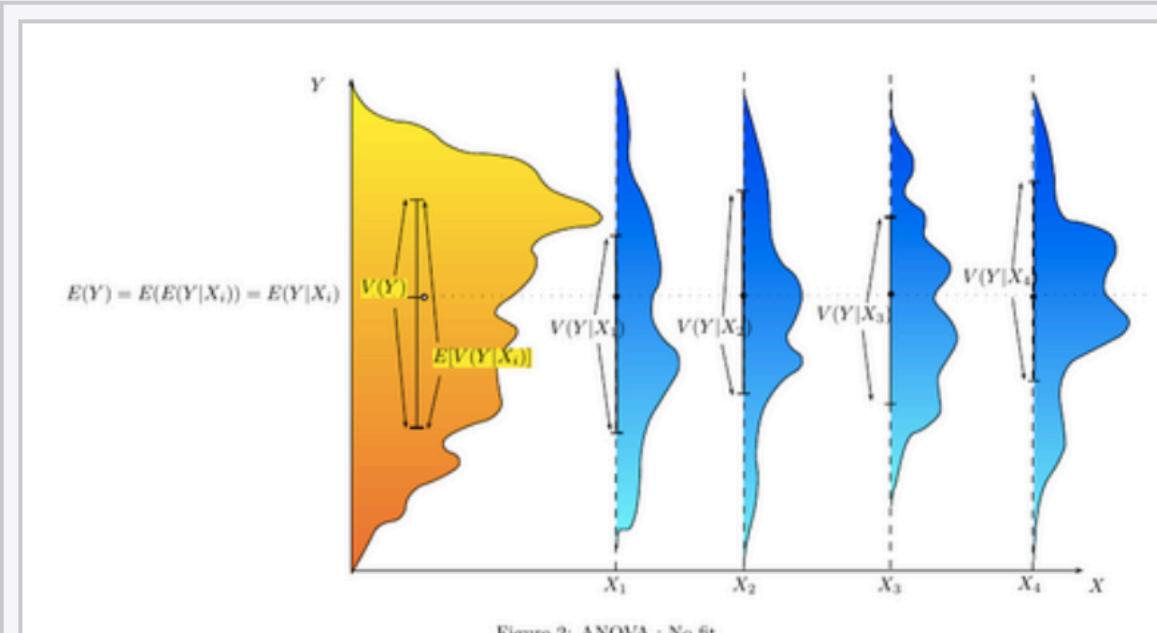
Fri Oct 31st Discussion lab 4

Statistical inference III

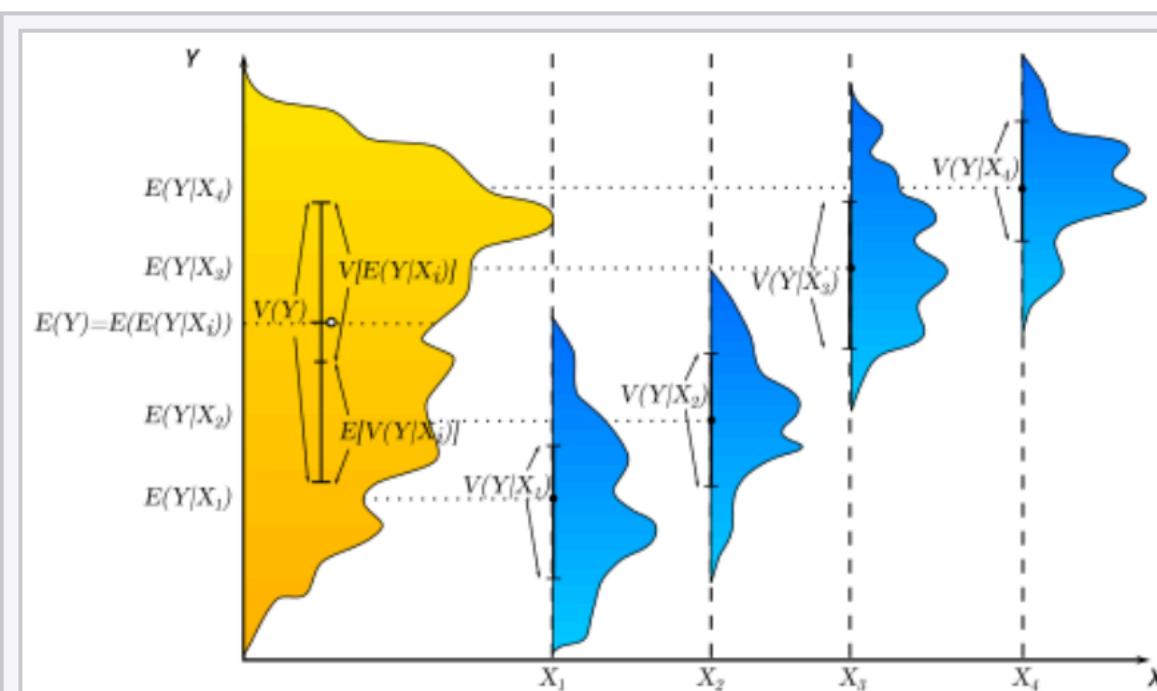
Data Science in Practice

Beyond Pairwise testing: ANOVA

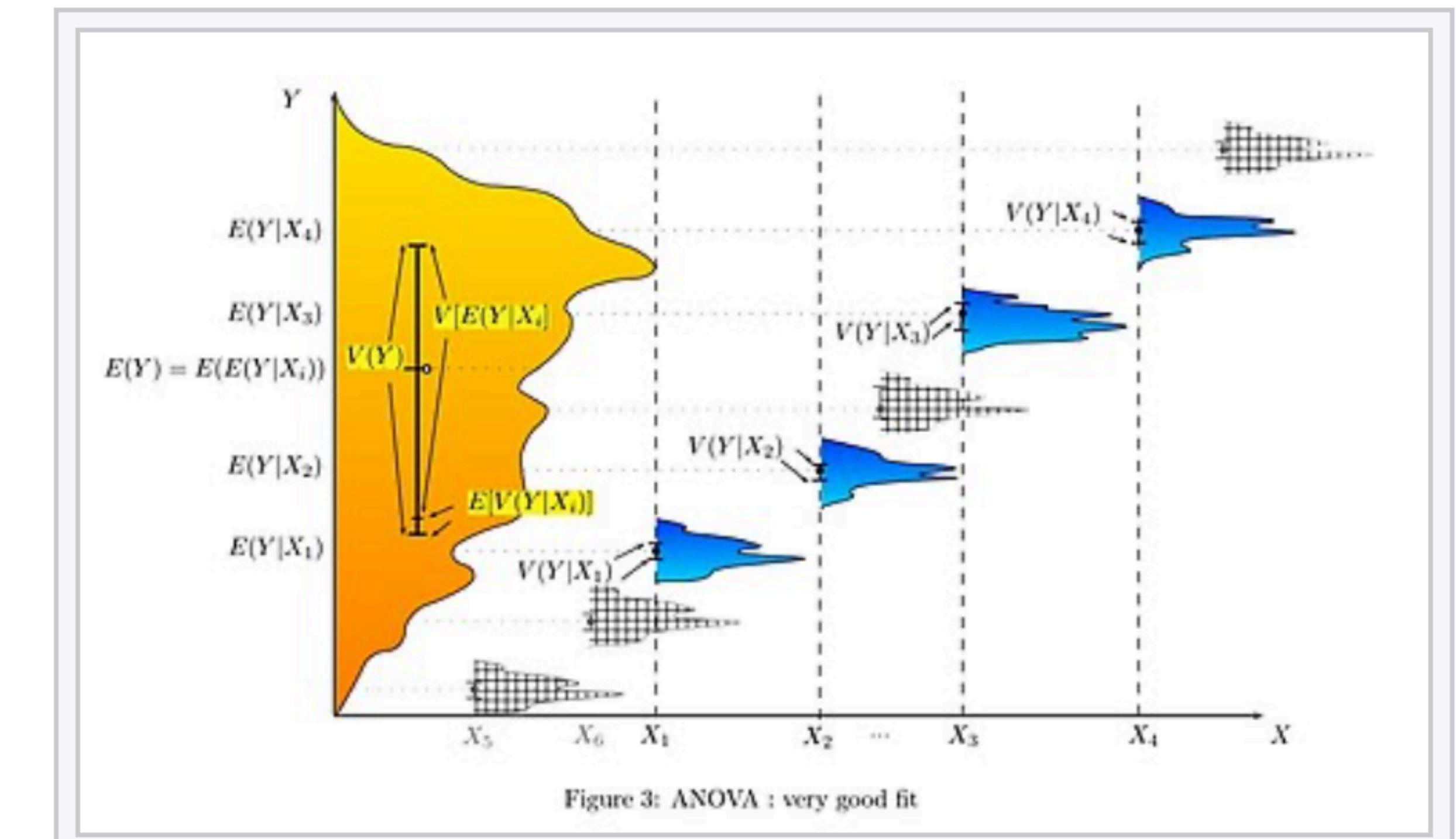
Same assumptions as t-test



No fit: Young vs old, and short-haired vs long-haired

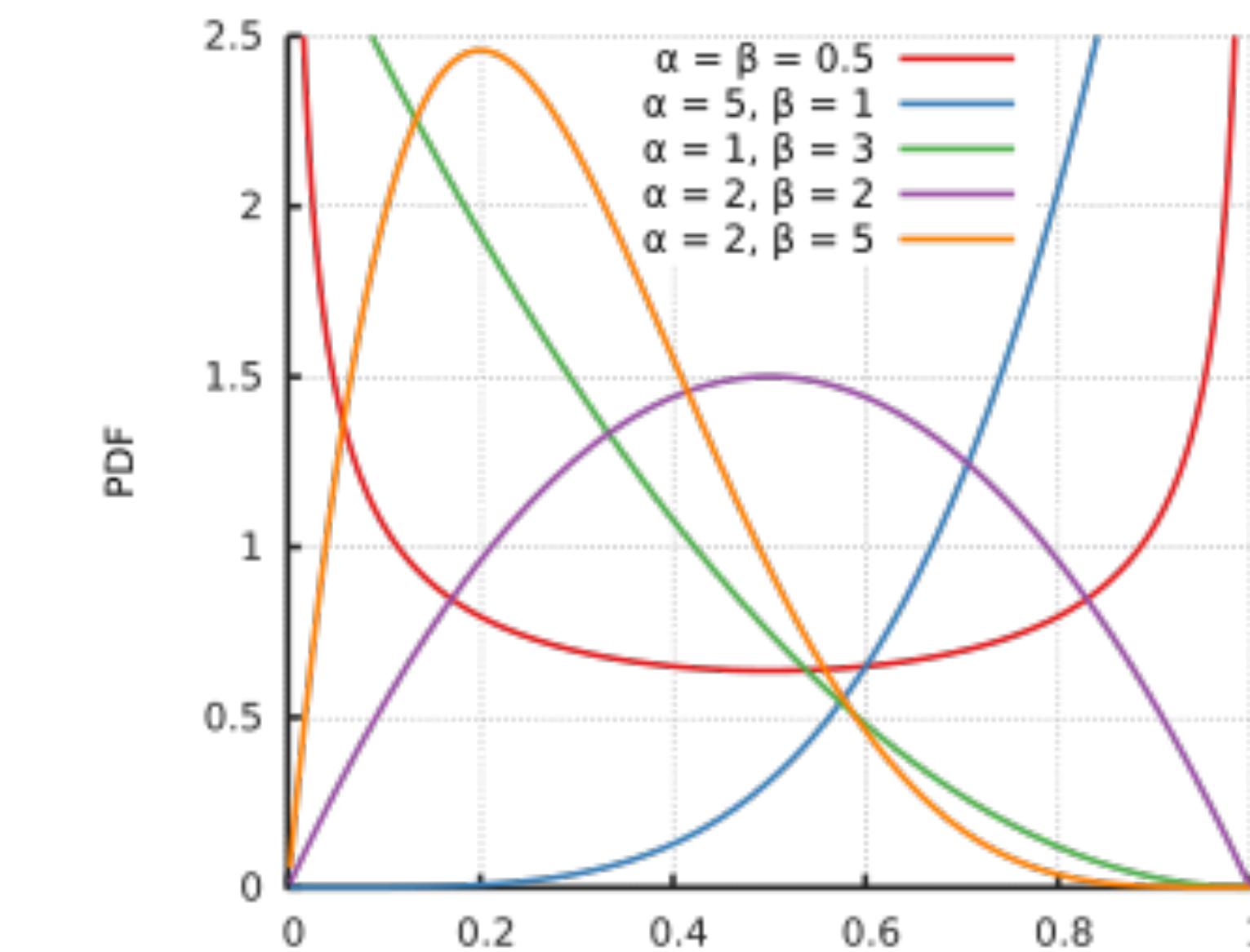
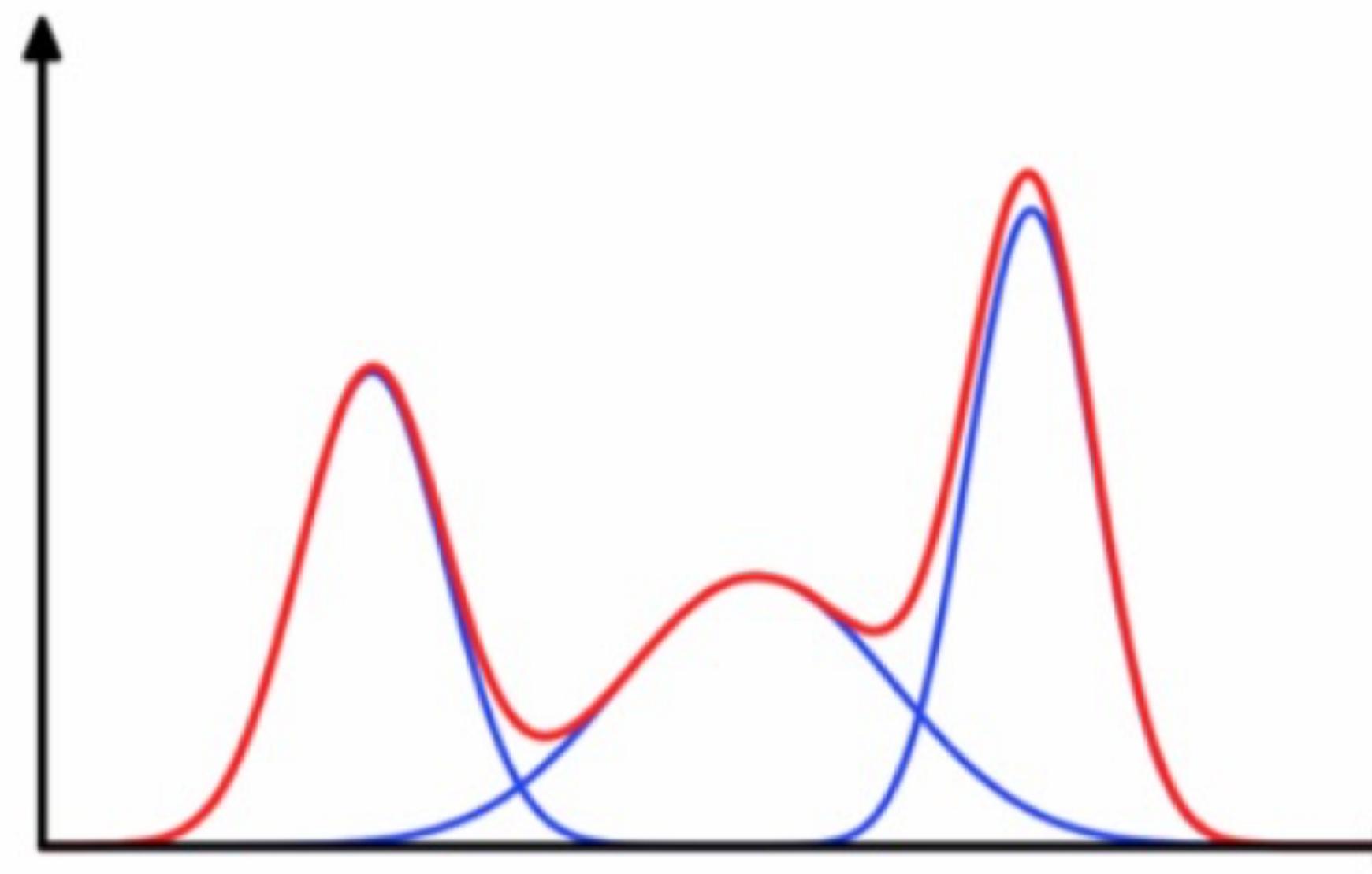


Fair fit: Pet vs Working breed and less athletic vs more athletic

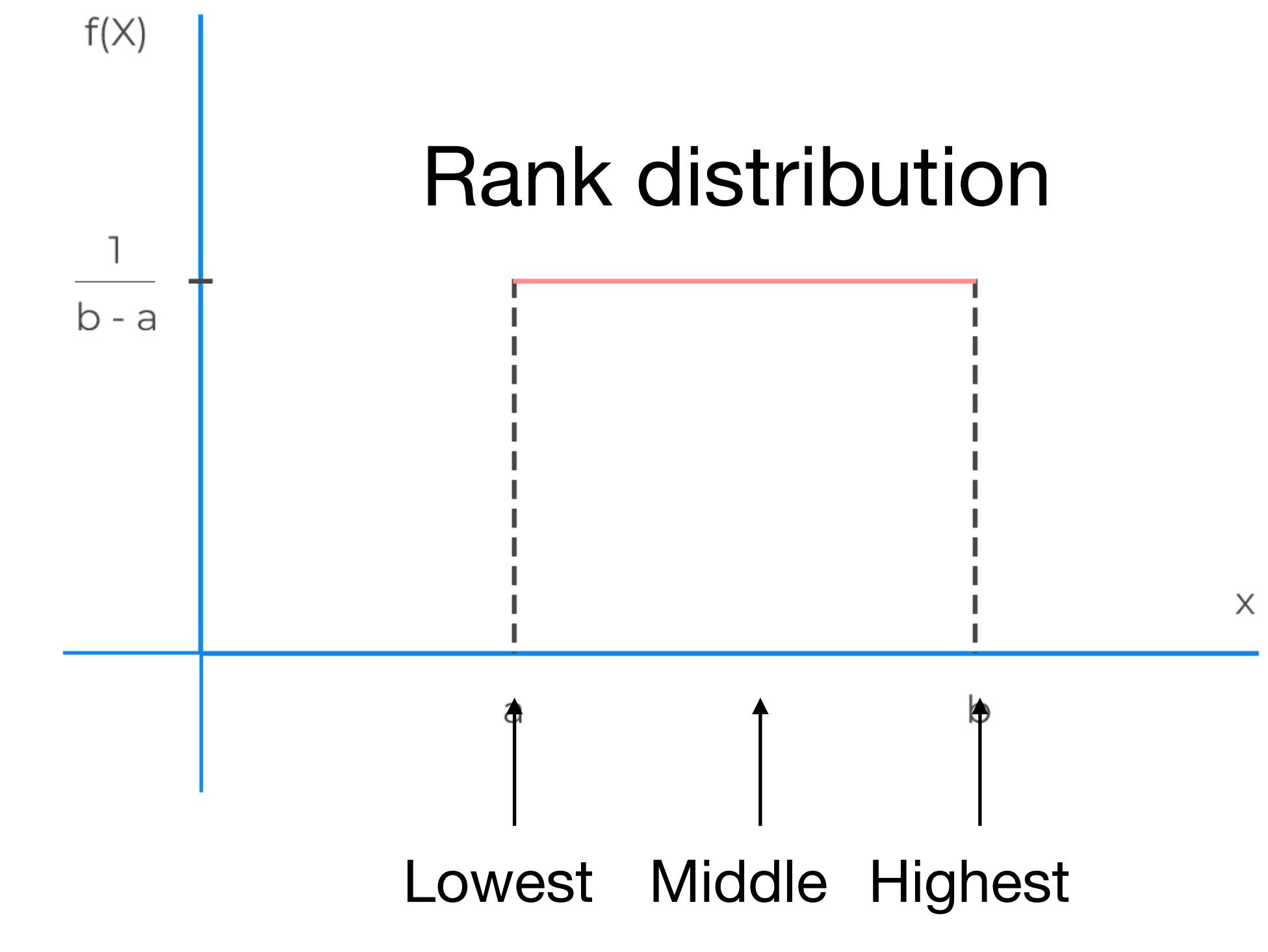
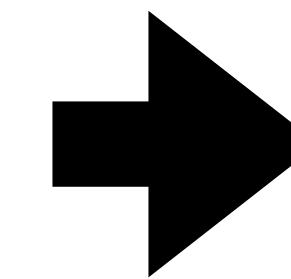
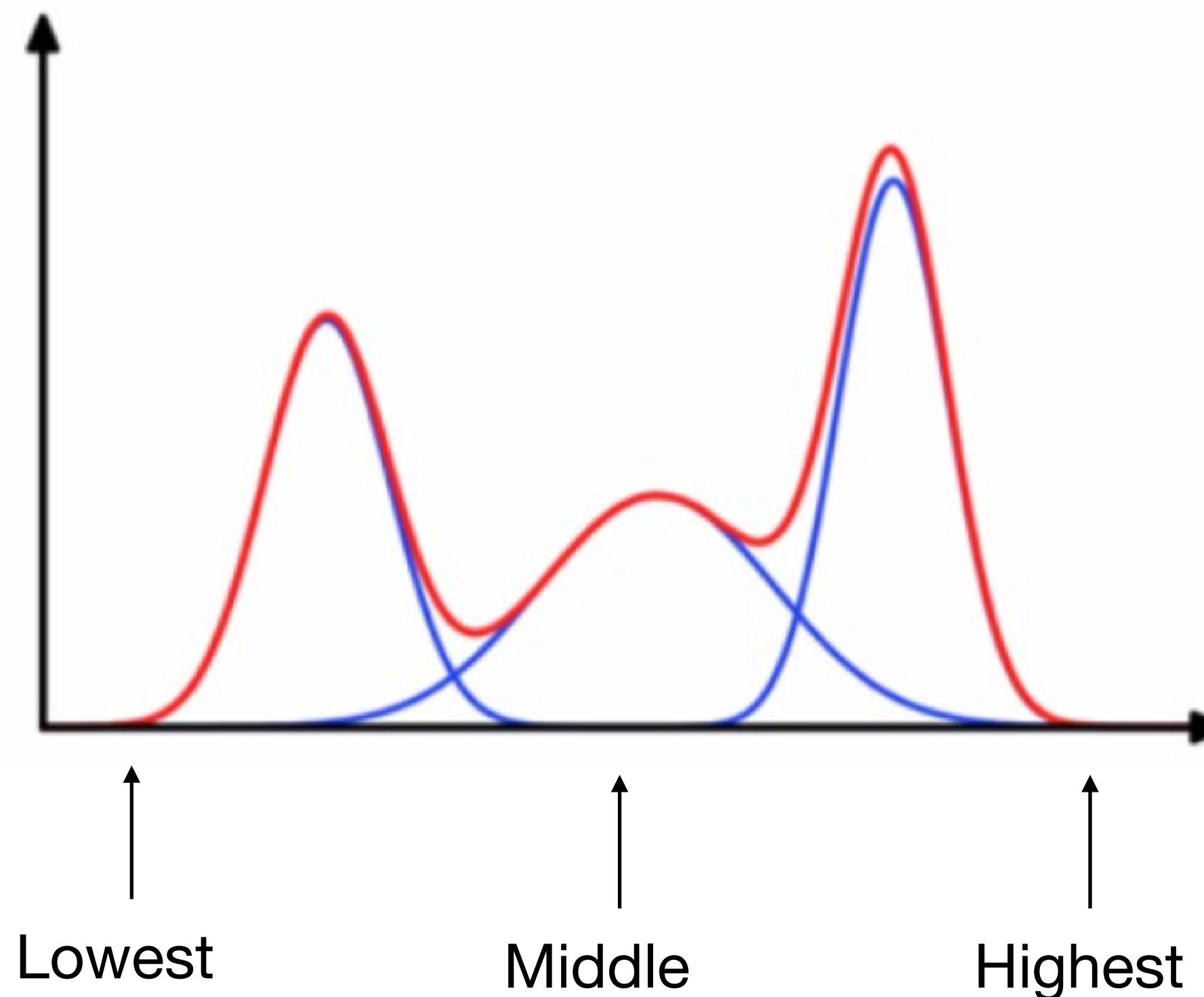


Very good fit: Weight by breed

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



Non-parametric Statistics: What if your distribution looks like this? Then rank transform!



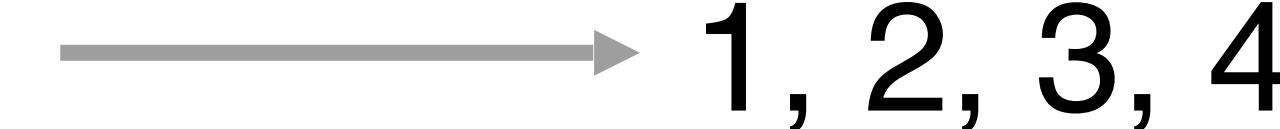
Rank Statistics

quantitative data

1, 4.5, 6.6, 9.2

ordinal data

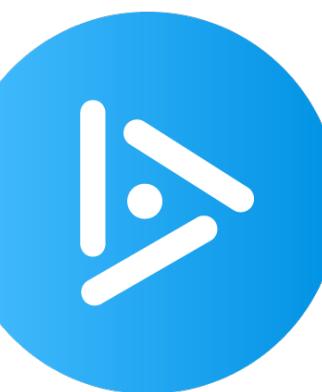
1, 2, 3, 4



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:

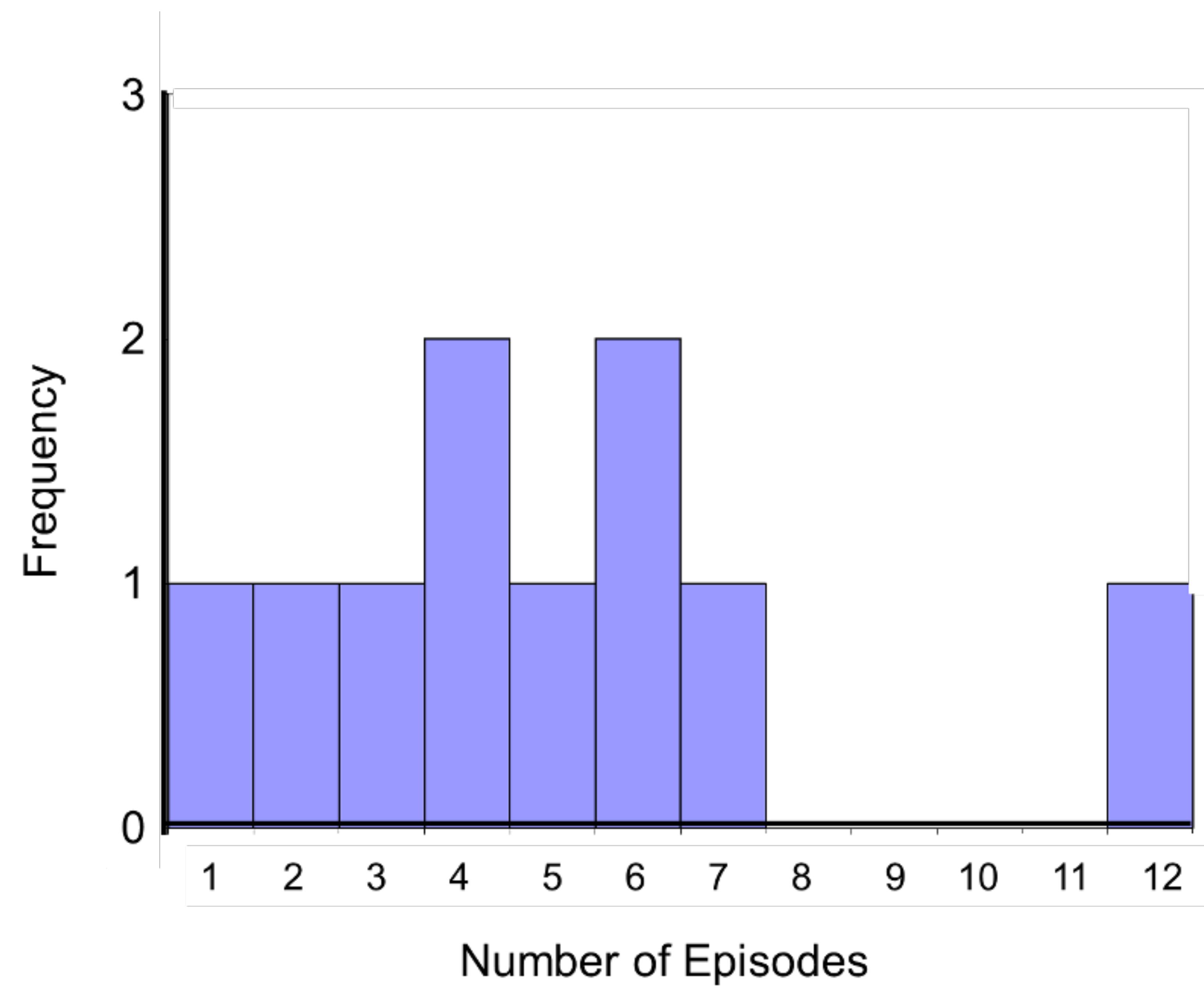
- Observations in each group are independent of one another
- Responses are ordinal

- H_0 : 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

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

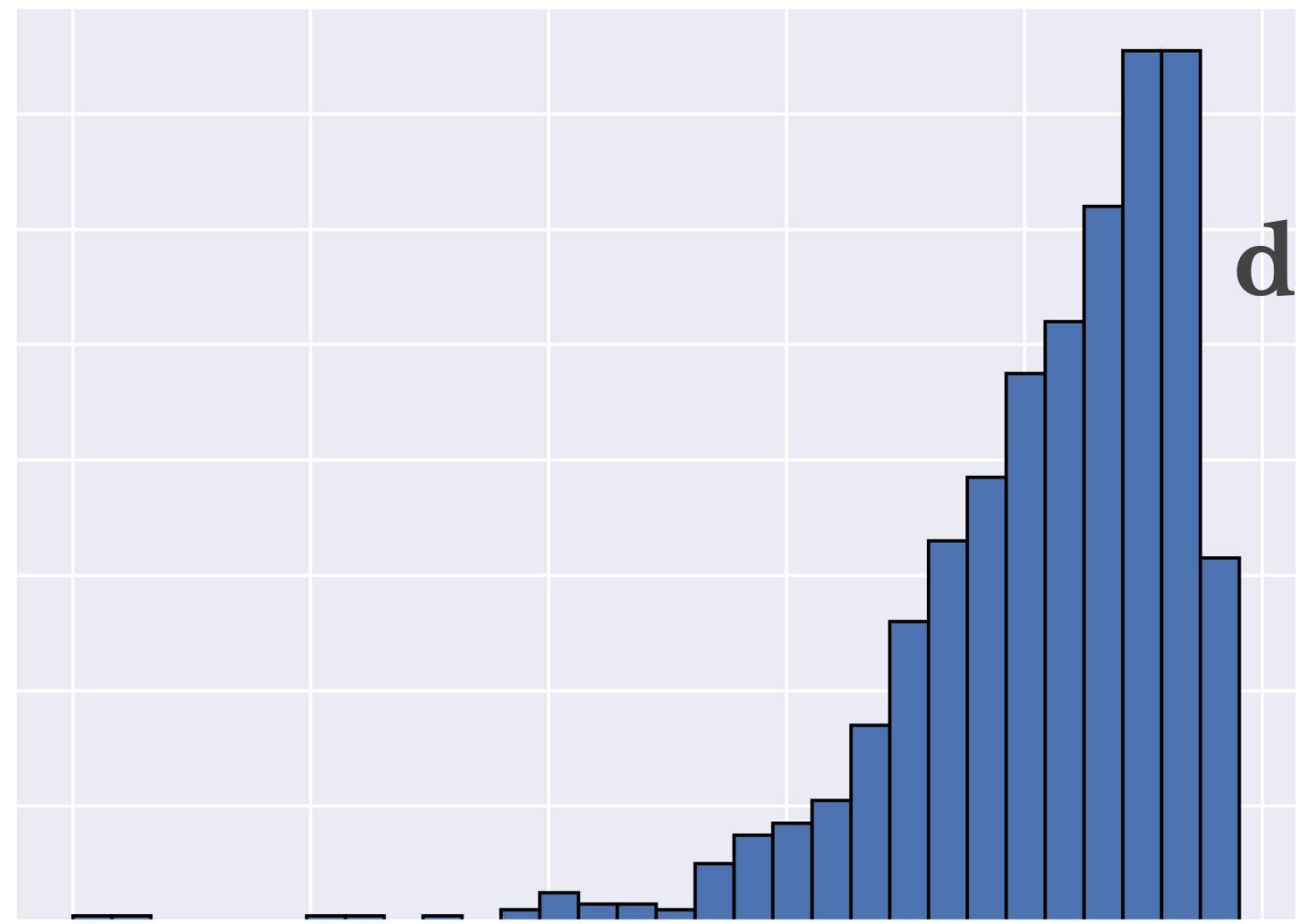
Complete separation \rightarrow no separation

We reject the null if U is small.

Anything can be transformed!

- Rank transforms are useful with any data which has ordinality!
- Why don't we use rank based statistics all the time, instead of wasting time with Student's t etc??
 - We lost statistical power by doing the transform and also from the lack of strong distributional assumptions in the test
- Other transforms can force normality on something non-normal and then we can use normal distribution assumptions and tests. Like, e.g. the log transform...

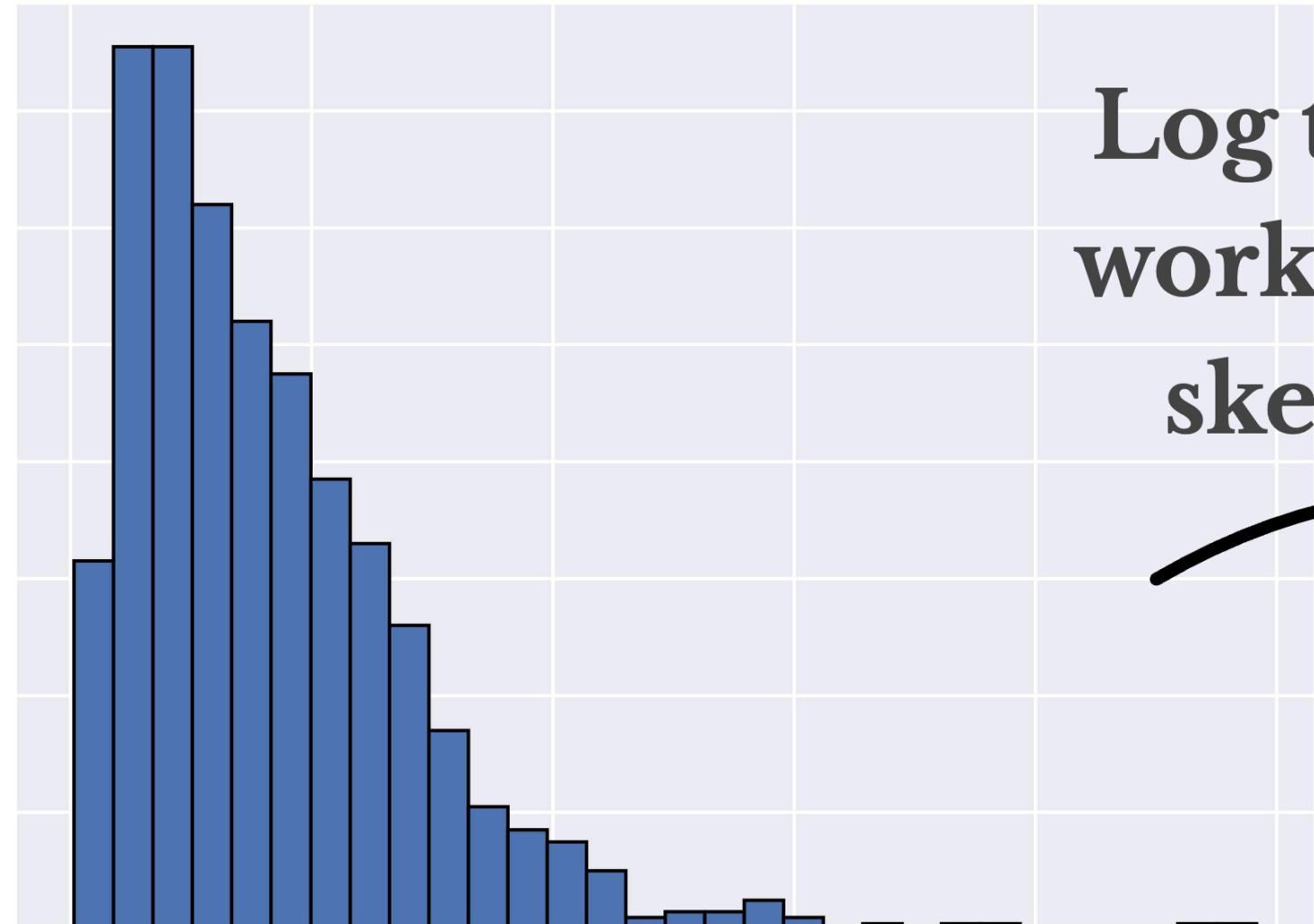
Left-skewed distribution



Log-transformed data

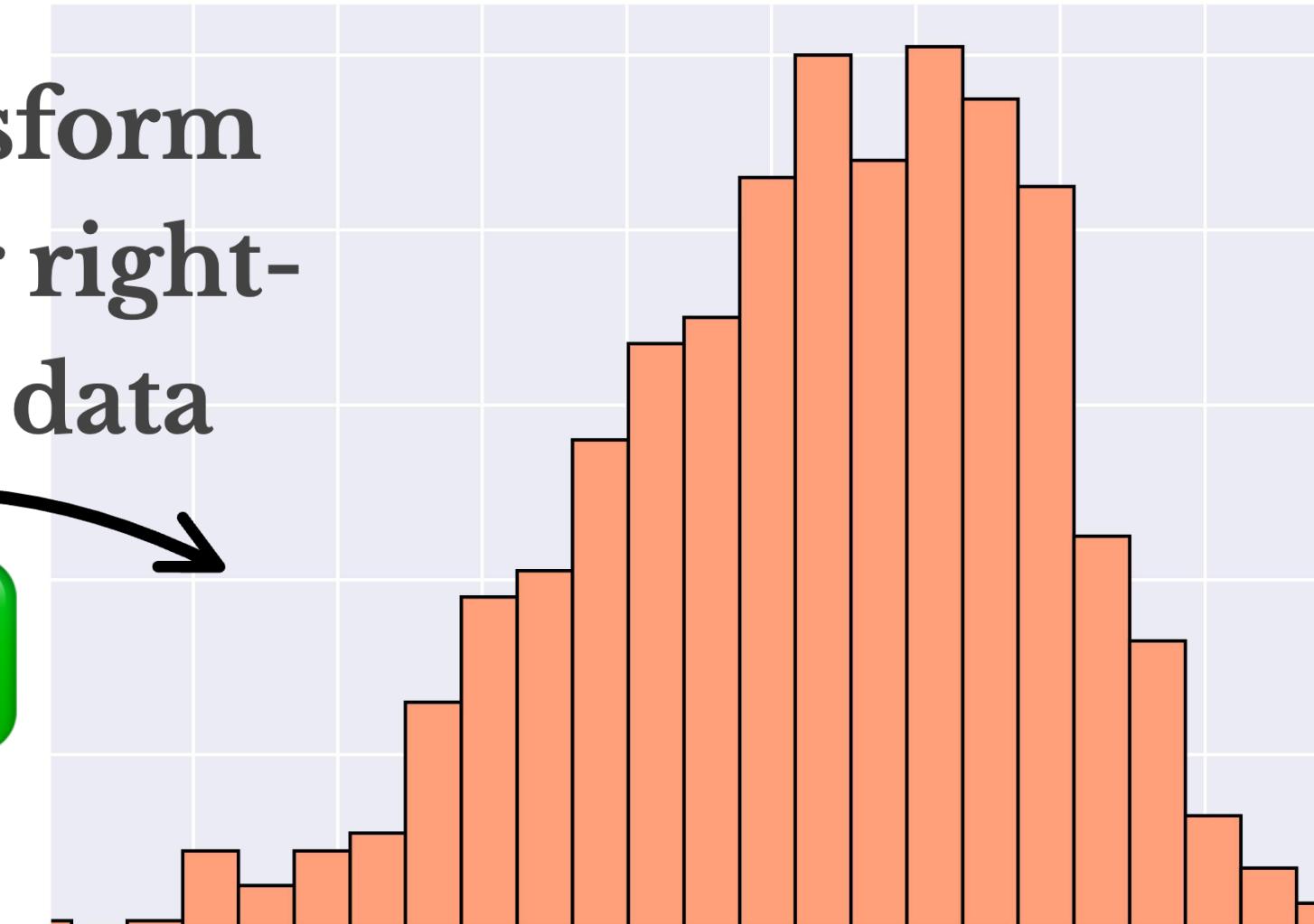


Right-skewed distribution



Log-transformed data

Log transform
works for right-
skewed data

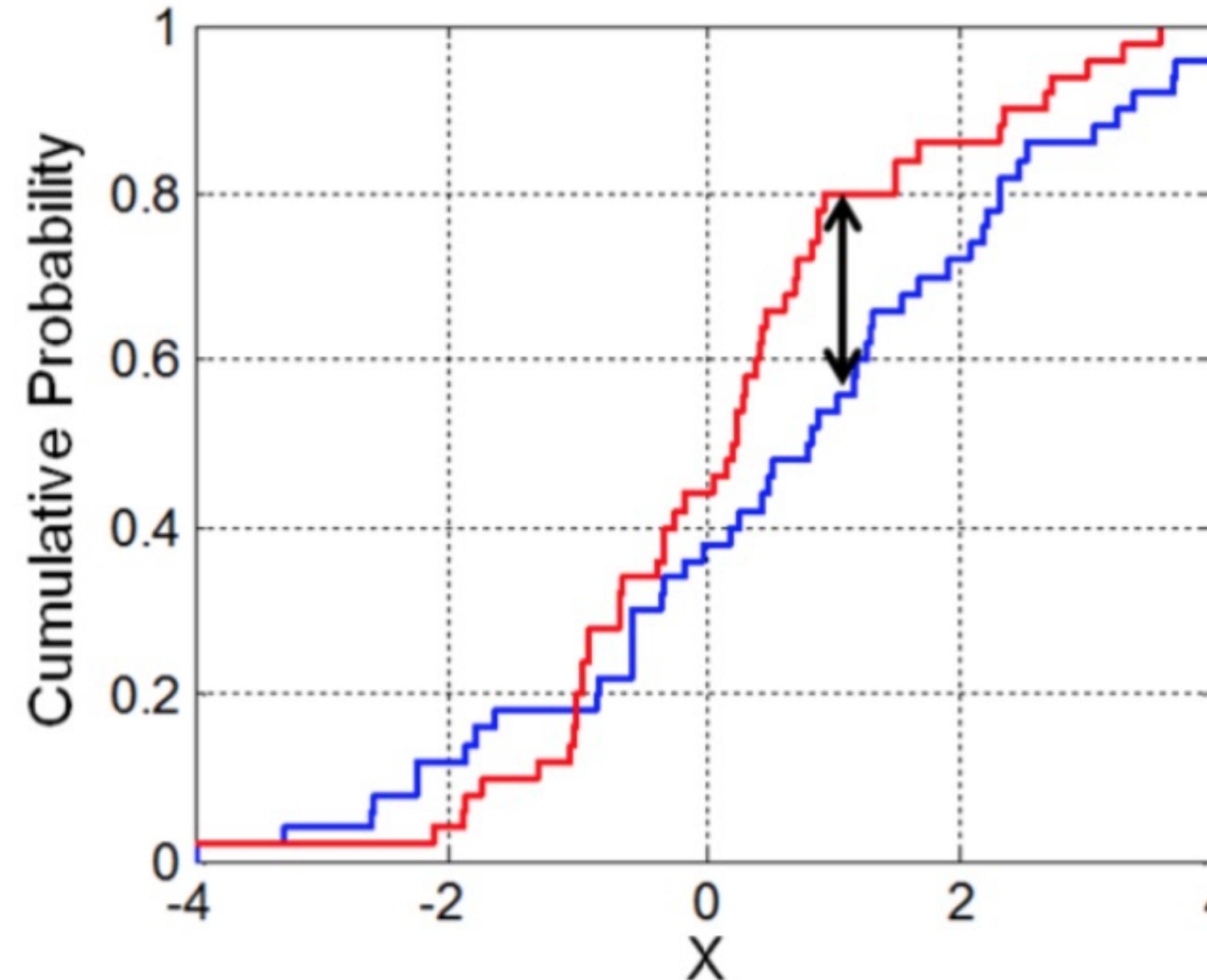


`log_data = np.log(data+1)`

Kolmogorov-Smirnov (KS) test

Comparing cumulative distributions regardless of their form

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

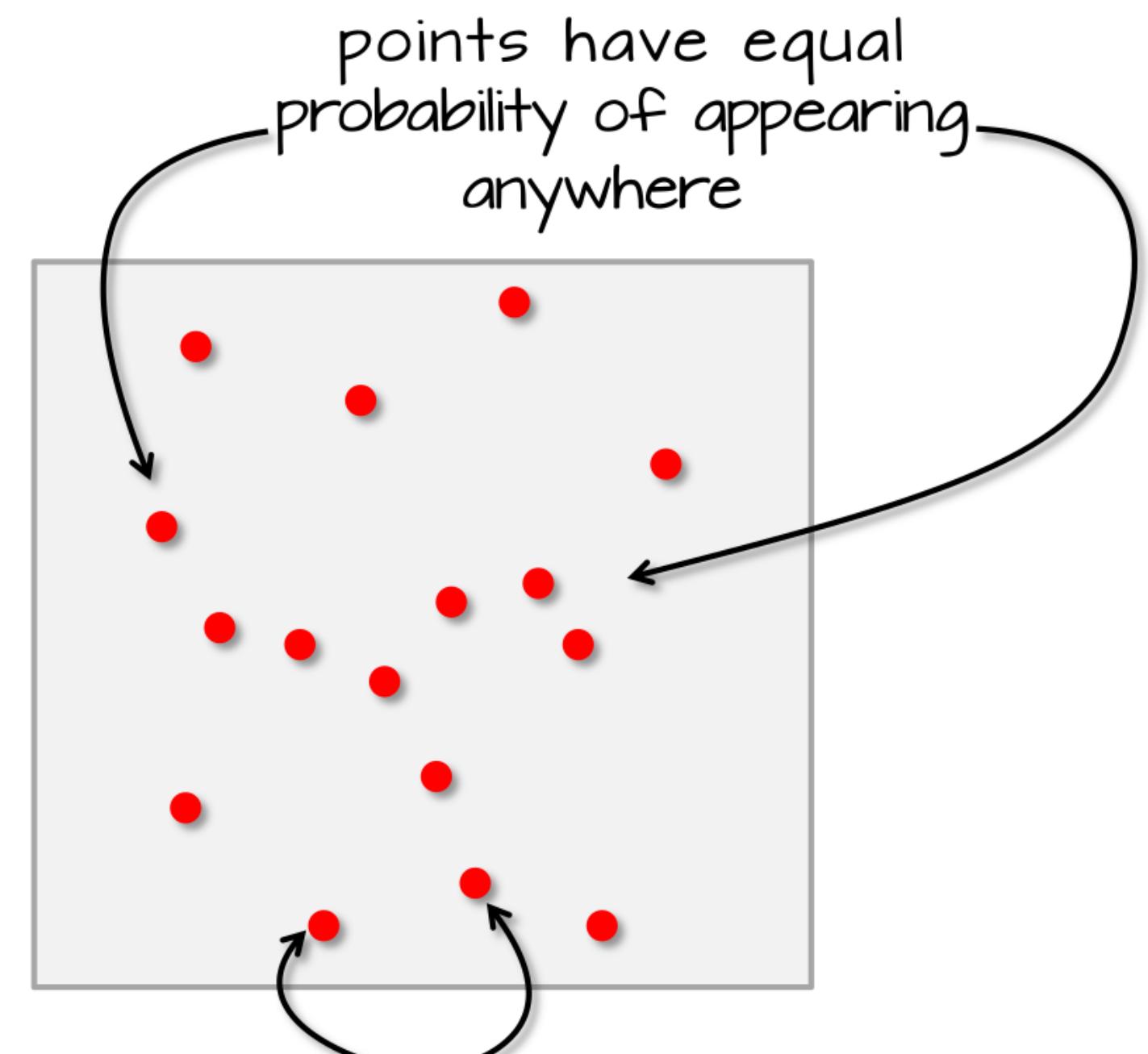
Geospatial statistical inference

What is a null distribution?

Compare observed point patterns to ones generated by an independent random process (IRP), aka complete spatial randomness (CSR).

CSR/IRP satisfy two conditions:

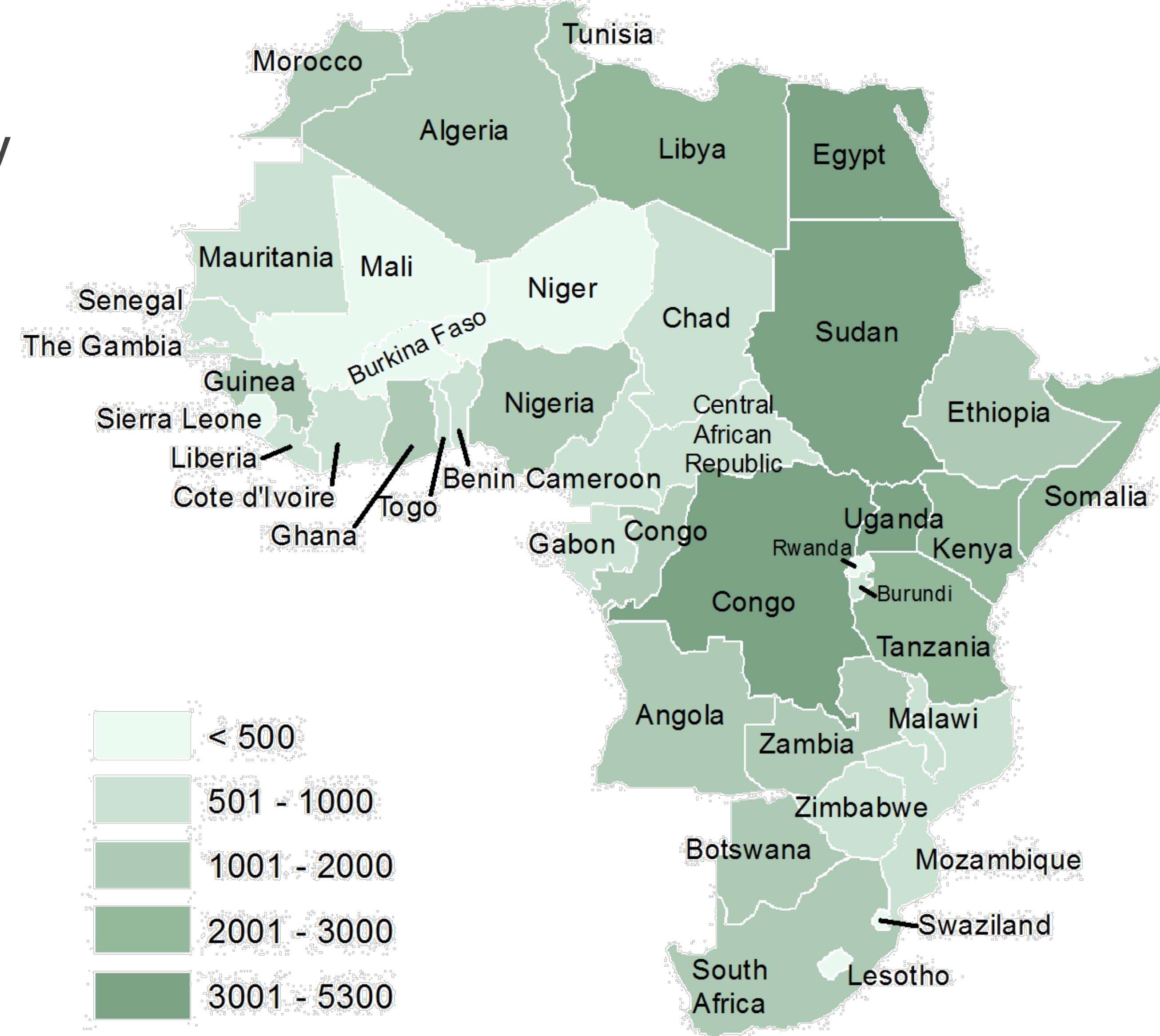
1. Any event has equal probability of being in any location, a 1st order effect.
2. The location of one event is independent of the location of another event, a 2nd order effect



points don't influence each other's location

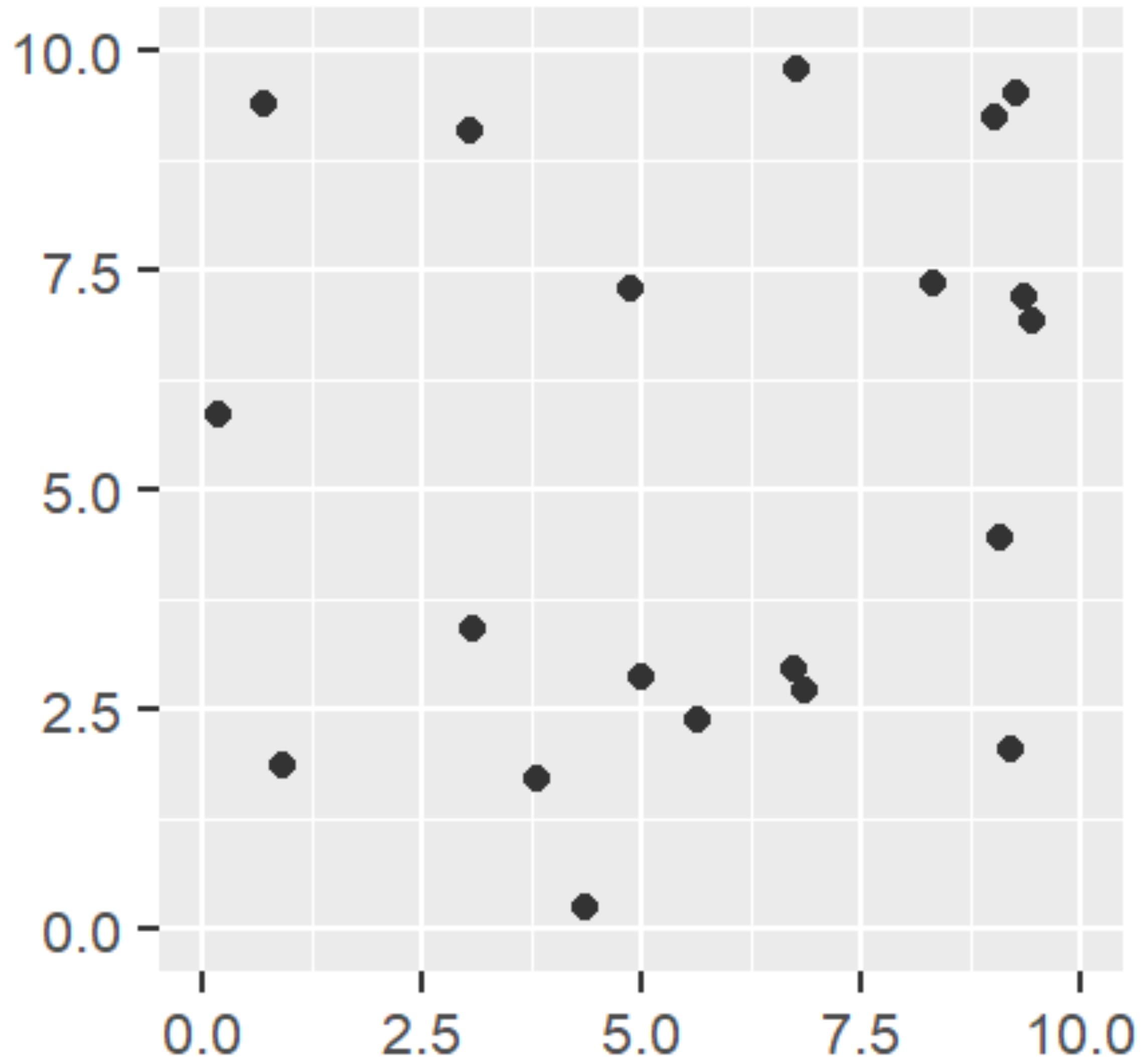
**What metrics are we doing inference
on in geospatial applications?**

Are countries with a high conflict index score geographically clustered?



Global Point Density

the ratio of observed
number of points to the
study region's surface area



Quadrat Density (local)

Surface is divided and then point density is calculated within quadrat

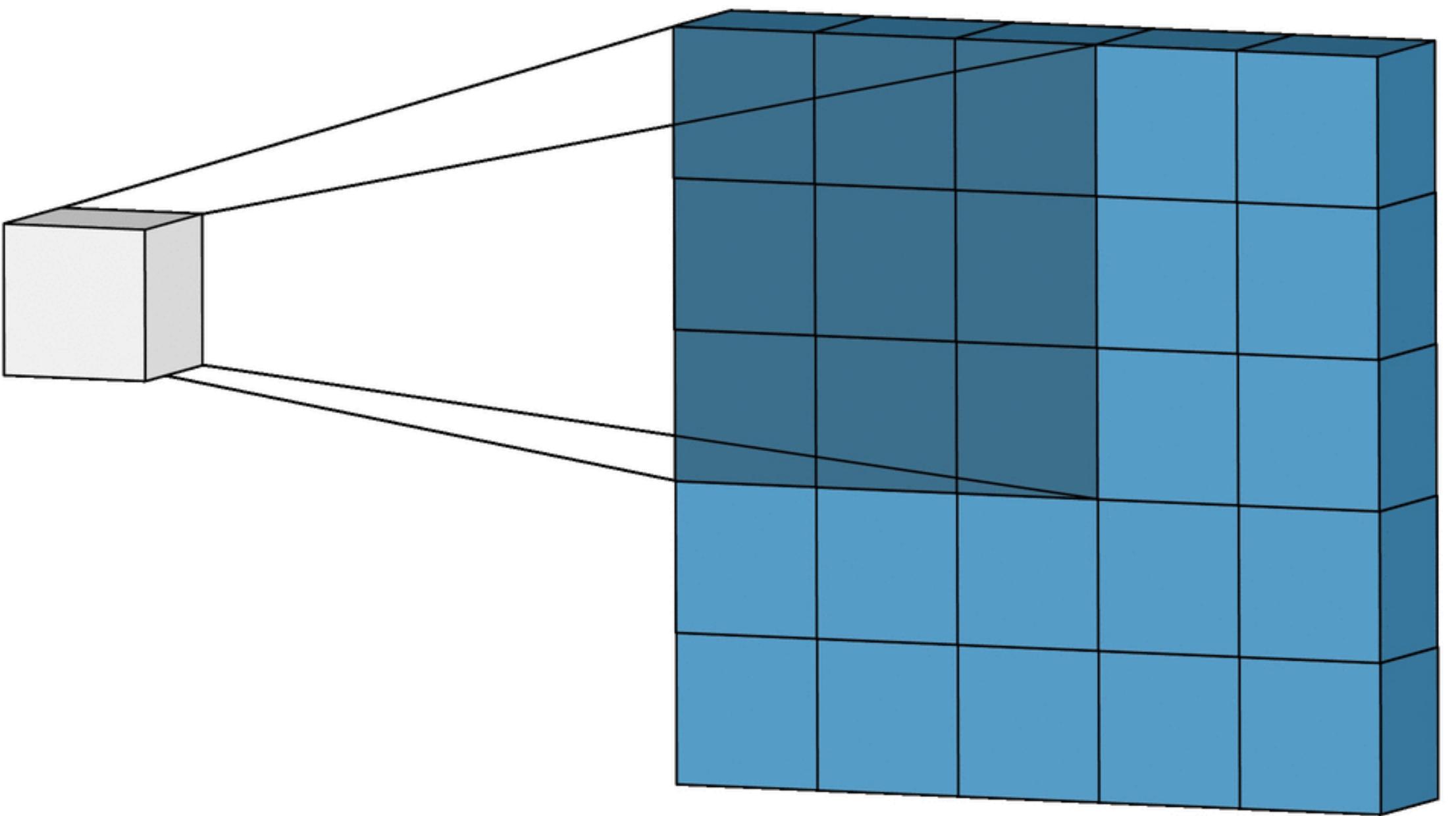
Note: quadrat number and shape will affect measurement estimate. Suffers from MAUP.



Kernel Density (local)

Point density is calculated within sliding windows (window size = kernel)

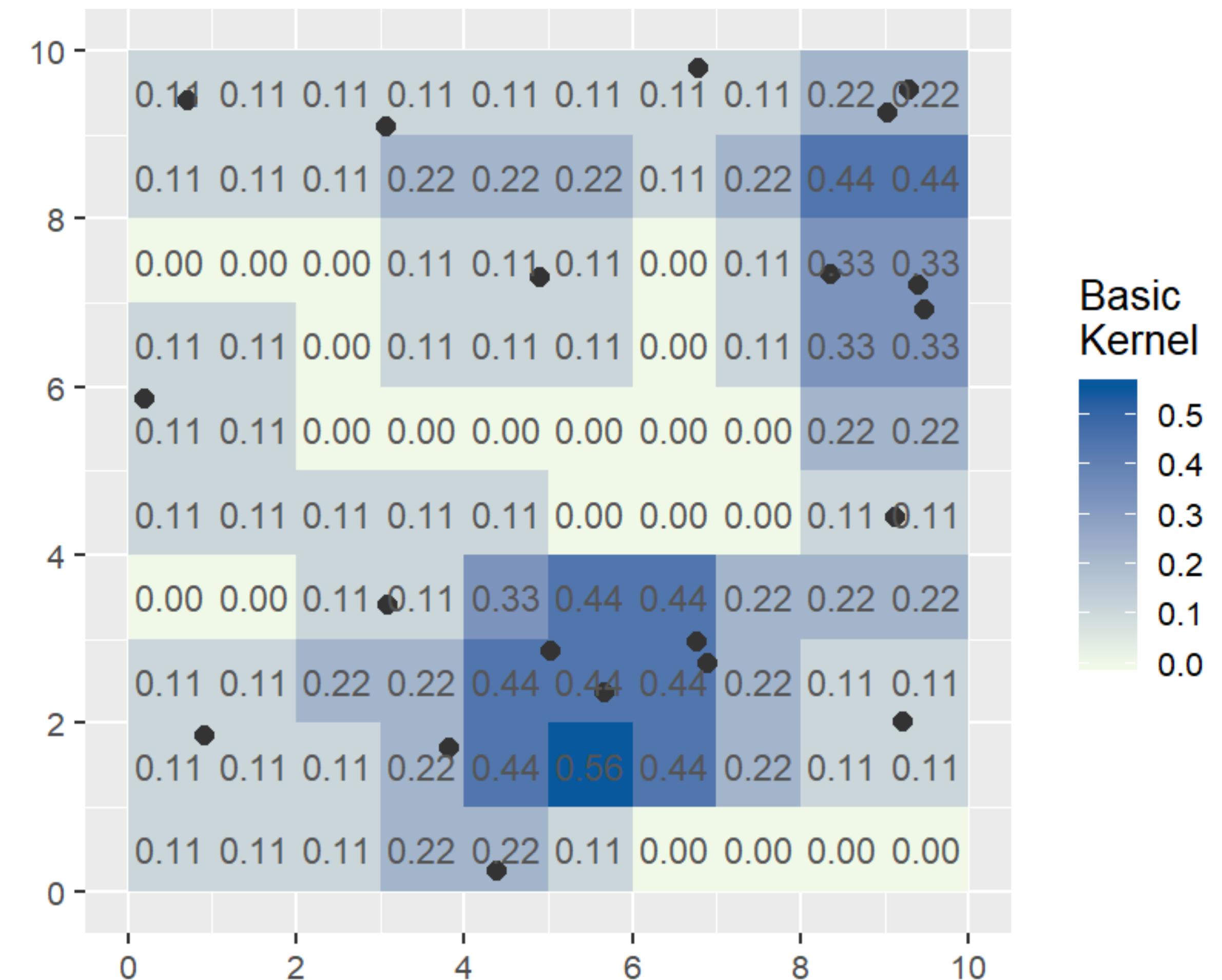
Note: kernel will affect measurement estimate, but this is less susceptible to MAUP.



Kernel Density (local)

Point density is calculated within sliding windows (window size = kernel)

Note: kernel will affect measurement estimate, but this is less susceptible to MAUP.



Modeling these data: Poisson Point Process

(Density-based Methods - - how the points are distributed relative to the study space)

$$\lambda(i) = e^{\alpha + \beta Z(i)}$$

$\lambda(i)$ is the modeled intensity at location i

e^α is the base intensity when the covariate is zero

e^β is the multiplier by which the intensity increases (or decreases) for each 1 unit increase in the covariate

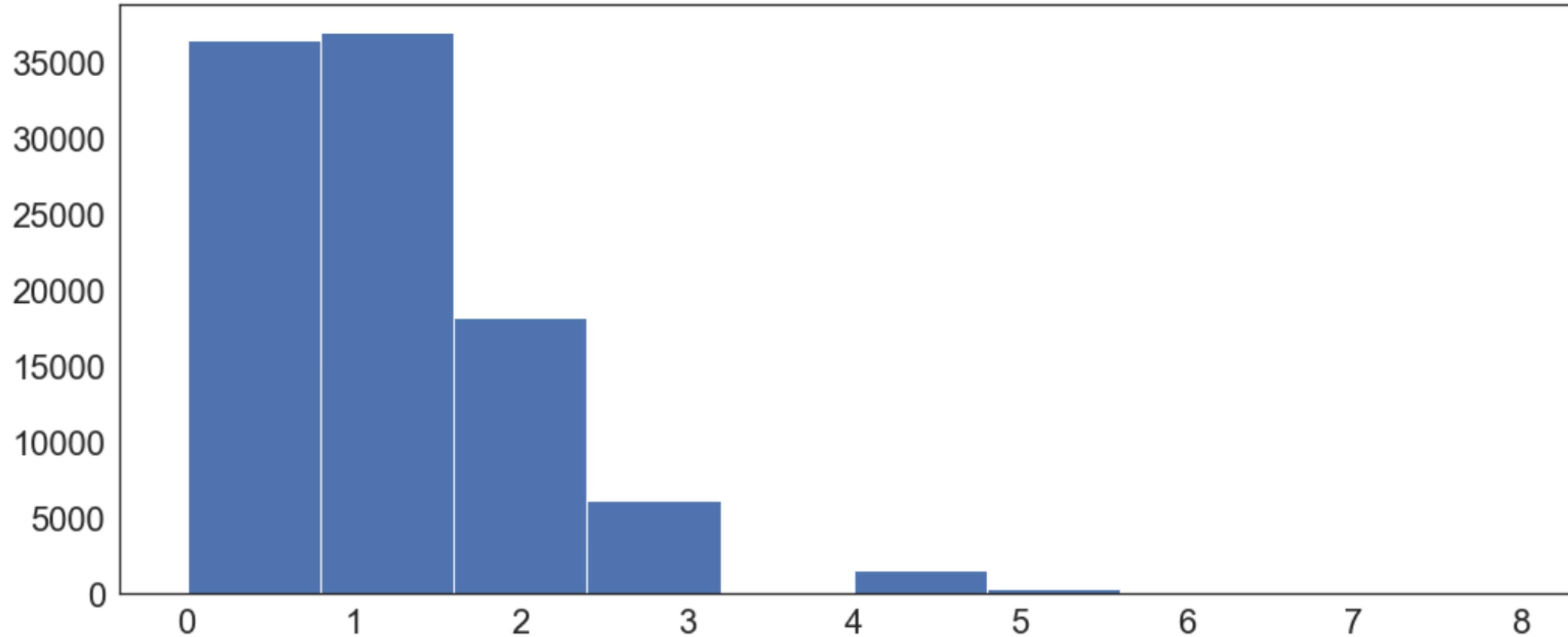
Poisson Distribution

The Poisson Distribution models events in fixed intervals of time, given a known average rate (and independent occurrences).

In [55]:

Slide Type Fragment ▾

```
dat = poisson.rvs(mu=1, size=100000)  
plt.hist(dat);
```

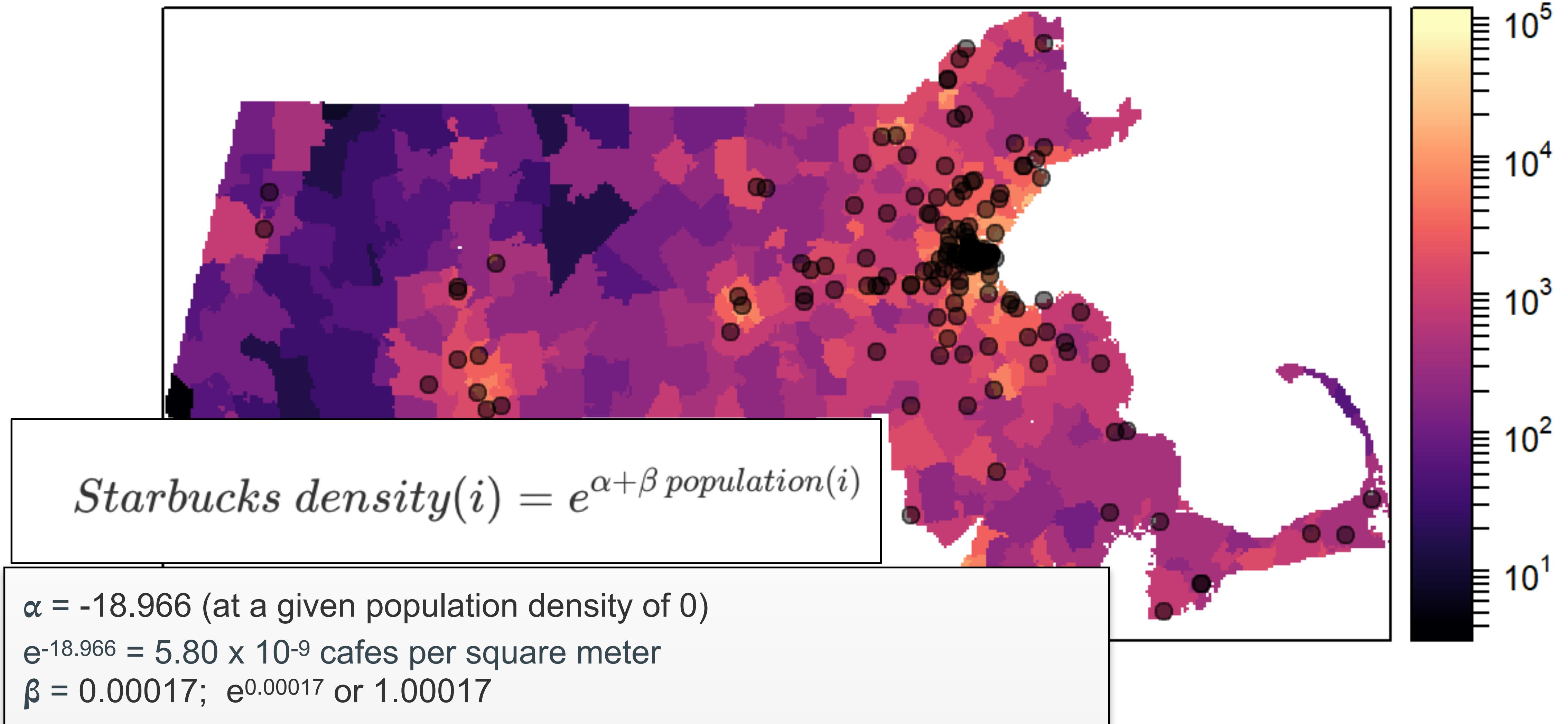


Slide Type Fragment ▾

The **number of visitors a fast food drive-through gets each minute** follows a Poisson distribution. In this case, maybe the average is 3, but there's some variability around that number.

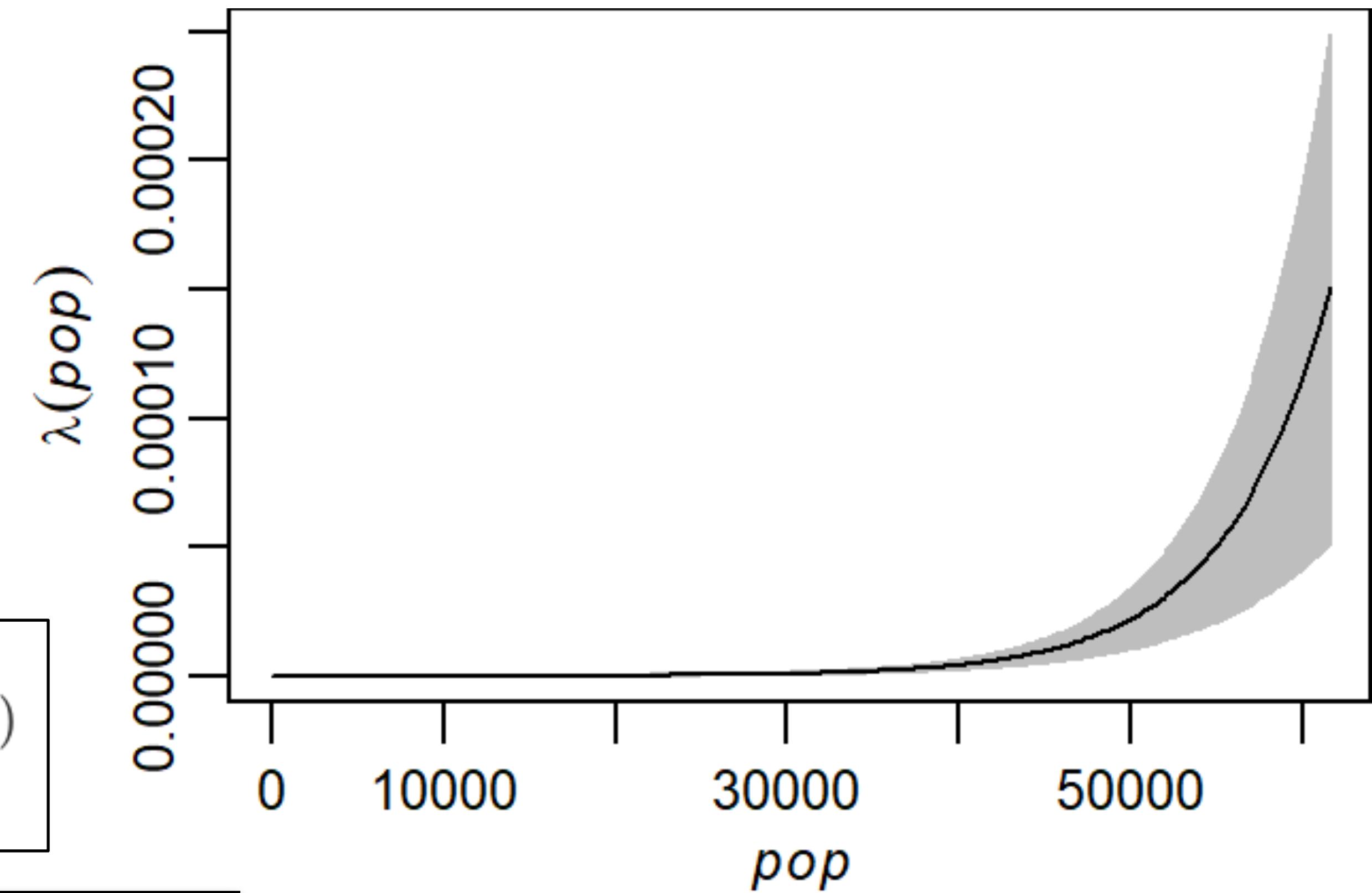
A Poisson distribution can help calculate the probability of various events related to customers going through the drive-through at a restaurant. It will predict lulls (0 customers) and flurry of activity (5+ customers), allowing staff to plan and schedule more precisely.

Location of Starbucks relative to population density in MA



Location of Starbucks relative to population density in MA

$$\text{Starbucks density}(i) = e^{\alpha + \beta \text{population}(i)}$$

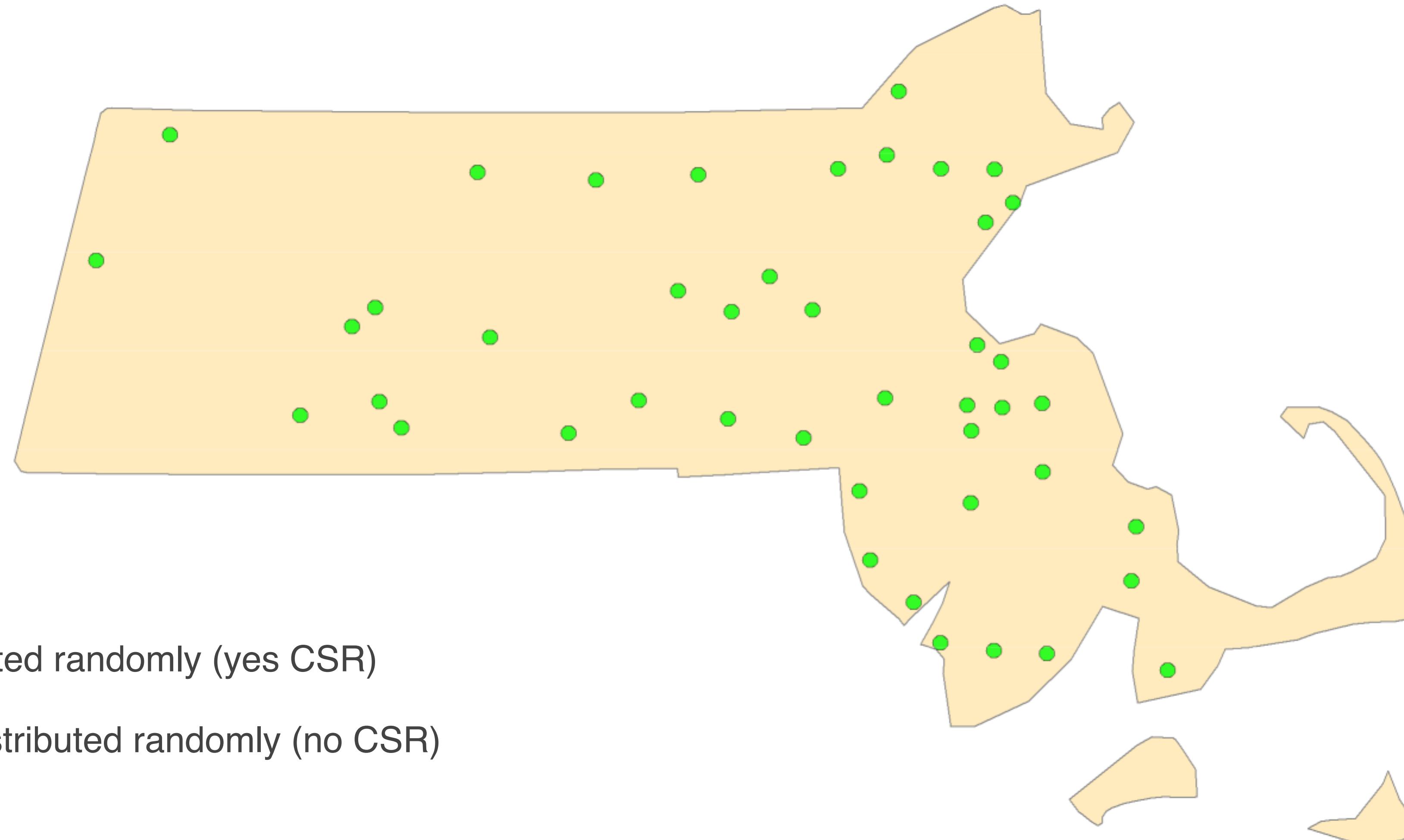


$\alpha = -18.966$ (at a given population density of 0)

$e^{-18.966} = 5.80 \times 10^{-9}$ cafes per square meter

$\beta = 0.00017$; $e^{0.00017}$ or 1.00017

Is this distribution of Walmarts in MA the result of CSR?



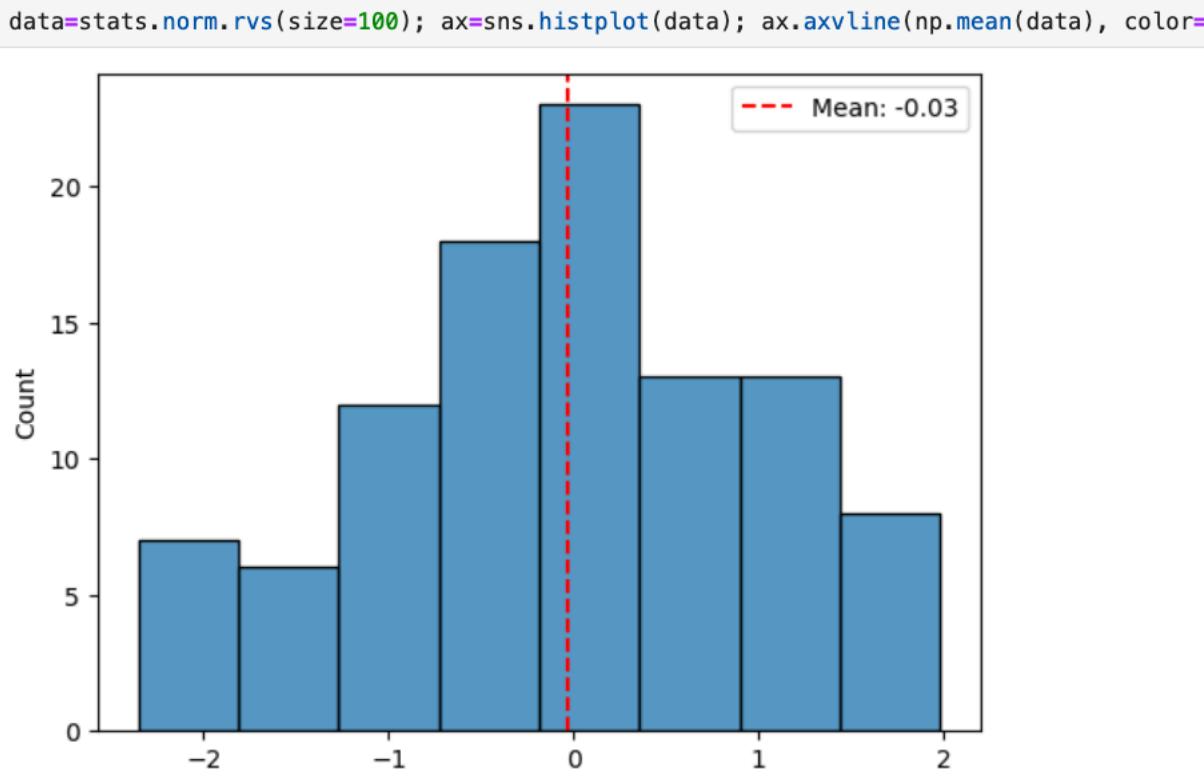
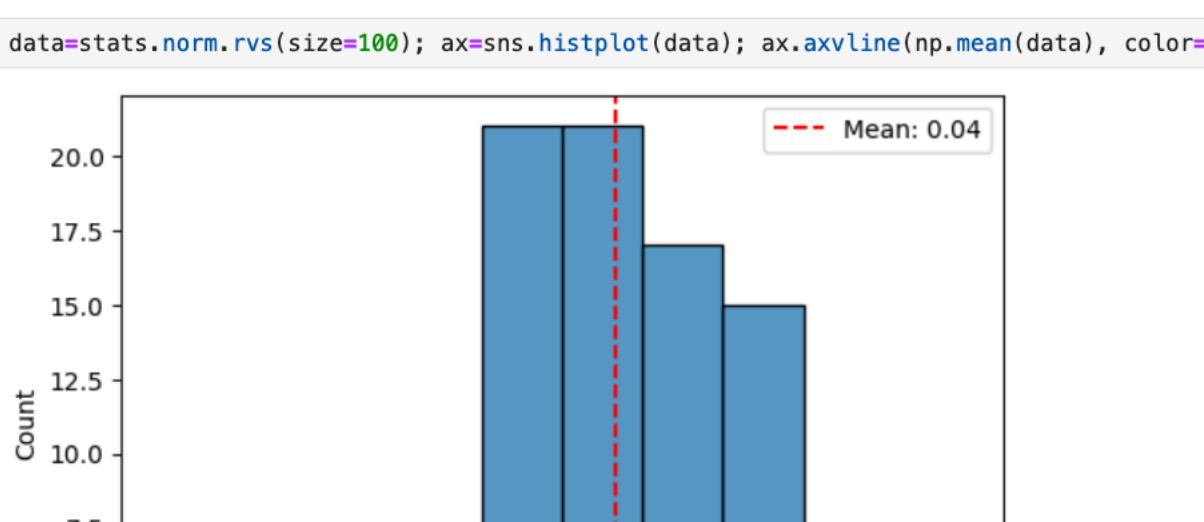
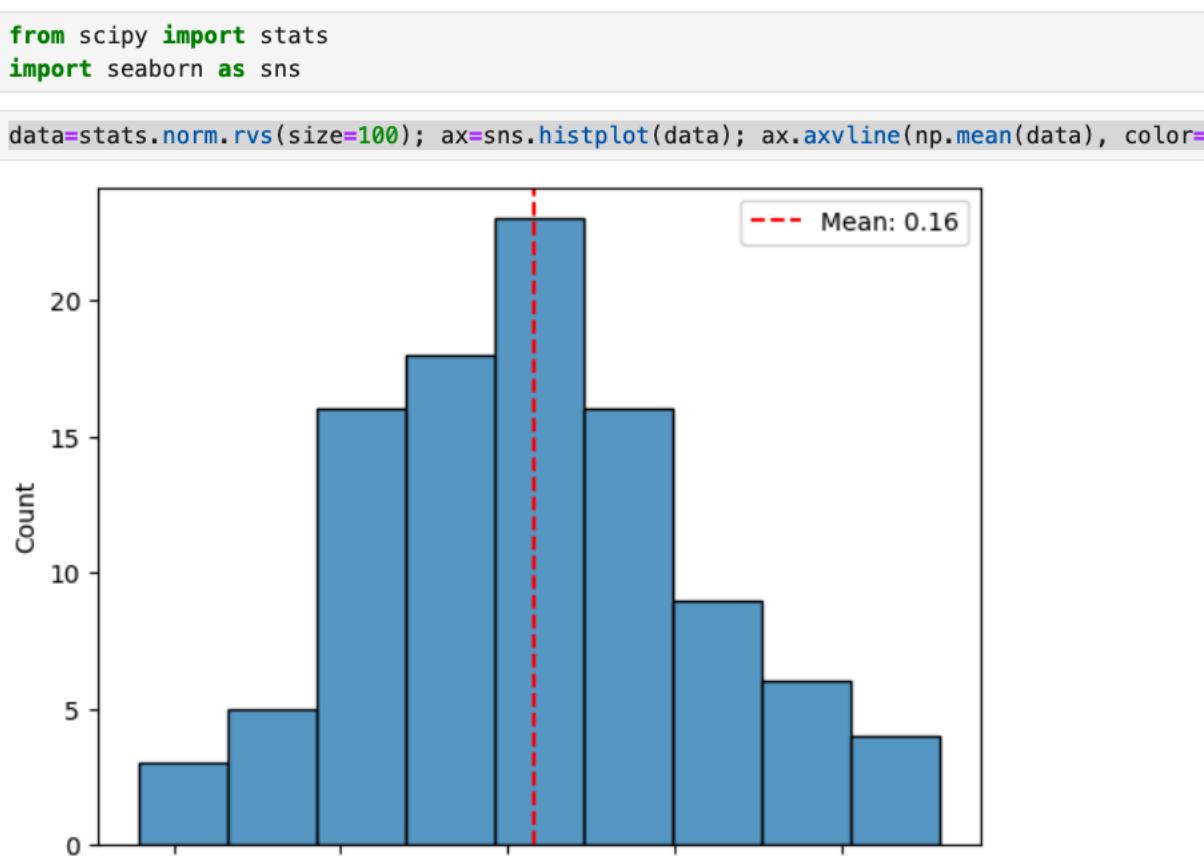
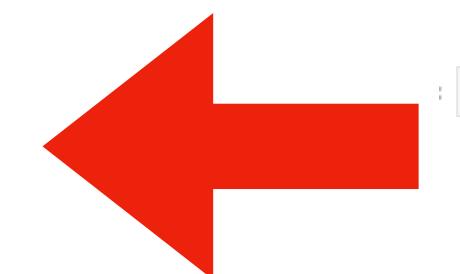
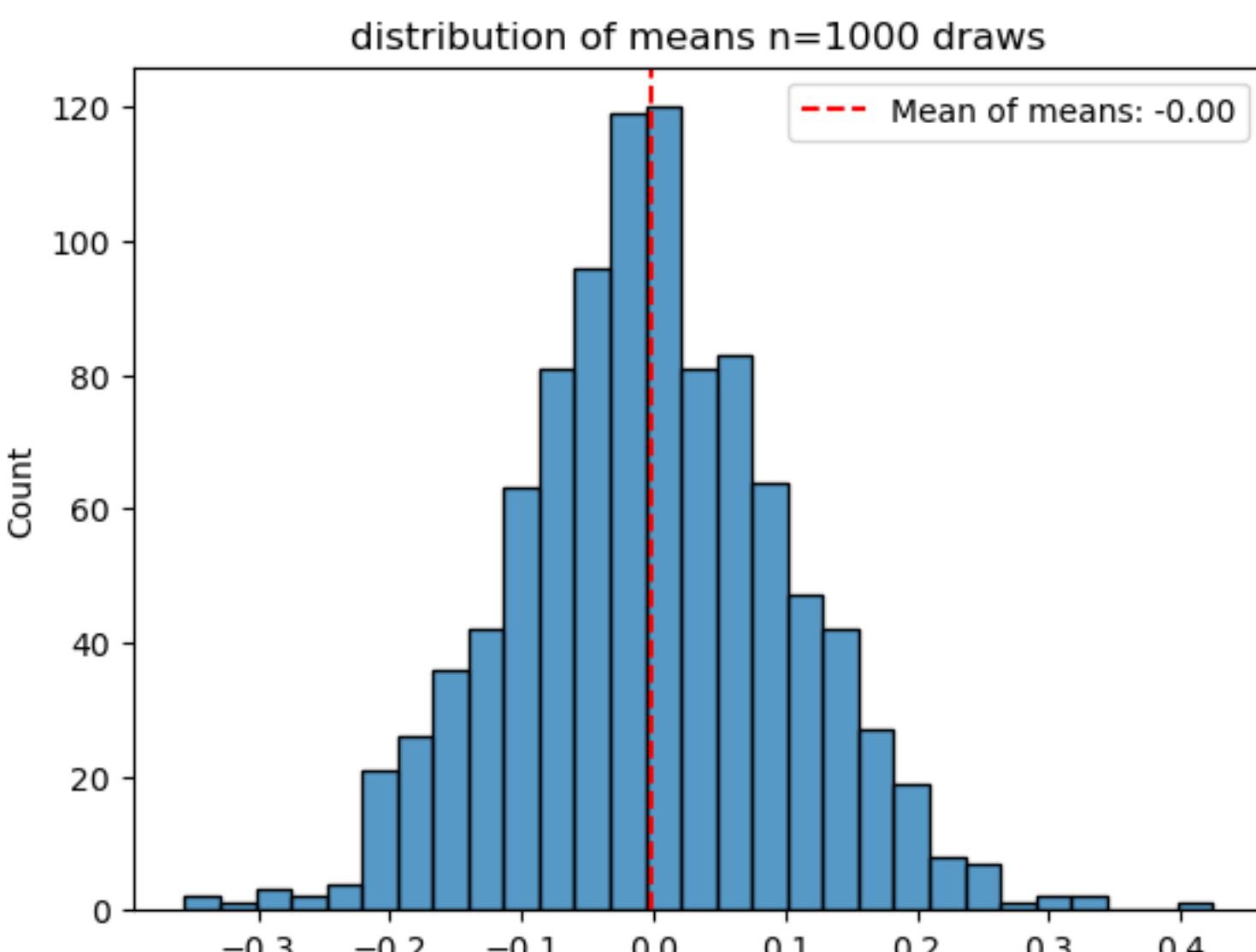
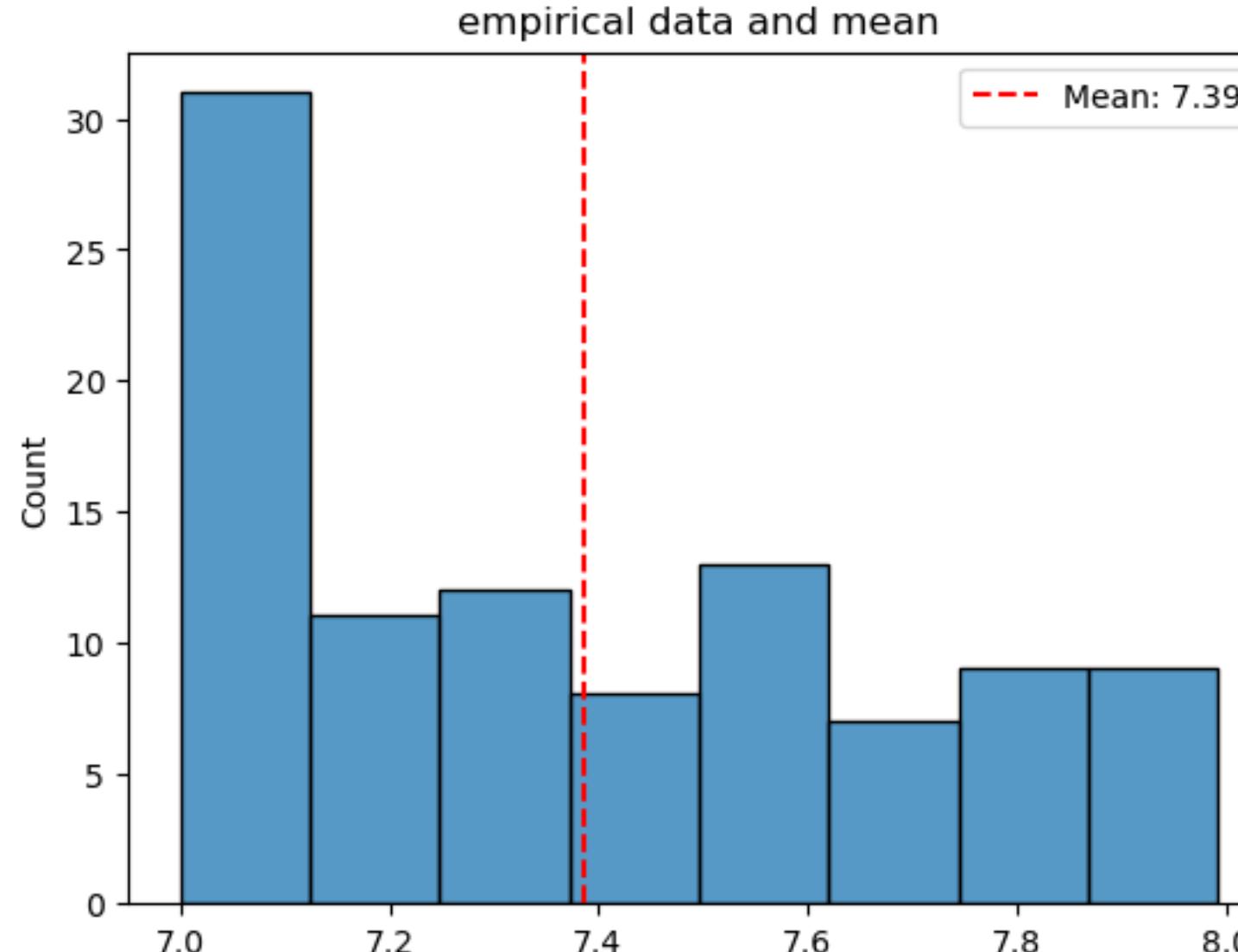
H_0 : Distributed randomly (yes CSR)

H_a : NOT distributed randomly (no CSR)

Monte Carlo simulation

Making an empirical null distribution

1. First, we postulate a null hypothesis H_0
2. Next, we simulate many draws from of our postulated process and compute a statistic on each one
3. Finally, we calculate the p-value of our real observed data in comparison to the simulated null distribution



MONTE CARLO simulations

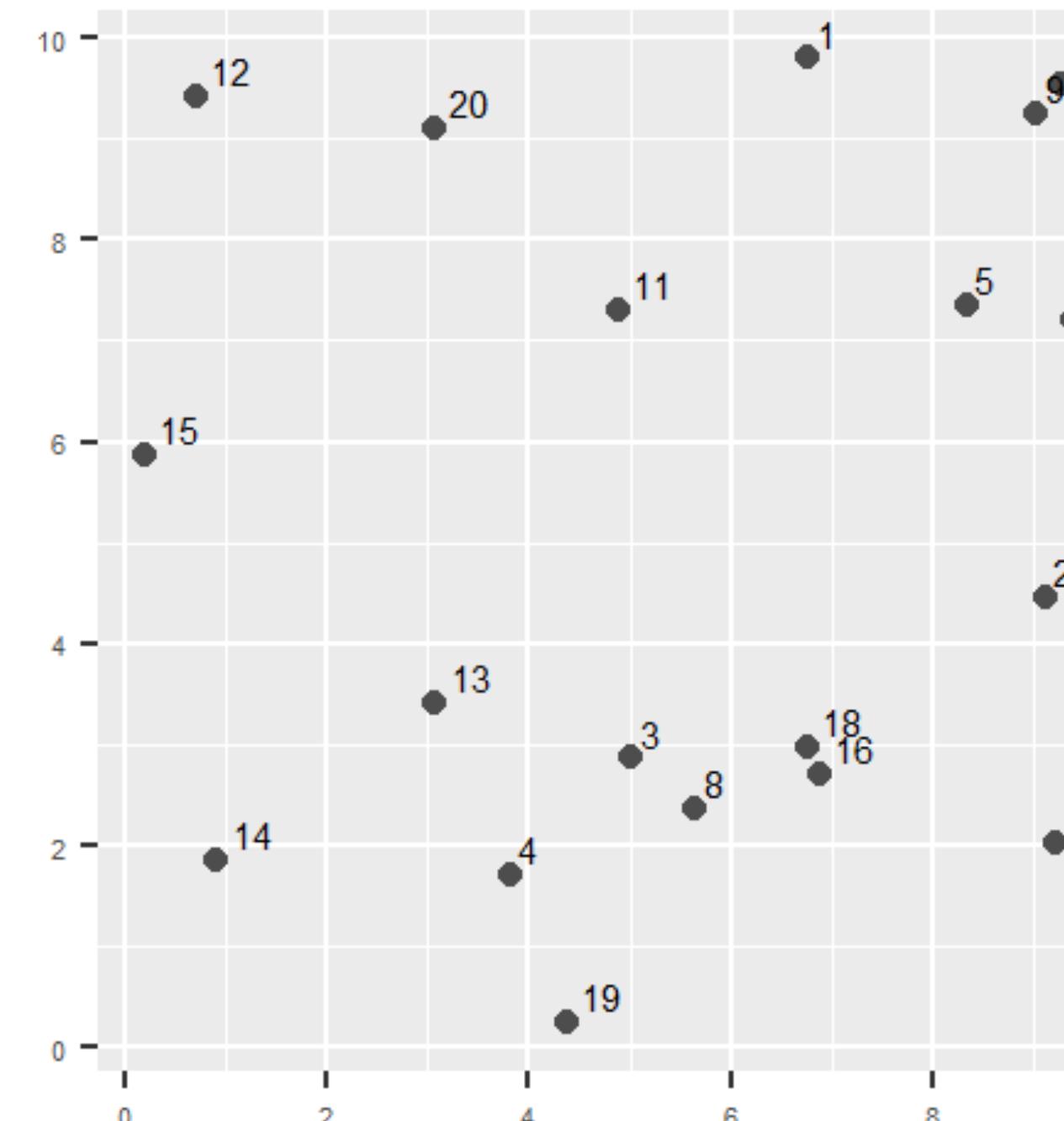
1. First, we postulate a process—our null hypothesis, H_0 . For example, we hypothesize that the distribution of Walmart stores is consistent with a completely random process (CSR).
2. Next, we simulate many realizations of our postulated process and compute a statistic (e.g., mean distance to nearest neighbor) for each realization.
3. Finally, we compare our observed data to the patterns generated by our simulated processes and assess (via a measure of probability) if our pattern is a likely realization of the hypothesized process.



<https://mgimond.github.io/Spatial/introGIS.html>

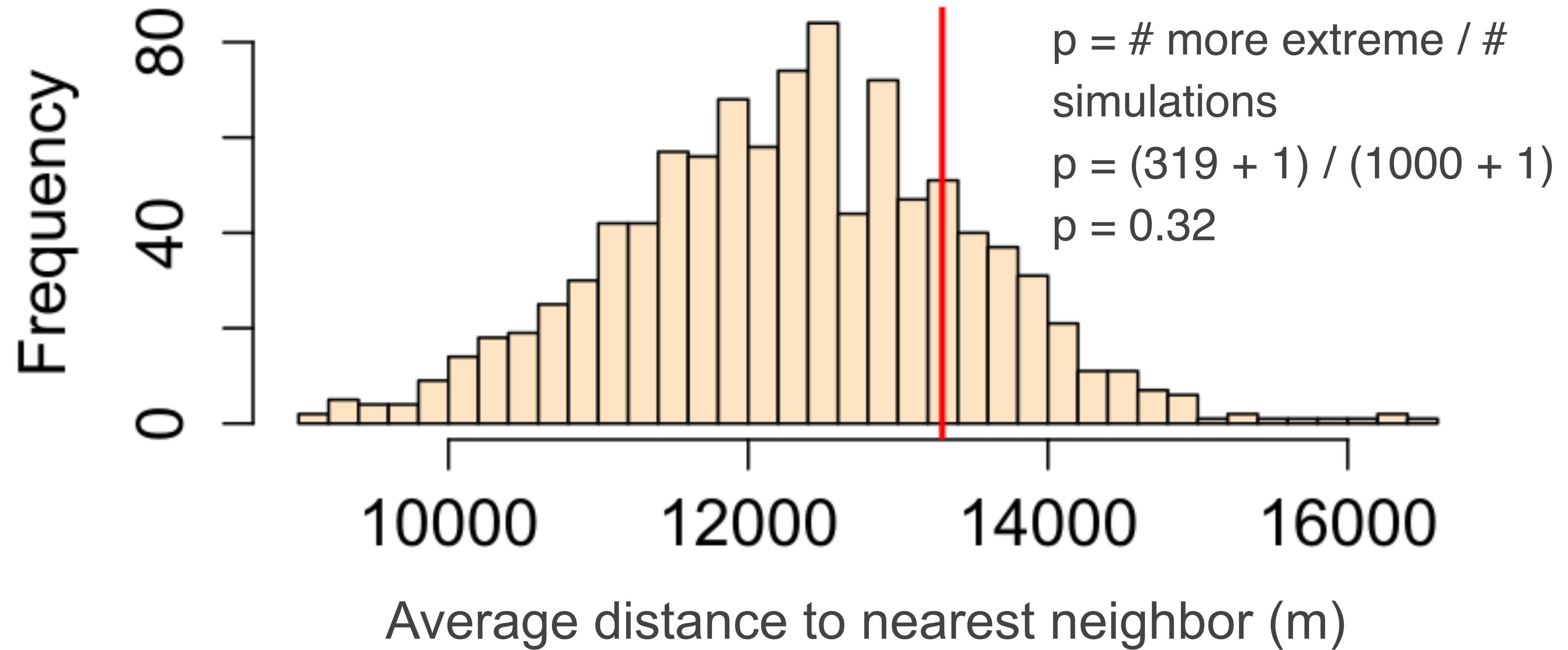
Modeling these data: Average Nearest Neighbor

(Distance-based Methods - how the points are distributed relative to one another)



From	To	Distance	From	To	Distance
1	9	2.32	11	20	2.55
2	10	2.43	12	20	2.39
3	8	0.81	13	4	1.85
4	19	1.56	14	13	2.67
5	6	1.05	15	12	3.58
6	7	0.3	16	18	0.29
7	6	0.3	17	9	0.37
8	3	0.81	18	16	0.29
9	17	0.37	19	4	1.56
10	2	2.43	20	12	2.39

$$\text{ANN} = 1.52 \text{ units}$$



What does the histogram tell us?



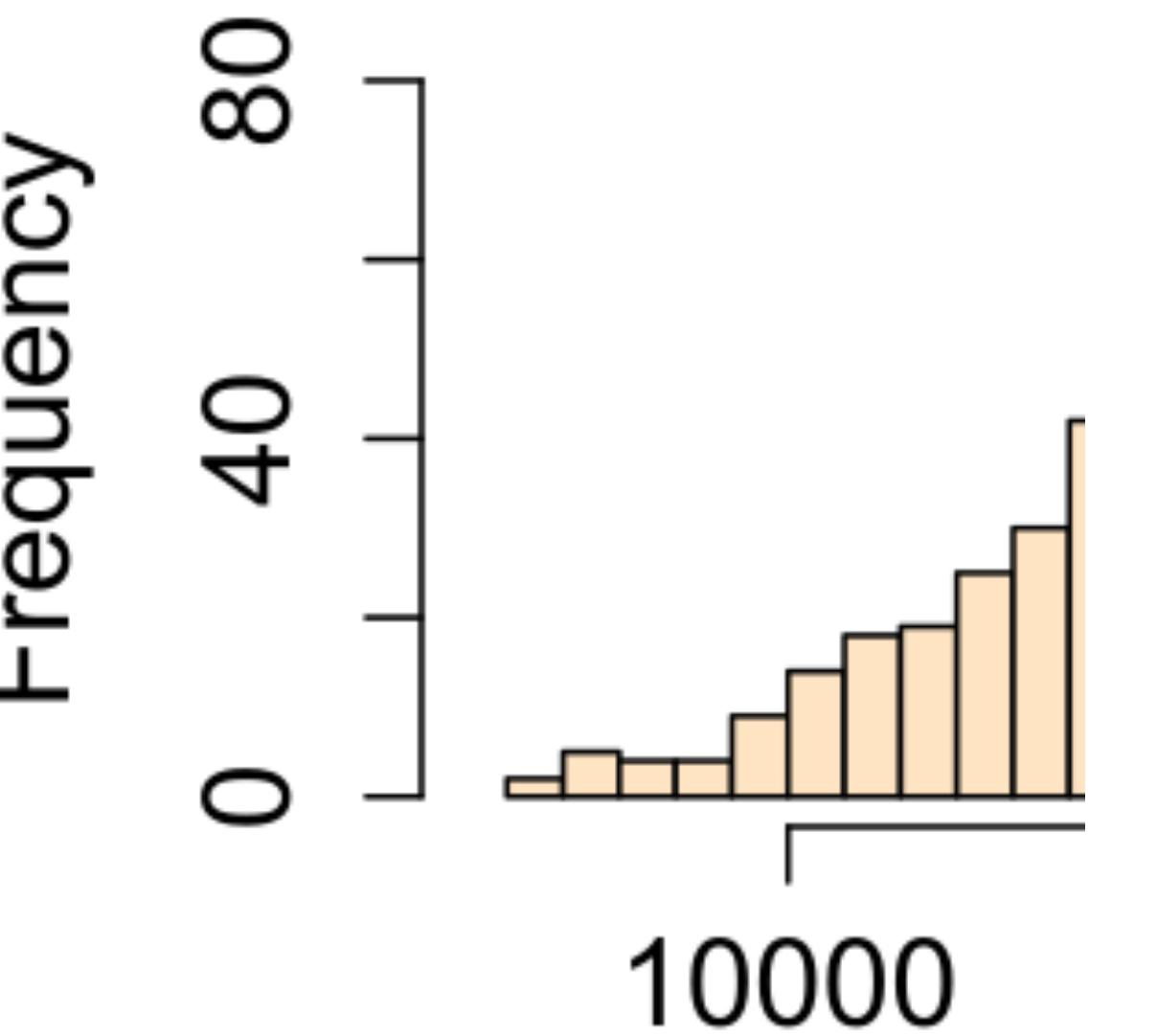
A

Alternative
Hypothesis



B

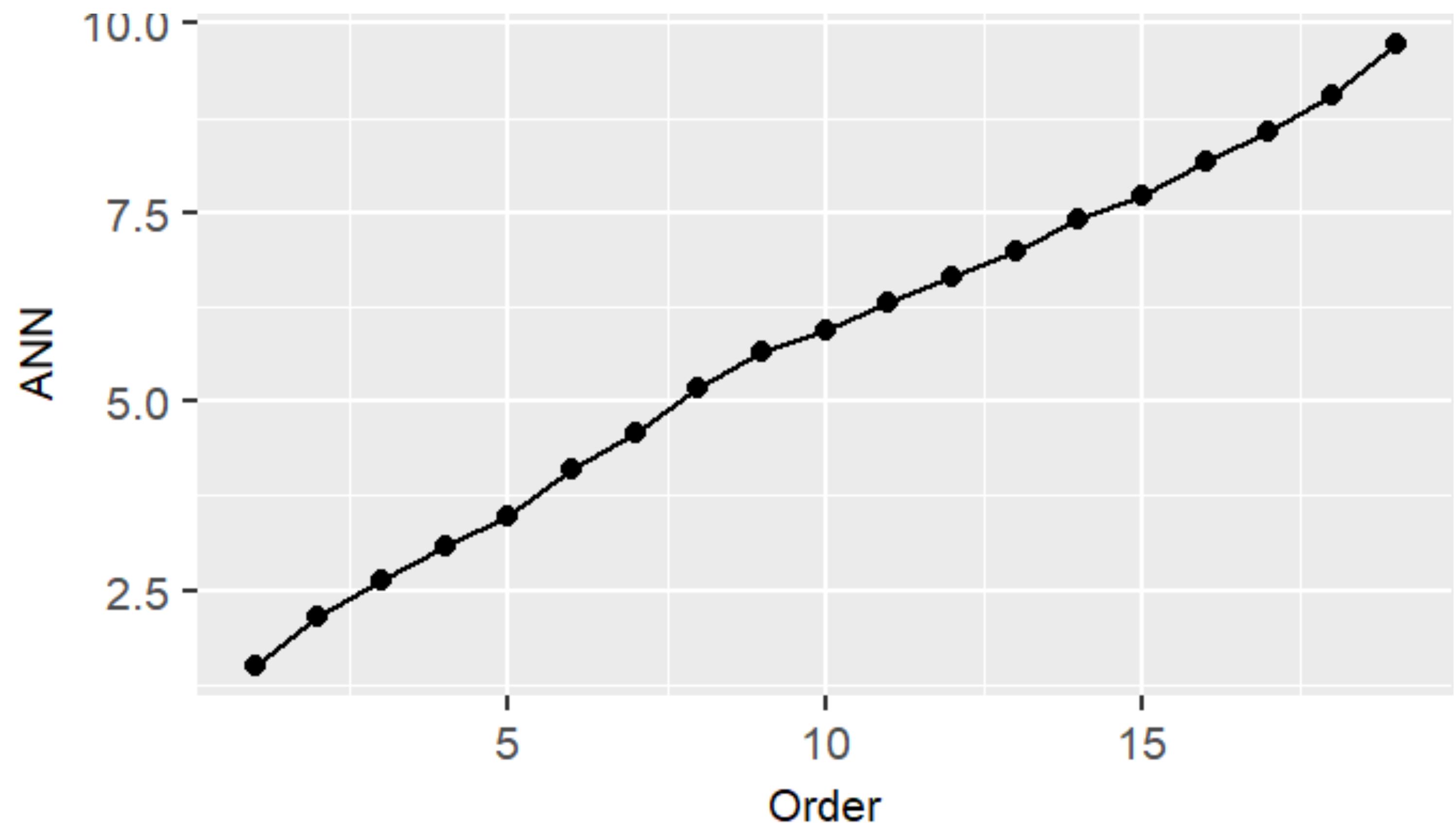
Null
Hypothesis



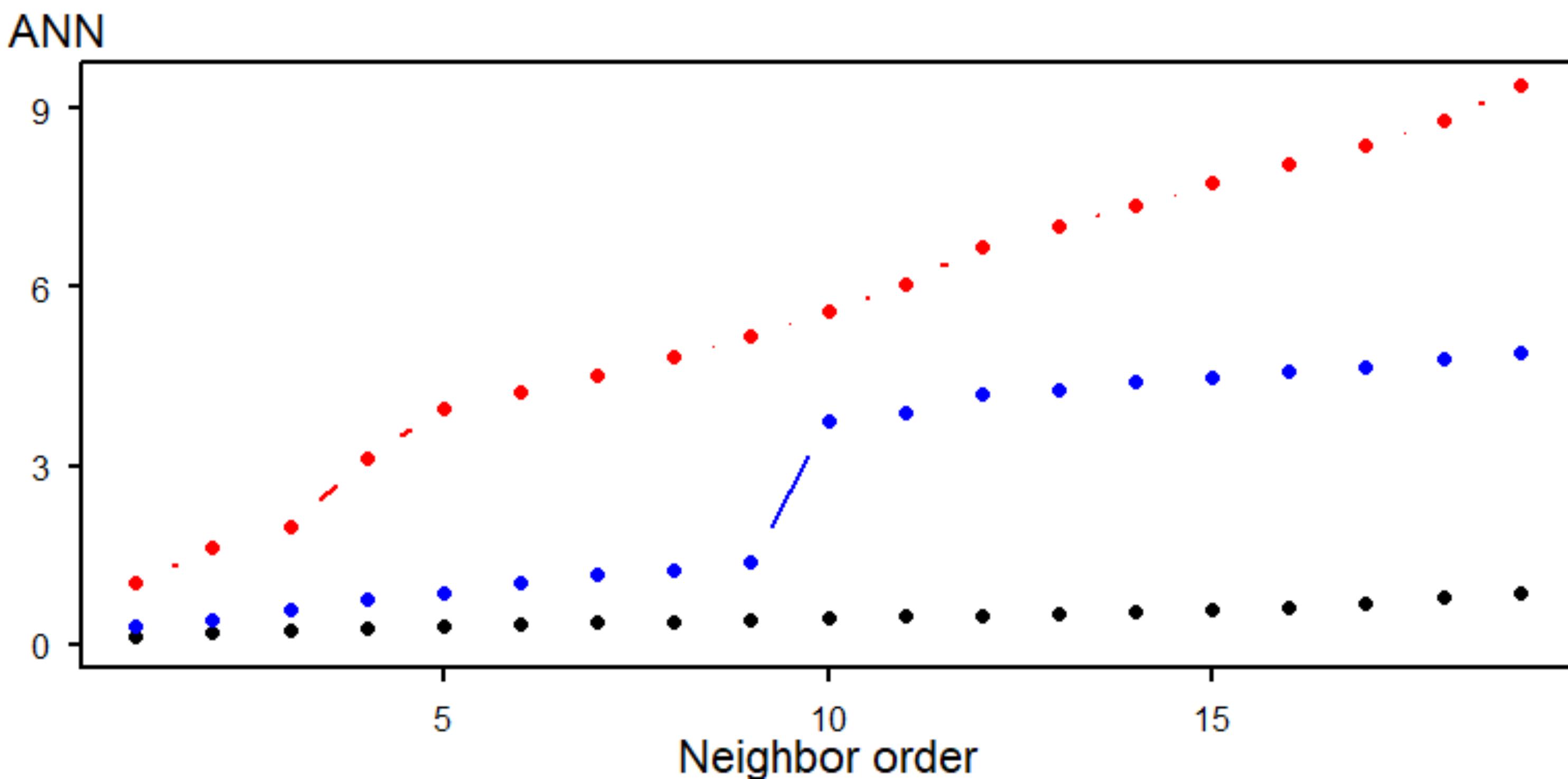
<https://forms.gle/5a2vacFEz3qLjU7K9>

We aren't just limited to “nearest neighbor” and calculating a statistic

plot the ANN values for different order neighbors, that is for the first closest point, then the second closest point, and so forth.



ANN vs neighbor order
offers insight into
underlying spatial
relationship



Note: study space definition
affects this measure

