Course Reminders

Due Dates:

- Dr. Brad Voytek guest lecture this Friday (5/31) - *no iclicker needed*
- A5 due Sunday after week 10 (11:59 PM)
- Final Project due Wed (6/12) of finals week at 11:59 PM

Course Notes:

- A3 grades posted & feedback sent
- Guest Lecture Attendance Posted
- **Friday 3PM Section** this week:
  - John (PCYNH 121) canceled
  - Phillip (MANDEB-150) still on
- No office hours today for Professor Ellis

Dr. Gina Merchant: Guest Lecture Review

- Behavioral scientist and data scientist and
- Background in exercise, social media/networks, and how to affect change in behavior around health
- Data: qualitative! and quantitative
    - Text (Facebook)
    - Accelerometer
    - Surveys
- 10 Pieces of Advice, including:
    - Know & love your data/measurement (and what you don't have!)
    - Know your contributions
    - Qualitative == Quantitative
    - Read!
    - Network
    - Sleep

Project Rubric

The grading rubric for the Final Project is as follows:

Category	Percentage of Project Grade
Introduction and Background	10%
Data Description	10%
Data Cleaning/Pre-processing	10%
Data Visualization	15%
Data Analysis & Results	25%
Privacy/Ethics Considerations	15%
Conclusion & Discussion	15%

Note: Individual grades *can* be adjusted based on the feedback provided in individual evaluations submitted. This means that team members in the same group can receive different scores from one another, if evaluations suggest that contributions were not evenly distributed. To avoid this, work together as a group and ensure that you're contributing to the project.

You can see rubric specifics on Gradescope

Rubric does not include grade adjustments for participation

Outline for **Final Project**

35 points total

Create questions and subquestions via the + buttons below. Reorder and indent questions by dragging them in the outline.

#	TITLE	POINTS
1	Introduction & Background	3.5
1.1	Overview	0.5
1.2	Research Question	0.5
1.3	Background & Prior Work	2
1.4	Hypothesis	0.5
2	Dataset(s)	3.5
3	Data Analysis	17.5
3.1	Data Cleaning / Pre-processing	3.5
3.2	Data Visualization	5.25
3.3	Data Analysis / Results	8.75
4	Privacy / Ethics Considerations	5.25
5	Conclusion & Discussion	5.25

Some final project notes:

1. Be sure that you've executed all code and output is as you want it to be. We will not be re-running.
2. Text including interpretations *must* be included throughout
   a. Explain data acquisition (i.e. web scraping outside of notebook)
   b. Explain data cleaning (or what checks you did to assure yourself your dataset was ready to go)
   c. Interpret visualizations - explain what is plotted *and* what is learned from it
   d. Interpret analysis - explain what you did and what conclusions you've drawn from the analysis
3. You can have more than three visualizations, but make sure all visualizations have axes labeled and are appropriate for the data being plotted
4. You *are* being graded on the report overall
   a. Do <u>not</u> over-explain; include necessary information; be concise
   b. It should tell a story
   c. Editing is important
5. After you convert to PDF, make sure all text/visualizations are visible. THIS document is what we'll grade. Looking over your work before submission for typos, clarity, and formatting matters.

A final word on your projects

I don’t care if your:

- R^2 is 0
- Your initial hypothesis was wrong
- Your predictive model is complete garbage

I do care that you:

- Do the analysis
- Interpret appropriately
- Discuss limitations and ways to improve

COGS108 Final Project Submission

1. Notebook Naming: **FinalProject_groupXXX.ipynb**
2. **PDF of notebook to Gradescope (*required*) - FinalProject_groupXXX.pdf**
3. Notebook to GitHub (*optional*)
    a. Fork https://github.com/COGS108/FinalProjects-Sp19 and submit your notebook as a PR to this repo
    b. Extra credit (1 pt)
    c. Upload gives consent for future use in class (PIDs removed)
    d. If you do NOT want it used as an example in the future, please add the name of your file here: http://bit.ly/not_example
4. **Fill out survey about team and individual teammates (*required*)**: http://bit.ly/COGS108_TeamEval

COGS108 Extra Credit Opportunities:

1. Final Project to GitHub (1 pt)
2. CAPE class response > 85% (1 pt to everyone)
3. End of course survey (1 pt)

Basic Geospatial Analysis: Summary

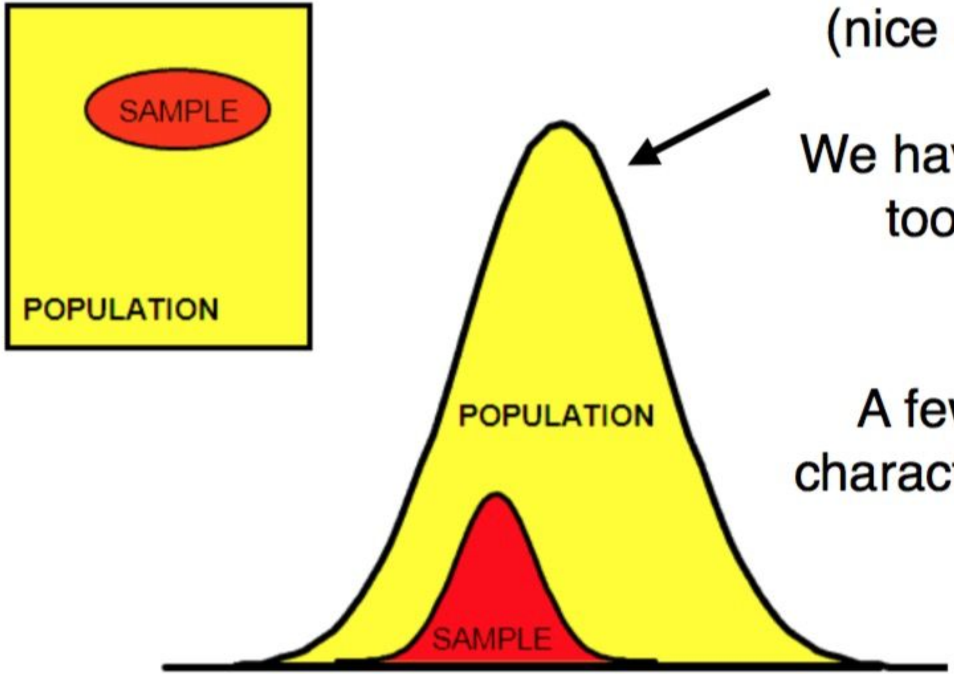
1. Considerations when visualizing spatial data important to conclusions drawn
    a. values to plot?
    b. map type?
    c. color scale?
2. Traditional statistics fail with geospatial data:
    a. Spatial autocorrelation
    b. MAUP
    c. Edge effects
    d. Ecological fallacy
    e. Nonuniformity of space
3. Analysis still possible
    a. Global Point Density, Quadrat Density, Kernel Density
    b. Poisson Point Process
    c. K-Nearest Neighbor (KNN)
    d. Comparison to a CRP (using simulation)

Non-parametric Statistics

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

Non-parametric Statistics: The Why

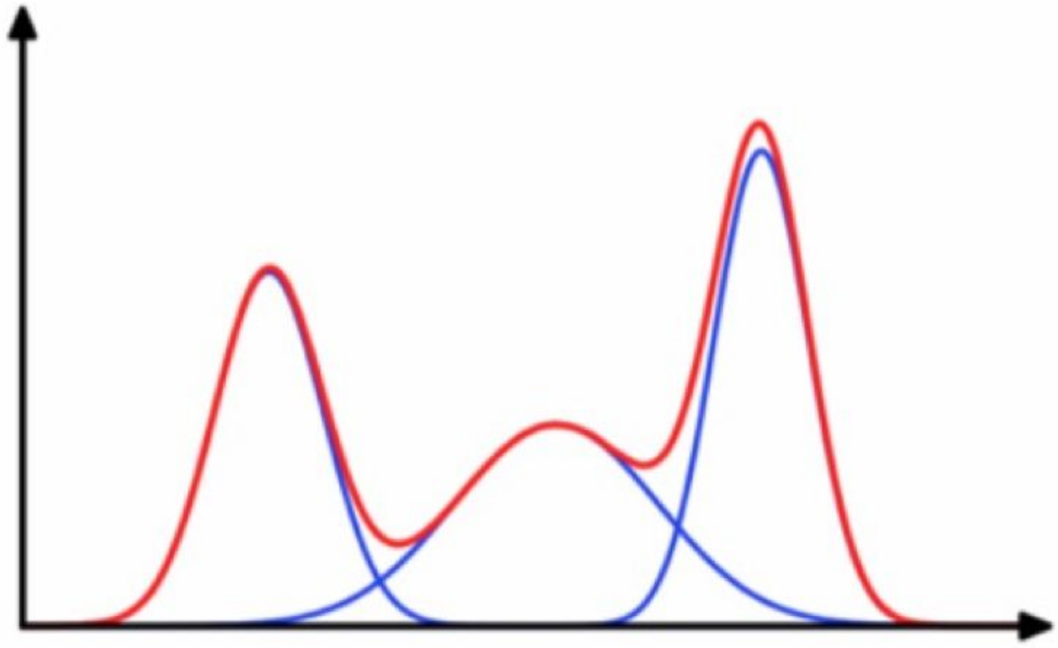


Normal distribution
(nice and friendly)

We have good math tools for this.

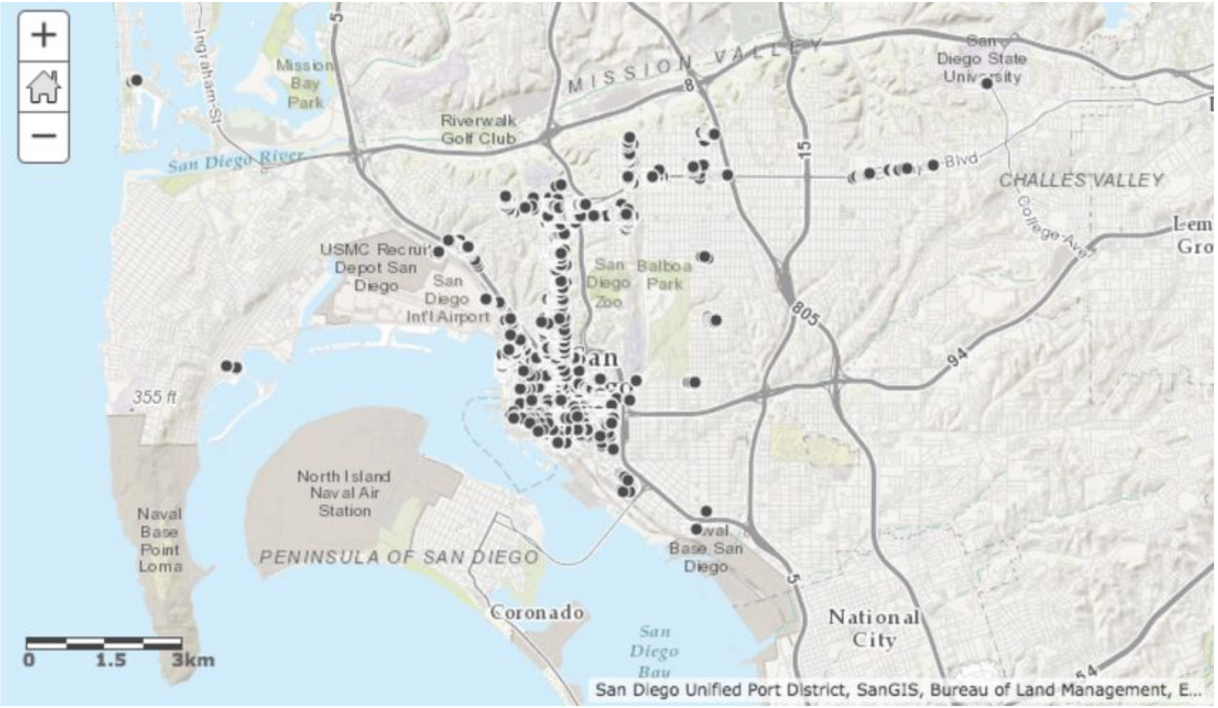
A few parameters **fully** characterize the distribution.

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



adapted from Brad Voytek

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

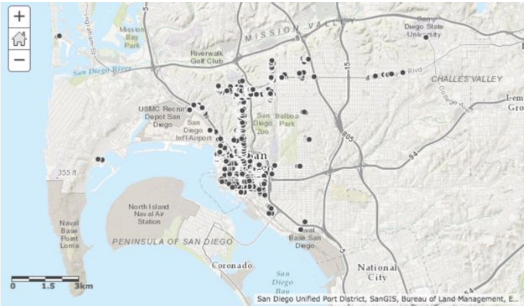


Parameters (like mean and variance) cannot fully and accurately capture this distribution!

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

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.

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 no assume data-generating process (like heights) result in, *e.g.*, normally-distributed data

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)

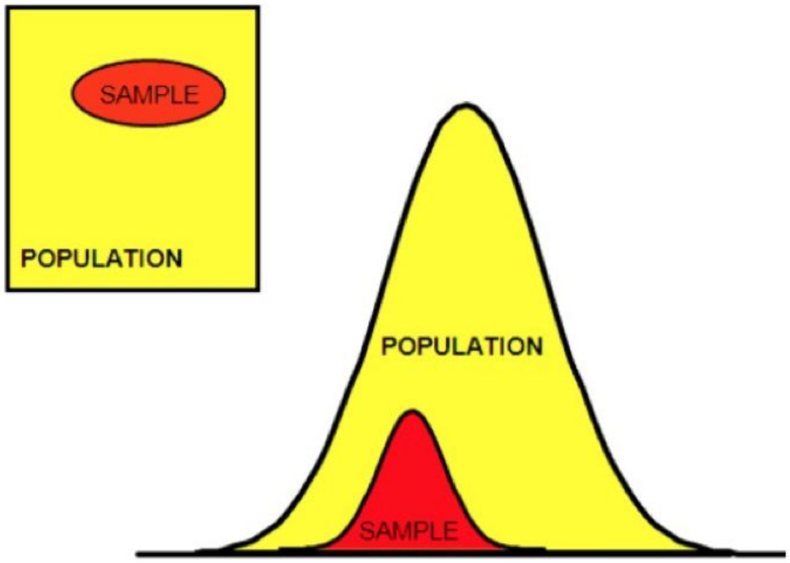
E
Human hair color (i.e. black, brown, red, blonde, etc.)

Resampling statistics: The What

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

1) Bootstrapping (resampling)

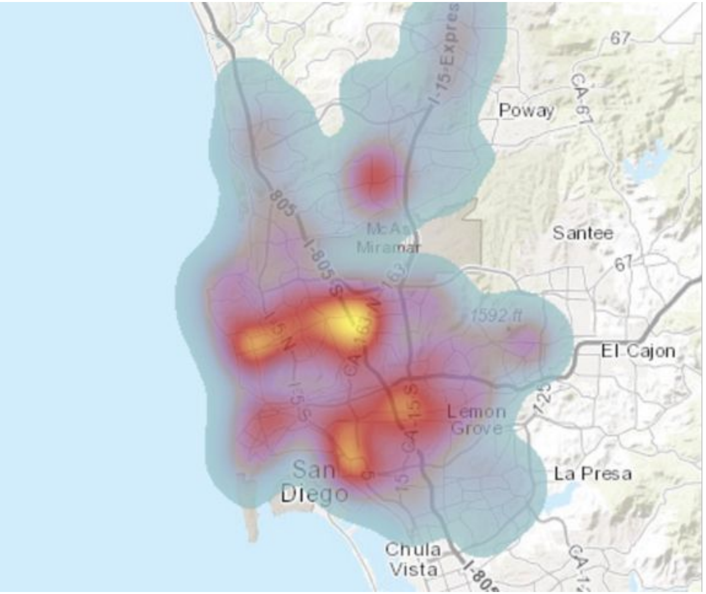
- How can we build a more realistic “null distribution” for the sample estimate without knowing the population it’s drawn from?

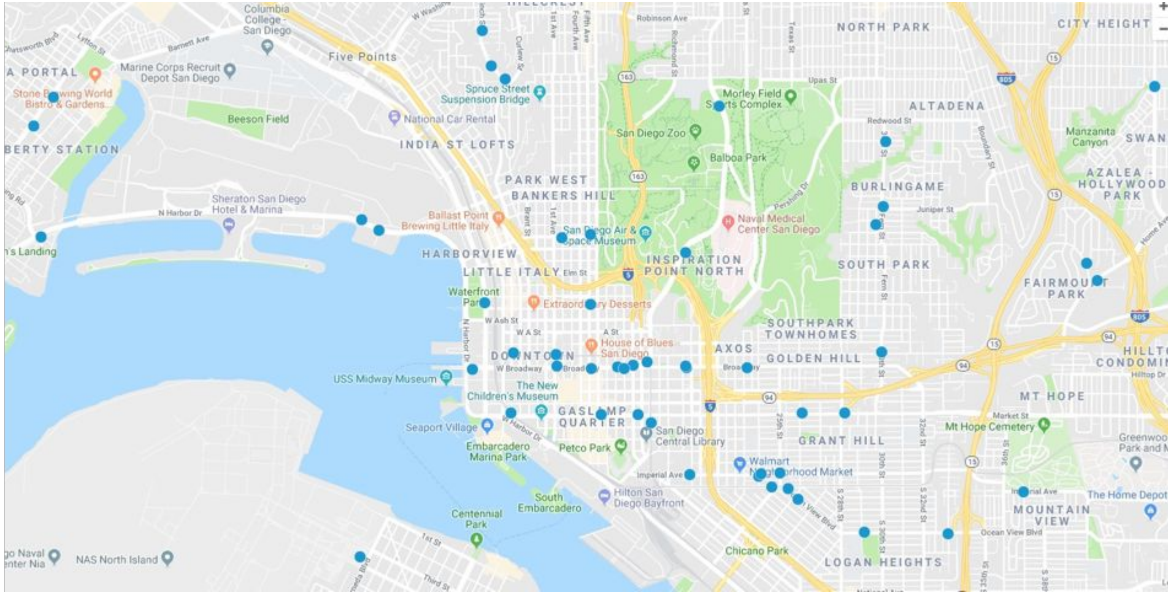


Bootstrapping (resampling)

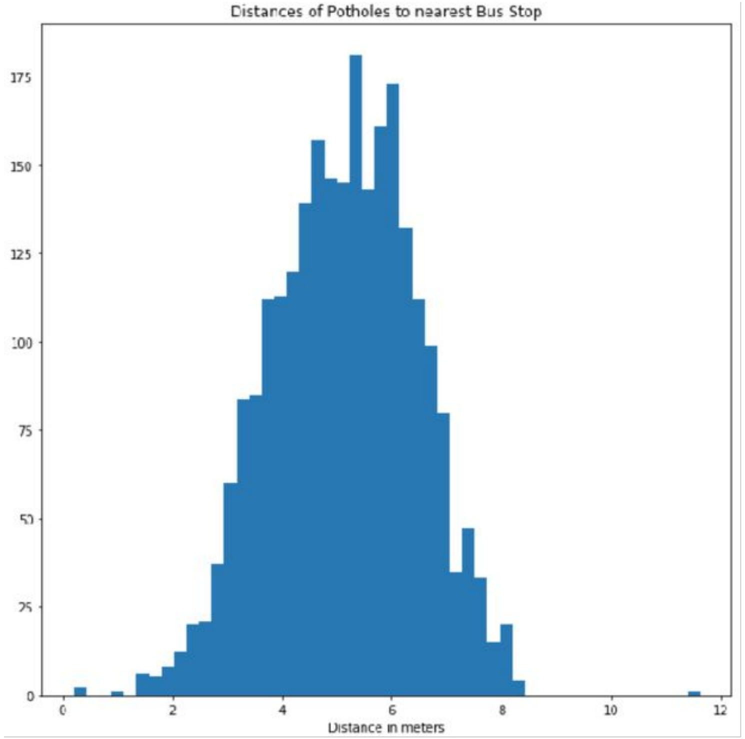
Example Question:

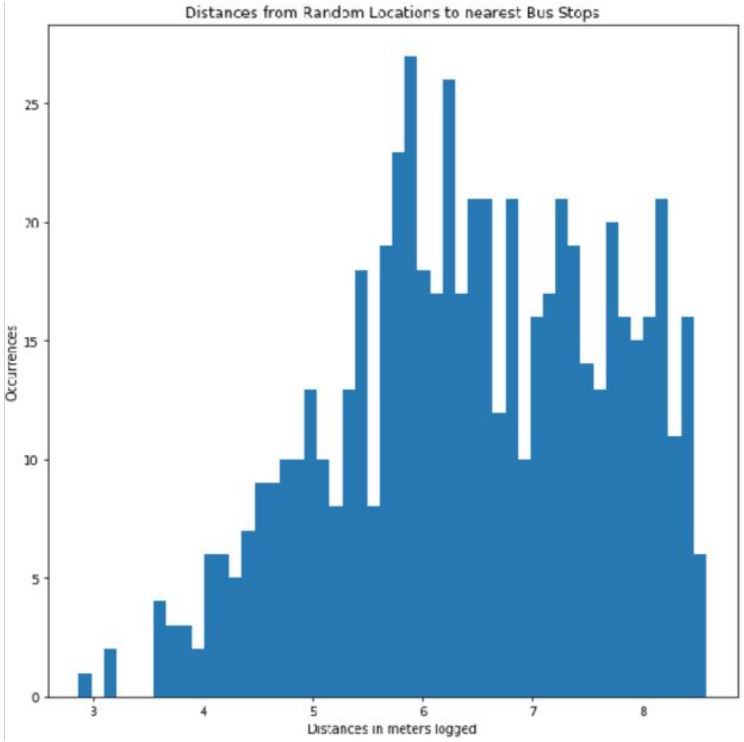
- Are San Diego's pot holes closer to bus stops than not?



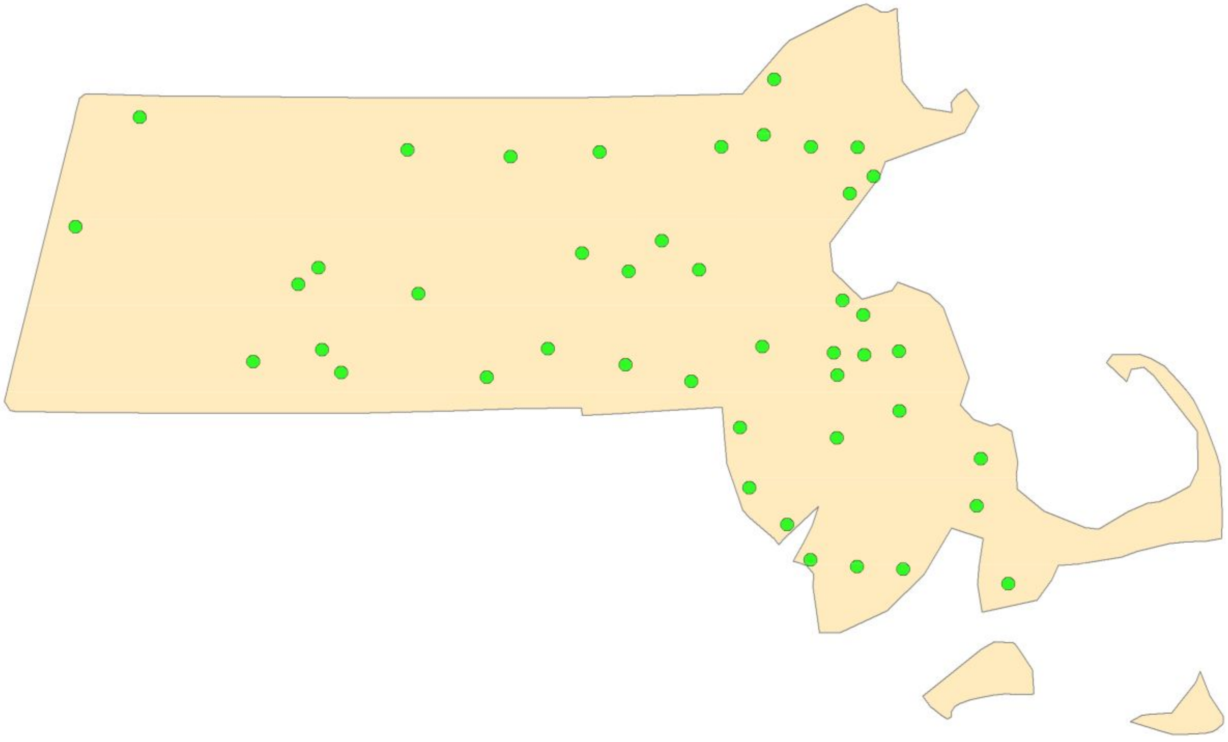


Bootstrapping (resampling)





Is this distribution of Walmarts in MA the result of a CRP?

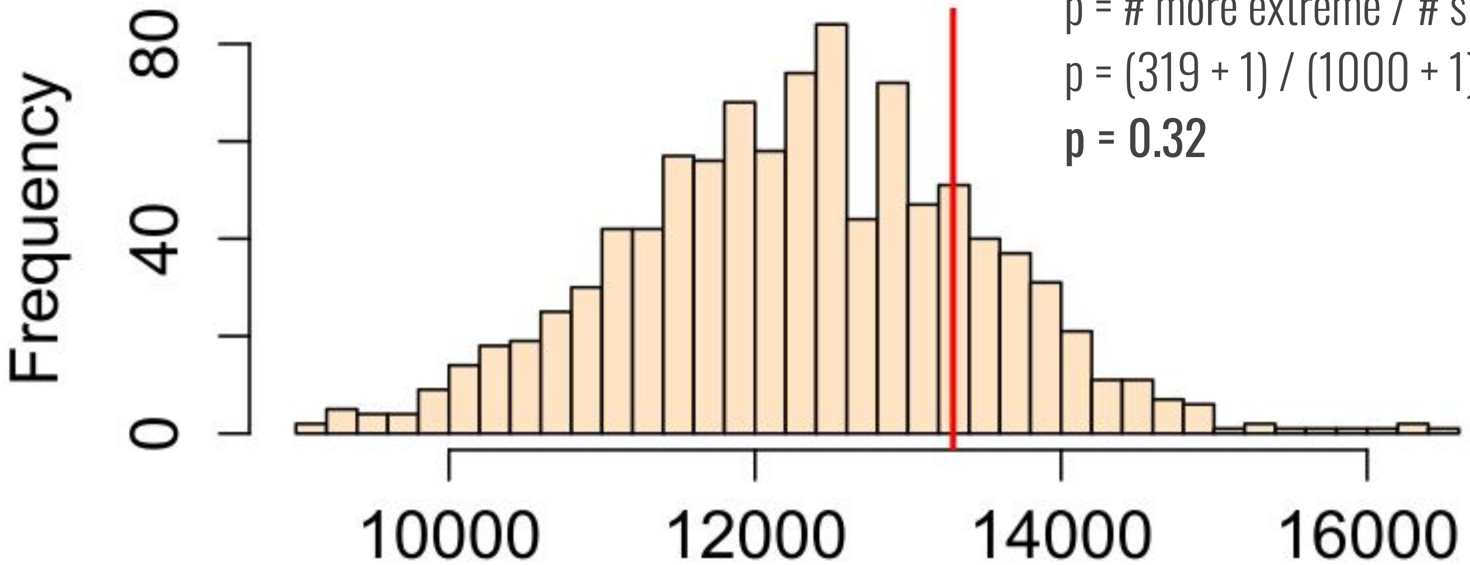


https://mgimond.github.io/Spatial/introGIS.html

Hypothesis Testing: A Monte Carlo Test

1. First, we postulate a process–**our null hypothesis, H_o.** 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. ANN) 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.

This is an example of bootstrapping!

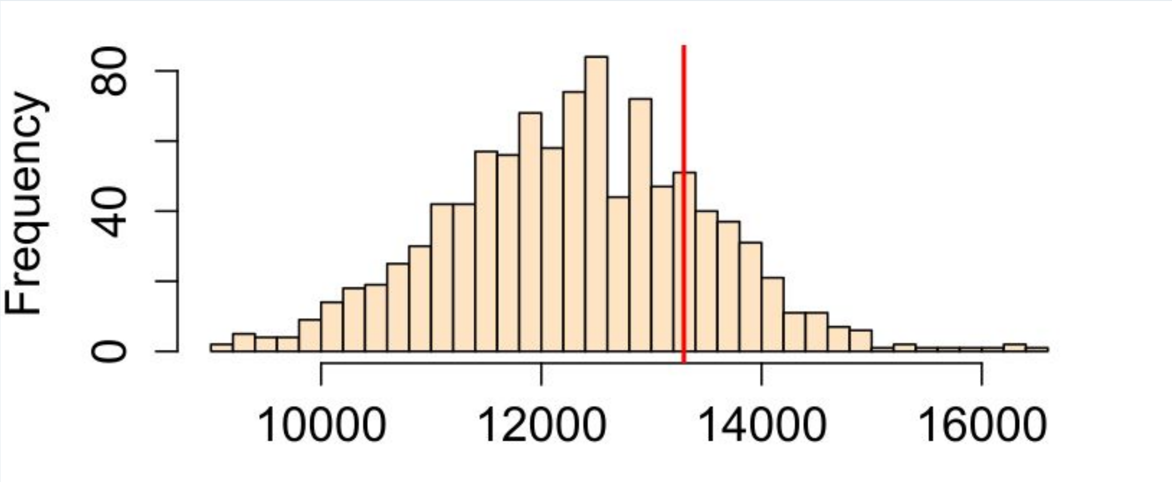


p = # more extreme / # simulations

p = (319 + 1) / (1000 + 1)

p = 0.32

Suggests that our results come from a CRP ANN



What does the **histogram** represent in this image?

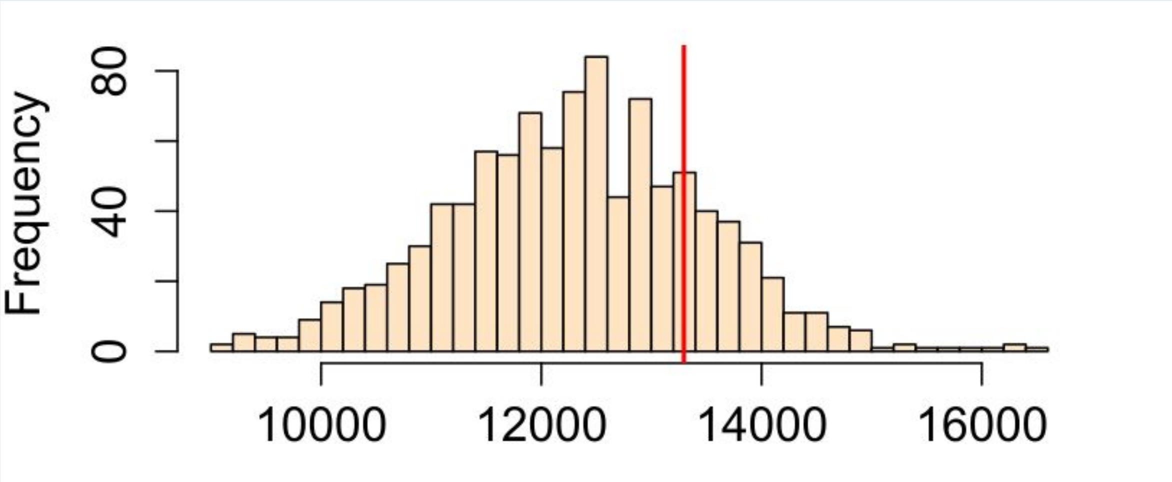
A
Alternative Hypothesis

B
Null Hypothesis

C
Distribution of ANN generated under the null

D
Distributi on of ANN generated under the alternative

E
I'm lost



What does the **red line** represent?

A
The observed data

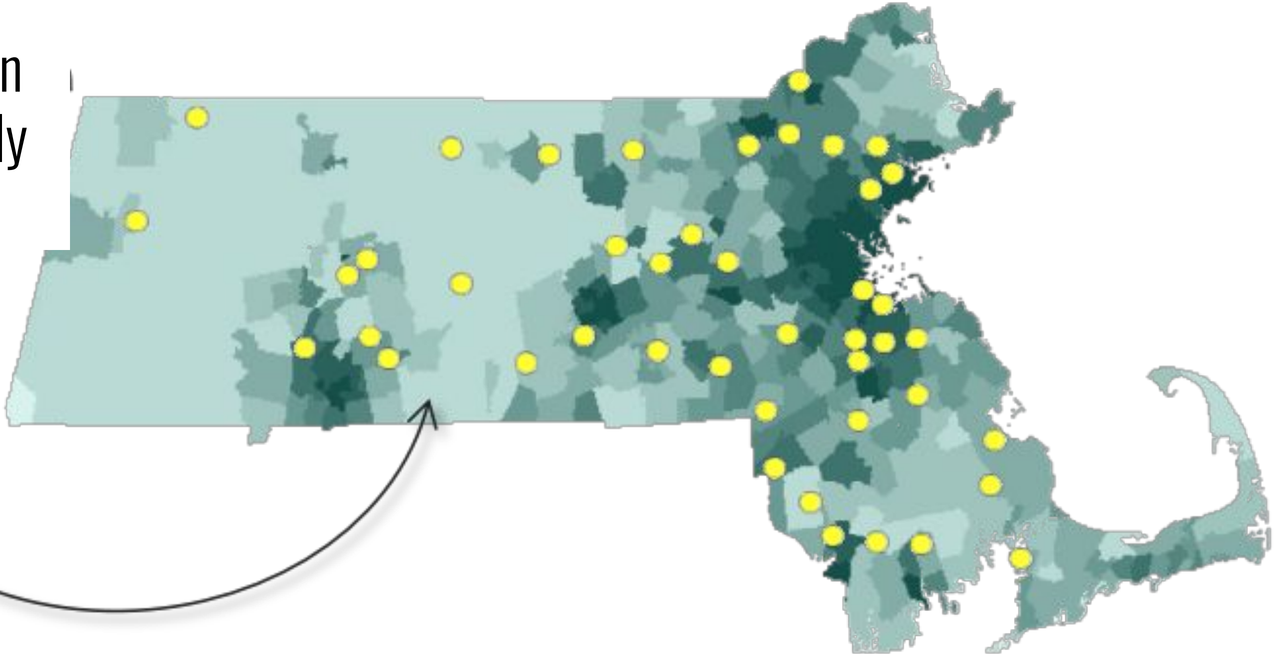
B
The hypothetical data

C
The null distribution

D
The alternative distribution

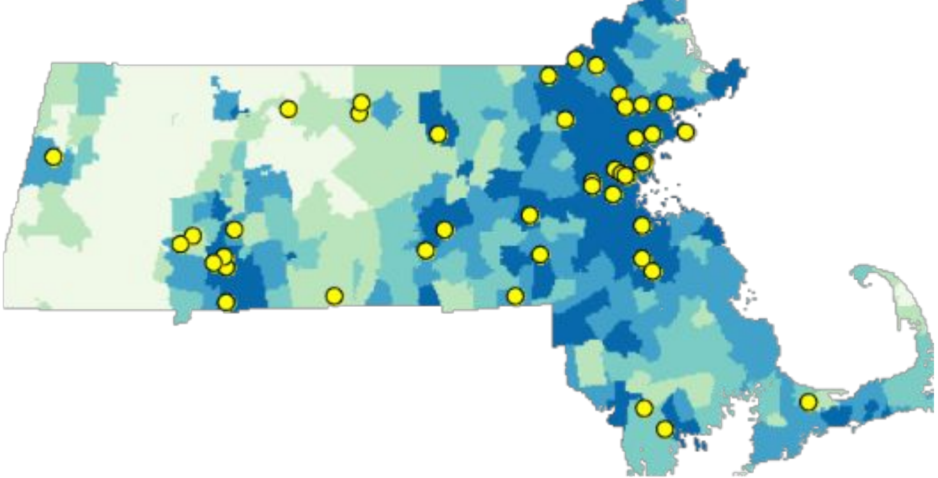
E
I'm lost

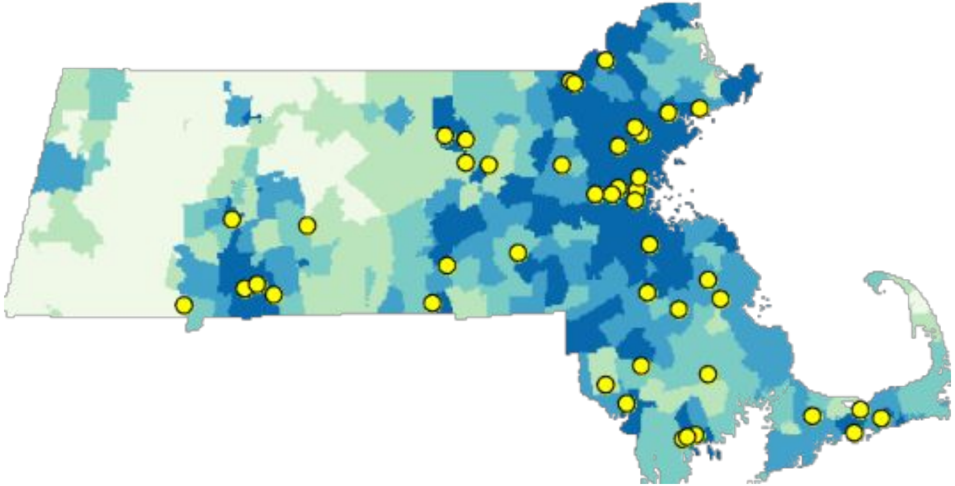
When controlling for population density, are Walmarts randomly distributed?



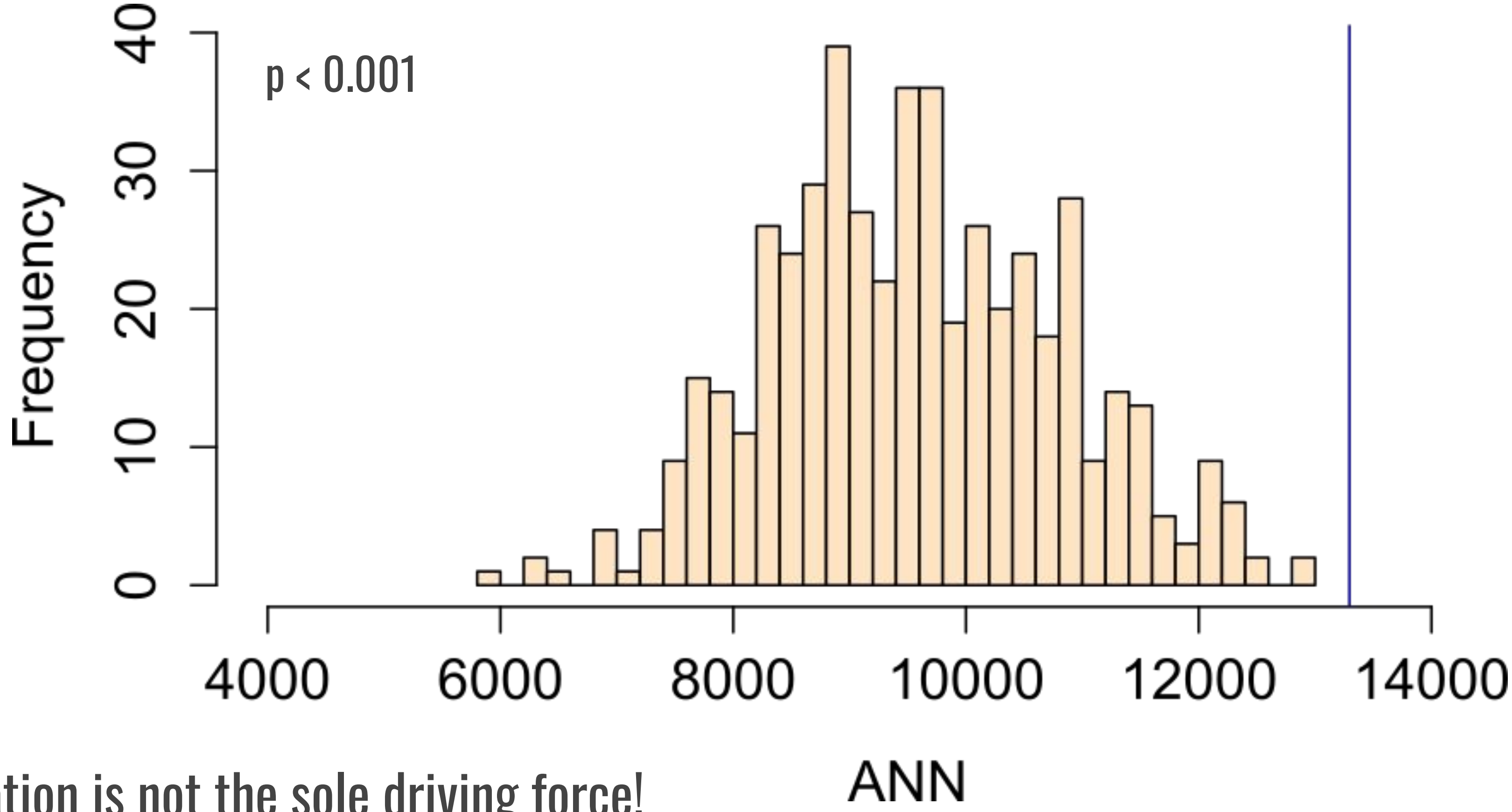
https://mgimond.github.io/Spatial/introGIS.html

two randomly generated point patterns using population density as the underlying process





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



Population is <u>not</u> the sole driving force!

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

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

ordinal data

1, 4.5, 6.6, 9.2 → 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 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

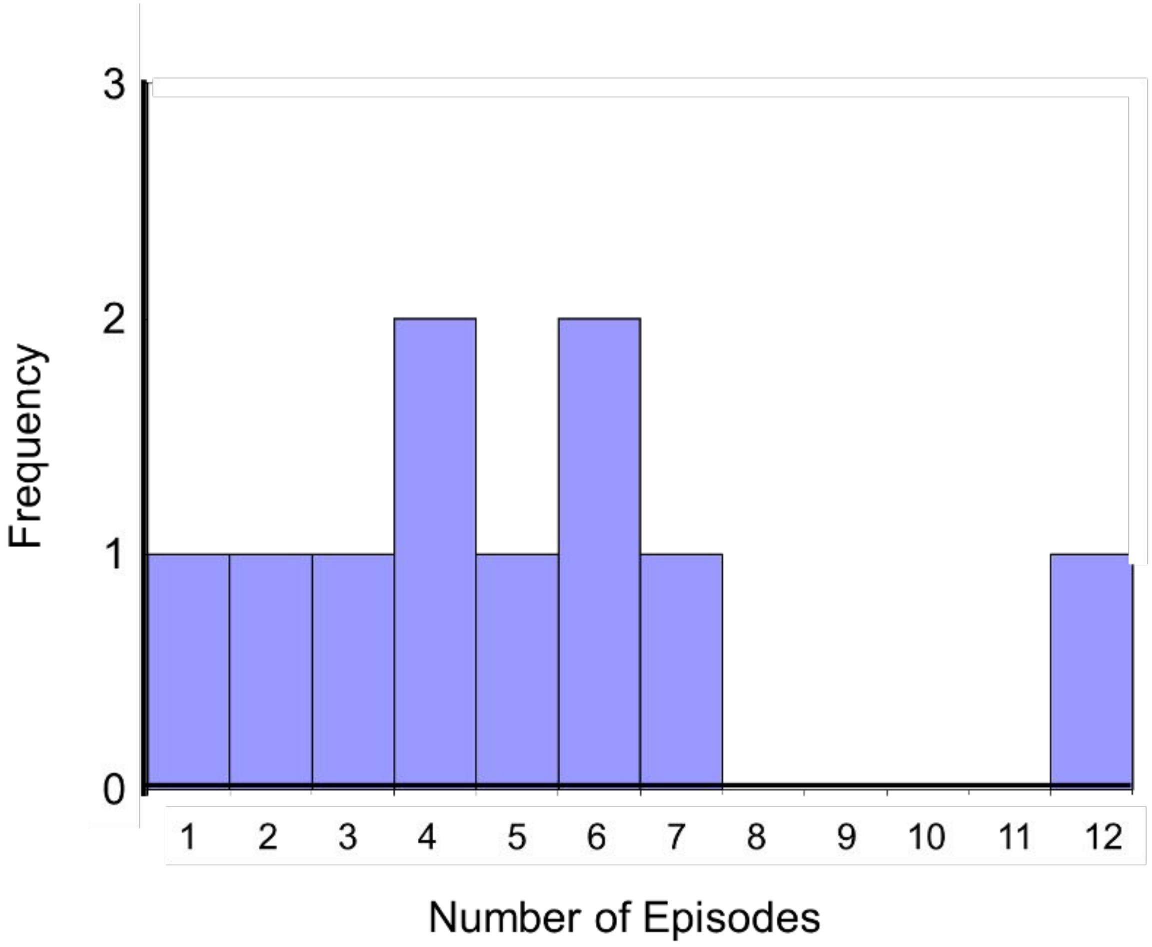
In a clinical trial, is there a difference in the number of episodes of shortness of breath between placebo and treatment?

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



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

		Total Sample (Ordered Smallest to Largest)		Ranks	
Placebo	New Drug	Placebo	New Drug	Placebo	New Drug
7	3		1		1
5	6		2		2
6	4		3		3
4	2	4	4	4.5	4.5
12	1	5		6	
		6	6	7.5	7.5
		7		9	
		12		10	

<u>Sum of ranks:</u>

Placebo = 37

New Drug = 18

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

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

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

The max possible value of TA

The observed sum of ranks for sample A

n_a = number of elements in group A
n_b = number of elements in group B

$U_{Placebo} = 3$

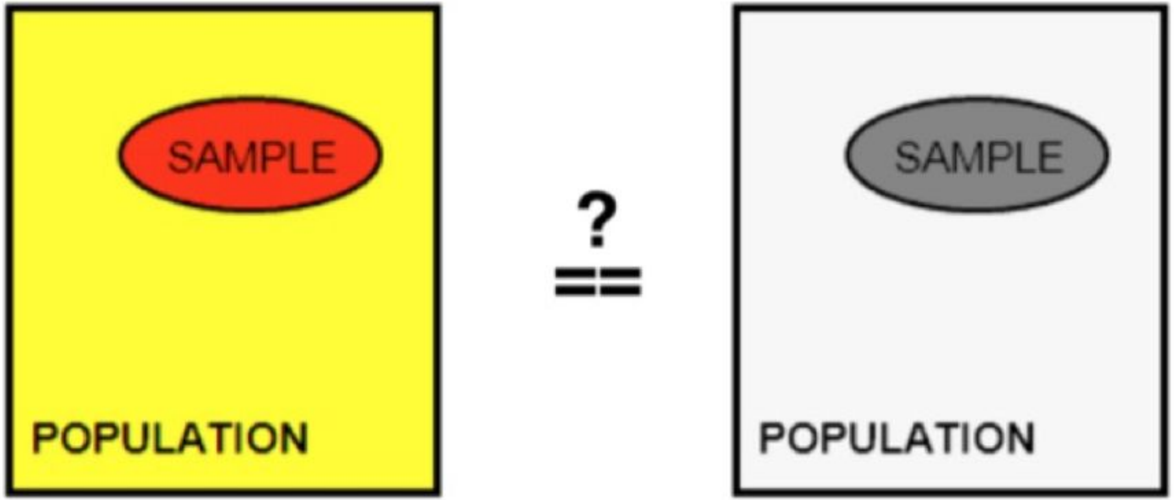
$U_{treatment} = 22$

$0 < U < n_1 * n_2$

Complete separation → no separation

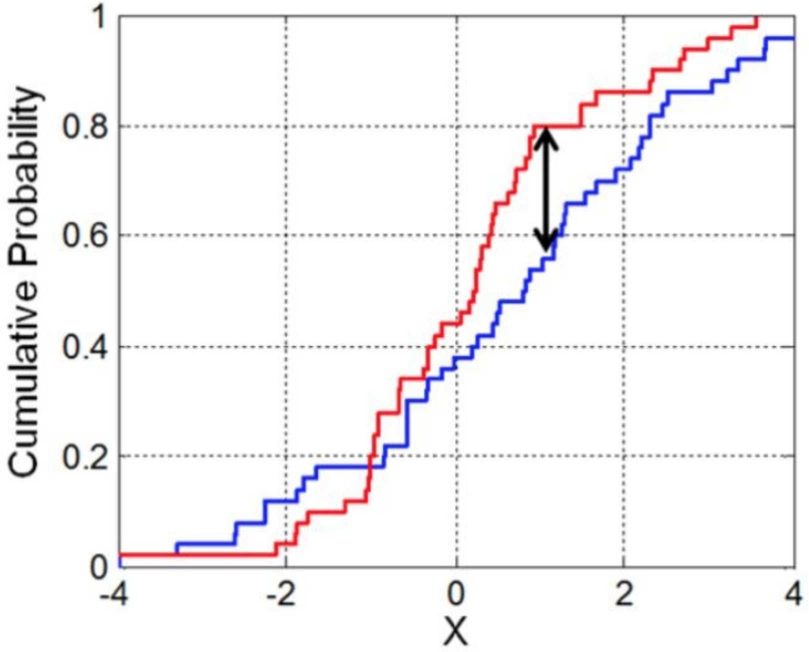
3) Kolmogorov-Smirnov (KS) test

- 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



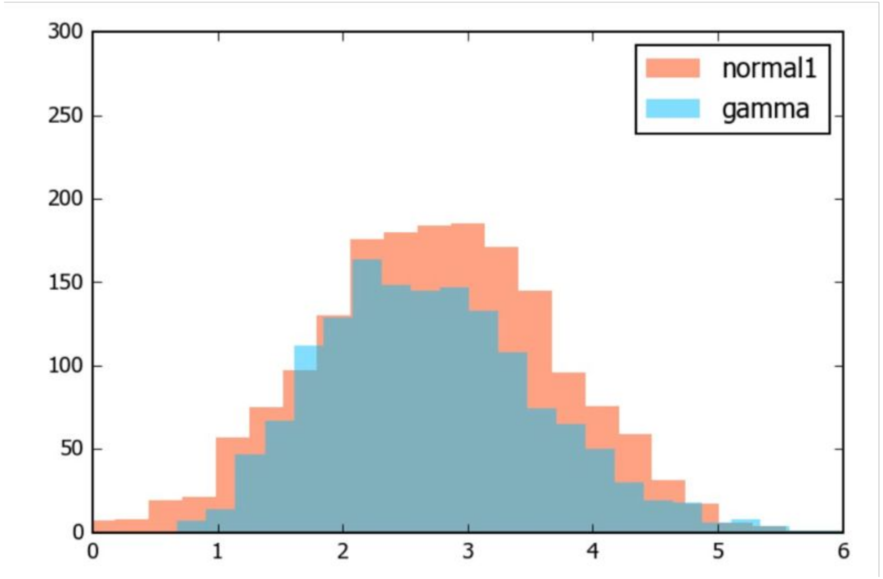
Find the maximum difference between the CDFs.

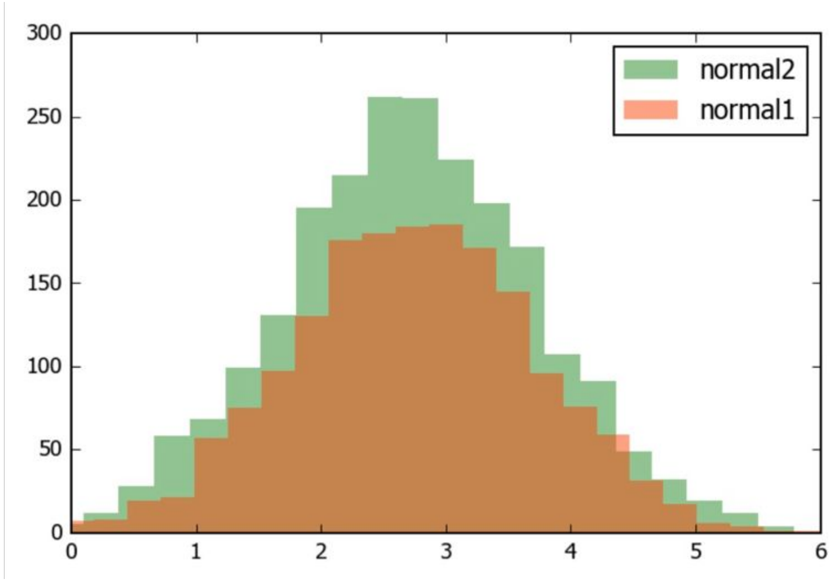
Tests:

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

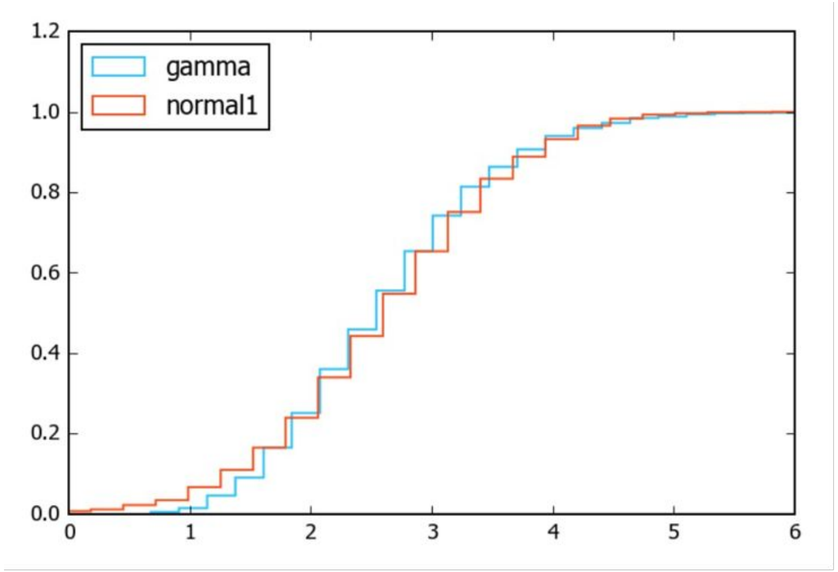
Kolmogorov-Smirnov (KS) test

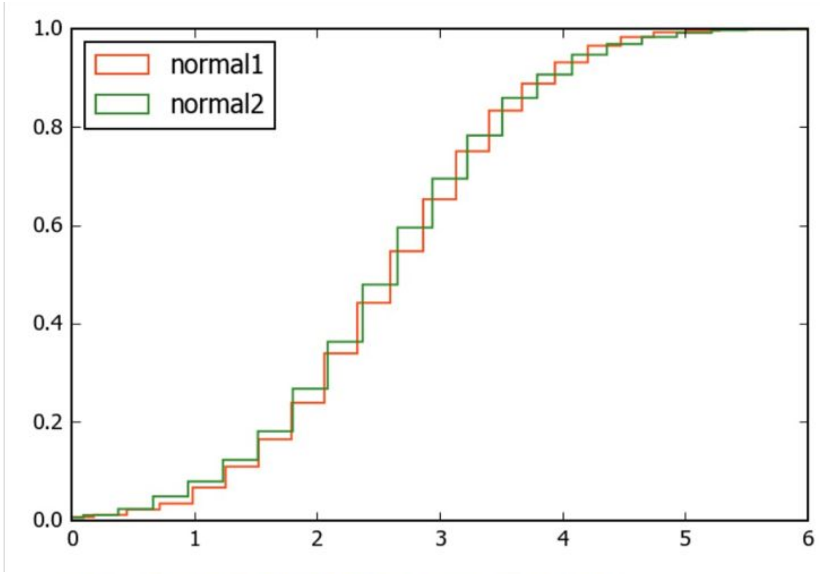
- Given (limited) samples from two populations, how do we quantify whether they come from the same distribution?





Kolmogorov-Smirnov (KS) test






```
gamma vs. normal1: p = 0.010680362841l
normal1 vs. normal2: p = 0.55073599824З
```

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