Course Reminders

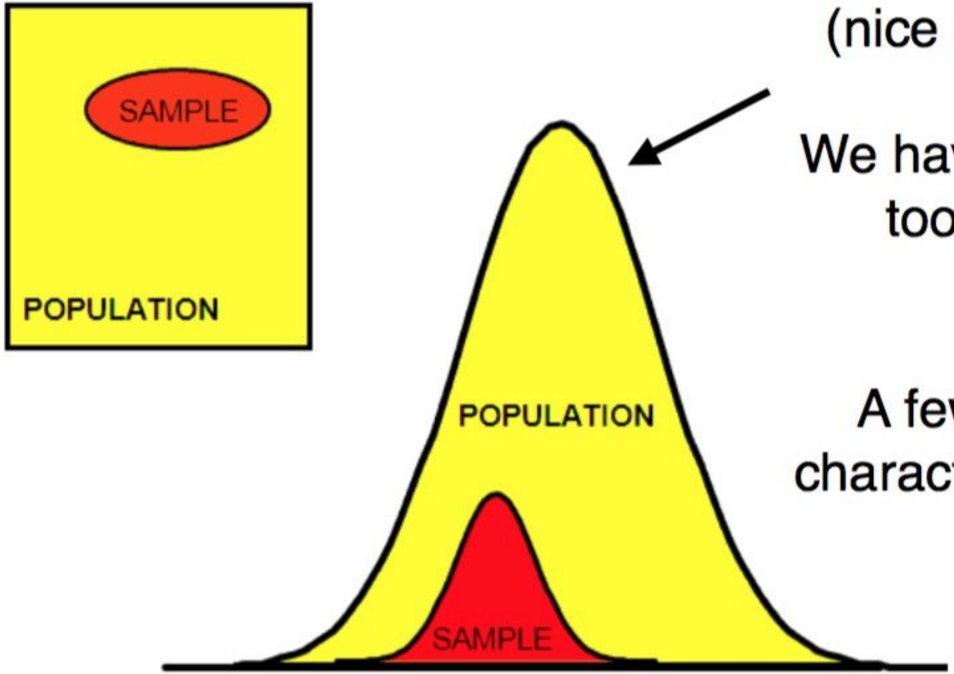
- A6 due Friday (11:59 PM)
- Week 8 quiz is available on Canvas
- Grades now available: Project Check-In, A4
  - A4 regrade: 1e - if you chopped off the newline character, also correct; 3f - if ‘0’ a string, totally correct
  - Project Check-In feedback
- Final Project Extra credit sign up (in-person): https://calendly.com/sellis-ucsd/cogs-108-project-ec
  - One time-slot per group
  - Mon 3/16; 8AM - 3PM (CSB 272)
  - Note: EC is either in-person OR video
- Please fill out CAPEs - feedback is super helpful!

Nonparametric Statistics

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

Non-parametric Statistics: The Why



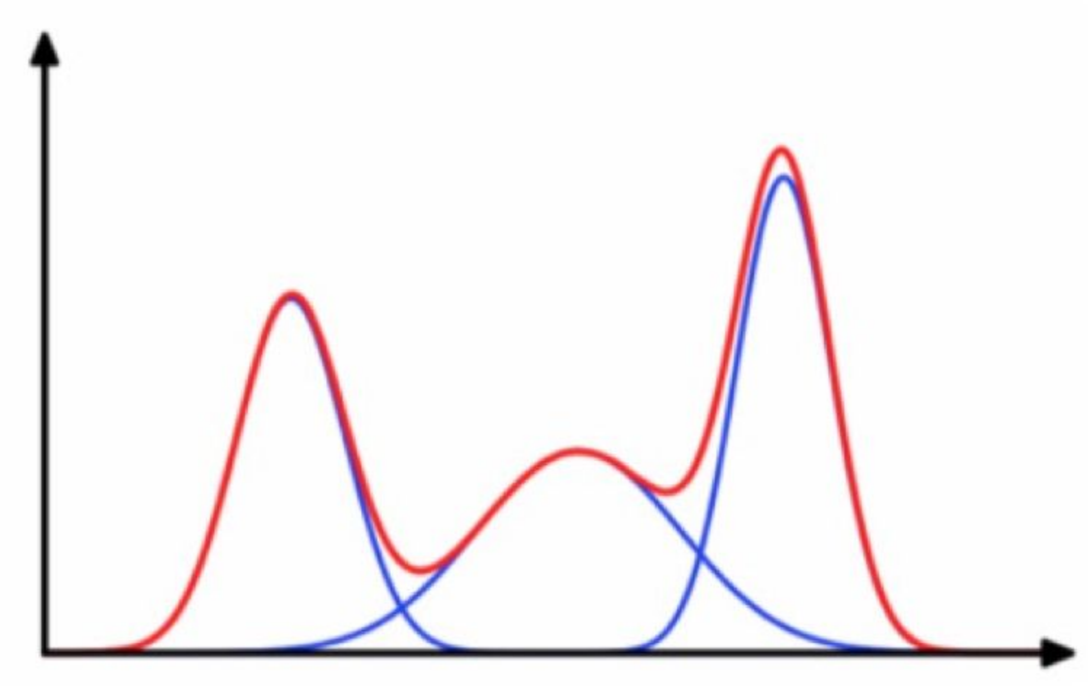
Normal distribution
(nice and friendly)

We have good math
tools for this.

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

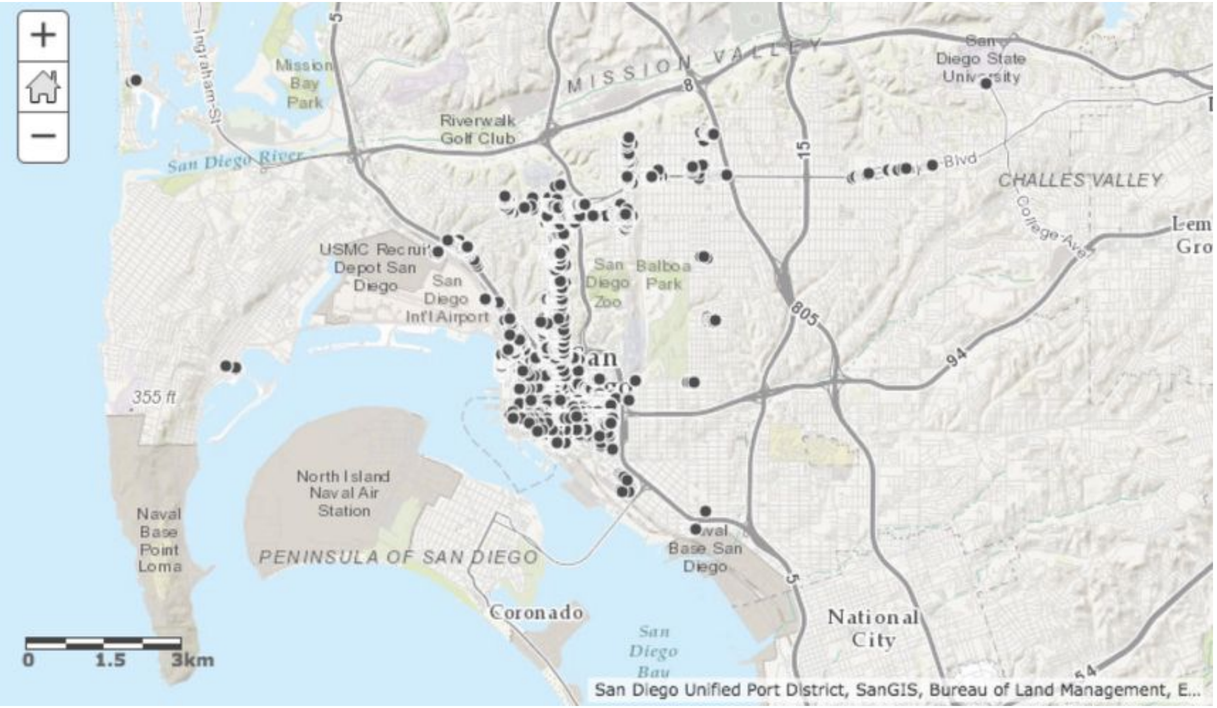
adapted from Brad Voytek

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



adapted from Brad Voytek

Non-parametric Statistics:
…or like this?



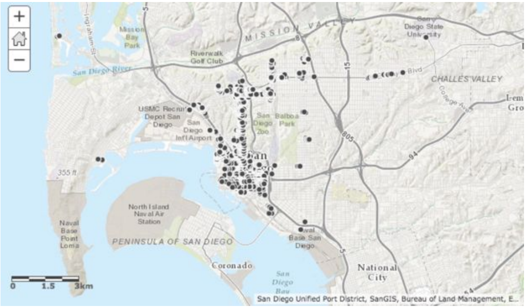
Parameters (like mean and variance) cannot fully and accurately capture this distribution!

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

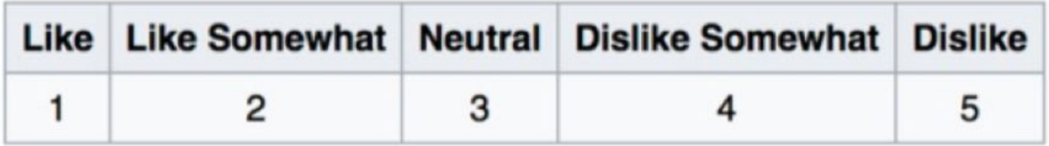
adapted from Brad Voytek

When to turn to non-parametric statistics...

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



- 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.

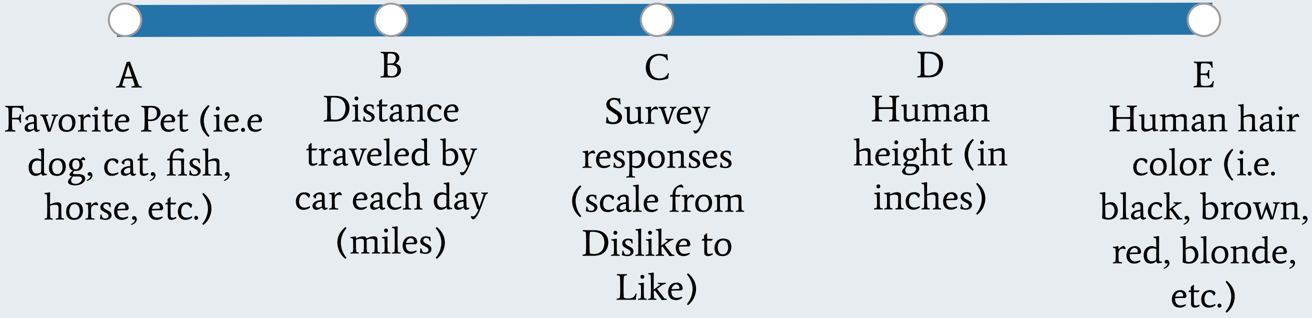
Non-parametric Statistics: distribution-free

- **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 not 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)

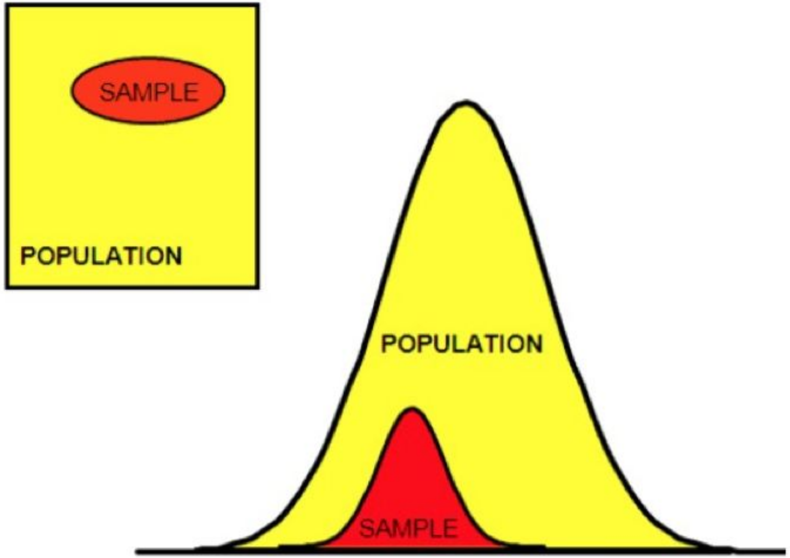
E
Human hair color (i.e. black, brown, red, blonde, etc.)

Resampling statistics: The What

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

1) Bootstrapping (resampling)

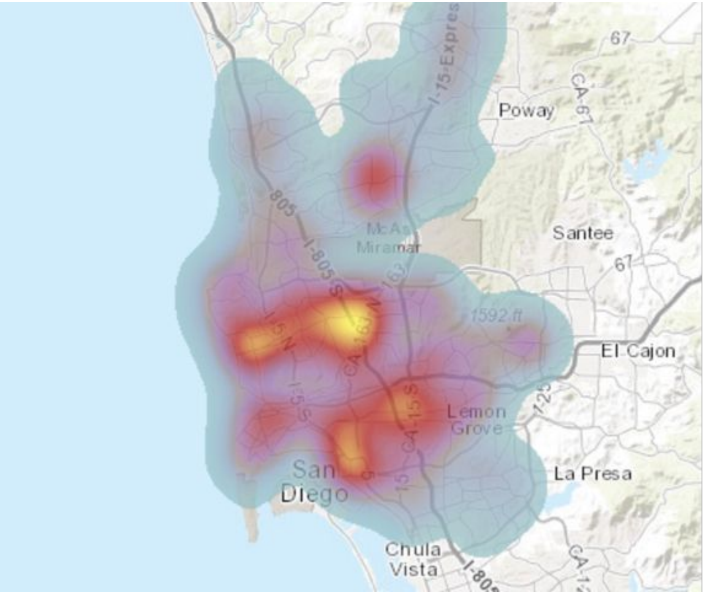
- How can we build a more realistic “null distribution” for the sample estimate without knowing the population it’s drawn from?

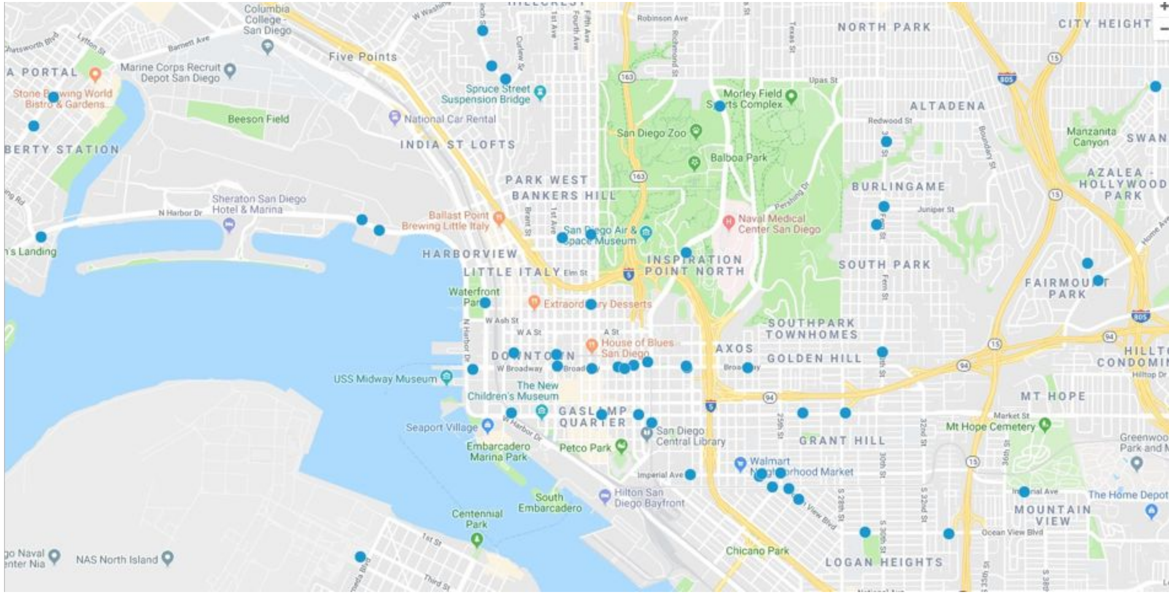


Bootstrapping (resampling)

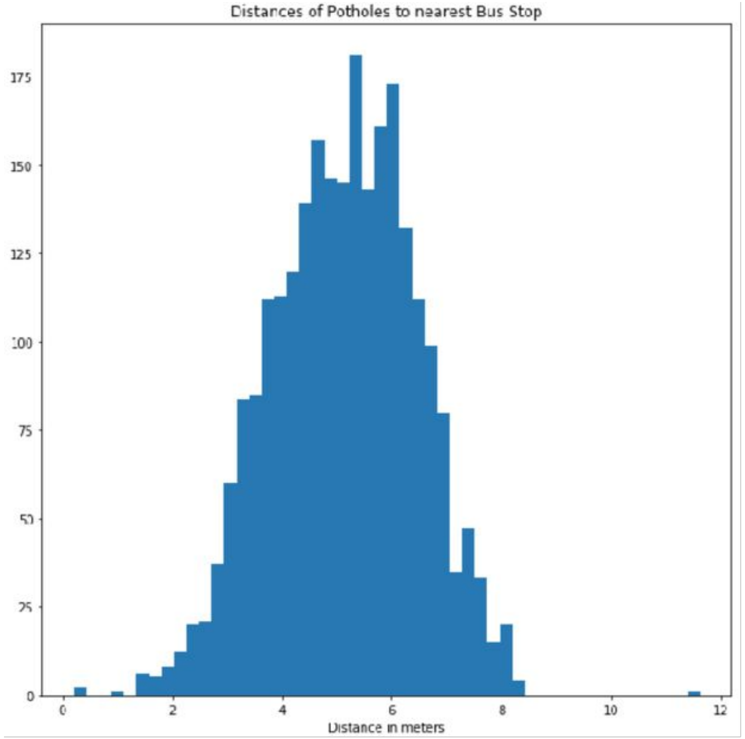
Example Question:

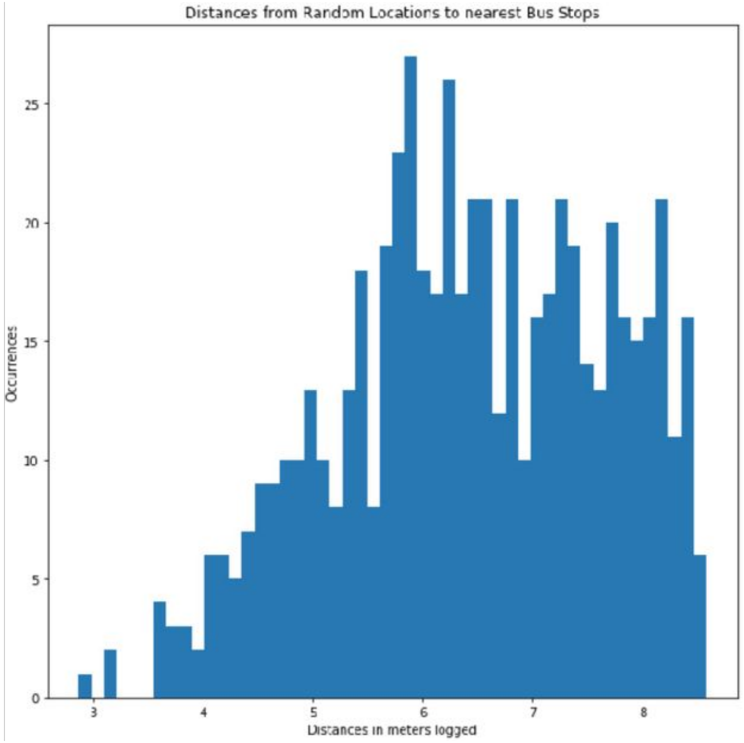
- Are San Diego's pot holes closer to bus stops than not?





Bootstrapping (resampling)





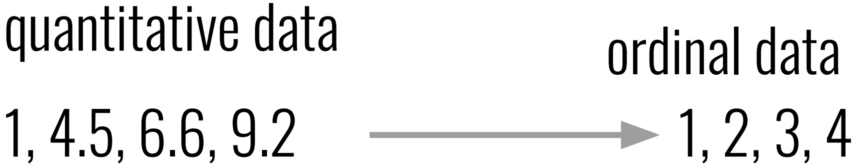
2) Rank Statistics

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

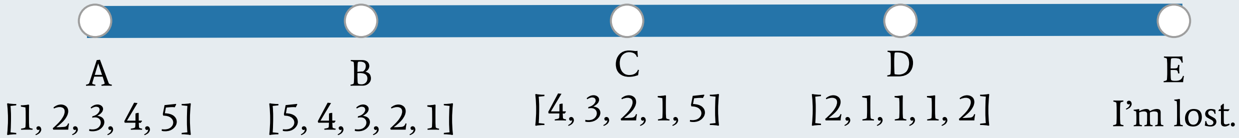
Ordinal data - categorical, where the variables have a natural order

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 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

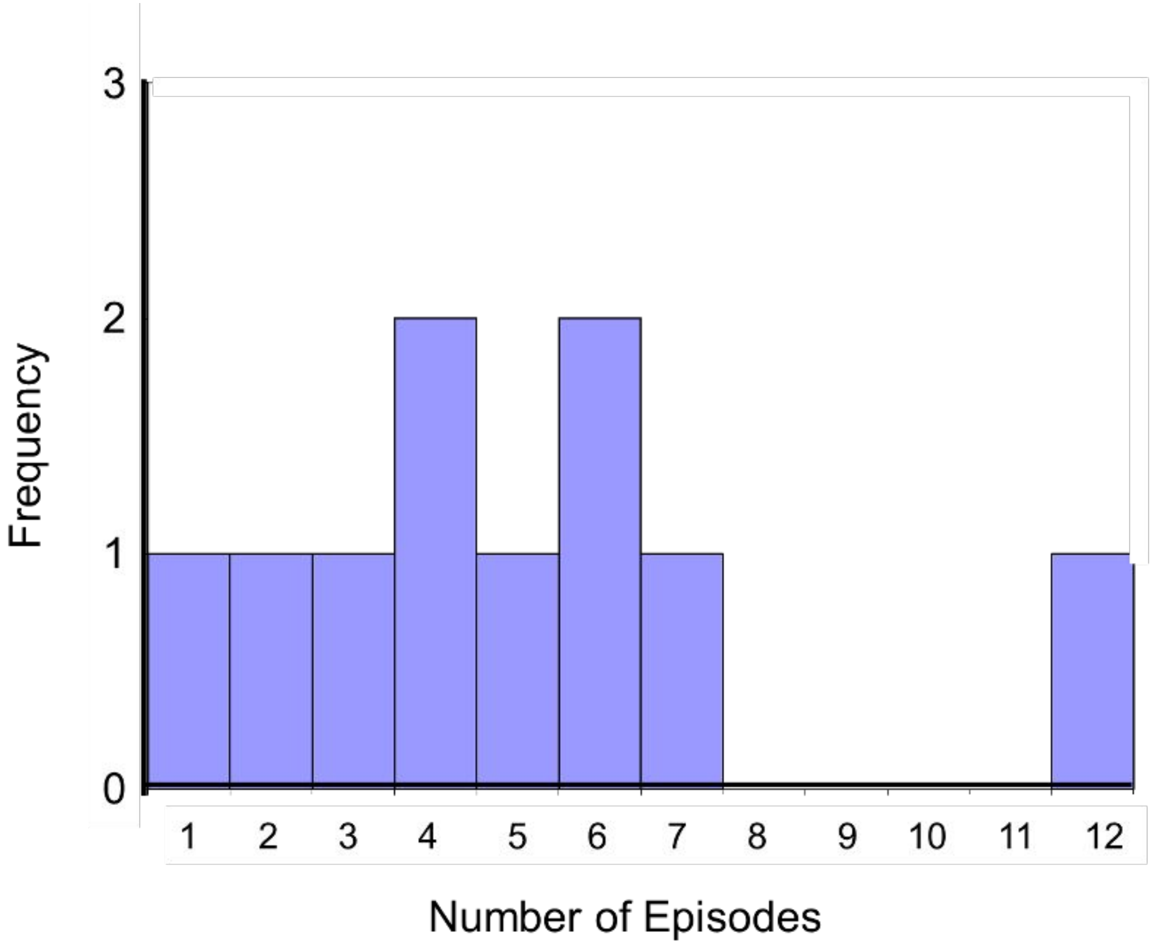
In a clinical trial, is there a difference in the number of episodes of shortness of breath between placebo and treatment?

Step 1: Participants record number of episodes they have.

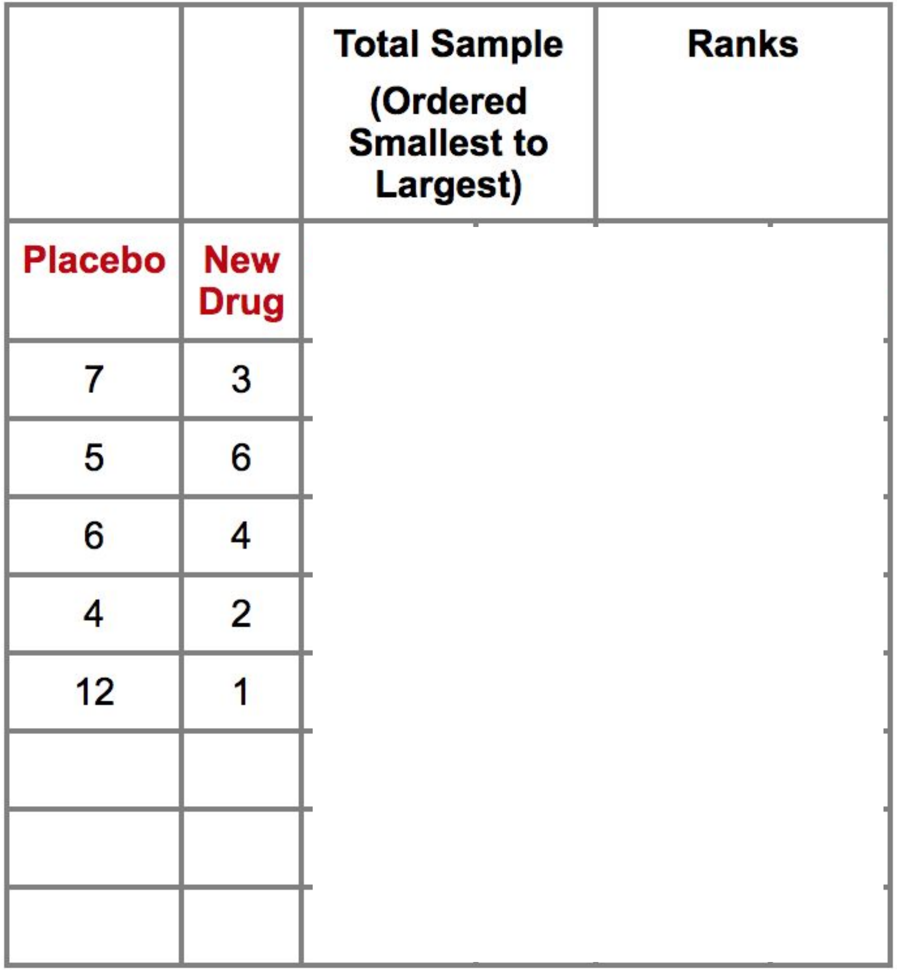
Step 2: Episodes from both groups are combined, sorted, and ranked

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				
7	3				
5	6				
6	4				
4	2				
12	1				

Sum of ranks:

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

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

$U_{Placebo} = 3$

$U_{treatment} = 22$

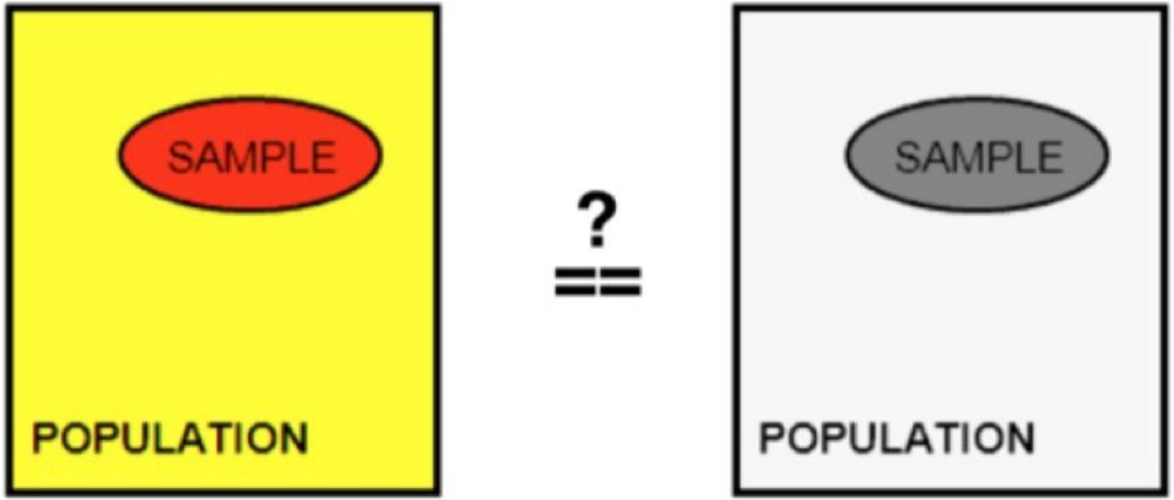
$0 < U < n_1 {*} n_2$

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?

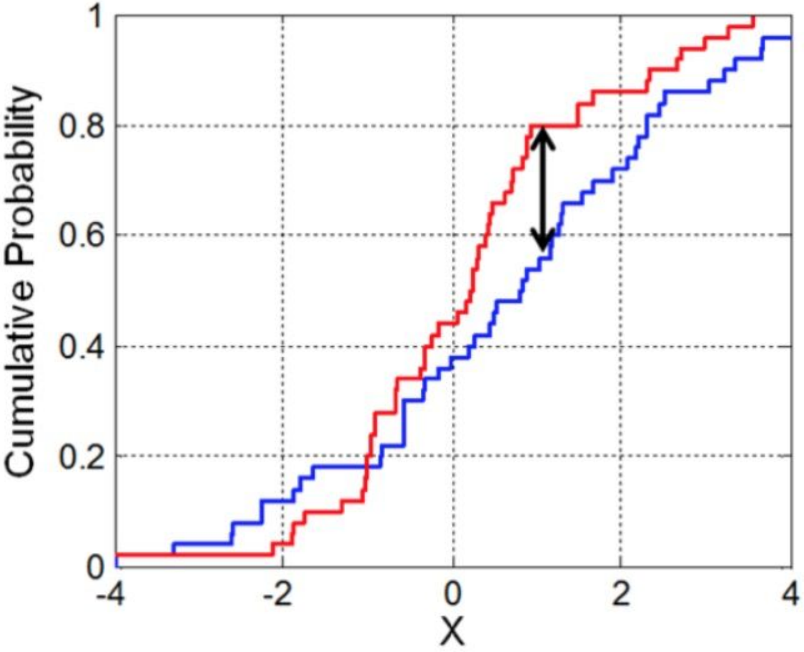


Kolmogorov-Smirnov (KS) test

Comparing cumulative distributions empirically

Tests:

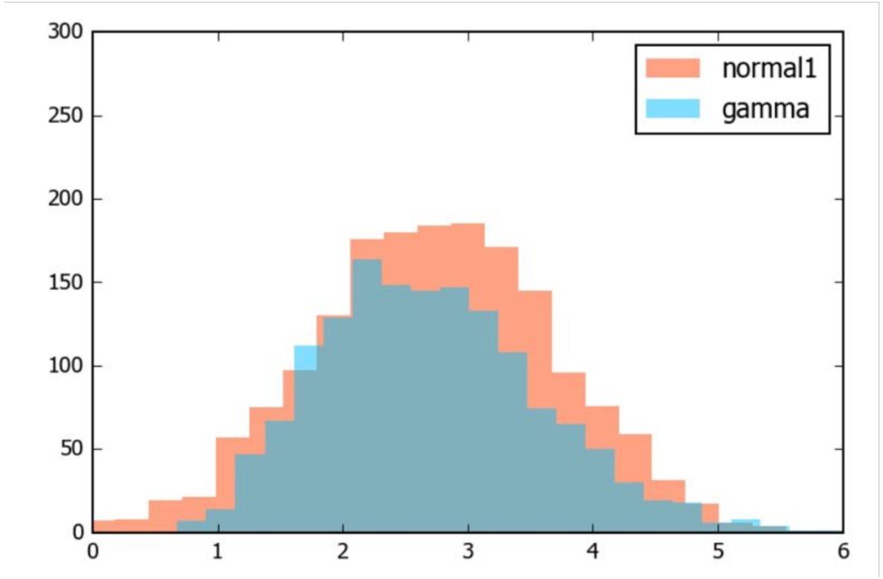
- whether a sample is drawn from a given distribution
- Whether two samples are drawn from the same distribution

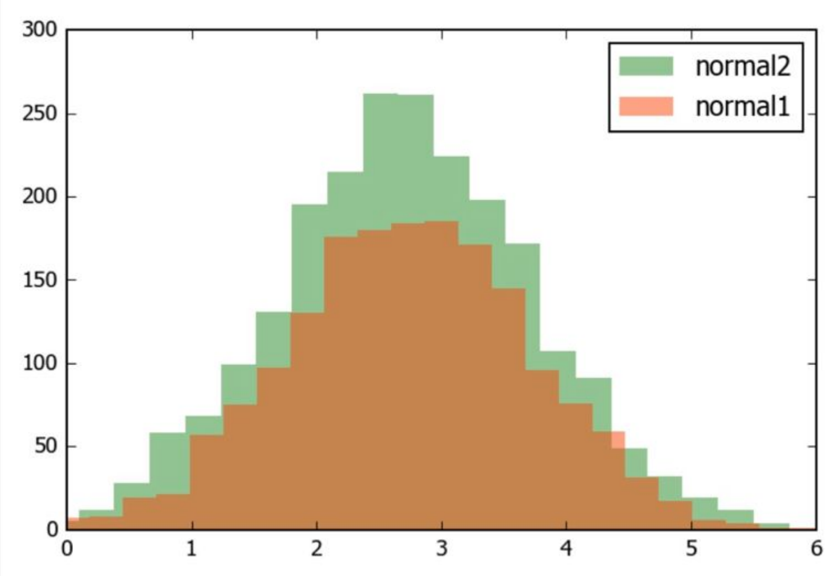


Find the maximum difference between the CDFs.

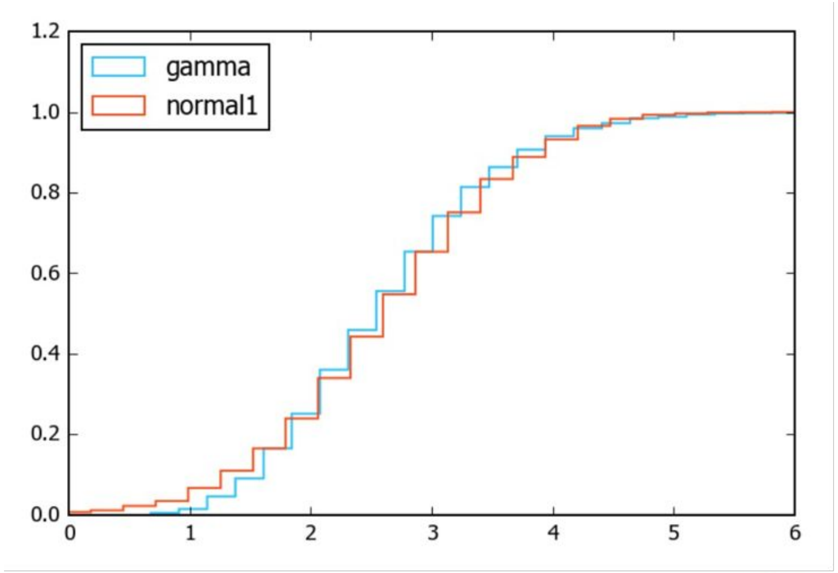
Kolmogorov-Smirnov (KS) test

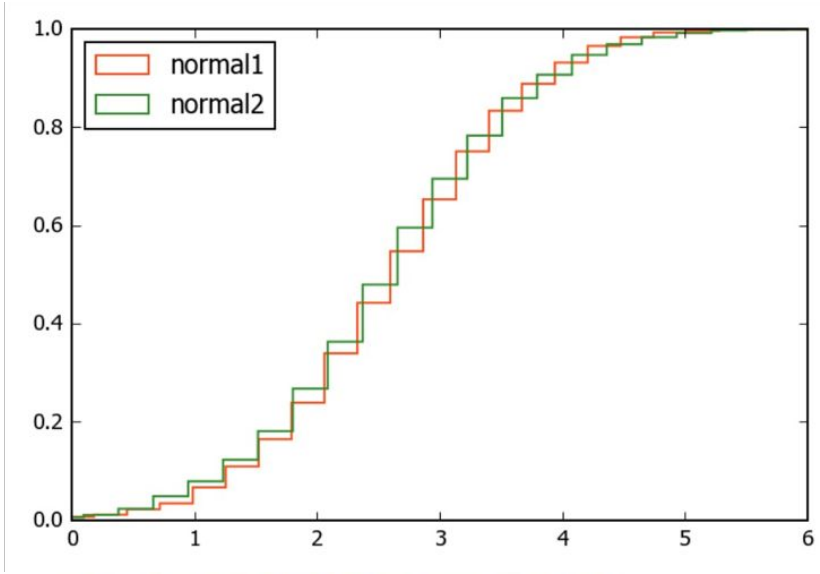
- Given (limited) samples from two populations, how do we quantify whether they come from the same distribution?





Kolmogorov-Smirnov (KS) test





```
gamma vs. normal1: p = 0.010680362841l
normal1 vs. normal2: p = 0.550735998243
```

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 (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