Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
@bradleyvoytek

COGS 108
Data Science in Practice

Quick note on correlations

UC San Diego

Simpson's Paradox

Source: Emil O.W. Kirkegaard

UC San Diego

Simpson's Paradox

Source: Emil O.W. Kirkegaard

UC San Diego

Simpson's Paradox

Source: Emil O.W. Kirkegaard

UC San Diego

Simpson's Paradox - Real world example

Example 2—Airlines on-time data: Here are numbers of flights on time and delayed for two airlines at five airports in June 1991. The table shows that Alaska Airlines outperforms America West at all 5 cities. If you collapse the table over city, it appears that America West outperforms Alaska.

	Alaska Airlines			America West Airlines		
	On time	**Delayed**	**Delay%**	**On time**	**Delayed**	**Delay%**
LA	497	62	11.1%	694	117	14.4%
Phoenix	221	12	5.4%	4840	415	7.9%
San Diego	212	20	8.6%	383	65	14.5%
San Fran.	503	102	16.9%	320	129	28.7%
Seattle	1841	305	14.2%	201	61	23.3%
Total	**3274**	**501**	**13.3%**	**6438**	**787**	**10.9%**

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive
  - You make a friendly wager: you'll both play a question-answering game twice, and whomever has the best record, wins

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive
  - You make a friendly wager: you'll both play a question-answering game twice, and whomever has the best record, wins

- **Day 1:**
  - you: 98/99 (98.99%)

- **Day 1:**
  - them: 1/1 (**100%**)

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive
  - You make a friendly wager: you'll both play a question-answering game twice, and whomever has the best record, wins

- **Day1:**
  - you: 98/99 (98.99%)
- **Day2:**
  - you: 0/1 (0%)

- **Day1:**
  - them: 1/1 (**100%**)
- **Day2:**
  - them: 1/99 (**1.01%**)

Simpson's Paradox - Intuition

- **Totally contrived example:**
  - You and a friend are both avid gamers
  - You are also both kind of competitive
  - You make a friendly wager: you'll both play a question-answering game twice, and whomever has the best record, wins

- **Day1:**
  - you: 98/99 (98.99%)
- **Day2:**
  - you: 0/1 (0%)
- **Total:**
  - you: 98/100 (**98%**)

- **Day1:**
  - them: 1/1 (**100%**)
- **Day2:**
  - them: 1/99 (**1.01%**)
- **Total:**
  - them: 2/100 (2%)

COGS 108
Data Science in Practice

Non-parametric statistics

UC San Diego

Non-parametric statistics

Central Limit reminder demo!

Non-parametric statistics - why?

Normal distribution
(nice and friendly)

We have good math tools for this.

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

Source: Richard Gao

Non-parametric statistics - why?

What if your population is distributed like this?

Source: Richard Gao

Non-parametric statistics - why?

Or like this?

Source: Richard Gao

UC San Diego

Non-parametric statistics - why?

Or like this?

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

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

Source: Richard Gao

UC San Diego

Resampling statistics

- **What & When?**
- **Kolmogorov-Smirnoff Test**
- **Rank Statistics**
- **Jackknife & Bootstrap**
- **Non-parametric prediction models**

Source: Richard Gao

Non-parametric statistics - when to use?

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

Source: Richard Gao

UC San Diego

Non-parametric statistics

- **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 a data-generating process (like heights) that result in, e.g., normally-distributed data

Source: Richard Gao

Kolmogorov-Smirnov test

Question:

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

Source: Richard Gao

Kolmogorov-Smirnov test

Question:

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

Source: Richard Gao

Kolmogorov-Smirnov test

Comparing cumulative distributions empirically

Find the maximum difference between the CDFs.

Source: Richard Gao

UC San Diego

Kolmogorov-Smirnov test

Very sensitive!

```
gamma vs. normal1: p = 0.0106803628411
normal1 vs. normal2: p = 0.550735998243
```

Source: Richard Gao

UC San Diego

Rank correlation

Spearman Correlation - Non-linear but monotonic relationships

Dampens effect of outliers!

Source: Richard Gao

UC San Diego

Bootstrapping (resampling)

Question:

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

Source: Richard Gao

Bootstrapping (resampling)

Example Question:

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

Source: Richard Gao

UC San Diego

Bootstrapping (resampling)

Source: Richard Gao

UC San Diego

Monte Carlo (also resampling) - π

- We have a circle of radius 0.5, enclosed by a 1×1 square
- The area of the circle is $\pi r^2 = \pi/4$
- The area of the square is 1
- If we divide the area of the circle by the area of the square we get $\pi/4$

Monte Carlo (also resampling) - π

- This means you can estimate π using random data!
- Count how many (uniformly) random dots appear *inside* the circle compared to how many are *outside* the circle

- $\pi \approx 4(N_{in}/N_{out})$

Monte Carlo (also resampling) - π

$n = 10^8$

```
3.1416079600000000 result
3.1415926535897931 real pi
```

Source: Wikipedia (Monte Carlo); *The Glowing Python*