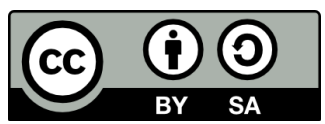
Applied Statistics

Why learn statistics?

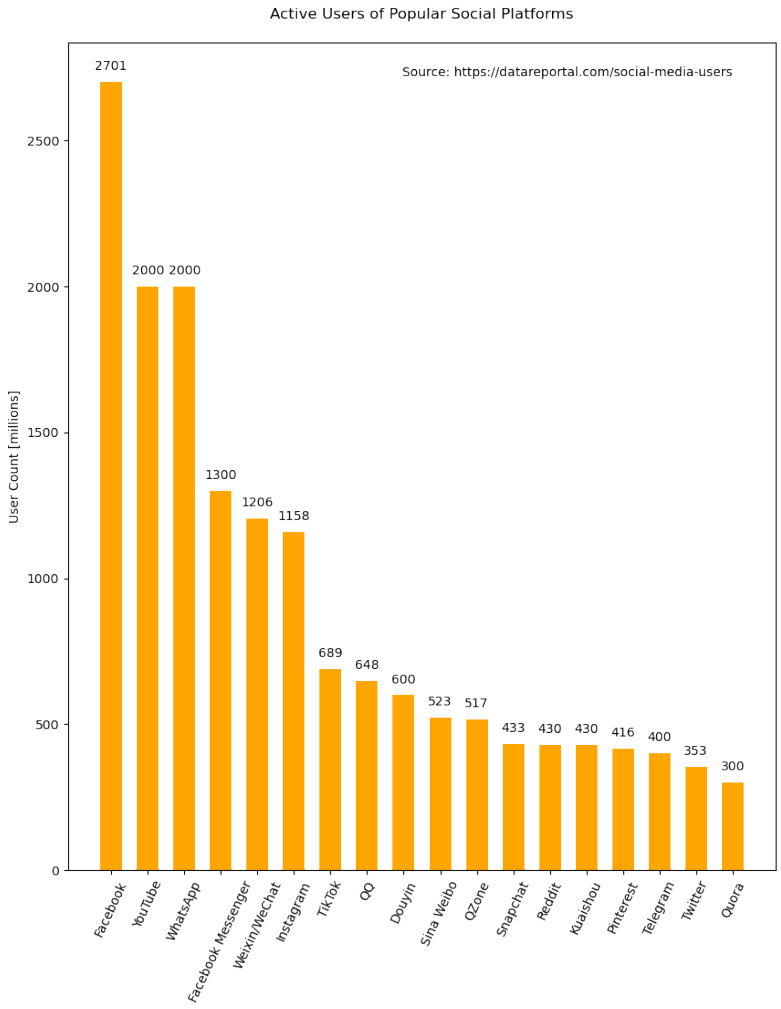
Cesar O. Aguilar
SUNY Geneseo

Portions of these notes were created from *Learning statistics with R* by Danielle Navarro and *Introduction to Statistical Thinking* by Benjamin Yakir.



Why learn statistics?

- On television and the internet, you will see statistical information
- There are statistics about crime, sports, social media, education, politics, real estate, etc.
- Typically, you are given **sample** information and you may make a decision about the correctness of a statement, claim, or fact
- Statistical methods can help you make the best educated guess



Why do statistics?

- We don't trust ourselves enough
- We are susceptible to all of the biases, temptations and frailties that humans suffer from
- Verbal arguments have to be constructed in language, and all languages have biases – some things are harder to say than others, and not necessarily because they're false
- Much of statistics is basically a safeguard from these biases
- Among the many things that psychologists have shown over the years is that we really do find it hard to be neutral, to evaluate evidence impartially and without being swayed by pre-existing biases
- A good example of this is the **belief bias effect** in logical reasoning
- If you ask people to decide whether a particular argument is logically valid, we tend to be influenced by the believability of the conclusion, even when we shouldn't

Logic vs belief conflict

- This was demonstrated in a classic study by Evans et. al (1983)[1] which focused on the alleged "belief bias" effect in reasoning
- The authors claimed that when presented with deductive arguments to evaluate (whether the arguments were logically valid), people will make judgments upon pre-existing beliefs rather than on the basis of logical argument
- Specifically, people will tend to endorse arguments whose conclusions they believe and reject arguments whose conclusions they disbelieve, irrespective of their actual logical validity
- For example, is this a valid logical statement?

  All addictive things are expensive
  Some cigarettes are inexpensive
  Therefore, some cigarettes are not addictive

[1] Evans, J.S.B.T., Barston, J.L. & Pollard, P. On the conflict between logic and belief in syllogistic reasoning. Memory & Cognition 11, 295–306 (1983)

Logic vs belief conflict

- In the study[2], it was found that when pre-existing biases (i.e., beliefs) were in agreement with the structure of the data, everything went the way you'd hope:

	conclusion believable	conclusion unbelievable
argument is valid	92% say "valid"	
argument is invalid		8% say "valid"

- But look what happens when our intuitive feelings about the truth of the conclusion run against the logical structure of the argument:

	conclusion believable	conclusion unbelievable
argument is valid	92% say "valid"	**46% say "valid"**
argument is invalid	**92% say "valid"**	8% say "valid"

[2]Evans, J.S.B.T., Barston, J.L. & Pollard, P. On the conflict between logic and belief in syllogistic reasoning. Memory & Cognition 11, 295–306 (1983)

Simpson's paradox

- **Simpson's paradox** is a phenomenon in statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined
- One of the best-known examples of Simpson's paradox comes from a study of gender bias among graduate school admissions to UC Berkeley

Simpson's paradox

- In 1973, UC Berkeley had some worries about the gender breakdown of their admissions:

	Number of applicants	Percent admitted
Males	8,442	44%
Females	4,321	35%
Total	12,763	41%

- A difference of 9% in admission rates between males and females is just way too big to be a coincidence

Simpson's paradox

- However, on a department by department basis, it turned out that most of the departments actually had a slightly *higher* success rate for female applicants than for male applicants:

	Males		Females	
Department	Applicants	Percent admitted	Applicants	Percent admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

- Remarkably, most departments had a *higher* rate of admissions for females than for males!
- Yet the overall rate of admission across the university for females was *lower* than for males
- How can this be? How can both of these statements be true at the same time?

Simpson's paradox

- First, notice that some departments (e.g., A, B) tended to admit a high percentage of the qualified applicants, whereas others (e.g., F) tended to reject most of the candidates, even if they were high quality:

	Males		Females	
Department	Applicants	Percent admitted	Applicants	Percent admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Simpson's paradox

- Next, notice that males and females tended to apply to different departments:

	Males		Females	
Department	Applicants	Percent admitted	Applicants	Percent admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

- If we rank the departments in terms of the total number of male applicants, we get **A**>**B**>D>C>F>E (the "easy" departments are in bold)

- On the whole, males tended to apply to the departments that had high admission rates

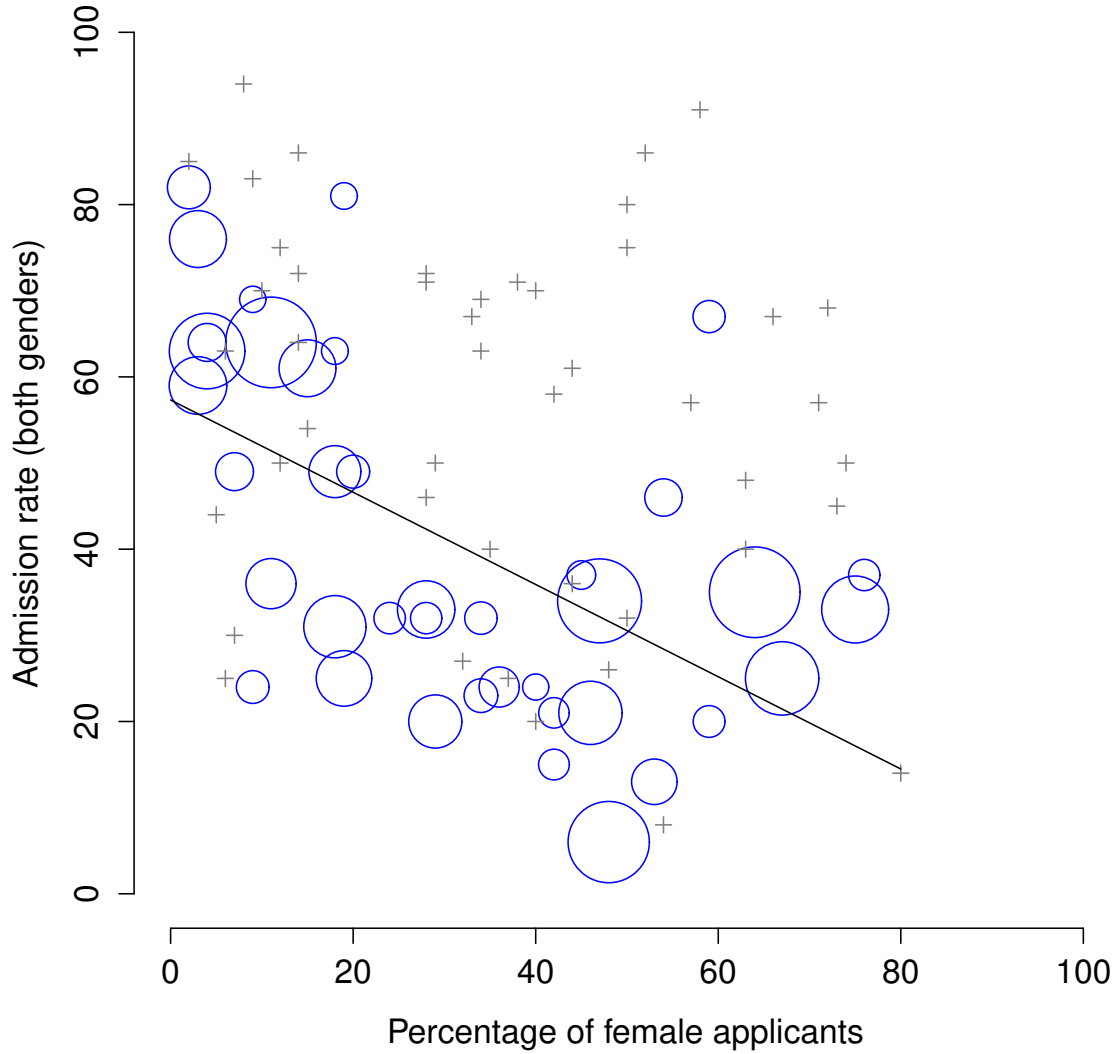
Simpson's paradox

- Ranking the departments in terms of the total number of female applicants produces a quite different ordering C>E>D>F>**A**>**B**:

	Males		Females	
Department	Applicants	Percent admitted	Applicants	Percent admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

- In other words, what these data seem to be suggesting is that the female applicants tended to apply to "harder" departments

Simpson's paradox



Admission rate per department as a function of the percentage of applicants that were female; the area of the circle is proportional to the total number of applicants.

Why do statistics?

- **Lesson**: Doing research is hard, and there are lots of subtle, counter-intuitive traps lying in wait for the unwary
- Statistics only solves part of the problem
- If we're interested in this from a more sociological and psychological perspective, we might want to ask why there are such strong gender differences in applications
- Why do males tend to apply to engineering more often than females, and why is this reversed for the English department?
- And why is it the case that the departments that tend to have a female-application bias tend to have lower overall admission rates than those departments that have a male-application bias?
- At least we are asking questions based on a more detailed analysis of the data

What is statistics?

- The science of **statistics** deals with the collection, analysis, interpretation, and presentation of data

Descriptive Statistics	Inferential Statistics
• Data is summarized from a **sample** using numerical values such as the mean or standard deviation	• Conclusions are drawn from the data that are subject to **random** variation
• Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population	• The goal of inferential statistics is to infer properties of a population, for example by testing hypotheses and deriving estimates; it is assumed that the observed data set is sampled from a larger population