

# Statistical Terminology

# Correlation

# What is correlation?

- A statistical measurement that shows the relationship between two variables.
- Example: Height & Weight

# Pearson's Correlation Coefficient

Pearson's  $r$ , which measures a 'normalized' covariance (how changes in one value are associated with those of another), has a value between -1 and 1

- 1 – perfect positive linear correlation
- 0 – no linear correlation
- -1 – perfect negative linear correlation

# Correlation types

- Positive correlation
  - An increase in one variable will lead to an increase in the other
- Negative correlation
  - An increase in one variable will lead to a decrease in the other
- Note: In System Dynamics, these are called Positive and Negative Feedback loops

# Knowledge check

- Height and weight
- Vehicle speed and travel time
- Gasoline prices and global oil production
- Caloric intake and weight
- Hours spent watching TV and school grades
- Car value and car mileage

# Case Study

In WWII, the US formed the Statistical Research Group to analyze situations like the following:

You don't want your planes shot down by enemy fighters, so you armor them. Armor makes the plane heavier, and heavier planes are slower and use more fuel. Too much armor and too little armor is bad. Where do you armor them?

## Case Study (cont)

When planes returned from missions, damage was unevenly distributed. The fuselage and fuel system would often have many bullet holes whereas the engines would have few. Should you put more armor on the fuselage?



# Case Study (cont)

Tip: Set a variable to zero to test the probability.

Ex.: By imagining that a plane is CERTAIN to be hit in the engine, the plane is CERTAIN to crash because planes can't fly without working engines.

Either German planes happen to hit every part of a plane but the engine, or the engine is a point of total vulnerability.

# Case Study

In WWII, bombing accuracy had a high positive correlation with fighter opposition. The more fighters, the better the bombing accuracy. Why?

# Case Study

In WWII, bombing accuracy had a high positive correlation with fighter opposition. The more fighters, the better the bombing accuracy. Why?

Cloud cover. If there are too many clouds, fighters aren't launched and bombers are inaccurate.

# Case Study

Statisticians often report that in assessing a child's likeliness to succeed at school, those children whose parents played classical music recordings for the unborn children will result in better grades. Why true?

# Case Study

Statisticians often report that in assessing a child's likeliness to succeed at school, those children whose parents played classical music recordings for the unborn children will result in better grades. Why true?

Adopting such a parental strategy indicates the parents are interested in the child's intelligence.

# Case Study

Economist recently announced that statistics prove the taller you are, the more you are likely to be paid. Why?

# Case Study

Economist recently announced that statistics prove the taller you are, the more you are likely to be paid. Why?

The lurking variable is more likely gender, as typically men are on average taller than women.

# Statistical Terminology



# Essential Terms

## Sample

- A portion of a population

## Stratified Sample

- The sample is chosen to reflect the population at large

## Random Sample

- The sample is chosen by chance

# Essential Terms

## Generalization

- Extending conclusions from the sample to the population. Only possible if sample is reflective.

## Causation

- When changes in one variable affect the other

## Elasticity

- How much a change in one variable affects the other

## Bias or Skew

## Margin of Error

# Essential Terms

## Mean

- The average after adding and dividing all data

## Median

- The middle number of a dataset

## Mode

- Number(s) appearing most often in a dataset

# Essential Terms

## Discrete variable

- A variable with a finite amount of values

## Continuous variable

- A variable with many different values in a range

# Variance

“The average of the squared differences from the mean” □ how different the data is

Ex.: [12, 12, 12, 12, 12]

- Variance = 0

Ex.: [12, 12, 12, 12, 13]

- Variance = 0.16

Ex.: [12, 12, 12, 12, 13013]

- Variance = 27,044,160

# Standard Deviation

The square root of the variance (more precise than variance) □ This is the main reason for variance

# Standard Deviation Example

Dataset of height of cats in cm:

[600, 470, 170, 430, 300]

Find the variance (Find the mean, calculate the difference of each datum from the mean, square, then average).

- 21,704

Find the standard deviation (square root of the mean).

- ~147

# Standard Deviation Example

If the standard deviation is 147, then a datum is “1 standard deviation from the mean”. A datum “2 standard deviations is 296” and so on...

HOWEVER...

This is has been a ‘population’ standard deviation where each possible value was considered.



# Sample Standard Deviation

How is the sample SD different from the population SD? How to correct the calculation?

Divide by 'n-1' instead of 'n' when finding both the variance and SD. Now find the sample SD and sample variance of the previous dataset.

Sample variance = 27,130

Sample SD = 164

# Essential terms

- Regression Analysis: estimates relationships between X and Y variables
- Null hypothesis: Assumes no significant difference (states alternative hypothesis false)
- P-value: indicates strong evidence against null hypothesis ( $x \leq 0.05$ ), or weak evidence ( $x > 0.05$ ) □ “Statistical significance”

# Descriptive & Inferential Stats

- Descriptive statistics
  - Describes what's happening in a dataset
- Inferential statistics
  - Generalizes sample findings to population

# Descriptive & Inferential Stats

- 50% of all Russian men are named Ivan.
- 20% of respondents are male
- From 2000 to 2005, 70% of the land cleared in the Amazon and recorded in Brazilian government data was transformed into pasture.
- Receive your college degree increases your lifetime earning by 50%.
- Teachers named Joshua demonstrate inferior intellect to teachers named Timmy.

# Question Design

# Open-ended vs. Closed-ended

- Open-ended  No response options provided
- Closed-ended  A list of options provided

## Fewer People Mention Economy in Open-Ended Version

What one issue mattered most to you in deciding how you voted for president?

	Open-ended <sup>1</sup>	Closed-ended <sup>2</sup>
<b>The economy</b>	<b>35%</b>	<b>58%</b>
<b>The war in Iraq</b>	5	10
<b>Health care</b>	4	8
<b>Terrorism</b>	6	8
<b>Energy policy</b>	*	6
<b>Other</b>	43	8
Candidate mentions	9	-
Moral values/social issues	7	-
Taxes/dist. of income	7	-
Other issues	5	-
Other political mentions	3	-
Change	3	-
Other	9	-
<b>Don't know</b>	7	2
	100	100

Data from Pew Research November 2008  
Post-election survey

<sup>1</sup> Unprompted first response to open-ended question.

<sup>2</sup> First choice from five options read to respondents.

# Open-ended vs. Closed-ended

- Open-ended used in pilot studies to determine most common options
- Subjective closed-ended  Fewer options
  - Satisfaction with economy
  - Fewer options avoids “recency effect”
  - Randomized order to ensure random bias
- Objective closed-ended  More options fine
  - Religious affiliation



# Closed-ended example

Form a question asking about a student's satisfaction with their high school education (hint  use ordinal categories).

How can you mitigate the recency effect?

# Question wording

Be aware of information and connotations in questions.

“Do you favor or oppose taking military action against Saddam Hussein?”

Favor = 68%; Oppose = 25%

# Question wording

Be aware of information and connotations in questions.

“Do you favor or oppose taking military action against Saddam Hussein even if it meant that U.S. forces might suffer thousands of casualties?”

Favor = 43%; Oppose = 48%

# Question Structure

- All reasonable responses included if closed.
- Responses shouldn't overlap.
- One question at a time.
  - Bad: “How much confidence do you have in Obama to handle domestic and foreign policy?”

# Question Structure

## Leading questions

- Do you think that the new cafeteria lunch menu offers a better variety of healthy foods?

## Neutral questions

- How do you feel about the new cafeteria lunch menu compared to the old one?

# Social desirability bias

- Sensitive issues lead to misreporting
- Understated alcohol/drug use, tax evasion
- Overstated donations, church attendance
- SDB higher when interviewer is present
- Include 'Prefer Not to Answer' option

# Question Order

- Contrast effects
  - When order results in greater differences in responses
- Assimilation effects
  - When responses are similar because of order

# Question Order

## More People Favor Civil Unions When Asked After Gay Marriage

<b>Asked first</b>	<i>Legal agreements</i>	%	<i>Gay marriage</i>	%
	Favor	<b>37</b>	Favor	33
	Oppose	55	Oppose	61
	Don't know	<u>8</u>	Don't know	<u>6</u>
		100		100
<b>Asked second</b>	<i>Gay marriage</i>		<i>Legal agreements</i>	
	Favor	30	Favor	<b>45</b>
	Oppose	58	Oppose	47
	Don't know	<u>12</u>	Don't know	<u>8</u>
		100		100
N		780		735



# Question Order

## More Overall Dissatisfaction When Asked After Bush Approval

<b>Asked first</b>	<i>Overall satisfaction</i>	<i>%</i>	<i>Bush approval</i>	<i>%</i>
	Satisfied	17	Approve	25
	Dissatisfied	<b>78</b>	Disapprove	67
	Don't know	<u>5</u>	Don't know	<u>8</u>
		100		100
<b>Asked second</b>	<i>Bush approval</i>		<i>Overall satisfaction</i>	
	Approve	24	Satisfied	9
	Disapprove	68	Dissatisfied	<b>88</b>
	Don't know	<u>8</u>	Don't know	<u>3</u>
		100		100
N		766		723

# Question Order

## More Endorse Working Together When Asked Second

<b>Asked first</b>	<i>Should Rep. leaders...</i>	%	<i>Should Dem. leaders...</i>	%
	Work with Obama	<b>66</b>	Work with Rep. leaders	<b>82</b>
	Stand up to Obama	28	Stand up to Rep. leaders	13
	Don't know	<u>6</u>	Don't know	<u>5</u>
		100		100
<b>Asked second</b>	<i>Should Dem. leaders...</i>		<i>Should Rep. leaders...</i>	
	Work with Rep. leaders	<b>71</b>	Work with Obama	<b>81</b>
	Stand up to Rep. leaders	21	Stand up to Obama	15
	Don't know	<u>8</u>	Don't know	<u>4</u>
		100		100
N		744		756

PEW RESEARCH CENTER Nov. 2008 Post-election survey.

# The Good, The Bad, & The Ugly

- How likely would you be to enroll in CookieDirect?
- How organized and interesting was the speaker?
- How helpful do you think our customer service representatives are?
- Should the government force you to pay higher taxes?
- How would you rate the career of legendary writer Dvlatov?

# The Good, The Bad, & The Ugly

- What do you like to do for fun?
- How dumb is President Trump at making America great again?
- Should teachers named Joshua offer pizza parties to obedient students?
- In your opinion, how would you rate the quality of your work?
- How do you feel about the following statement?  
We should reduce military spending.

# Create your own survey

- Select a dependent variable and at least two independent variables to test. Examples:
  - Playing Online Multiplayer games vs Multicultural Awareness & Open-mindedness
  - Amount of time spent on social media vs level of English & extroverted/introverted
  - Reading habits vs social media habits & cultural activities