Math 9 Lesson 7: Statistics

- Statistics in society
- Population versus sample, bias, ethics, sampling techniques, misleading stats
- Analyzing a given set of data (and/or its representation) and identifying potential problems related to bias, use of language, ethics, cost, time and timing, privacy, or cultural sensitivity
- Using First Peoples data on water quality, Statistics Canada data on income, health, housing, population

1. What is the mean of 2, 4, and 9?
   The word "mean" is equivalent to "average"
   $\frac{2+4+9}{3} = 5$

2. What is the average of the numbers $10, 20, 80, 30$.
   $\frac{10+20+80+30}{4} = 35$

3. $2, 5, k$
   Given the mean is 6, find k
   $\frac{2+5+k}{3} = 6$
   $2 + 5 + k = 18$
   $7 + k = 18$
   $k = 18 - 7 = 11$

4. Find the median of the following numbers:
   a. 1, 2, 4, 10, 50, 1000, 999999
      10

   b. 2, 10, 20, 5000
      $\frac{10+20}{2} = 15$

   c. $-20, 0, 100, 50, -1000$
      $-1000, -20, 0, 50, 100$
      0

5. Find the range of the numbers $10, -20, 5, 100$
   $100 - (-20) = 100 + 20 = 120$

6. Find the average of the numbers $0.5, \frac{3}{2}, -8$, and 2000%
   $0.5, 1.5, -8, 20$
   $\frac{0.5+1.5-8+20}{4} = 3.5$ or $\frac{7}{2}$

7. YouTube: Political media's bias in a single chart
   What is the criteria that defines a "respectable" news source?

    a. Accuracy:
        i. Fact-checking: Reputable sources rigorously verify the facts in their reporting.
        ii. Corrections: They promptly and transparently correct errors when they occur.

    b. Editorial Independence:
        i. Freedom from Bias: While complete neutrality is often unattainable, high-quality sources strive for impartiality, clearly separating news from opinion.
        ii. No Undue Influence: They avoid conflicts of interest, especially financial ones, that could sway editorial content.

    c. Transparency:
        i. Source Disclosure: They reveal their sources when possible, or explain why sources remain anonymous.
        ii. Funding: Clear disclosure of ownership and funding to ensure there's no hidden agenda.

    d. Ethical Standards:
        i. Journalistic Integrity: Adherence to a code of ethics, which includes respect for privacy, avoiding sensationalism, and not manipulating content misleadingly.
        ii. Accountability: They are accountable for their reporting, responding to public queries or criticisms.

    e. Quality of Reporting:
        i. Depth and Context: Stories are not just reported but are placed in context, offering background and implications.
        ii. Diverse Perspectives: They aim to represent a broad spectrum of views, especially on complex or controversial issues.

    f. Editorial Process:
        i. Professional Journalism: Employment of trained journalists who follow established journalistic practices.
        ii. Editing: Thorough editing to ensure clarity, coherence, and adherence to journalistic standards.

    g. Audience Engagement:
        i. Responsiveness: They engage with their audience, often providing platforms for feedback or clarification.
        ii. Educational Mission: A commitment to informing the public, not just entertaining or provoking.

    h. Reputation:
        i. Peer Recognition: Often recognized or awarded by peers for excellence in journalism.
        ii. Longevity: A history of consistent, quality journalism can be a marker of respectability.

    i. Legal and Regulatory Compliance:

        i. Adherence to Laws: Operates within the legal frameworks of journalism, including libel laws, privacy rights, etc.

    j. Adaptability to New Media:
        i. Innovation in Delivery: Embraces new technology and platforms to disseminate news effectively while maintaining journalistic integrity.

However, the perception of a news source's respectability can also vary significantly based on cultural, political, or personal biases. Therefore, while these criteria provide a framework, the exact application can differ. One might also consult media watchdog organizations like Reporters Without Borders, the Committee to Protect Journalists, or local equivalents for more insight into the credibility of news sources globally or in specific regions.

Remember, even with all these checks, it's crucial for consumers to be media literate, cross-check facts, and consume news from multiple sources to get a well-rounded view.

8. YouTube: Ted X – How statistics can be misleading
    a. How can statistics be misleading?

    Statistics can be misleading in several ways, often due to manipulation, misinterpretation, or inherent biases in data collection and analysis. Here are some common ways statistics can mislead:

        i. Cherry-Picking Data:
            1. Selective Reporting: Only presenting data that supports a particular narrative while ignoring data that contradicts it. This can give a false impression of the situation.
            2. Time Frame Selection: Choosing a specific time frame where the data shows the desired trend, which might not be representative of the long-term trend.

        ii. Misleading Visual Representation:
            1. Graph Manipulation: Altering scales on graphs (like a y-axis that doesn't start at zero), using 3D effects that distort proportions, or choosing inappropriate chart types can exaggerate or minimize differences.
            2. Data Visualization Choices: Using visuals that are not suited to the data type, leading to misinterpretation (e.g., pie charts for data that should be shown in a bar chart for clarity).

        iii. Sampling Bias:
            1. Non-Representative Samples: If the sample is not representative of the population, the statistics can be misleading. For example, conducting a survey in one city and extrapolating results to the entire country.
            2. Self-Selection Bias: When participants choose to be in the survey, leading to results that might not reflect the broader population's views or behaviors.

        iv. Causation vs. Correlation:
            1. Confusing Correlation for Causation: Just because two variables are correlated does not mean one causes the other. This can lead to misleading conclusions about cause and effect.

        v. Misuse of Averages:
            1. Mean vs. Median: Using the mean (average) when the median would be more appropriate can skew perceptions, especially with skewed data where outliers significantly affect the mean.

2. Ignoring Distribution: Focusing on one measure of central tendency without considering the spread or skewness of the data can hide important insights.

vi. Data Dredging (P-hacking):
   1. Multiple Comparisons: Conducting numerous statistical tests until something significant appears by random chance, then only reporting that result without context.

vii. Overgeneralization:
   1. Extrapolation: Making predictions or generalizations far beyond the data collected. This can lead to assumptions that are not supported by the actual data.

viii. Lack of Context:
   1. Without Baseline: Presenting statistics without a baseline for comparison can make changes appear more dramatic than they are.
   2. Ignoring External Factors: Not considering external variables that might affect the data, leading to an oversimplified or incorrect interpretation.

ix. Rounding Errors or Precision:
   1. Over-precision: Reporting numbers with more decimal places than the data collection or measurement warrants, giving a false sense of accuracy.

x. Framing Effects:
   1. Contextual Framing: How data is framed (e.g., "95% fat-free" vs. "contains 5% fat") can change how it's perceived even though the information is identical.

To combat misleading statistics, consumers should:
Look for the source of data and methodology.
Check for peer review or corroboration by other sources.
Understand the context of the data, including sample size, methodology, and potential biases.
Be skeptical of overly precise or dramatic claims.
Learn basic statistical literacy to recognize common tricks in data presentation.

Ultimately, being an informed consumer of statistics involves critical thinking and sometimes seeking out additional data or expert analysis to interpret the numbers accurately.

b. What is Simpson’s Paradox?

Simpson's Paradox, also known as the Yule-Simpson effect, is a phenomenon in probability and statistics where trends that appear in different groups of data can reverse or disappear when these groups are combined. Here's how it typically manifests:

Definition: Simpson's Paradox occurs when:

i. Trends within individual groups go in one direction (e.g., a positive correlation or one group performing better than another), but
ii. When these groups are combined, the overall trend goes in the opposite direction or the relationship disappears.

Example:
Consider a simplified scenario about the success rate of treatments in two hospitals:

Hospital A:

- Treatment 1: 5/5 success in males, 20/40 success in females.
- Treatment 2: 10/30 success in males, 40/60 success in females.

If you look at each gender separately:

- Males: Treatment 1 has a higher success rate (100%) than Treatment 2 (33.3%).
- Females: Treatment 1 has a lower success rate (50%) than Treatment 2 (66.7%).

However, when you combine the data for both genders:

- Treatment 1: (5+20)/(5+40) = 25/45 ≈ 55.6% success rate.
- Treatment 2: (10+40)/(30+60) = 50/90 ≈ 55.6% success rate.

In the combined data, both treatments appear equally effective, which masks the gender-specific differences. Here, Simpson's Paradox has made it look like there's no difference between treatments when, in fact, for each gender, there is a clear preference for one treatment over the other.

Causes:

- Confounding Variables: Often, there's an underlying variable (like gender in the above example) that affects the outcome. When not accounted for, it can lead to paradoxical results.
- Data Aggregation: When data from different groups or under different conditions are pooled together, the nuances in each subgroup can be lost.

Implications:

- Misleading Conclusions: Without a careful breakdown by subgroup, one might make incorrect decisions or draw wrong conclusions from data.
- Importance of Stratification: This paradox underscores the need to analyze data at multiple levels (stratified analysis) rather than just looking at the totals.
- Policy and Decision Making: In fields like medicine, economics, or education, ignoring this phenomenon can lead to suboptimal policies or treatments.

When dealing with statistical data, it's crucial to:

- Check for Confounding Variables: Always look for and control for variables that might affect your outcome differently in different groups.
- Perform Stratified Analysis: Examine trends within subgroups before looking at the overall data.
- Be Cautious with Aggregated Data: Remember that combined statistics might not tell the whole story.

Understanding Simpson's Paradox helps in making more informed decisions by revealing hidden trends that might be obscured by aggregate data.

9. YouTube: “This is How Easy it is to lie with statistics”
How can statistics be used to deceive others?

ex. Bringing in a mathematician to prove the guilt of the couple
Black man with beard 1 in 10
Man with mustache 1 in 4
White woman with pony tail 1 in 10
White woman with blonde hair 1 in 3
Yellow motor car 1 in 10
Interracial couple in car 1 in 1000
Multiply resulting in 1 in 12 million chance that this random couple who just happen to fit all these descriptions (resulting in a guilty verdict)
This is the prosecutor’s fallacy
People assume $P(A|B) = P(B|A)$
Given an animal has 4 legs, what is the probability that it is a dog?

vs. Given a dog, what is the probability it has 4 legs?

Given innocent couple, the probability of fitting all descriptions is 1 in 12 million
Now Given couble fits all descriptions, suppose 10 people in the entire city fit these descriptions
Then there is a 1 in 10 chance of being guilty

1996 Sudden Infant Death Syndrome
Again a year later other son died
Chance of back to back SIDS about 1 in 73 million (Sally Clark found guilty)
Assumed that first and second death were independent (but actually the deaths could be genetic)

In 2007 Colgate claimed that 80% of dentists recommend Colgate
However 100% recommend Crest (misleading) – dentists could choose more than one brand

If high school dropout rate goes from 5% to 10% in one year is that a 5% increase or a 100% increase?

Number of blood clots go up by 100% (when in actuality 1 in 7000 became 2 in 7000)

Head lice is good for people?
Ex. People who were sick rarely had head lice (but lice is very sensitive to body temperature)
Correlation vs. Causation

Third-Cause Fallacy
Ice Cream Sales does not cause Heatstrokes nor do Heatstrokes cause Ice Cream Sales
But rather How weather causes both

Zoomed in bar chart

10. YouTube: Ted-Ed How can you spot a misleading graph?
    How can graphs be misleading?
    Ex. Zooming in

11. When conducting a survey, why do we poll a sample instead of the entire population?
    limited resources

12. Give an example of unethical statistical practice.
    Producing false data

13. Define the following sampling techniques
    a. Convenience sampling
       Convenience Sampling is a non-probability sampling technique where members of the target population are selected for the sample based on their accessibility and proximity to the researcher.

    b. Simple random sampling
       Simple Random Sampling is a probability sampling method where each member of the population has an equal and independent chance of being selected for the sample.

    c. Stratified sampling
       Stratified Sampling is a probability sampling method that involves dividing the population into distinct subgroups, or strata, and then randomly selecting individuals from each stratum. The goal is to ensure that each subgroup is adequately represented in the sample, thereby improving

the accuracy and representativeness of the results.

14. Explain how the following can be misleading
    a. Manipulating scales
       Zooming in

    b. Cherry-picking data
       Picking what supports your arguments

    c. Small samples
       Not reliable

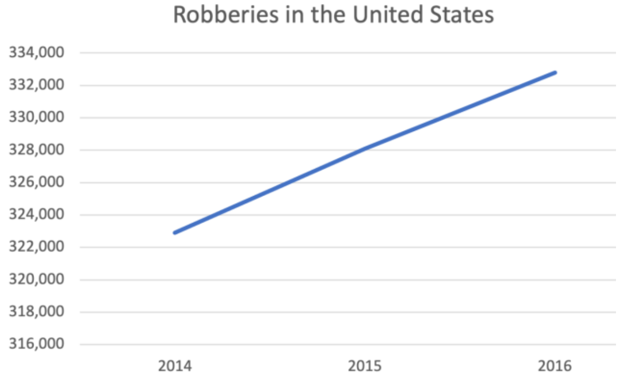
    d. Non-representative samples
       Taking free throw percentage of NBA athletes vs. average high school athletes
    e. Selective data reporting
       We don't get the whole truth

    f. Showing averages
       Averages are pulled towards the "whales"

    g. Poorly written surveys
       ex. Quebec referendum

15. Give an example of how correlation does not imply causation.
    ex. Ice cream sales and home invasions

16. What is misleading about the following graph?



    The scale exaggerates the growth

17. What is the median household income in BC, Canada?
    $82,000

18. What is the median household income in Canada in US dollars?
    How does this amount compare to the median household income in the US?
    $48,222 USD vs. $80,610

19. What is the average cost of a home in BC?
    979K
    1.3 Million in Greater Vancouver
    The median cost of a detached home in Surrey, BC, is around $1.2 million (sounds low)

20. What is the population of Canada?
    41.5 million

21. What percent of BC students graduate with a Dogwood Diploma?
    90%

22. What percentage of Canadians (aged 25 to 64) have a
    post-secondary degree or diploma?
    58%

23. If the national debt were divided equally among Canadians,
    how much would each person owe?
    $98K (vs. USA 107K)

24. Was the 1995 Quebec referendum question well-written?
    No, it favored separation (49.4% in favor of separation)

    The Quebec referendum question in 1995 was:
    "Do you agree that Quebec should become sovereign, after having made a formal offer to Canada for a new economic and political partnership, within the scope of the Bill respecting the future of Quebec and of the agreement signed on June 12, 1995?"

25. What percent of First Nations communities are under long-term boil water advisories?
    According to the most recent data, approximately 10% of First Nations communities in Canada are under long-term boil water advisories. (close to 0% in the lower mainland)