Variables and Data Presentation

Richard Ssekitoleko
Yale University

Follow this and additional works at: https://escholarship.umassmed.edu/liberia_peer

Part of the Clinical Epidemiology Commons, Epidemiology Commons, Family Medicine Commons, Infectious Disease Commons, Medical Education Commons, and the Statistics and Probability Commons

Repository Citation
https://escholarship.umassmed.edu/liberia_peer/6

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in PEER Liberia Project by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Variables and Data presentation

Richard Ssekitoleko
Department of Global Health
Yale University
Objectives

By the end of this session you should be able to:

• Recognise different types of variables
• Explain how different types of variables are described
  • Using graphs
  • Using descriptive statistics
  • Understand associations between variables
Variables

• Variable
  • A characteristic that is observed or manipulated
  • Can take on different values e.g. height, weight, blood pressure
Types of variable

• Independent and dependent variables
  - Does smoking cause lung cancer
• Numerical
  • continuous
  • discrete (counts)
• Categorical
  • ordered categorical (ordinal)
  • unordered categorical (nominal)
    • dichotomous / binary
Ordered categorical

Occupational Social Class

Unordered categorical

Ethnicity

Dichotomous

Gender
Continuous variables

Discrete variables

Height
Temperature
Weight

1 2 3
What type of variable?

- Age
- Whether HIV +
- Size of tumour
- Stage of disease
- Number of siblings
Levels of measurement

• There are 4 levels of measurement
  • Nominal, ordinal, interval, and ratio

1. Nominal
  • Data are coded by a number, name, or letter that is assigned to a category or group
  • Examples
    • Gender (e.g., male, female)
    • Treatment preference (e.g., manipulation, mobilization, massage)
Levels of measurement (cont.)

2. Ordinal
   • Is similar to nominal because the measurements involve categories
   • However, the categories are ordered by rank
   • Examples
     • Pain level (e.g., mild, moderate, severe)
     • Medical rank (e.g., undergraduate, Intern, medical officer, Registrar, Consultant)
Levels of measurement (cont.)

• Ordinal values only describe order, not quantity
  • Thus, severe pain is not the same as 2 times mild pain
• The only mathematical operations allowed for nominal and ordinal data are counting of categories
  • e.g., 25 males and 30 females
Levels of measurement (cont.)

3. Interval
   • Measurements are ordered (like ordinal data)
   • Have equal intervals
   • Does not have a true zero
   • Examples
     • The Fahrenheit scale, where 0° does not correspond to an absence of heat (no true zero)
     • In contrast to Kelvin, which does have a true zero
Levels of measurement (cont.)

4. Ratio
   • Measurements have equal intervals
   • There is a true zero
   • Ratio is the most advanced level of measurement, which can handle most types of mathematical operations
Levels of measurement (cont.)

• Ratio examples
  • Range of motion
    • No movement corresponds to zero degrees
    • The interval between 10 and 20 degrees is the same as between 40 and 50 degrees
  • Lifting capacity
    • A person who is unable to lift scores zero
    • A person who lifts 30 kg can lift twice as much as one who lifts 15 kg
Levels of measurement (cont.)

- NOIR is a mnemonic to help remember the names and order of the levels of measurement
  - Nominal
  - Ordinal
  - Interval
  - Ratio
Collecting observations

*Data set:*

- Collection of observations on a variable
- Typical data set is often represented with a matrix of information.
- Each row represents an individual or unit, while each column represents a variable

<table>
<thead>
<tr>
<th>No</th>
<th>age</th>
<th>sex</th>
<th>Ht</th>
<th>wt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13</td>
<td>f</td>
<td>138</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>M</td>
<td>142</td>
<td>40</td>
</tr>
</tbody>
</table>
Summarizing Data

• Descriptive statistics
  • deal with the enumeration, organization, and graphical representation of data.

• Inferential statistics
  • deal with reaching conclusions from incomplete information, that is, generalizing from the specific sample
Descriptive Studies

• Do not examine associations

• Just summarize outcomes or exposure

• Exposure or outcomes described in terms of time, place or person
Descriptive statistics

• A way to summarize data from a sample or a population

• Illustrate the *shape*, *central tendency*, and *variability* of a set of data
  • The shape of data has to do with the frequencies of the values of observations
DSs (cont.)

- Central tendency describes the location of the middle of the data
- Variability is the extent values are spread above and below the middle values
  - a.k.a., Dispersion
- DSs can be distinguished from inferential statistics
  - DSs are not capable of testing hypotheses
Descriptive statistics for continuous data

• Measures of central tendency
  • Mean
  • Mode
  • Median

• Measures of dispersion
  • SD
  • IQR
  • Range
We will explore these measures using the following data on the number of days spent in hospital by 19 patients following an operation:

3 4 4 6 7 8 8 8 10 10 12 14 14 17 20 25 27 37 42
Measures of Central Tendency

• **Mean** ($X$): The sum of all the values in a set of observations divided by the number of observations ($\Sigma x/n$)

• **Median**: The middle value when values are arranged in order

• **Mode**: The most frequently occurring value
Measures of Central Tendency

Bar Graph of Length of Hospital Stay

- Mode (8)
- Median (10)
- Mean (14.5)
Measures of dispersion

• **Standard deviation (SD):** A measure of the spread of observations around the mean

\[
SD = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}
\]
Measures of dispersion

• **Inter-quartile range:**
  - Measure of dispersion around a median
  - The range from the first (25%) to the third (75%) quartiles of a distribution

• **Range:** The difference between the largest and smallest values in a distribution
  - Problem = over influenced by ‘outliers’

• Reference range = mean \(-1.96 \times SD\) to mean \(+1.96 \times SD\) (in a normal distribution 95% of values lie within this range)
Road accident casualties, Harrow, 1985
Histogram

• Allows the inspection of the data for its underlying distribution

• Normal distribution, outliers, skewness,

• Unlike a bar chart, there are no "gaps" between the bars (although some bars might be "absent" reflecting no frequencies).

• This is because a histogram represents a continuous data set, and as such, there are no gaps in the data
Graphical Presentation of categorical data

Bar Chart

No. of malaria cases by age in 5 sentinel sites

District

Pie Chart

Death from Infectious Disease
The Normal Distribution

- Distribution of data are symmetrical around the mean

- Mean=Median=Mode

- 68.3% of observations lie within 1SD of $X$ ($X\pm1SD$)

- 95.4% lie between $X\pm2SD$

- 99.7% lie between $X\pm3SD$

- Standardization of data makes all normal distributions the same
The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean:

- 68% of the data falls within one standard deviation of the mean.
- 95% of the data falls within two standard deviations of the mean.
- 99.7% of the data falls within three standard deviations of the mean.
• Normal distribution also called a bell curve

• The mean and standard deviation completely specify a normal distribution

• Occurs naturally in many situations.

• Many groups follow this type of pattern e.g. Heights of people, Measurement errors, Blood pressure, IQ Scores, Salaries.
• The curve is symmetric at the center (i.e. around the mean, $\mu$).

• Exactly half of the values are to the left of center and exactly half the values are to the right.

• The total area under the curve is 1.

• An observation in a distribution with mean $\mu$ and standard deviation $\delta$ can be standardized to get a Z-score $(X - \mu)/\delta$

• Z-score tells how many standard deviations the original observation falls away from the mean and in which direction
Skewed distributions

• If one tail is longer than another, the distribution is skewed. Skewness differentiates extreme values in one versus the other tail.

• Also called asymmetric or asymmetrical distributions.

• Symmetry means that one half of the distribution is a mirror image of the other half.

• Kurtosis measures extreme values in either tail.
  • Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean).
  • Distributions with low kurtosis exhibit tail data that is generally less extreme than the tails of the normal distribution.
• Clue to skewed data from derived statistics
  - Mean and the median differ considerably.

• Better to describe a skewed distribution by means of a median and Interquartile range

• Sometimes a transformation will convert a skewed distribution into a symmetrical one.

• Methods of transformation
  - Square root transformation
  - Logarithmic transformation
  - Quintile regression
Skewed Distributions

Positively Skewed
- long tail to right
- mode < median < mean

Negatively Skewed
- long tail to left
- mean < median < mode
Positively skewed data

• Many blood tests e.g. HDLc, Triglycerides, CRP

• Arithmetic mean (SD) not ideal for describing data
  • e.g. Mean HDLc (SD) = 2.80 (2.97) mmol/l
  • What is wrong with above e.g.?

• May cause problems in regression models because of skewed residuals
Log Transformations

- Positively skewed distribution
- Log Transformation
- More symmetrical distribution
Analyses with skewed data

• Descriptive analyses
  • Geometric mean = log all values, calculate mean of logged values, exponentiate this mean
  • 95%CI of geometric mean or SD of logged mean

• Median (IQR)

• For right skewed data Median will be closer to geometric mean than it will arithmetric mean
Descriptive statistics for categorical data

Proportions and Percentages

• Proportions: Describe the share of one value for a variable in relation to a whole.

• Calculated by dividing the number of times a particular value for a variable has been observed, by the total number of values in the population.
• Percentage: **Expresses a value for a variable in relation to a whole population as a fraction of one hundred.**

Percentage total of an entire dataset should always add up to 100

• Calculated by dividing the number of times a value for a variable has been observed, by the total number of observations in the population, then multiplying this number by 100.
Patients on each ward in JFK

<table>
<thead>
<tr>
<th>Ward</th>
<th>Number</th>
<th>Proportion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicine</td>
<td>400</td>
<td>0.4 (40)</td>
</tr>
<tr>
<td>Paediatrics</td>
<td>200</td>
<td>0.2 (20)</td>
</tr>
<tr>
<td>Surgery</td>
<td>150</td>
<td>0.15 (15)</td>
</tr>
<tr>
<td>Obstetrics</td>
<td>200</td>
<td>0.2 (20)</td>
</tr>
<tr>
<td>ENT</td>
<td>20</td>
<td>0.02 (2)</td>
</tr>
<tr>
<td>TB</td>
<td>30</td>
<td>0.03 (3)</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
<td>1 (100)</td>
</tr>
</tbody>
</table>
Exposures and outcomes

• For any study there is usually a hypothesis linking an exposure (e.g. blood sugar) with an outcome (e.g. risk of heart disease)

• There are various ways of displaying these associations depending on the types of variables

• Remember an exposure can also be treated itself as an outcome
What are exposure & outcome in each of the following questions

• Does smoking cause lung cancer?

• Is HAART effective for reducing the risk of conversion to AIDS in HIV positive patients?

• Do younger individuals smoke more than older individuals?

• Is AIDS a risk factor for TB?
Displaying exposure and outcomes

• Continuous exposure & continuous outcome
  • Scatter plot with regression line, correlation

• Categorical exposure & continuous outcome
  • Box and whisker plot
  • Table of means (SD) by categories

• Categorical exposure versus categorical outcome
  • Contingency table
Box and whisker plot

Heights of males and females (m)
Box and Whisker Plot

• Graphically presents groups of numerical data through their quartiles

• Bottom and top of the box are always the 25th and 75th Percentiles

• Band in the middle is always the Median or 50th Percentile

• They are non-parametric. Display variation in the samples of a statistical distribution without making assumptions of the underlying statistical distribution.
## Table of Means

<table>
<thead>
<tr>
<th></th>
<th>Training group (n = 10)</th>
<th>Control group (n = 9)</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Age (years)</td>
<td>21.6</td>
<td>1.4</td>
<td>21.7</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>171.8</td>
<td>5.8</td>
<td>173.3</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>65.2</td>
<td>4.7</td>
<td>68.8</td>
</tr>
</tbody>
</table>

*: p < 0.05, M: mean, SD: standard deviation, ES: effect size
Scatter Plots

Weight against Height in Boys aged 2-14 Years
• Scatter plots are used to plot data points on a horizontal and a vertical axis.

• Show how much one variable is affected by another.

• The relationship between two variables is called their correlation.

• Correlation may be high, positive, negative low or zero.

• Correlation may seem to be present, but this might not always be the case.

• Both variables could be related to some third variable, explaining the variation or chance might cause an apparent correlation.
State the type of correlation for the scatter graphs below and write a sentence describing the relationship in each case.

1. Positive
   - Maths test scores vs. Physics test scores

2. Negative
   - Petrol consumption (mpg) vs. Car engine size (cc)

3. None
   - Height vs. KS 3 Results

4. Negative
   - Heating bill (£) vs. Outside air temperature

5. Positive
   - Sales of Sun cream vs. Daily hours of sunshine

6. Negative
   - Value of car (£) vs. Age of car (years)

The older the car the less its value.
Linear regression

- Regression line describes how a response variable $Y$ changes as an explanatory variable changes
- The outcome is always on the Y axis
- The exposure is on the X-axis
Interpretation of correlation and regression

• Used mainly for linear relationships

• The regression line is affected by outliers

• The relationship between 2 variables could be explained by a 3rd unknown variable

• Association does not imply causation
Contingency table

- Crosstabs or two-way tables
- Summarize the relationship between several categorical variables.
- Type of frequency distribution table, where two variables are shown at ago.

![Contingency Table]

<table>
<thead>
<tr>
<th></th>
<th>infected</th>
<th>not infected</th>
</tr>
</thead>
<tbody>
<tr>
<td>inoculated</td>
<td>3</td>
<td>276</td>
</tr>
<tr>
<td>not inoculated</td>
<td>66</td>
<td>473</td>
</tr>
<tr>
<td></td>
<td></td>
<td>279</td>
</tr>
<tr>
<td></td>
<td></td>
<td>539</td>
</tr>
<tr>
<td></td>
<td>69</td>
<td>749</td>
</tr>
<tr>
<td></td>
<td></td>
<td>818</td>
</tr>
</tbody>
</table>

Cholera Inoculation Study, 1894-96
• Used by statisticians when they need to make sense of data that has more than one variable.

• Contingency tables are displayed in matrix, or grid, form.

• The numbers displayed give the frequency of each data point.

• The table allows one to better understand the data using probability and relative frequencies.

• Can use table to calculate odds ratios and risk ratios.

• Chi2 tests can be used to compare the association of different categorical variables
Data Summary

• Distributions
  • Center (mean, median, mode)
  • Spread (variance & SD, IQR)
  • Shape (skewness)
  • Density models (normal: 67-95-99.7% rule)

• Association
  • Correlation (interpretation, pitfalls)
  • Regression (interpretation, pitfalls)
  • Chi2 test
  • Association vs causation
References

- Introduction to the field of statistics: Moore, McCabe and Craig
Acknowledgements

• Dr. Sarah Lewis: University of Bristol
Thanks
??