

Module: Introduction to Statistics

Module Outline

- Introduction
- Where do data come from?
- Types of variables
- Visual representations of data

Statistics: Introduction

- Social science questions:

Normative: "What ought to be"

Positive: "What is"



Statistics: Introduction

- Social science questions:

Normative: "What ought to be"

Positive: "What is"

- Answering "What is" question:

(a) *theory*

(b) *empirical* i.e. using data i.e. **statistics**

Why statistics?

- *Tells us about the world:* answer "What is" questions



Why statistics?

- *Tells us about the world: answer "What is" questions*
- *Test theories / impacts of policies and actions*



Why statistics?

- *Tells us about the world:* answer "What is" questions
- *Test theories / impacts of policies and actions*
- *Make predictions:* can provide competitive advantage



Why statistics?

- *Tells us about the world:* answer "What is" questions
- *Test theories / impacts of policies and actions*
- *Make predictions:* can provide competitive advantage
- *Age of "big data"*



Where do data come from?

- **Collection/observation**

Where do data come from?

- **Collection/observation**

- *Examples:*

What web-pages is a consumer viewing?

Police records: number of crimes in a city

OECD/World Bank Macroeconomic data

TV Ratings.....



Where do data come from?

- Collection/observation
- **Surveys**

Where do data come from?

- Collection/observation

- **Surveys**

- *Examples:*

Canadian Labour Force Survey

Current Population Survey (US)

many, many COVID-related surveys

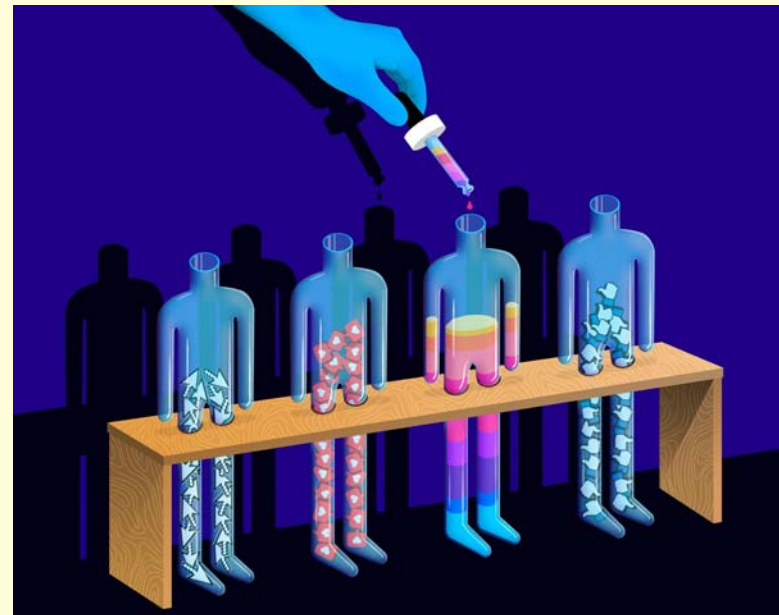
Electoral Polls (e.g. 338canada.com.....)

Opinion surveys (e.g. Pew Research Center.....)



Where do data come from?

- (i) Collection/observation (ii) Surveys
- (iii) **Experiments**



Where do data come from?

- (i) Collection/observation (ii) Surveys

- (iii) **Experiments**

- *Examples:*

Lab experiments: In the lab.

Field experiments: In real life settings

Natural experiments: Takes advantage of a natural/policy occurrence

Variables and data

- *Variable X*: Some measurable characteristic

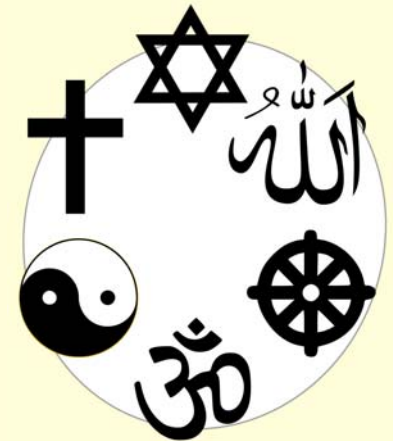
Variables and data

- *Variable X*: Some measurable characteristic
- *Examples*:

Daily value of a stock, monthly income, unemployment rate....

Amount of rainfall, snowfall, water level,....

Number of protests, votes, crime rate,.....



Variables and data

- *Variable X*: Some measurable characteristic

- *Examples*:

Daily value of a stock, monthly income, unemployment rate....

Amount of rainfall, snowfall, water level,....

Number of protests, votes, crime rate,.....

- *3 elements*:

- What is it? *description*

- Unit of measurement? *cm / feet, \$ (CAD / US), number,....*

- Unit of observation? *per day, per month, city, village,.....*

Data

- Datum/data: value(s) of the variable

Data

- Datum/data: value(s) of the variable
- *Example:* X = number of new Covid cases in Ontario per day



Data

- Datum/data: value(s) of the variable
- *Example:* X = number of new Covid cases in Ontario per day
- The 3 elements?

Data

- Datum/data: value(s) of the variable
- *Example*: X = number of new Covid cases in Ontario per day
- The 3 elements?
- Data (*over past 7 days*: Dec. 7, 2021):

780	959	1031	1053	1184	887	928
-----	-----	------	------	------	-----	-----

Types of Variables

- | Quantitative (numeric) | Qualitative (categorical) |
|------------------------------|-------------------------------|
| Data is numeric | Data is categorical |
| <i>Examples:</i> | <i>Examples:</i> |
| income, rainfall, cases,.... | religion, major, opinion,.... |

Types of Quantitative variables

- **Discrete:** some specific, discrete values (nothing in-between)

Types of Quantitative variables

- **Discrete:** some specific, discrete values (nothing in-between)
- *Examples:*
 - Number of children
 - Number of protests, job vacancies, ...



Types of Quantitative variables

- **Discrete:** some specific, discrete values (nothing in-between)
- *Examples:*
 - Number of children
 - Number of protests, job vacancies, ...
- **Continuous:** any value, including all in-between values (fractions/decimals)

Types of Quantitative variables

- **Discrete:** some specific, discrete values (nothing in-between)

- *Examples:*

Number of children

Number of protests, job vacancies, ...

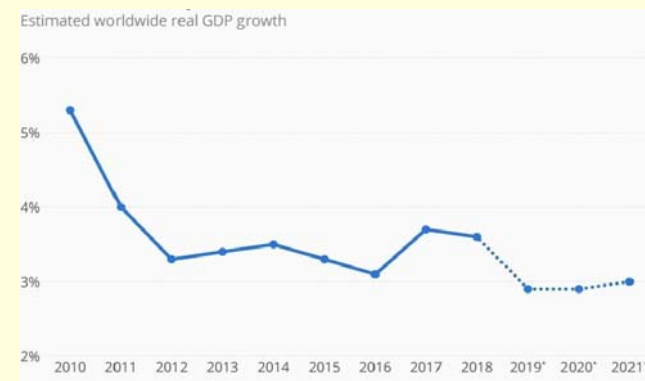
- **Continuous:** any value, including all in-between values (fractions/decimals)

- *Examples:*

Amount of rainfall

GDP growth rate

Fraction of the population vaccinated.....



Types of Quantitative variables

- **Discrete:** some specific, discrete values (nothing in-between)
- *Examples:*
 - Number of children
 - Number of protests, job vacancies, ...
- **Continuous:** any value, including all in-between values (fractions/decimals)
- *Examples:*
 - Amount of rainfall
 - GDP growth rate
 - Fraction of the population vaccinated.....
- *Discrete to continuous:* e.g. number of crimes and crime-rate

Types of Qualitative variables

- **Ordinal:** there is some natural ordering of the values that the variable can take

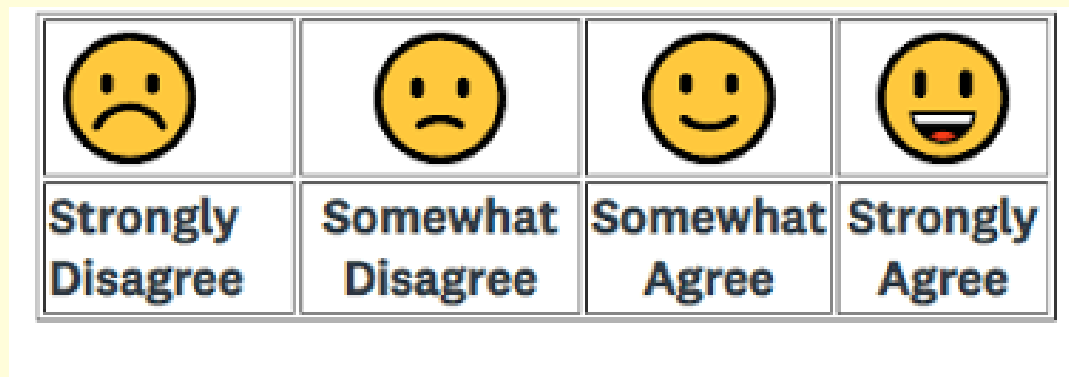
Types of Qualitative variables

- **Ordinal:** there is some natural ordering of the values that the variable can take

- *Examples:*

Wealth level: Poor, Lower middle-income, Upper middle-income, Rich

Opinion: Strongly disagree, Disagree, Agree, Strongly Agree



Types of Qualitative variables

- **Ordinal:** there is some natural ordering of the values that the variable can take
- *Examples:*
 - Wealth level:* Poor, Lower middle-income, Upper middle-income, Rich
 - Opinion:* Strongly disagree, Disagree, Agree, Strongly Agree
- **Nominal:** no ordering of the values that the variable can take

Types of Qualitative variables

- **Ordinal:** there is some natural ordering of the values that the variable can take

- *Examples:*

Wealth level: Poor, Lower middle-income, Upper middle-income, Rich

Opinion: Strongly disagree, Disagree, Agree, Strongly Agree

- **Nominal:** no ordering of the values that the variable can take

- *Examples:*

Religion: Catholic, Muslim, Jewish, Hindu,.....

Major: Biology, Economics, Geography, Sociology,.....



Distinction between Population and Sample

- *Population* : entire unit of study

Sample : a subset of the population

Distinction between Population and Sample

- *Population* : entire unit of study

Sample : a subset of the population

- *Examples:*

Population: Census, Covid case data, all NBA/NHL games in a season,.....



Distinction between Population and Sample

- *Population* : entire unit of study

Sample : a subset of the population

- *Examples:*

Population: Census, Covid case data, all NBA/NHL games in a season,.....

- Question: how to present/analyze the data?

Distinction between Population and Sample

- *Population* : entire unit of study

Sample : a subset of the population

- *Examples:*

Population: Census, Covid case data, all NBA/NHL games in a season,.....

- Question: how to present/analyze the data?

- *Sample:* (most of the time).....

Distinction between Population and Sample

- *Population* : entire unit of study

Sample : a subset of the population

- *Examples:*

Population: Census, Covid case data, all NBA/NHL games in a season,.....

- Question: how to present/analyze the data?

- *Sample*: (most of the time).....

- Question: how to present/analyze the data? + how to make **inferences about the population?**.

Getting a sense of the data visually

- X is a **qualitative / categorical variable** e.g. vaccination status of Covid cases in ON on Dec. 7, 2021



Getting a sense of the data visually

- X is a **qualitative / categorical variable** e.g. vaccination status of Covid cases in ON on Dec. 7, 2021
- *Frequency distribution:*

Unvaccd.	Partially vaccd.	Fully vaccd.	Unknown status
424	26	401	77



Getting a sense of the data visually

- X is a **qualitative / categorical variable** e.g. vaccination status of Covid cases in ON on Dec. 7, 2021
- *Frequency distribution:*

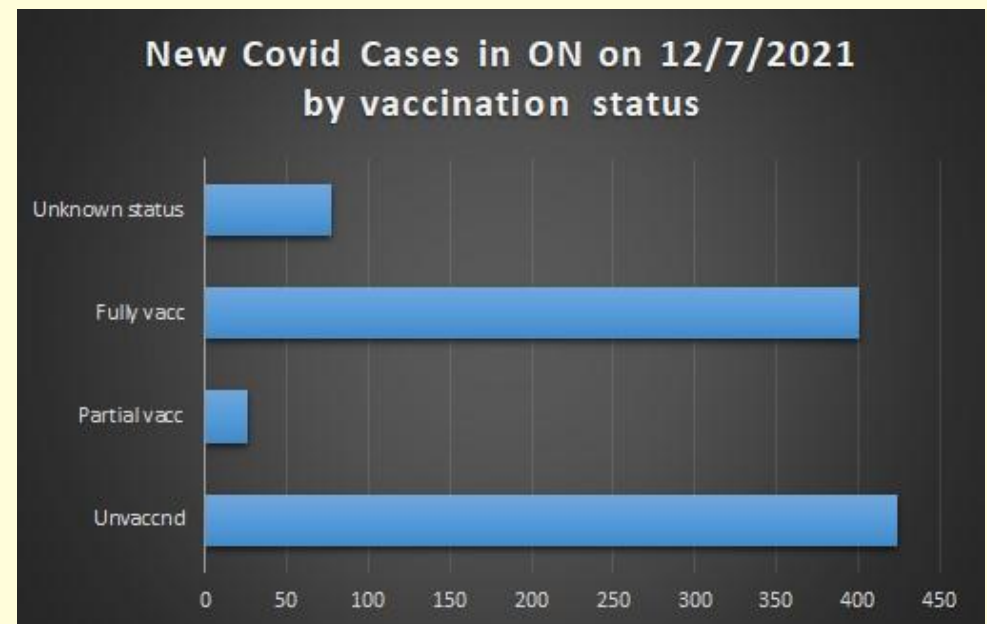
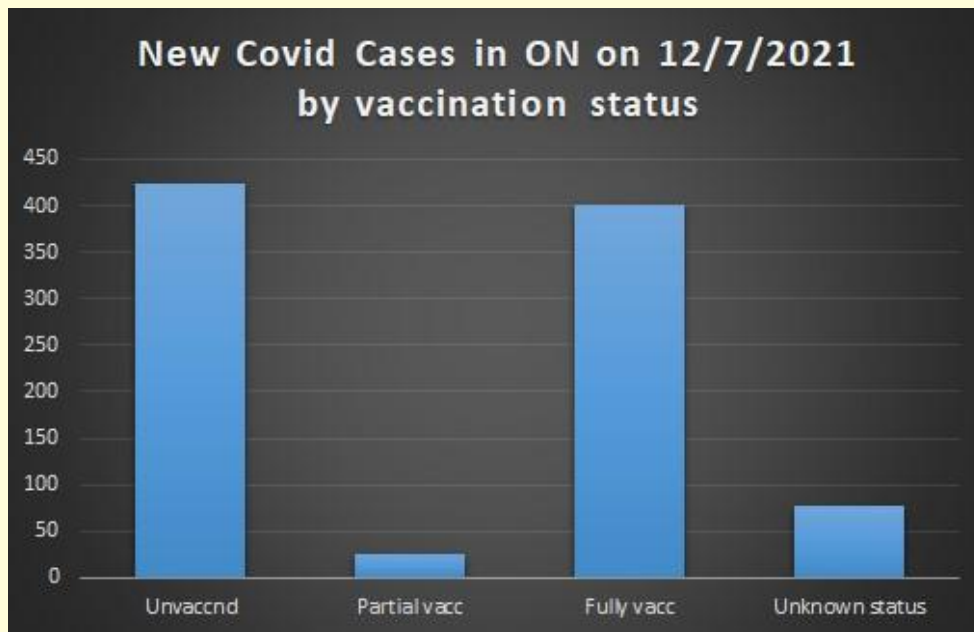
Unvaccnd.	Partially vaccnd.	Fully vaccnd.	Unknown status
424	26	401	77

- Bar graph / Column chart

Getting a sense of the data: Column chart / Bar Graph

- *Frequency distribution:*

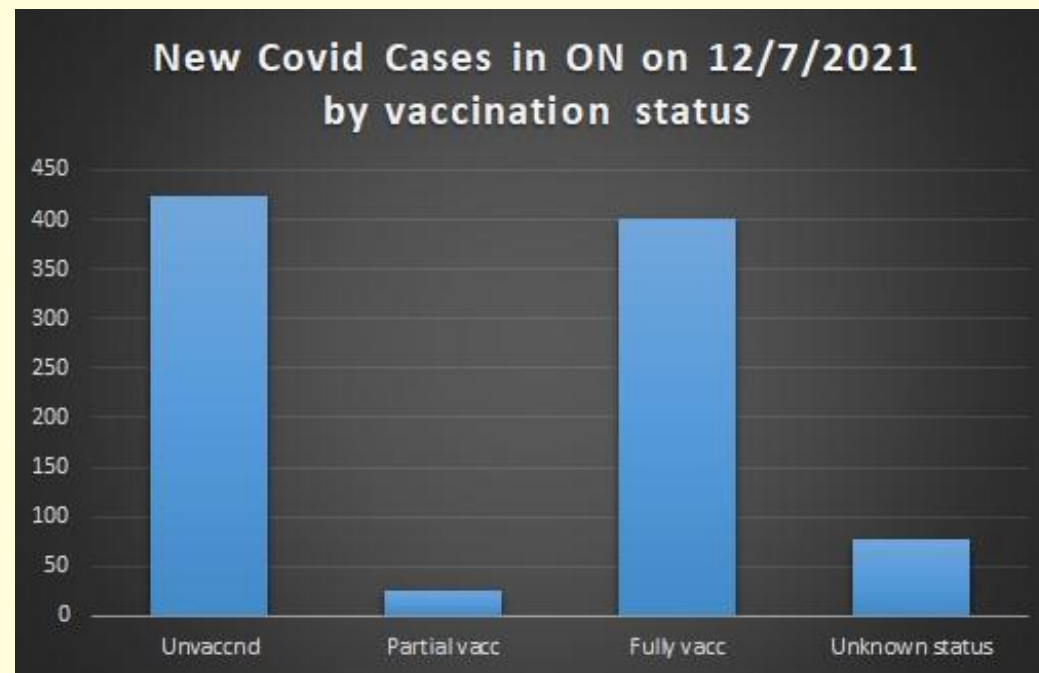
Unvaccd.	Partially vaccd.	Fully vaccd.	Unknown status
424	26	401	77



Warning: Be careful with statistics

-

Unvaccnd.	Partially vaccnd.	Fully vaccnd.	Unknown status
424	926	401	77



Warning: Be careful with statistics



Unvaccnd.	Partially vaccnd.	Fully vaccnd.	Unknown status
424	926	401	77

- *Frequency distribution of cases per 100,000:*

Unvaccnd.	Partially vaccnd.	Fully vaccnd.
14.35	6.17	3.91

Warning: Be careful with statistics

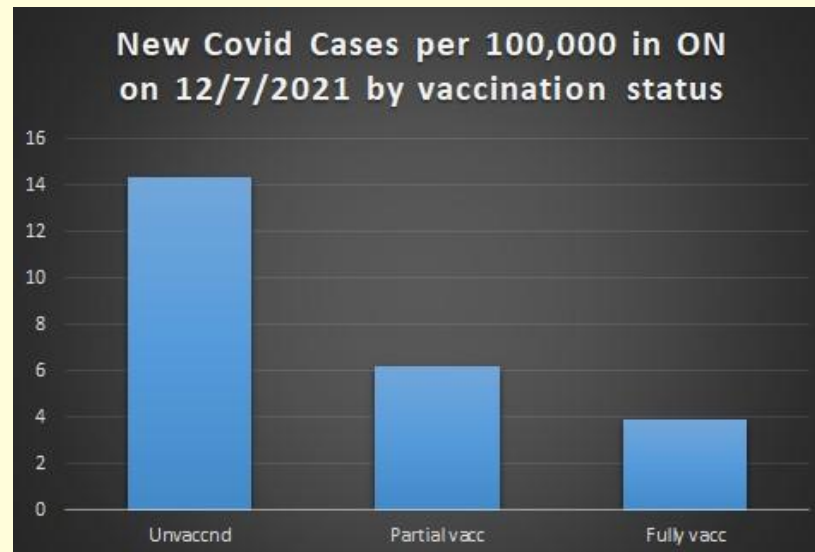
-

Unvaccnd.	Partially vaccnd.	Fully vaccnd.	Unknown status
424	926	401	77

- *Frequency distribution of cases per 100,000:*

Unvaccnd.	Partially vaccnd.	Fully vaccnd.
14.35	6.17	3.91

- Column chart



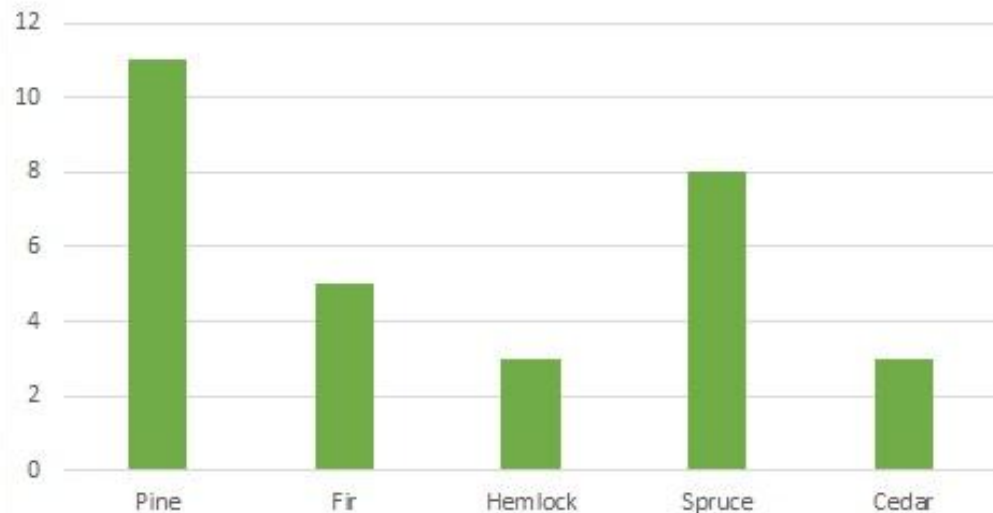
How to create a Column chart / Bar Graph using Excel?

- 3 examples:
 - Frequency distribution already given
 - Create a frequency distribution from data
 - *Pie chart*: shows composition as percentages

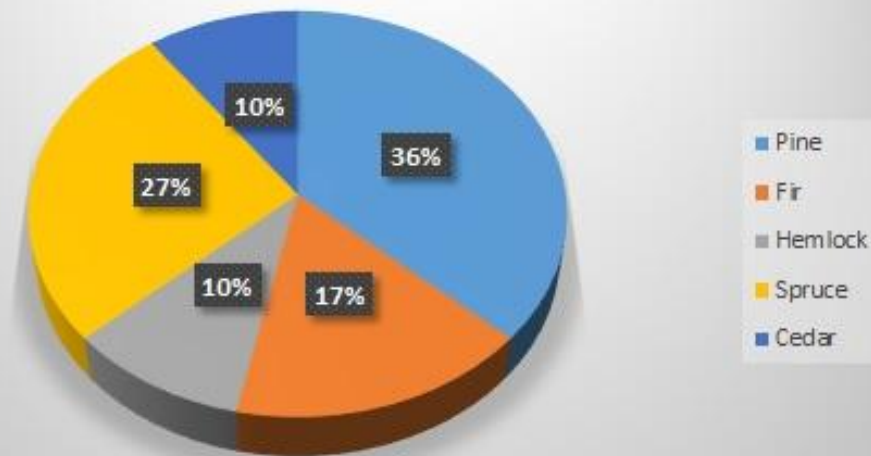
Ways of visually presenting data on a qualitative variable

- X is a **qualitative / categorical variable**
- *Bar graph / Column chart*: Frequency distribution
- *Pie chart*: Percentage composition

Frequency distribution of tree types



Percentage composition of tree types



Ways of visually presenting data on a qualitative variable

- X is a **qualitative / categorical variable**
- *Bar graph / Column chart*: Frequency distribution
- *Pie chart*: Percentage composition
- What if 2 or more qualitative variables? *e.g. data on different types of crimes, by province*

Ways of visually presenting data on a qualitative variable

- X is a **qualitative / categorical variable**
- *Bar graph / Column chart*: Frequency distribution
- *Pie chart*: Percentage composition
- What if 2 or more qualitative variables? *e.g. data on different types of crimes, by province*
- Use **grouped bar/column charts** or **stacked bar/column charts**

Ways of visually presenting data on a Quantitative variable

- So far: X is a **qualitative / categorical variable**

Examples: X = type of tree, type of crime, subject major

- Now: **quantitative / numeric variable**

Examples: X = a person's income

X = a country's GDP per capita

X = height of a tree, X = rainfall level

Ways of visually presenting data on a Quantitative variable

- *Visual tools:* depends on objective

Ways of visually presenting data on a Quantitative variable

- *Visual tools:* depends on objective
- *Example:* X = number of new Covid cases in Ontario per day



Ways of visually presenting data on a Quantitative variable

- *Visual tools:* depends on objective
- *Example:* X = number of new Covid cases in Ontario per day
- *Data (over past 7 days: Dec. 7, 2021):*

780	959	1031	1053	1184	887	928
-----	-----	------	------	------	-----	-----

Ways of visually presenting data on a Quantitative variable

- *Visual tools:* depends on objective
- *Example:* X = number of new Covid cases in Ontario per day
- Data (*over past 7 days:* Dec. 7, 2021):

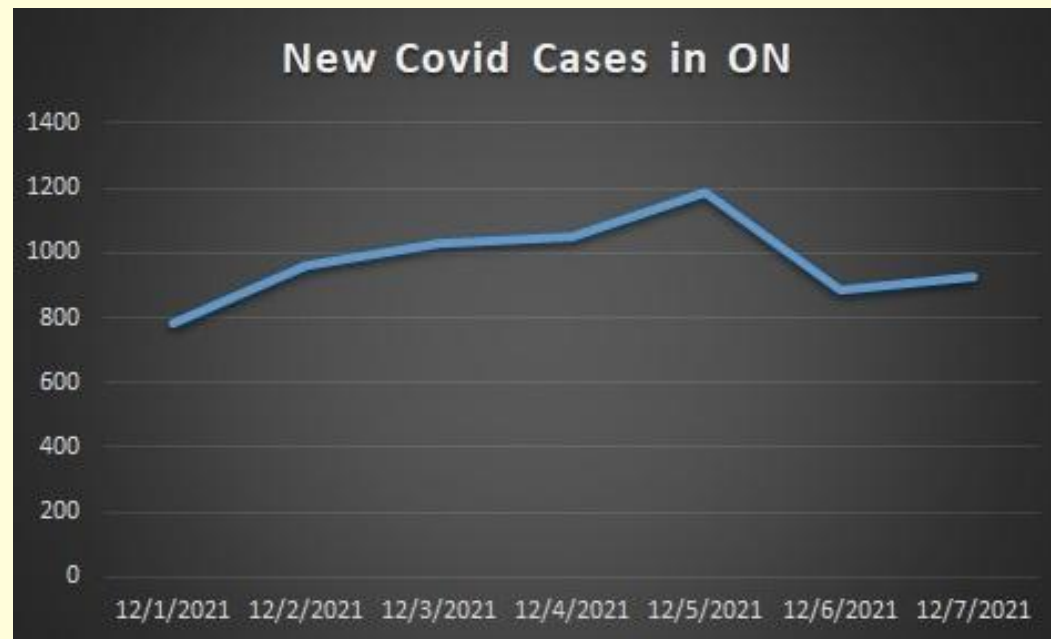
780	959	1031	1053	1184	887	928
-----	-----	------	------	------	-----	-----

- **Q:** *Trend?* Plot over time

Quantitative variable: Time trends

- X = number of new Covid cases in Ontario per day

12/1	12/2	12/3	12/4	12/5	12/6	12/7
780	959	1031	1053	1184	887	928

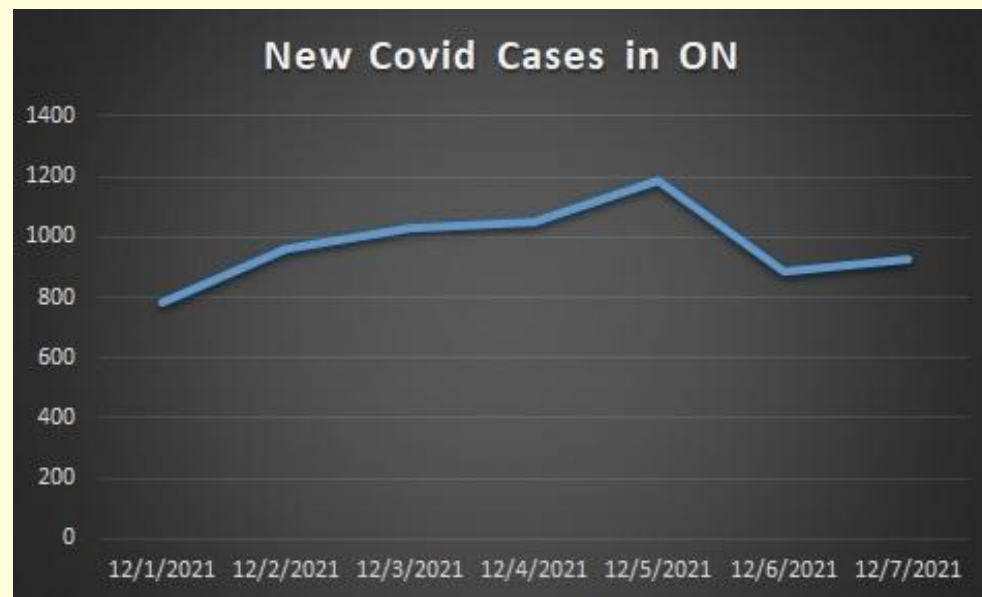


Quantitative variable: Time trends

- X = number of new Covid cases in Ontario per day

12/1	12/2	12/3	12/4	12/5	12/6	12/7
780	959	1031	1053	1184	887	928

- **Q: Trend** : Relationship between time and X



Quantitative variable: Time trends

- X = number of new Covid cases in Ontario per day

12/1	12/2	12/3	12/4	12/5	12/6	12/7
780	959	1031	1053	1184	887	928

- **Q:** *Trend* : Relationship between time and X

- *Other examples of trends:*

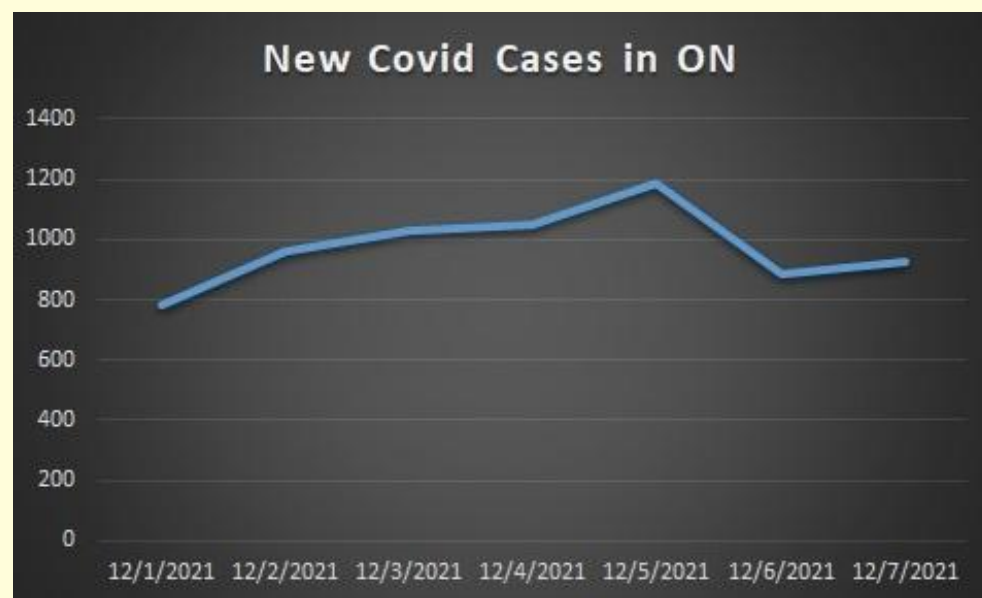
GDP, population, immigration,.....over the years

Sales, profits,.....over days / months / quarters

Plotting Time trends in Excel

- | 12/1 | 12/2 | 12/3 | 12/4 | 12/5 | 12/6 | 12/7 |
|------|------|------|------|------|------|------|
| 780 | 959 | 1031 | 1053 | 1184 | 887 | 928 |

Use **Line Plot**



Plotting Time trends in Excel

- | | | | | | | |
|------|------|------|------|------|------|------|
| 12/1 | 12/2 | 12/3 | 12/4 | 12/5 | 12/6 | 12/7 |
| 780 | 959 | 1031 | 1053 | 1184 | 887 | 928 |

Use **Line Plot**

- If Excel does not recognize a Date variable:*

Use "Select Data" and change "Horizontal Axis Labels"

Relationship between two Quantitative variables

- Example: X = yearly GDP per capita

Y = yearly CO_2 emissions per capita

Q: Is there a relationship?

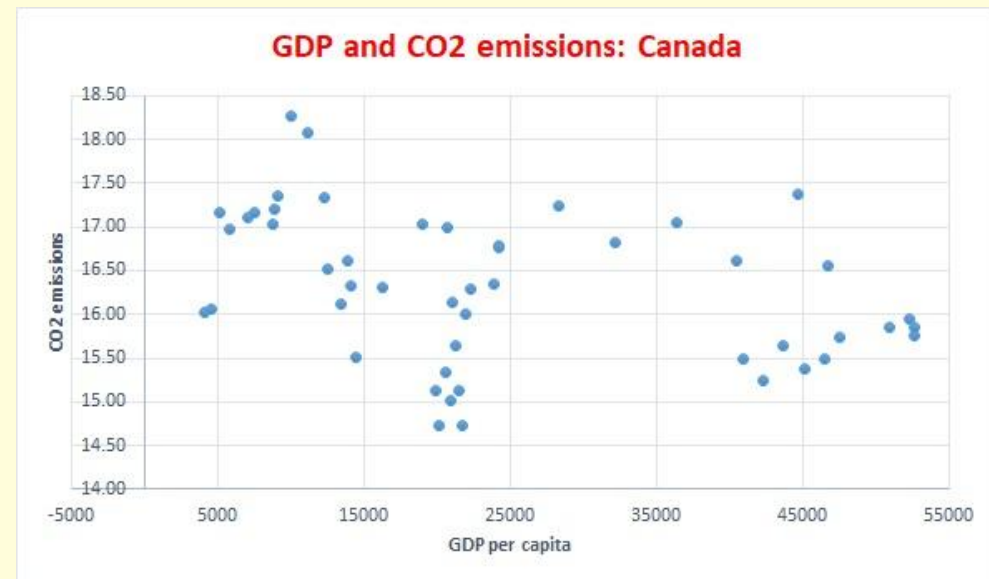
Relationship between two Quantitative variables

- Example: X = yearly GDP per capita

Y = yearly CO_2 emissions per capita

Q: Is there a relationship?

- *Scatter plot:* (X_t, Y_t) $t = \text{year}$



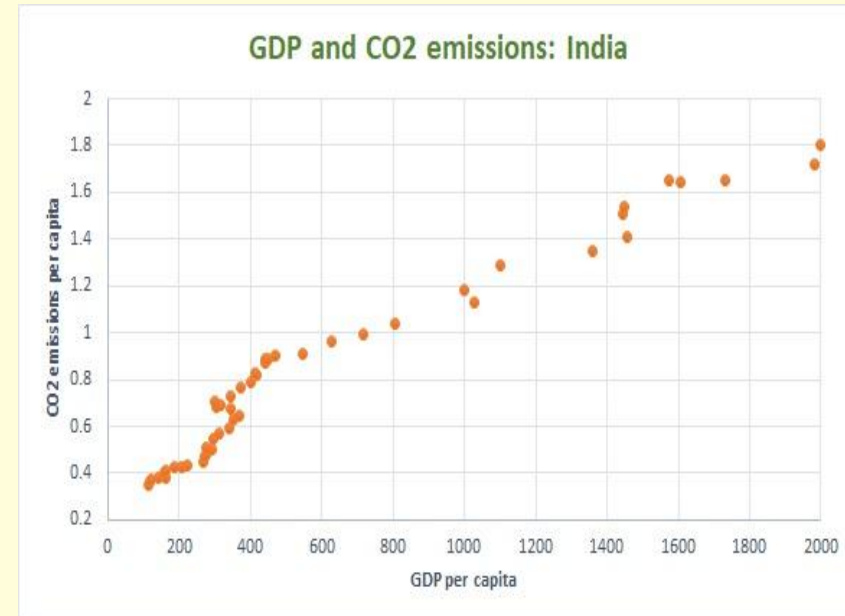
Relationship between two Quantitative variables

- Example: $X =$ yearly GDP per capita

$Y =$ yearly CO_2 emissions per capita

Q: Is there a relationship?

- *Scatter plot:* (X_t, Y_t) $t = \text{year}$
- Excel: Scatter Chart



Relationship between two Quantitative variables

- Example: X = yearly GDP per capita

Y = yearly CO_2 emissions per capita

Q: Is there a relationship?

- *Scatter plot:* (X_t, Y_t) $t = \text{year}$
- Excel: Scatter Chart
- Adjust to make the plot easier to analyze

Quantitative variable: Showing its distribution

- *Example:* Marks in a class, tree heights in a forest, CO_2 emissions by countries, incomes in a country

Q: How are these distributed?

Quantitative variable: Showing its distribution

- *Example:* Marks in a class, tree heights in a forest, CO_2 emissions by countries, incomes in a country

Q: How are these distributed?

- *Histogram:*

Create *equal sized* bins,

and then look at the frequency in each bin

Creating a Histogram in Excel

- *Histogram*: automatic command OR

Creating a Histogram in Excel

- *Histogram*: automatic command OR
- Create *equal sized* bins (not too many, not too few)

Creating a Histogram in Excel

- *Histogram*: automatic command OR
- Create *equal sized* bins (not too many, not too few)
- Get frequencies for each bin, and then plot bars relative to the frequency of each bin (no gaps)

Module Conclusion

- Introduction to data and variables
- Types of variables: *Qualitative (categorical)*
OR *Quantitative (numeric)*
- Visual representations of data:
Categorical variable: bar/column chart (grouped/stacked),
pie chart (for percentages)
- *Numeric variable*: Line plot, scatter plot, Histogram