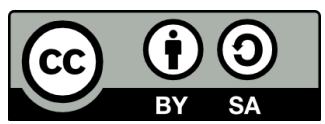
Applied Statistics

Research Design

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.



Measurements

- Before we can begin to analyze data, we need to first collect data
- Collecting data usually involves making **measurements**
- Making a measurement is the process of assigning a number, label, or other well-defined value to a characteristic of an object or event
- For example:
  - My **age** is *33 years*.
  - I *do not* **like anchovies**.
  - My **chromosomal gender** is *male*.
  - My **self-identified gender** is *female*.

Measurements

- Before we can begin to analyze data, we need to first collect data
- Collecting data usually involves making **measurements**
- Making a measurement is the process of assigning a number, label, or other well-defined value to a characteristic of an object or event
- For example:
  - My **age** is *33 years*.
  - I *do not* **like anchovies**.
  - My **chromosomal gender** is *male*.
  - My **self-identified gender** is *female*.

characteristic	**measurement**
age	33 years
like anchovies	do not
chromosomal gender	male
self-identified gender	female

Operationalisation

- **Operationalisation** is the process by which we take a meaningful but somewhat vague concept and turn it into a precise measurement
- The proccess can involve several different things:
  - Being precise about what you are trying to measure, e.g. does **age** mean *time since birth* or *time since conception*?
  - Determining what method you will use to measure, e.g., will you use self-report to measure age, ask a parent, or birth certificate?
  - Defining the set of allowable values that the measurement can take, e.g., is age measured in years, years and months, or years, months, and days?

Operationalisation

Key terms in the process of operationalisation:

- **A theoretical construct**: This is the thing that you're trying to take a measurement of, like "age", "gender" or an "opinion". A theoretical construct can't be directly observed, and often they're actually a bit vague.
- **A measure**: The measure refers to the method or the tool that you use to make your observations. A question in a survey, a behavioural observation or a brain scan could all count as a measure.
- **A variable**: A variable is what we end up with when we apply our measure to something in the world. That is, variables are the actual "data" that we end up with in our data sets.

Scales of measurement

- Variables come in different qualitative types and we therefore need to be able to distinguish the types when working with data
- One simple reason is that not every statistical operation can be performed on every type of variable
- The classification of the qualitiative type of a variable is done using the concept of **scale of measurement** or **level of measurement**[1]
- There are four widely used scales of measurement:
  - **nominal**
  - **ordinal**
  - **interval**
  - **ratio**
- The scales of measurement are ordered from lowest to highest in terms of the amount of the mathematical/statistical operations that can be performed on the data

[1] Stevens, S. S. (7 June 1946). "On the Theory of Scales of Measurement". Science. 103 (2684): 677–680.

Nominal scale

- A **nominal scale** variable, also referred to as a **categorical** variable, is one in which there is no particular relationship between the different possibilities
- For example, a nominal variable is "eye colour" with possible values black, brown, blue, green, etc.
- Or "university major": biology, math, chemistry, finance, etc.
- Other examples: categories, colors, names, labels, favorite foods, etc.
- With nominal data, it doesn't make any sense to say that one of them is "bigger" or "better" than any other one
- It doesn't make any sense to order them
- It doesn't make any sense to average them
- The only thing you can say about the different possibilities in nominal data is that they are different

Ordinal scale

- An **ordinal scale** variable is one in which there is a natural, meaningful way to order the different possibilities, but you can't do anything else
- An ordinal variable allows for rank order (1st, 2nd, 3rd, etc.) by which data can be sorted but still does not allow for a relative degree of difference between them
- For example, "finishing position in a race" can be ordered in a meaningfuly way: 1st $>$ 2nd $>$ 3rd $> \cdots$
- But it doesn't make sense to say that the difference between 1st and 2nd is the same as the difference between 2nd and 3rd
- All that can be said is that one position is higher or lower on the scale than another, but more precise comparisons cannot be made
- Opinion scales are also ordinal: "agree" $>$ "neutral" $>$ "disagree"

Interval scale

- An **interval scale** variable is one in which the numerical value is genuinely meaningful
- In particular, the differences between the numbers are interpretable, but the variable doesn't have a "natural" zero value
- A good example of an interval scale variable is measuring temperature in **degrees celsius** or **degrees fahrenheit**
- For instance, if it was 15° yesterday and 18° today, then the 3° difference between the two is genuinely meaningful
- Moreover, that 3° difference is exactly the same as the 3° difference between 7° and 10°
- However, notice that the 0° does not mean "no temperature at all"; it means "the temperature at which water freezes"
- Because there is no natural zero, multiplication and division do not make sense; 20° is not twice as hot as 10°

Ratio scale

- In a **ratio scale** variable, zero really means zero, and it's okay to multiply and divide
- As with an interval scale variable, addition and subtraction are both meaningful for ratio scale variables
- Examples include mass, length, duration, plane angle, energy and electric charge
- Informally, many ratio scales can be described as specifying "how much" of something (i.e. an amount or magnitude)
- Most measurements in the physical sciences and engineering are done on ratio scale

Scales of measurement

Below is a summary of the different scales of measurement and the operations that can be performed on them:

Level	Measure property	Mathematical operators	Other operations
Nominal	Classification, membership	$=, \neq$	Grouping
Ordinal	Ranking, comparison	$<, >,$	Sorting
Interval	Difference	$+, -$	Comparison to a standard
Ratio	Magnitude, amount	$\times, \div$	Proportions, ratios, percentages

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
3. The colors of crayons in a 24-crayon box
4. Incomes measured in dollars
5. The distance in miles to the closest grocery store
6. Political outlook: extreme left, left-of-center, right-of-center, extreme right
7. Time of day on an analog watch
8. Common letter grades: A, B, C, D, and F

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**

2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**

2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**

2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**

3. The colors of crayons in a 24-crayon box

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**

2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**

3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars $\Longrightarrow$ **RATIO**

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars $\Longrightarrow$ **RATIO**
5. The distance in miles to the closest grocery store

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars $\Longrightarrow$ **RATIO**
5. The distance in miles to the closest grocery store $\Longrightarrow$ **RATIO**

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars $\Longrightarrow$ **RATIO**
5. The distance in miles to the closest grocery store $\Longrightarrow$ **RATIO**
6. Political outlook: extreme left, left-of-center, right-of-center, extreme right

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars $\Longrightarrow$ **RATIO**
5. The distance in miles to the closest grocery store $\Longrightarrow$ **RATIO**
6. Political outlook: extreme left, left-of-center, right-of-center, extreme right $\Longrightarrow$ **NOMINAL**

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars $\Longrightarrow$ **RATIO**
5. The distance in miles to the closest grocery store $\Longrightarrow$ **RATIO**
6. Political outlook: extreme left, left-of-center, right-of-center, extreme right $\Longrightarrow$ **NOMINAL**
7. Time of day on an analog watch

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars $\Longrightarrow$ **RATIO**
5. The distance in miles to the closest grocery store $\Longrightarrow$ **RATIO**
6. Political outlook: extreme left, left-of-center, right-of-center, extreme right $\Longrightarrow$ **NOMINAL**
7. Time of day on an analog watch $\Longrightarrow$ **INTERVAL**

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars $\Longrightarrow$ **RATIO**
5. The distance in miles to the closest grocery store $\Longrightarrow$ **RATIO**
6. Political outlook: extreme left, left-of-center, right-of-center, extreme right $\Longrightarrow$ **NOMINAL**
7. Time of day on an analog watch $\Longrightarrow$ **INTERVAL**
8. Common letter grades: A, B, C, D, and F

Examples: Scales of measurement

What type of measure scale is being used? Nominal, ordinal, interval or ratio?

1. High school soccer players classified by their athletic ability: Superior, Average, Above average $\Longrightarrow$ **ORDINAL**
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300 $\Longrightarrow$ **INTERVAL**
3. The colors of crayons in a 24-crayon box $\Longrightarrow$ **NOMINAL**
4. Incomes measured in dollars $\Longrightarrow$ **RATIO**
5. The distance in miles to the closest grocery store $\Longrightarrow$ **RATIO**
6. Political outlook: extreme left, left-of-center, right-of-center, extreme right $\Longrightarrow$ **NOMINAL**
7. Time of day on an analog watch $\Longrightarrow$ **INTERVAL**
8. Common letter grades: A, B, C, D, and F $\Longrightarrow$ **ORDINAL**

Continous and discrete variables

- In addition to the four scales of measurement, variables can also be distinguished by whether they are **continous** or **discrete**
- A **continuous** variable is one in which, for any two values, it's always logically possible to have another value in between
- For example, time, mass, distance, temperature, concentration, and volume are continuous variables
- A **discrete** variable is, in effect, a variable that isn't continuous
- For a discrete variable it's sometimes the case that there's nothing in the middle between two values
- Examples of discrete variables:
  - the number of children in a family
  - the number of cars in a parking lot
  - the year you graduated from high school

Continous and discrete variables

Below is the relationship between the scales of measurement and the discrete/continuity distinction:

	continuous	discrete
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family
2. Temperature in degrees Fahrenheit
3. Bank account balance
4. The distance to the closest grocery store
5. The number of spelling errors on an essay
6. The time of day on an analog watch
7. The number of sand grains on a beach

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance $\Longrightarrow$ **DISCRETE**

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance $\Longrightarrow$ **DISCRETE**
4. The distance to the closest grocery store

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance $\Longrightarrow$ **DISCRETE**
4. The distance to the closest grocery store $\Longrightarrow$ **CONTINUOUS**

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance $\Longrightarrow$ **DISCRETE**
4. The distance to the closest grocery store $\Longrightarrow$ **CONTINUOUS**
5. The number of spelling errors on an essay

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance $\Longrightarrow$ **DISCRETE**
4. The distance to the closest grocery store $\Longrightarrow$ **CONTINUOUS**
5. The number of spelling errors on an essay $\Longrightarrow$ **DISCRETE**

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance $\Longrightarrow$ **DISCRETE**
4. The distance to the closest grocery store $\Longrightarrow$ **CONTINUOUS**
5. The number of spelling errors on an essay $\Longrightarrow$ **DISCRETE**
6. The time of day on an analog watch

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance $\Longrightarrow$ **DISCRETE**
4. The distance to the closest grocery store $\Longrightarrow$ **CONTINUOUS**
5. The number of spelling errors on an essay $\Longrightarrow$ **DISCRETE**
6. The time of day on an analog watch $\Longrightarrow$ **CONTINUOUS**

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance $\Longrightarrow$ **DISCRETE**
4. The distance to the closest grocery store $\Longrightarrow$ **CONTINUOUS**
5. The number of spelling errors on an essay $\Longrightarrow$ **DISCRETE**
6. The time of day on an analog watch $\Longrightarrow$ **CONTINUOUS**
7. The number of sand grains on a beach

Examples: Continous vs discrete variables

Is the variable continuous or discrete?

1. The number of children in a family $\Longrightarrow$ **DISCRETE**
2. Temperature in degrees Fahrenheit $\Longrightarrow$ **CONTINUOUS**
3. Bank account balance $\Longrightarrow$ **DISCRETE**
4. The distance to the closest grocery store $\Longrightarrow$ **CONTINUOUS**
5. The number of spelling errors on an essay $\Longrightarrow$ **DISCRETE**
6. The time of day on an analog watch $\Longrightarrow$ **CONTINUOUS**
7. The number of sand grains on a beach $\Longrightarrow$ **DISCRETE**

Predictors and outcome variables

- In some research studies, we are interested in the relationship between two variables, and in particular, if and to what extent one variable affects the other
- In this situation, the variables are classified into a **predictor** variable (or **independent** variable) and an **outcome** variable (or **dependent** variable), usually denoted by X and Y respectively
- The predictor variable is the variable that we think might affect the outcome variable, or can be used to explain the outcome variable
- For example, in a study of the relationship between the amount of exercise and the amount of weight loss, the amount of exercise is the predictor variable, and the amount of weight loss is the outcome variable
- In a study of the relationship between the amount of time spent studying and the grade on an exam, the amount of time spent studying is the predictor variable, and the grade on the exam is the outcome variable

Types of research

- There are two main types of research studies, **experimental** and **non-exprimental** (or **observational**)
- In **experimental research**, the researcher controls all aspects of the study, especially what participants experience during the study
- The researcher manipulates or varies the predictor variables but allows the outcome variable to vary naturally
- To eliminate the effect of other variables on the outcomes, everything else is kept constant or is in some other way "balanced"
- This is hard in practice and so **randomization** is often used to minimize the effect of other variables
- In randomization, we randomly assign the subjects of the study to different groups and assign them different values of the predictor variables

Example: Does smoking marijuana cause lung cancer?

- Suppose you wanted to find out if smoking marijuana causes lung cancer

- You could gather people who smoke marijuana and people who don't and investigate if smokers have a higher rate of lung cancer
- This is not a proper experiment, since the researcher doesn't have a lot of control over who is and isn't a marijuana smoker
- For instance, it might be that people who choose to smoke marijuana also tend to have poor diets, or maybe they tend to work in asbestos mines, or whatever
- So it might be that the higher incidence of lung cancer among marijuana smokers is caused by something else, and not necessarily by smoking

Example: Does smoking marijuana cause lung cancer?

- Instead we should introduce randomization so that we minimize the effect of the potential differences between smokers and non-smokers
- The solution is to control who smokes and who doesn't
- We would randomly divide young non-smokers into two groups and ask half of them to become marijuana smokers
- In this way, it's very unlikely that the groups will differ in any respect other than the fact that half of them smoke marijuana
- If our smoking group gets cancer at a higher rate than the non-smoking group, we can feel pretty confident that smoking marijuana does cause cancer
- Moreover, since this research study would be unethical, we would probably get fired, sued and/or put in jail

Non-experimental research

- **Non-experimental research** is a broad term that covers "any study in which the researcher doesn't have as much control as they do in an experiment"
- A non-experimental study arises in situations in which you can't or shouldn't try to obtain that control
- In the silly marijuana study example, it would be unethical and certainly criminal to force people to smoke marijuana
- In non-experimental research, you do not manipulate any variables or randomly assign participants to a control or treatment group
- Instead you are either describing a situation or phenomenon simply as it stands
- Or you are describing a relationship between two or more variables
- In all cases, no **interferences** are made by the researcher to explain the phenomenon or relationship