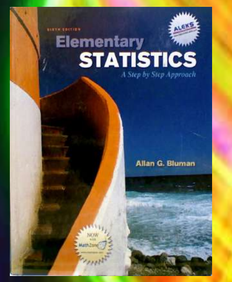
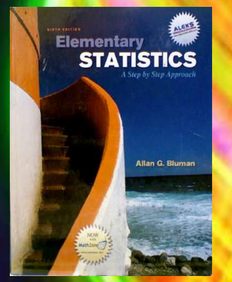
Chpt 1

The Nature of Probability and Statistics



1-2/3 Terminology, Variables and types of data.

Chpt 1-2/3



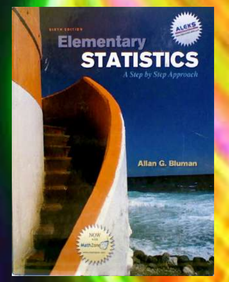
Homework

1-2/3 Read pages 1-9

p9 Applying the Concepts

p26 1-10

Chpt 1-2/3



Objectives

Introduction to Statistical Terminology
Identify Data types
Identify Measurement Level of a variable

Introduction to Statistics

- **Statistics** is the science of conducting studies to:
  - Collect
  - Organize
  - Summarize
  - Analyze
  - Infer from

  } **Data**

"Data don't make any sense,
we will have to resort to statistics."

Data

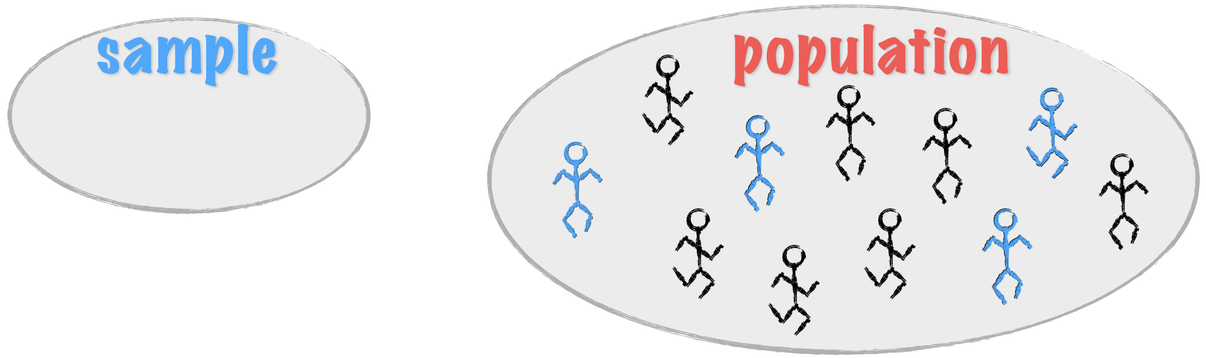
- So what is **Data**?
  - Data – Information that is collected, recorded, and organized.
    - Political Approval Rating
    - Sports Performance Data
    - Market Research
    - Medical Research (Vaccine research)
    - Student Performance
  - Until information is collected, recorded, and organized, that information is just background noise.

Descriptive and Inferential Statistics

- It all starts with **Data**
  - A **variable** is a characteristic or attribute that can assume different values (not simply the familiar x & y).
    - Height
    - Strength of Depression
    - Incidence of Covid-19
    - Gender
  - The **values** that a variable can assume become **data**.
    - A **data set** is the collected values of a variable.
    - Each value in the data set is a **datum** (one data value).

Descriptive and Inferential Statistics

- From whence comes the data?
  - A **population** consists of all subjects (human or otherwise) that are to be studied.
  - A **sample** is a subset (subgroup) of the population of interest.



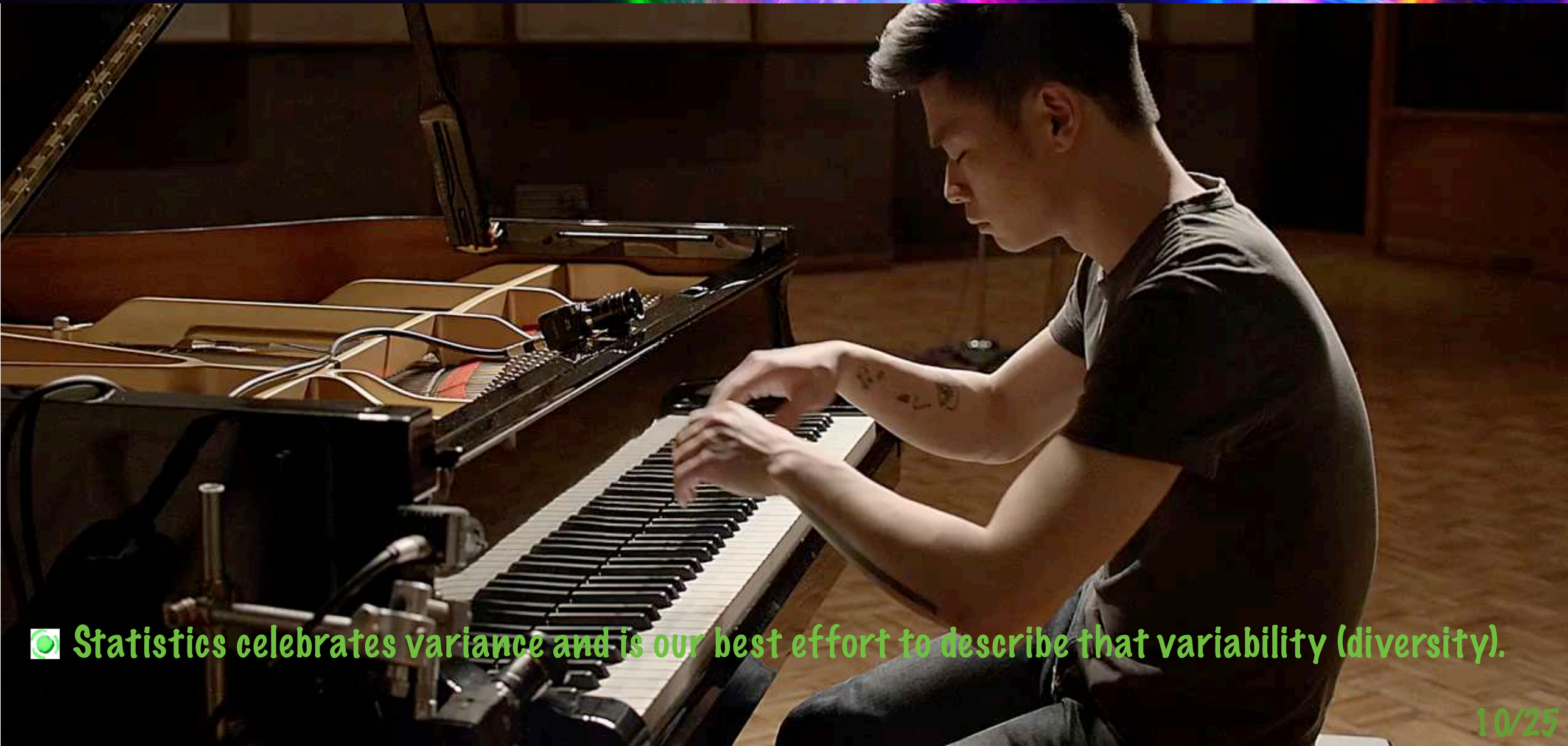
Descriptive Statistics

- Descriptive statistics does exactly what is says. Descriptive statistics describe a **distribution** (collection) of data.
  - It is important that you understand that statistics only derive from a collection of data, a **distribution** of several data values.
- Descriptive statistics consist of ...
  - collecting and recording data.
  - organizing data.
  - summarizing data.
  - and presenting the summary.

Inferential Statistics

- Inferential statistics take the process one step farther. In addition to collecting, recording, organizing, summarizing and reporting data, inferential statistics infer the results to a larger distribution or into the future.
    - generalizing from samples to much larger populations
    - estimating,
    - testing hypotheses,
    - suggesting relationships between variables.
    - making predictions

Metaphor for Life (and Statistics)



- Statistics celebrates variance and is our best effort to describe that variability (diversity).

What is Statistics Really About?

- Statistics is about variation.
  - As we proceed, remember that all measurements are imperfect. Your height and weight are not exactly the same every day. All measurements have variation that is unavoidable or unknown to us.
  - Statistics provide methods that are our best attempt to understand the real, imperfect, varying, diverse world in which we live.

How Do We Do Statistics?

- This will be the procedure for how we do statistics in this class (and in the real world).
  - 1. Decide what you want to **know** and **draw a picture** of the data.
  - 2. Determine the appropriate **test statistic** that will assess the data.
  - 3. Calculate values of the **test statistic, a statistic from the data,** and the **likelihood (probability) of that data statistic.**
  - 4. **Decide** what the statistics tell you.
  - 5. Write out your **conclusion**.
- As you can see, statistics is much more than simple calculations like finding the mean.
  - Let's get started.

Variables

- A random variable is a variable whose value cannot be predetermined. The value of a given datum is determined entirely by chance.

- For example: Your height is primarily predetermined by your genetics, with a small contribution by the environment.

- However, if we choose a person at random from the population we cannot know, apriori, what that person's height will be. Thus height is a random variable.

Qualitative Variables

- There are two major categories of variables; **qualitative** and **quantitative**.

- A **qualitative variable** (also **categorical** variable) is one whose data are attributes or characteristics. The values can be placed into distinct **categories**.
  - Color
  - Success or Failure
  - Status
  - Age grouping (infant, toddler, pre-school, etc.

- A **categorical variable** can have numeric values if the numbers simply identify **categories**.
  - Groups 1, 2, 3, 4

Quantitative Variables

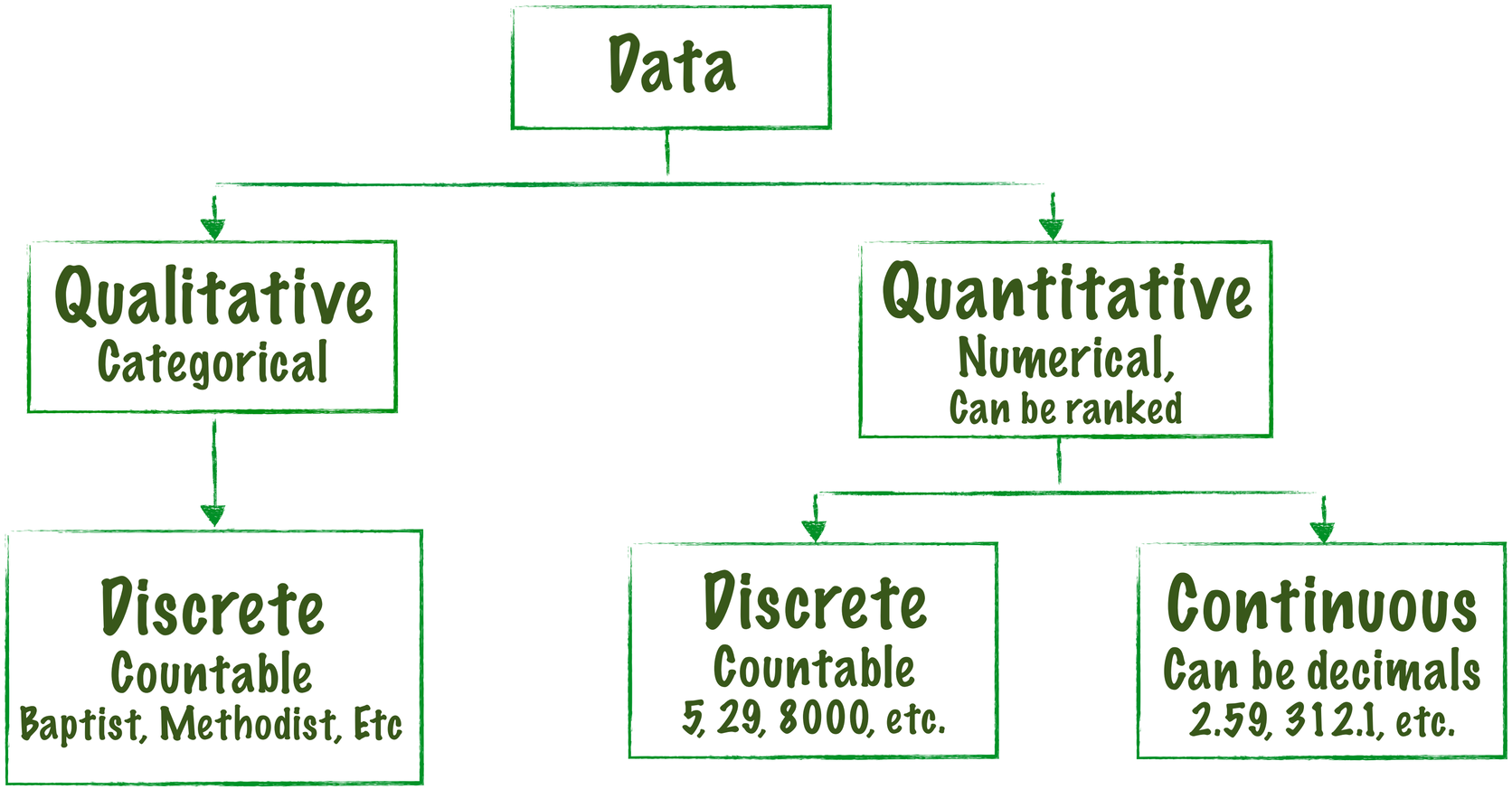
- A quantitative variable is a variable whose data (values) are numbers. It is necessary but not sufficient that the data be numbers, the numbers must also have meaning and order (rank).

- Height
- Age
- Proportion of cancer victims
- Attendance in school
- Number of cancer victims
- Time spent texting

Continuous and Discrete Variables

- Qualitative (categorical) variables are, by their nature, discrete variables.
- Quantitative variables, howsomever, can be continuous or discrete variables.
  - Continuous variables have data that have an infinite number of values between any two values (dense). Continuous variable values are determined by measuring.
    - Height
    - Age
    - Proportion of Covid-19 cases
  - Discrete variables have data such that there are values that do not have values between them. Discrete variable values can be determined by counting.
    - Number of Covid-19 cases
    - Length of your streak
    - Number of twits tweeting
    - Attendance at school

Continuous and Discrete Variables



Boundaries

- Continuous data are nearly always recorded as discrete, but the values recorded actually **represent intervals**.
  - Age is recorded as 18 but represents a value between 18.0 and 19.0, not including 19.0. [18.0, 19.0)
  - Temperature is recorded as 100° representing a value between 99.5° and 100.5°. [99.5, 100.5)
- When a datum value represents an interval, the **boundaries** of the interval are the values at each end of the interval. (18.0 & 19.0 or 99.5 & 100.5)
  - **Boundaries**, by definition, are **NOT** possible data values.

Boundaries

- If you are told that the normal human body temperature is 98.6°, what are the boundaries of the interval represented by 98.6°?

[98.55°, 98.65°)

- The average weight of an ostrich egg is 2 1/2 pounds, what are the boundaries of the interval?

[2.45, 2.55)

- Time to run the 40-yard sprint is 5.44 seconds, boundaries?

[5.435, 5.445)

Measurement Scales

- Variables are classified by the level of measurement of the data values.

    1. **Nominal** – categorical (names)
    2. **Ordinal** – nominal, plus can be ranked (order)
    3. **Interval** – ordinal, plus intervals are consistent
    4. **Ratio** – interval, plus ratios are consistent, true zero

- Distinguishing between Interval and ratio data is not very productive. We will collapse both into Interval/Ratio level.

Nominal

- The nominal level of measurement classifies data into mutually exclusive (non-overlapping), exhaustive categories in which no order or ranking can be imposed on the data. (Categories only, no implicit order, no value or measurement.)

  - State of Birth - (California, Utah, Georgia, ...)
  - Color of eyes - (Green, Brown, Blue, Hazel, ...)
  - Political Persuasion - (Progressive, Liberal, Moderate, Conservative, ...)
  - Religion - (Druze, Muslim, Jewish, Christian, ...)
  - High School Club - (Key Club, Tree Hugger, Black Student Union, ...)

Ordinal

- The ordinal level of measurement classifies data into categories that can be ranked (ordered); measurement of differences between the ranks do not provide useful information. Value of category implies order, implicit order, but only order.

  - School Class (Frosh, Soph, Jr,, Sr.)
  - Shirt Size (XXXL, XXL, XL, L)
  - Letter Grade in Statistics (A, B, C, D, F)
  - Age category (Infant, Toddler, ..., Ancient Decrepit Old Person)

Interval

- The interval level of measurement ranks data. Data is ordered and difference is constant and meaningful. Precise differences between units of measure do exist; but there is no meaningful zero. (No true zero, arbitrary 0, no real starting place.)
  - CST Scores
    - 400 is 5 more than 395, but not twice as good as 200
  - API Scores (for ranking schools)
    - 800 is wonderful, but not twice as good as 400
  - Temperature in degrees Fahrenheit or Celsius
    - The difference between 90° and 100° is 10°, the difference between 50° and 60° is also 10°. Do they feel the same?

Ratio

- The ratio level of measurement possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist for the same variable. Ordered, Difference is constant and meaningful, True, meaningful zero. True, meaningful ratios.
  - Length of Infant at birth
    - No child is born 0", but 0 does represent a meaningful value.
  - Age
    - 8 is twice 4 years old.
  - Score on Statistics Test
    - A 90 is 10 more than 80, and 90 is twice 45, the ratio is meaningful.

Level of measurement

Variable	Nominal	Ordinal	Interval	Ratio	Level
Hair Color					
Zip Code					
Letter Grade					
SAT Score					
Height					
Age					
Temperature					

interval/Ratio