Statistics: Basics – Epidemiology & Biostatistics

When people think about epidemiology, they often confuse it with statistics. Some even believe that epidemiologists are statisticians—but that's not accurate. While we use statistics as a tool in our work, epidemiology is not the same as statistics.

I'll admit, I'm not the biggest fan of statistics myself and sometimes feel I'm not great at it—even though I teach it and hold a Ph.D. in biostatistics. This should reassure you that even if you're not confident in this topic, it's okay to stumble through. You'll be fine! Statistics may seem daunting, but with practice, it becomes manageable.

Today, we'll cover some foundational concepts:

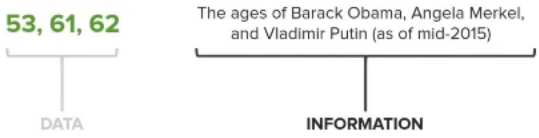
- **P-values, confidence intervals, and the null hypothesis** — cornerstones of frequentist statistics.
- The difference between **data and information** and how context turns one into the other.
- Types of variables, which determine the appropriate statistical tests to use.

Data vs. Information

Numbers, on their own, lack meaning. For example: **53, 61, 62**. These numbers might seem meaningless unless I provide context. If I tell you, these are the ages of Barack Obama, Angela Merkel, and Vladimir Putin in mid-2015, the numbers suddenly transform into **information**.

In essence:

- **Data** are raw numbers.
- **Information** is data given context.



For instance, you could calculate the average age of these leaders (about 58.7), which now has meaning because of the context.

Understanding Variables

A **variable** is essentially a placeholder for an idea. Variables allow us to perform mathematical or statistical functions to derive insights about broader concepts.

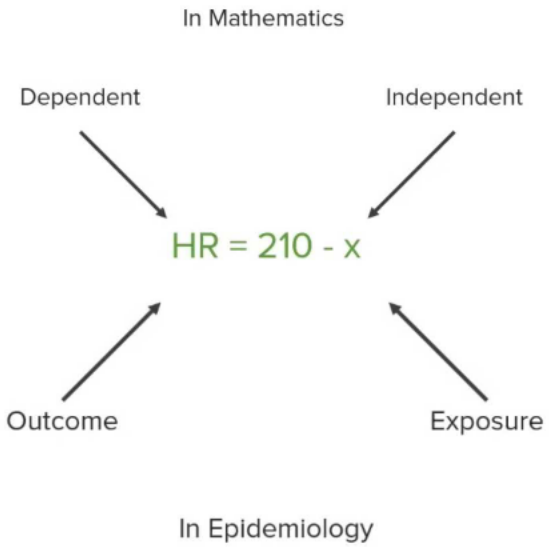
Depending on the field:

- In **mathematics**, a variable represents a value that changes within a problem.
- In **research**, a variable can represent characteristics like age, gender, or location.
- In **computers**, a variable might store a number or string of text.

In epidemiology, we often categorize variables as **exposures** (independent variables) and **outcomes** (dependent variables). For example:

- If we're examining the relationship between smoking rates (exposure) and cancer rates (outcome), we analyze how changes in smoking rates influence cancer rates.

Math	A value that may change within the scope of a problem or situation (vs. a "constant").
Research	A logical set of attributes (gender, age, etc.).
Computers	A symbolic name given to an unknown quantity.

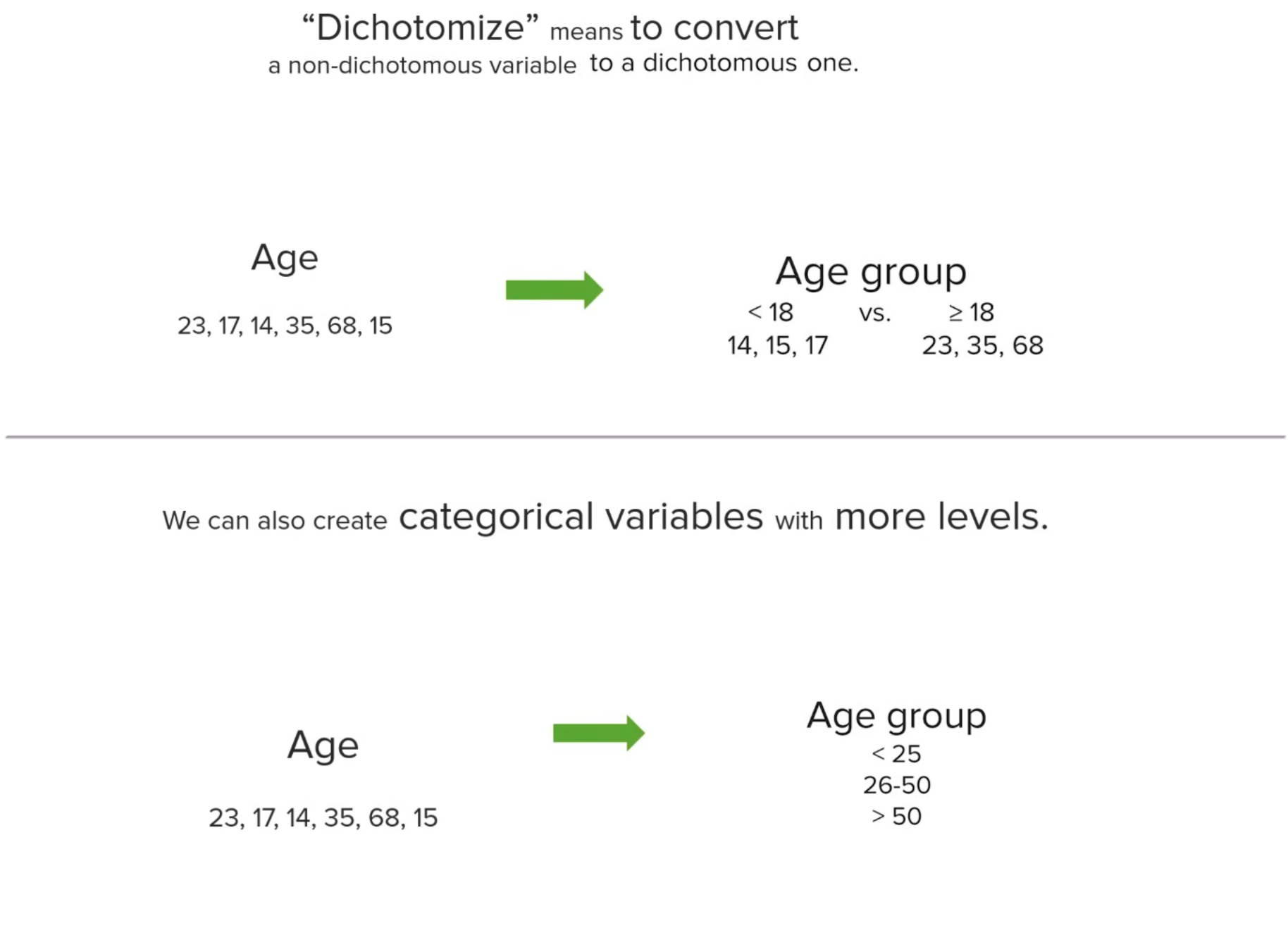


Types of Variables: Continuous vs. Categorical

Variables can be broadly categorized as:

1. **Continuous Variables** — Values that exist on a spectrum, like age, height, or temperature.
    - Example: 25.5 years or 98.6°F.
2. **Categorical Variables** — Defined categories, like gender or employment status.
    - Subtype: **Dichotomous Variables** — Variables with only two levels (e.g., male/female, employed/unemployed).

Sometimes, continuous variables are converted into categorical ones, a process called **dichotomization**. For instance, categorizing ages into “under 18” and “18 or over.” While this can simplify analysis, it often sacrifices information.



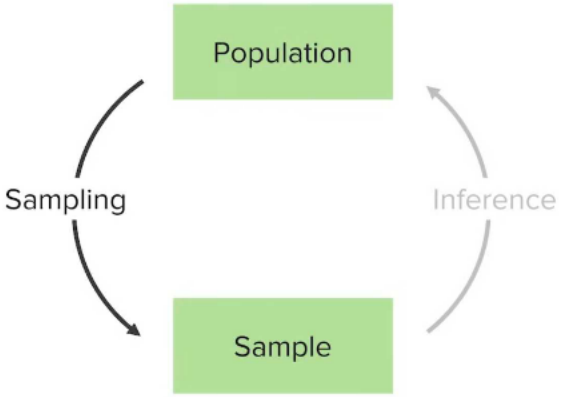
Populations and Sampling

In research, we often study a **sample** to make inferences about a larger **population**. It's crucial that the sample is representative of the population to avoid biases.

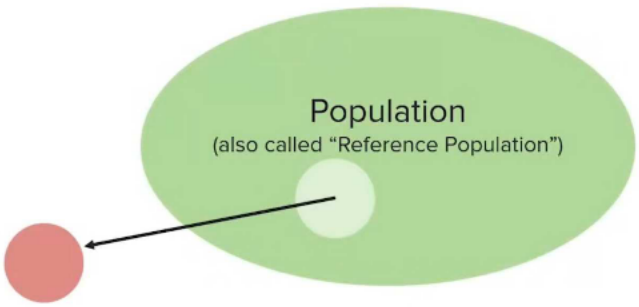
Example:
If we're studying back pain prevalence in the U.S. via a telephone survey:

- **Reference population:** All U.S. adults.
- **Accessible population:** Adults with telephones.
- **Sampling frame:** People with listed phone numbers.
- **Sample:** Those who answer the phone and agree to participate.

Bias may occur if the sample doesn't reflect the broader population (e.g., landline owners may not represent all adults).



The statistics and epidemiologic approaches we use
are affected by the assumptions of the sampling strategies used.



Null Hypothesis and P-Values

The **null hypothesis** (H_0) states there's no relationship between variables being tested. It's the baseline assumption we aim to reject. For example:

- H_0: The mean blood pressure in a treatment group equals that of a control group.

We use **p-values** to determine the likelihood that our results occurred due to chance under the null hypothesis. A smaller p-value suggests that it's unlikely the null hypothesis is true, prompting us to reject it.

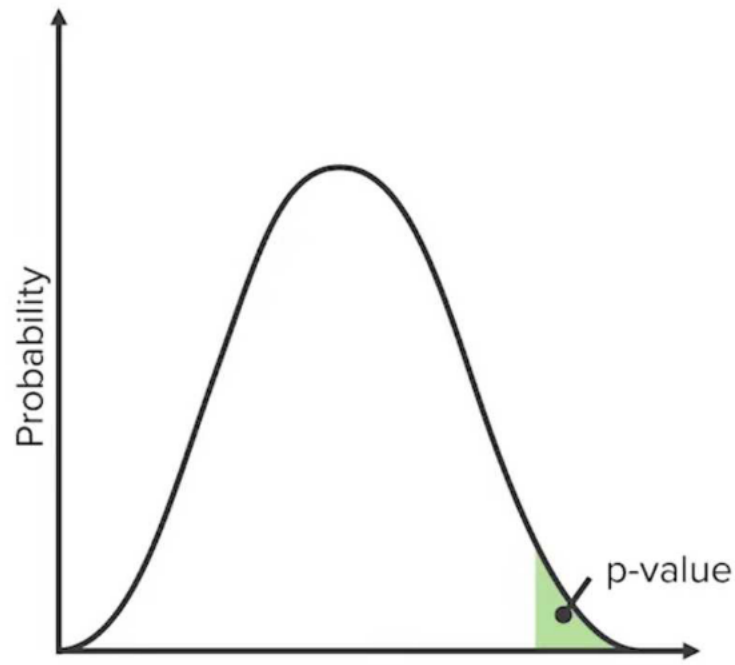
Why do we care?

Statistical tests allow us to either "reject" or "fail to reject" the null hypothesis.

$$H_o: \mu_1 = \mu_2$$

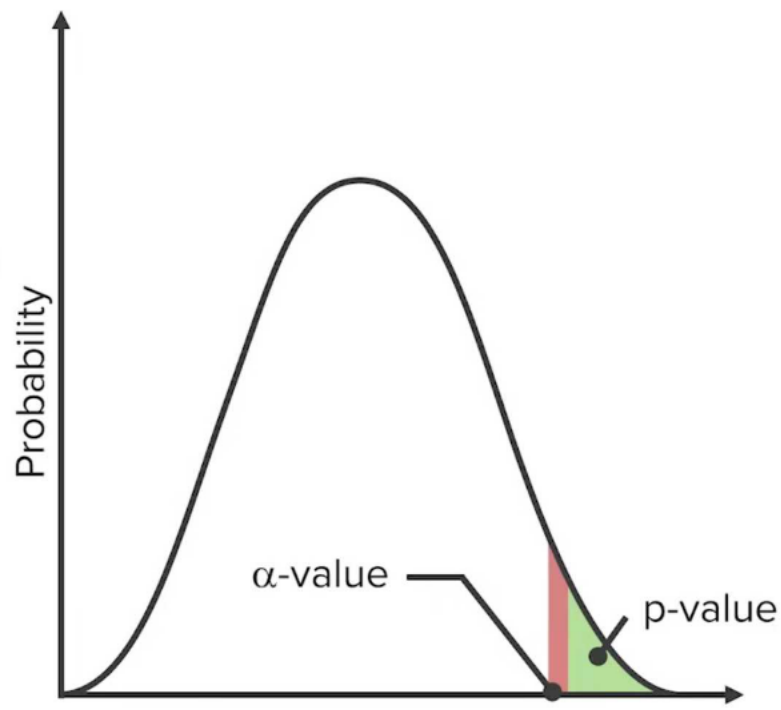
H_o: the average number of subjects getting better in the test group is no different from the average number of subjects in the placebo group.

A “p-value” is computed from a statistical test. It tells us whether we **should reject the null hypothesis.**

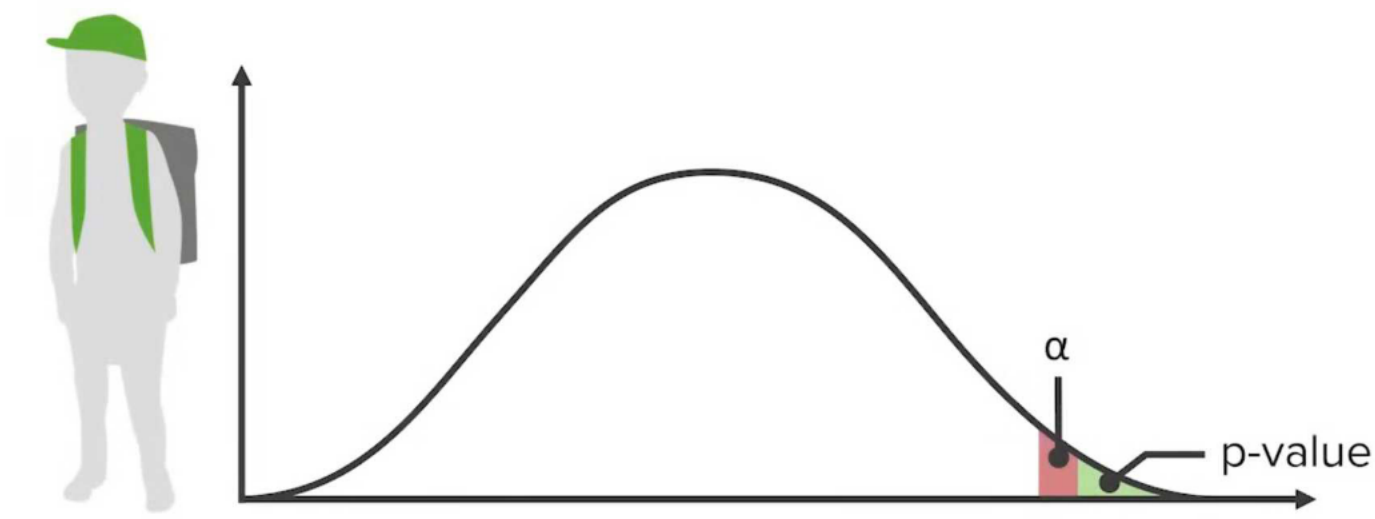


Whether or not we reject the null is determined by whether the p-value is **below a certain cut-off**, which we call the alpha value.

Traditionally, we tend to set alpha at either 0.05 or 0.01.



For example, if we are testing whether the average heights of two groups of children are different, and perform a t-test to produce a p-value of 0.02, setting $\alpha = 0.05$, we can conclude that **null hypothesis is rejected** and that the two groups do indeed have different average heights.



Confidence intervals offer an alternative way to express statistical significance without relying on p-values. A confidence interval provides a range within which the true value likely lies, based on the data. In modern science, confidence intervals are often preferred over p-values, and researchers typically report one or the other, not both.

How Confidence Intervals Work

A confidence interval contains:

1. **Point Estimate**: The specific value being measured, such as the mean age of university students (e.g., 21 years).
2. **Range**: The interval in which the true value is likely to be found. For example, a 95% confidence interval means that 95% of the time, the true value will fall within this range. The 95% comes from an alpha level of 0.05 (1 - 0.05 = 0.95).

Confidence intervals give us a sense of certainty and variability in our estimates, making them a valuable tool for interpreting data.

Common Statistical Tests

- **T-Test**: Compares the means of two groups.
- **Chi-Square Test**: Examines relationships between categorical variables.
- **ANOVA (Analysis of Variance)**: Extends the t-test to compare means across three or more groups.
- **Correlation**: Measures the relationship between two continuous variables (e.g., height and age or income and lifespan).
- **Regression**: Determines the influence of one or more variables on an outcome. Regression is particularly popular in epidemiology for modeling relationships and predicting outcomes.

Key Takeaways

You've learned several foundational concepts in statistics:

1. **P-Values vs. Confidence Intervals**: Two approaches to expressing statistical significance.
2. **Null Hypothesis**: A starting assumption that there is no effect or relationship.
3. **Data vs. Information**: Numbers gain meaning only when interpreted within context.
4. **Variable Types**: Understanding continuous, discrete, dichotomous, and multi-categorical variables.

Statistics is more than number-crunching; it's about interpretation, expertise, and adding value to the data. Mastering these concepts will enhance your skills as an epidemiologist and clinical researcher.