Biostatistics – Spring 2026
Lecture 01: Course Overview and Basic Data Summaries

Dr. Zaid T. Al-Khaledi
Department of Statistics and Informatics
University of Mosul

February 9, 2026

Introduction

Biostatistics is the use of statistical thinking in problems related to health and biology. In this course, we learn how to describe data clearly, how to compare groups, and how to draw correct conclusions from data. Many real studies are not perfect: data can be unbalanced, outcomes can be categorical, and sometimes the event time is not fully observed. That is why biostatistics has its own common tools and study designs.

We focus on foundations: what the course is about, what kinds of variables we meet in biostatistics, and how to summarize data using counts, proportions, and percentages.

What Is Biostatistics?

Biostatistics is a branch of statistics that analyzes data from biomedical and population studies. The statistical ideas are the same as in general statistics (sampling, variability, estimation, inference), but the data types and questions appear again and again in health problems.

A typical biostatistical question is about a real outcome such as "disease yes/no", "test positive/negative", "number of infections", or "time until death". For example, a hospital may ask: "What proportion of patients have diabetes?" or "Is the proportion different between males and females?" A public health office may ask: "Has mortality decreased after a policy change?" A clinical study may ask: "Do treated patients survive longer than untreated patients?"

In this course, we learn how to translate such questions into statistical objects: variables, tables, proportions, and later tests and models.

Scope of This Course (What You Will Learn)

This course is designed for undergraduate students in statistics. So our goal is not only to compute results, but also to understand **why** a method is suitable and **how** to interpret it correctly.

The course is organized around three common data situations:

1. **Categorical outcomes** such as (disease: yes/no). Here we use contingency tables and methods for comparing proportions.
2. **Rate-based summaries** such as prevalence, incidence, and mortality-related measures. Here the denominator and the time scale are very important.

3. **Time-to-event outcomes** such as survival time. Here we meet censoring and learn life tables and survival curve comparison.

Later in the course, we also study association measures such as Relative Risk and Odds Ratio. These measures help us summarize the strength of association between an exposure and an outcome.

Types of Variables (Very Important)

Before any statistical method, we must understand what type of variable we have. A wrong variable type usually leads to a wrong method.

Categorical Variables

A categorical variable describes **membership in a group**. Its values are labels, not numbers that represent size.

Nominal (no order). The categories do not have a natural order. Examples: sex (male/female), blood group (A/B/AB/O), smoking status (smoker/non-smoker), test result (positive/negative). If we code "male=1" and "female=2", this does not mean female is "larger" than male; it is only a coding.

Ordinal (ordered categories). The categories have a natural order, but the distance between levels is not numeric. Examples: disease severity (mild/moderate/severe), pain scale (low/medium/high), cancer stage (I/II/III/IV). Ordinal variables are still categorical because the difference between stage I and II is not necessarily the same as between II and III.

Numerical Variables

A numerical variable represents a measurable quantity.

Discrete (counts). Discrete variables are counts: 0,1,2,3,... Examples: number of hospital visits in one year, number of infections in a family, number of deaths in a month. Counts are numerical, but they are not continuous.

Continuous (measurements). Continuous variables can take any value in an interval. Examples: age, weight, blood pressure, time until recovery. Even if we record age in years (like 20, 21, 22), age is truly continuous in reality.

Example: Classifying Variables

Consider the following study variables:

1. "Diabetes status" (Yes/No): categorical nominal.
2. "Cancer stage" (I/II/III/IV): categorical ordinal.
3. "Number of cigarettes per day": numerical discrete (count).
4. "Cholesterol level" (mg/dL): numerical continuous.

5. "Survival time" (months): numerical continuous (time-to-event).

Correct classification helps you decide how to summarize the data. For categorical data we summarize with counts and proportions. For numerical data we summarize with mean/median and spread (later lectures).

Counts, Proportions, and Percentages

Many biostatistics results start with a simple idea: "How many?" and "Out of how many?"

Counts

A **count** is the number of observations in a category. If we have 200 individuals and 48 are infected, then 48 is the count of infected cases.

Counts alone can be misleading if the total size is not shown. For example, "48 infected" is very different if the total is 200 versus 20,000.

Proportions

A **proportion** is a count divided by the total number of observations:

$$\text{Proportion} = \frac{\text{Number in category}}{\text{Total number}}.$$

If 48 out of 200 are infected, then

$$\hat{p} = \frac{48}{200} = 0.24.$$

We read this as: "the observed proportion is 0.24".

Percentages

A **percentage** is simply a proportion multiplied by 100:

$$0.24 \times 100 = 24\%.$$

Percentages are easier to communicate, but they do not add extra information.

Example 1 (Single group)

A clinic screens 500 people for a disease and finds 75 positives.

$$\hat{p} = \frac{75}{500} = 0.15 = 15\%.$$

A correct sentence is:

> "In this sample, 15% tested positive."

We must say "in this sample" because we have not yet discussed inference. The sample proportion is an estimate of the population proportion, but it includes sampling variability.

Example 2 (Why the denominator matters)

Hospital A reports "40 positive cases" and Hospital B reports "60 positive cases". Which hospital has higher positivity? We cannot answer without totals.

Suppose Hospital A tested 200 people and Hospital B tested 600 people:

$$\hat{p}_A = \frac{40}{200} = 0.20 = 20\%, \qquad \hat{p}_B = \frac{60}{600} = 0.10 = 10\%.$$

Even though Hospital B has a larger count, Hospital A has a higher proportion. This example shows why proportions (and denominators) are essential.

Example 3 (Two groups; same proportion)

Two groups are tested:

Group	Positive	Total
A	30	200
B	45	300

$$\hat{p}_A = \frac{30}{200} = 0.15, \qquad \hat{p}_B = \frac{45}{300} = 0.15.$$

The correct interpretation is:

"The observed proportions are equal in the two samples (both 0.15)."

Later, we will learn how to test whether a difference is statistically meaningful.

Example 4 (Two groups; different proportions)

Now consider:

Group	Positive	Total
A	40	200
B	30	300

$$\hat{p}_A = \frac{40}{200} = 0.20, \qquad \hat{p}_B = \frac{30}{300} = 0.10.$$

A correct descriptive statement is:

"Group A has a higher observed positivity (20%) than Group B (10%)."

This is descriptive. Later we will learn inference: does this difference likely reflect a true population difference or only random sampling?