

# Statistical Methods in Epidemiology

## Lab 2

### Rates in Follow-up Studies

#### Introduction and Rationale

The purpose of **Lab 2 – Rates in Follow-up Studies** is to extend the concepts introduced in **Lecture 2 (Analysis of Cohort Studies)** and to apply them using *Stata* on real follow-up data.

While Lab 1 focused on *risks* and *risk ratios*, this session introduces *rates* and *rate ratios*—the fundamental measures of disease frequency and association in cohort studies with variable follow-up time.

Specifically, this lab aims to:

- Familiarize students with the Stata survival-time framework (*stset*, *strate*, *stmh*) used in the analysis of cohort data.
- Demonstrate the estimation of *incidence rates* and *rate ratios* ( $RR = \lambda_1 / \lambda_0$ ).
- Explore how continuous exposures can be grouped or treated as metric variables to evaluate *linear trends* in disease rates.
- Introduce methods for *confounding control* and *effect modification* using stratified (Mantel–Haenszel) analysis.

The session emphasizes interpretation of epidemiologic meaning—beyond statistical significance—linking person-time concepts, exposure comparisons, and adjusted rate ratios.

#### Structure and Flow

The practical follows a logical sequence mirroring the analytic process in cohort studies and the theoretical progression of Lecture 2.

Section	Theoretical correspondence	Practical focus
<b>1.1 Data exploration</b>	Defining the population, exposure, and outcome	Understanding variables, CHD cases, and time variables
<b>1.2 Setting up survival-time data</b>	Concept of time at risk; origin, entry, exit	Declaring survival-time data with stset and defining time scales
<b>1.3 Preliminary analysis</b>	Estimation of incidence rates	Using strate to calculate and visualize crude CHD rates
<b>1.4 Rate ratios</b>	Comparison of disease rates between exposure groups	Estimating and interpreting rate ratios via stmh
<b>1.5 Controlling for confounding</b>	Adjustment by stratification; Mantel–Haenszel estimator	Assessing the effect of high energy intake adjusted for job and height

### Link to Lecture 2

Lecture content	Practical component	Conceptual goal
Slides 5–15: <i>Definition of rate (<math>\lambda = D/Y</math>) and person-time</i>	<code>strate</code>	Compute and interpret incidence rates
Slides 16–25: <i>Rate ratio and its 95 % CI</i>	<code>stmh hieng</code>	Quantify relative rate of disease between exposure groups
Slides 26–36: <i>Continuous exposures and trend testing</i>	<code>stmh eng3</code> , <code>stmh energy</code>	Evaluate linear effect of exposure levels
Slides 37–56: <i>Confounding and Mantel–Haenszel adjustment</i>	<code>stmh hieng job</code> , <code>stmh hieng, by(job htgrp)</code>	Adjust for confounders and check effect modification
Slides 57–63: <i>Interpretation and limitations</i>	Discussion questions	Epidemiologic interpretation of results

**Objectives**

The objective of this practical is to introduce statistical methods and commands for analyzing follow-up studies in epidemiology, emphasizing the calculation and interpretation of rates and rate ratios. This session covers techniques for handling survival-time data, conducting preliminary analyses, and exploring associations between dietary factors, personal characteristics, and Coronary Heart Disease risk.

**1. The Diet Dataset**

**File:** diet

**Description:** This dataset, derived from a prospective cohort pilot study, comprises dietary and health information from 337 adult male subjects, collected over a two-week period. The primary outcome of interest is Coronary Heart Disease (CHD), represented by the binary variable 'chd', where 1 indicates the presence of CHD and 0 its absence.

**1.1. Data Exploration**

1. Read the data into Stata and examine the variables.

```
. use diet, clear  
. desc
```

2. Determine the number of CHD cases in the dataset. Record this number for future reference.

```
. tab chd
```

3. Examine the date variables

```
. list id doe dox chd in 1/20
```

- doe: date of entry
- dox: date of exit

**Note:** The time variables are stored in Stata format as days since January 1, 1960. For calculations, these dates are treated as numbers of days, but for output, they are displayed in standard date format.

## 1.2. Setting up Survival-Time Data

1. Set the survival-time (st) variables

```
. stset dox, failure(chd) origin(doe) scale(365.25)
```

2. Examine the newly created variables

```
. list id _t0 _t _d _st in 1/20
```

3. Change the time scale from time-since-entry to age.

```
. stset dox, failure(chd) origin(dob) enter(doe) scale(365.25)  
. list id _t0 _t _d _st in 1/20
```

## 1.3. Preliminary Analysis

1. Examine the distribution of the high energy intake (hieng) variable.

```
. tab hieng
```

2. Calculate CHD rates using the `strate` command.

```
. strate, per(1000)
```

3. Create a new categorical variable called `htgrp` for height groups using the cut points (150, 170, 175, 180, 195 cm).

```
. egen htgrp = cut(height), at(150, 170, 175, 180, 195) icodes
```

4. Calculate and compare CHD rates across the different height groups (`htgrp`) using the `strate` command.

```
. strate htgrp, per(1000)
```

5. Create an appropriate graph to visualize how CHD rates change across height groups.

```
. strate htgrp, per(1000) graph
```

### 1.4. Rate Ratios

1. Calculate rate ratios for high energy intake

```
. stmh hieng
```

2. Create a new variable with multiple energy intake levels

```
. egen eng3 = cut(energy), at(1.5, 2.5, 3.0, 4.5) icodes  
. tab eng3
```

3. Analyze rates for different energy levels

```
. strate eng3, per(1000)  
. strate eng3, per(1000) graph  
. strate eng3, per(1000) graph ylog
```

4. Compare CHD rates between different levels of energy intake

```
. stmh eng3, c(1,0)  
. stmh eng3, c(2,0)
```

5. Investigate the effect of high energy intake on CHD rate

```
. stmh hieng  
. stmh eng3  
. stmh energy
```

### 1.5 Controlling for Confounding

1. Examine the effect of high energy intake on CHD rate, controlling for job.

```
. stmh hieng job
```

2. Does the effect of high energy intake on CHD rate differ across job categories?

```
. stmh hieng, by(job)
```

3. Investigate the relationship between high energy intake and CHD rate, controlling for both job and height group.

```
. stmh hieng, by(job htgrp)
```

### **Discussion Questions**

- 1.** What does your analysis suggest about the association between high energy intake and the rate of CHD?
- 2.** How does the CHD rate change across different height groups?
- 3.** What can you conclude about the association between different energy intake levels (eng3) and the rate of CHD?
- 4.** Interpret the association between height and CHD rate based on your analysis.
- 5.** Consider the Mantel-Haenszel estimate when controlling for job. Do you think the underlying assumption of this estimate is relevant in this context? Why or why not?
- 6.** What are your overall conclusions about the relationship between high energy intake and CHD rate when controlling for both job and height group?