

# Statistical Methods in Epidemiology

## Lab 1

### Measures of Disease and Effect

#### Introduction and Rationale

The purpose of *Lab 1 – Measures of Disease and Effect* is to consolidate the core epidemiologic concepts introduced in Lecture 1 through practical application in **Stata**. Specifically, this lab aims to:

- Familiarize students with Stata commands used to compute key epidemiologic measures—*cumulative incidence*, *incidence rate*, *prevalence*, and both *absolute* and *relative* measures of effect (*risk ratios*, *rate ratios*, *odds ratios*).
- Interpret these measures in realistic analytic contexts (cohort and case–control designs).
- Translate theoretical definitions (e.g., *incidence vs prevalence*, *absolute vs relative effects*, *2×2 tables*, *person-time*) into operational calculations using real data.

#### Structure and Flow

The lab follows a progressive structure that mirrors the conceptual development of the theory.

Section	Theoretical correspondence	Practical focus
<b>1.1 Descriptive analysis</b>	Introduction – population under study	Understanding the population at risk and the definition of exposure/outcome
<b>1.2 Risk analysis (Whitehall)</b>	Cumulative incidence and risk ratio	Quantifying risk and comparing exposed vs unexposed groups
<b>1.3 Incidence rate analysis</b>	Incidence rate, person-time, IRR	Transition from static risk to dynamic rate measures
<b>2. Case–control analysis (Mwanza)</b>	Odds ratio; homogeneity and trend tests	Applying odds-based measures and understanding study-design differences

**Link to Lecture 1**

The practical directly reinforces the main lecture topics:

Lecture content	Practical component	Conceptual goal
Slides 10–22: <i>Incidence and population at risk</i>	<code>cs</code> , <code>csi</code> commands	From theoretical “risk” to estimated probability via 2×2 tables
Slides 17–20: <i>Incidence rate (IR)</i>	<code>ir</code> , <code>iri</code> , <code>tabrate</code> , <code>mhrate</code>	Understanding person-time and the $CI \approx IR \times T$ relation
Slides 43–51: <i>Absolute measures (Risk Difference, Attributable Risk)</i>	Interpretation of <code>cs</code> / <code>ir</code> outputs	Public-health meaning: excess cases attributable to exposure
Slides 52–58: <i>Relative measures (RR, OR)</i>	Risk, rate and odds ratios	Quantifying strength of association and causal inference
Slides 37 & 58–63: <i>Choice of measure by study design</i>	Separate datasets: cohort vs case–control	Why RR cannot be estimated from a case–control study

## Objectives

The aim of this practical is to introduce the Stata commands used to calculate the main measures of disease and effect used in epidemiology—rates, rate ratios, odds, and odds ratios.

## 1. Analysis of cohort studies

**File:** Whitehall study (wha1110)

**Description:** This study reports a 20-year follow-up of civil servants. Subjects were recruited into the study over a 2-year period. Exposures of interest included smoking status and the grade of work, classified as high (administrative, professional or executive) and low (clerical or other).

### 1.1. Descriptive analysis

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

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

2. How many deaths (from all causes) occurred?

```
. tab all
```

3. How many deaths were due to coronary heart disease?

```
. tab IHD all, col
```

4. How many individuals were in each smoking category?

```
. tab smok
```

5. How many were in each working-grade group?

```
. tab work
```

**1.2. Risk Analysis**

1. Display the number of deaths (and survivors) by smoking status.

```
. tab smok all
```

2. Calculate the risk of death according to smoking status by adding row percentages to the table.

```
. tab smok all, row
```

3. Interpret the row percentage for all = 1 (deaths).

4. Assess the risk of death by working group and calculate the risk ratio for the high-grade vs low-grade groups. Interpret the results.

```
. cs all work
```

5. Repeat the previous analysis using the "exact" option. What is the difference between this command and the previous one?

```
. cs all work, exact
```

6. Suppose you already know the tabular counts shown below. Perform a "tabular" risk analysis to assess the risk of death by working group.

	Exposed	Unexposed	Total
Cases	a	b	a + b
Non-cases	c	d	c + d
Total	a + c	b + d	

```
. csi 181 97 55 66
```

**1.3. Incidence Rate Analysis**

1. Examine the effect of work grade on mortality using incidence-rate analysis.

```
. ir all work y
```

2. What is the mortality rate in the total sample?

3. Does the mortality rate differ significantly between the two work-grade groups?
4. What does the incidence-rate difference represent?
5. Perform a "tabular" incidence-rate analysis for work grade and total mortality using the `iri` command with the table below.

	Exposed	Unexposed	Marginal Totals
Disease Count	A	B	A + B
Person-years	C	D	C + D
Totals	A + C	B + D	

```
. iri 181 97 18647.61 12812.45
```

6. Examine the effect of smoking status on risk of death using incidence rates and rate ratios for exposures with more than two levels. Use the `tabrate` and `mhrate` commands (`findit tabrate`).

```
. gen y1 = y * 100000
. tabrate all smok, e(y1) per(100000000)

. mhrate all smok, e(y) c(1,0)
. mhrate all smok, e(y) c(2,1)
. mhrate all smok, e(y) c(3,2)
```

7. Plot a graph of death rates by smoking status.

```
. tabrate all smok, e(y1) per(100000000) graph border xlab(0(1)3)
```

8. What does the rate ratio represent when using the `mhrate` command across all smoking categories?

```
. mhrate all smok, e(y)
```

9. Create a new variable, **smok2**, taking the values 0 for never/ex-smokers and 1 for current smokers. Perform incidence-rate analyses using both the `ir` command and `tabrate/mhrate` commands and compare the results.

## 2. Analysis of case-control studies

**File:** mwanza

**Description:** Cases were all women found HIV-positive in a cross-sectional survey of 12 communities in Mwanza, Tanzania. Controls were randomly selected from HIV-negative women. Risk factors of interest included educational level and the presence of skin incisions or tattoos..

1. Calculate the odds of HIV according to educational level. Do the odds vary with duration of education?

```
. tab case ed
```

2. Examine the odds of disease using the `tabodds` command. What does the output show?

```
. tabodds case ed
```

3. What are the null hypotheses for the test of homogeneity and the score test for trend? How do these tests differ?

4. Calculate the odds ratio using the `mhodds` command. What does this show?

```
. mhodds case ed
```

5. Using the `cc` command, examine whether the odds of HIV vary by the presence of skin incisions or tattoos. How should you handle missing values in the skin variable?

```
. recode skin 9 = .  
. cc case skin
```

6. Calculate the odds ratio of HIV (with 95% CI's) using the `cc`, `exact` option. What do you observe?

```
. cc case skin, exact
```

7. Repeat the previous command, but calculate 90% CI's instead.

```
. cc case skin, level(90) exact
```

8. If you know the tabular cell counts, use the `cci` command to assess the odds of HIV according to skin incisions or tattoos.

```
. cci 76 97 73 152, exact
```

9. Create a new variable, `ed2`, taking the value 0 for women with no formal education and 1 for those with any education. Perform odds analysis using both the `cc` and `tabodds/mhodd`s commands. Compare and interpret the results.

```
. gen ed2 = (ed > 0)
```