

## Module 7: Data Analysis



© 2007. The World Bank Group. All rights reserved.



### *Learning Objectives*

At the end of this module, participants should understand:

- basic data analysis concepts
- the relationship among types of data, types of samples, and data analysis techniques
- the use of data analysis in monitoring



2

Europe and Central Asia Region and World Bank Institute Evaluation Group



## *Data Analysis Strategy*

### **Key Choice:**

- Quantitative Analysis
- Qualitative Analysis

3

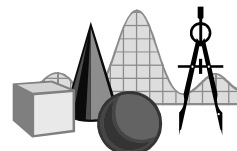
Europe and Central Asia Region and World Bank Institute Evaluation Group



## *Quantitative Analysis*

### **Three Basic Types**

- Descriptive Methods
- Associational Methods
- Deterministic Methods



4

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Variables

A **variable** is a characteristic or attribute that varies or changes over time or among individuals or groups

Examples: age, gender, agricultural production, miles of paved roads, number of children who are undernourished, hectares of national parks, etc.

Independent variable: the intervention or explanatory variable

Dependent variable: what we expect to change as a result of changes in the independent variable



5

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Examples

Independent variable: education

Dependent variable: income

Independent variable: access to skilled birth attendants

Dependent variable: maternal deaths



6

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Types of Variables

### Discrete versus Continuous:

- **discrete** variables are measured in units that cannot be subdivided, e.g., number of books on my shelf, number of children in a school
- **continuous** variables are measured in units that can be subdivided, e.g., temperature, time

### Nominal versus Ordinal:

- **nominal** (categorical) variables assign a label to categories, e.g., male, female; single, married, divorced
- **ordinal** variables also assign names to each possible response category but the categories can be ranked, e.g., level of satisfaction with training (unsatisfied to satisfied)

### Interval/ratio variables:

- **interval** scale uses equidistant measurement but zero point is not meaningful (e.g., celsius)
- **ratio** has a meaningful zero point (i.e., zero indicates absence of what is being measured) e.g., income, years of schooling, birth rates, kilometers of paved roads

7

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Quantitative Descriptive Methods

### Applied to one variable

- **Frequency/Percent Distribution**
  - A chart or table showing how often each value or range of values of a variable appear in a data set.
- **Central Tendency**
  - A measure of location of the middle or the center of a distribution.
  - Central tendency can refer to a **Mean, Median, or Mode**
- **Dispersion**
  - Describes how much the observations vary around the central tendency.
  - **Range and Standard Deviation**

8

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Frequency Distributions

How many males and females are in the program?

Distribution of Respondents by Sex

Sex	Number	%
Male	100	33%
Female	200	67%
Total	300	100%

Narrative: Of the 300 people in this program, 67 percent are women and 33 percent are men.

9

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Describing Distributions

### Central Tendency:

- What are the typical characteristics?
  - Example: What is the average age of graduates?
  - Example: What is the average income in rural areas?

### Dispersion:

- How dissimilar or concentrated are cases on a characteristic?
  - Example: How much variation in ages?

10

Europe and Central Asia Region and World Bank Institute Evaluation Group



## **Measures of Central Tendency**

**The 3-Ms:** Mode, Median, Mean

**Mode:** most frequent response

**Median:** midpoint of the distribution

**Mean:** arithmetic average

11

Europe and Central Asia Region and World Bank Institute Evaluation Group



## **Measures of Central Tendency: Number of vehicles per hour**

9	31	
17	34	mode =
19	38	median =
23	41	mean =
23	151	
28	Sum = 414	

12

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Measure of Dispersion: Range and Standard Deviation

**Range:** the difference between the largest and the smallest values

**Standard deviation:** measures the dispersion of scores – the distance from the mean

- Small standard deviation: not much dispersion; most of the data or “scores” are close to the mean
- Large standard deviation: lots of dispersion and many scores are far from the mean

13

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Measure of Dispersion: Hours of television watched per month

	11	3	
	16	4	<b>Which distribution</b>
	18	6	<b>has the larger</b>
	19	12	<b>standard</b>
	<u>21</u>	<u>60</u>	<b>deviation? Why?</b>
<b>Sum =</b>	85	85	
<b>Mean =</b>	17	17	
<b>Median =</b>	?	?	

14

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Standard Deviation

As deviations grow large, so too does the variance – potentially on a different magnitude from the data in the distribution we’re examining.

We only square the deviations to keep the sum from being zero (a property of the mean). Now that we have a non-zero number, take the square-root to get a statistic that is back in the metric in which we started out:

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}, \quad \text{so} \quad s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

s is the sample standard deviation.

15

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Variance

Deviation defined:  $(X_i - \bar{X})$

One way to measure all deviation in a distribution is to add the deviations from each observation:  $\sum(X_i - \bar{X})$

But we know that one of the special properties of the mean is that it is always equal to zero, so its not particularly useful in understanding dispersion.

By squaring each deviation, we can produce a value that will always be positive.

If we divide the sum by the number of observations, we obtain the *variance*.

$$\frac{\sum(X_i - \mu)^2}{N} = \text{variance of the population distribution} = \sigma^2$$
$$\frac{\sum(X_i - \bar{X})^2}{n-1} = \text{variance of the sample distribution} = s^2$$

16

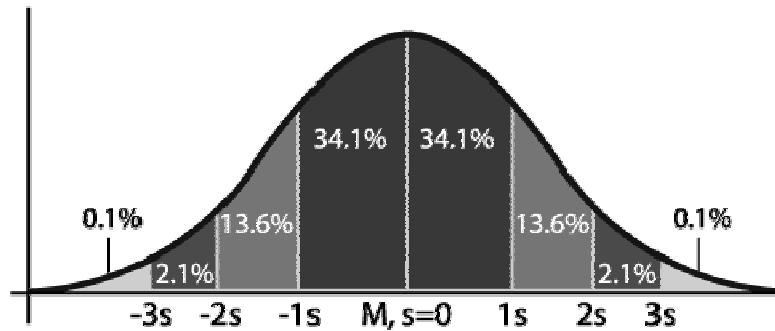
Europe and Central Asia Region and World Bank Institute Evaluation Group



## Measure of Dispersion: Standard Deviation

Normal Distribution: Bell-shaped curve

- 68.26% of the variation is within 1 standard deviation of the mean
- 95.44% of the variation is within 2 standard deviations of the mean



Note: this is not to scale.

17

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Descriptive Statistics

Applied to two or more variables

- Comparison of Means
- Cross-tabulation

18

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Comparison of Means

Do males earn more than females?

Or, is sex related to income differences?

Independent Variable=Gender

Dependent Variable=Income

Gender	Mean Income
Male	\$924
Female	\$798

19

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Cross-Tabulation

- Used when working with nominal and ordinal data
- Can be used with interval/ratio data that has been categorized

Teachers	Introduced new methods into classroom	Have not introduced new methods into classroom	Total
Received 1 week of training on using modern pedagogy in the classroom	25%	75%	100%
Did not receive training, but received book on using modern pedagogy in the classroom	15%	85%	100%

### Interpretation:

- Teachers trained in modern pedagogy are somewhat more likely (25%) to introduce modern pedagogical methods into their classrooms as compared to teachers who received only books about modern pedagogy (15%).
- Appears to be some relationship, but how strong is it?

20

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Associational Statistics

### Strength and Direction

#### How strong is the association?

- Several different measures of association
- Some measures of association range from zero to 1
- Others range from -1 to +1

Association does not prove causation!

21

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Establishing Causality

*Causality:* In impact evaluations, our ultimate goal often is to identify the **causal** relationships among phenomena we study

There are three factors necessary for causal inference:

1. The cause must precede the effect. Changes in the independent variable must occur before changes in the dependent variable.
2. The cause and effect must be related (i.e., correlated).
3. Other explanations of the cause-effect relationship must be eliminated (i.e., rule out spurious or confounding factors)

22

Europe and Central Asia Region and World Bank Institute Evaluation Group

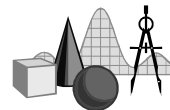


## Deterministic Methods

Deterministic statistical methods can be used to build upon descriptive analyses to explore causal relationships

Types include...

- **Bivariate (simple) regression**
- **Multivariate (multiple) regression**
- **Approaches using categorical data**



23

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Deterministic Methods

### Bivariate Regression Model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i, \quad i = 1, \dots, n$$

- $X_1$ , is an *independent variable (regressor)*
- $\beta_0$  = unknown population intercept
- $\beta_1$  = effect on  $Y$  of a change in  $X_1$
- $u$  = "error term" (omitted factors)

### Population Multiple Regression Model:

Consider the case of two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, \dots, n$$

- $\beta_1$  = effect on  $Y$  of a change in  $X_1$ , holding  $X_2$  constant
- $\beta_2$  = effect on  $Y$  of a change in  $X_2$ , holding  $X_1$  constant

24

Europe and Central Asia Region and World Bank Institute Evaluation Group



## *Data Analysis in Monitoring Plans*

- Data Analysis in monitoring utilizes more basic methods
- Need to consider:
  - Use of data subsets
  - Need for comparisons

25

Europe and Central Asia Region and World Bank Institute Evaluation Group



## *Identify Data Subsets in Monitoring Plans*

### **Be Cautious of Overly Aggregated Data!!**

Data for each outcome measure should be broken out (disaggregated) to show outcomes for different sub-groups or subunits.

26

Europe and Central Asia Region and World Bank Institute Evaluation Group



## **Possible Data Subsets for Monitoring: Demographic Characteristics**

### **Demographic Characteristics**

- By household income (or proxy for this)
- By gender
- By age group
- By race/ethnicity
- By geographical area, such as rural versus urban locations, by district, by municipality

27

Europe and Central Asia Region and World Bank Institute Evaluation Group



## **Possible Data Subsets for Monitoring: Service Characteristics**

### **Service Characteristics**

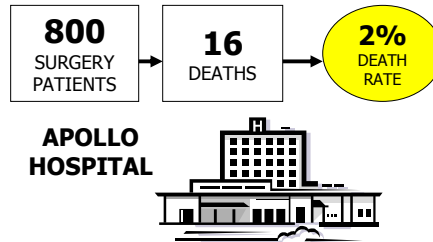
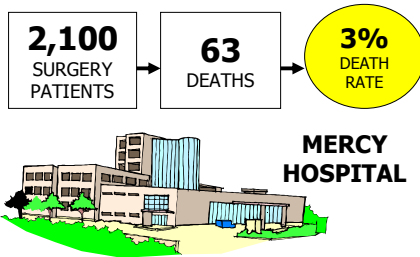
- By organizational unit, if the service is provided in more than one facility (such as different health clinics, schools, parks, water bodies, or districts)
- Type of procedure used by service provider
- Amount or level of service
- By customer needs

28

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Discussion: Which Hospital Would You Choose?

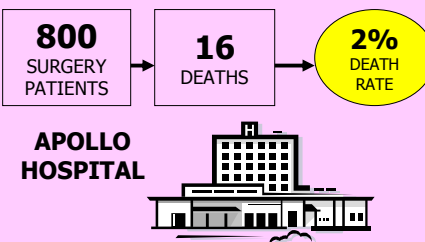
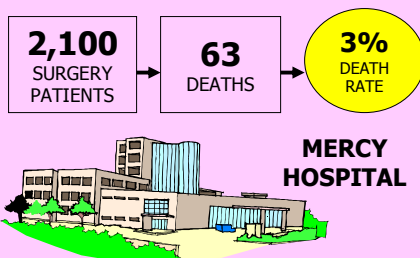


29

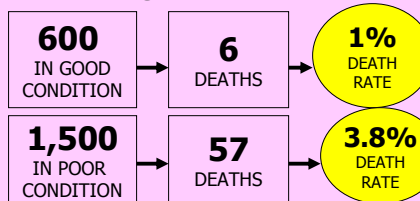
Europe and Central Asia Region and World Bank Institute Evaluation Group



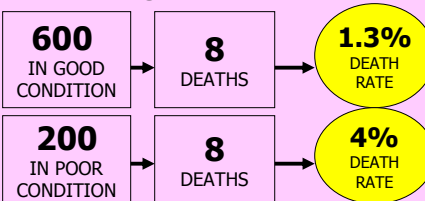
## Discussion: Which Hospital Would You Choose?



**BUT...**



**BUT...**



30

Europe and Central Asia Region and World Bank Institute Evaluation Group



## ***Comparisons for Interpreting Outcome Data for Monitoring***

- Across time (such as previous versus current year, or current month versus same month last year)
- Against targets set by agency
- Across demographic characteristics
- Across service delivery characteristics
- With other similar programs
- With other cities, countries, or regions

31

Europe and Central Asia Region and World Bank Institute Evaluation Group



## ***Qualitative Data Analysis***

- Data from narrative documents, open-ended
- Interviews, focus groups, unstructured
- Observations
- Methods for Analysis
  - ✓ Inductive Analysis
  - ✓ Logical Analysis
  - ✓ Synthesis

32

Europe and Central Asia Region and World Bank Institute Evaluation Group



## *Qualitative Data Analysis*

- Identify common words, ideas, themes
- Develop spreadsheet or write on cards
- Identify “quotable quotes”

33

Europe and Central Asia Region and World Bank Institute Evaluation Group



## *Qualitative Data Analysis*

Greatest Risk: Bias

- Hard to recognize things you don't expect

Have a second person do the analysis

- Compare results
- Work out differences

34

Europe and Central Asia Region and World Bank Institute Evaluation Group



## *Qualitative Data Analysis*

### Writing about results

- Feature major themes
- “A number of participants said”
- Highlight interesting perspectives even if only said by one or two people
- Do not report numbers or percentages

35

Europe and Central Asia Region and World Bank Institute Evaluation Group



## *Evaluation Plan*

### *Case Projects*

- For each question or subquestion in the evaluation plan, determine the best analysis strategies for answering each question or subquestion.
- Insert information in the Data Analysis column

36

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Evaluation Plan

General Questions	Specific Sub-Questions	Type of Question	Type of Design	Indicators & Measures	Data Sources	Data Collection & Sampling	Data Analysis

37

Europe and Central Asia Region and World Bank Institute Evaluation Group



## Group Project Exercise

### Your project – for evaluation

- For each question or subquestion, determine the best analysis strategy

38

Europe and Central Asia Region and World Bank Institute Evaluation Group