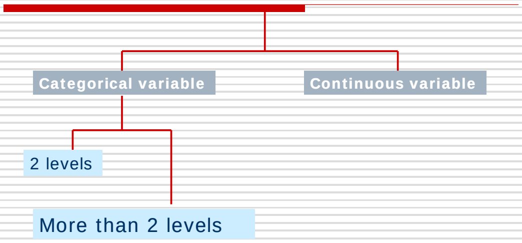
Outline

- Scales of measurement
- Descriptive statistics
- Inferential statistics
  - Confidence Interval
  - Hypothesis testing
  - Bivariate versus multivariate analysis

Definition of some basic terms

- **Variable:** A characteristic of the objects under observation that takes on different values for different cases, example: age, gender, diastolic blood pressure
- **Parameter:** descriptive measure computed from data of a population
- **Statistic:** descriptive measure computed from data of a sample

Scales of measurements

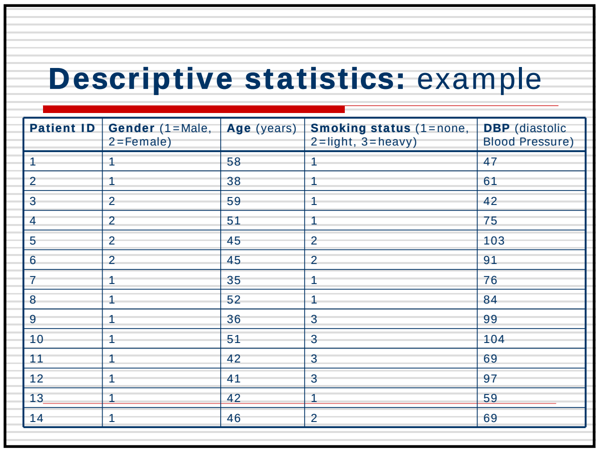


Categorical variable

Continuous variable

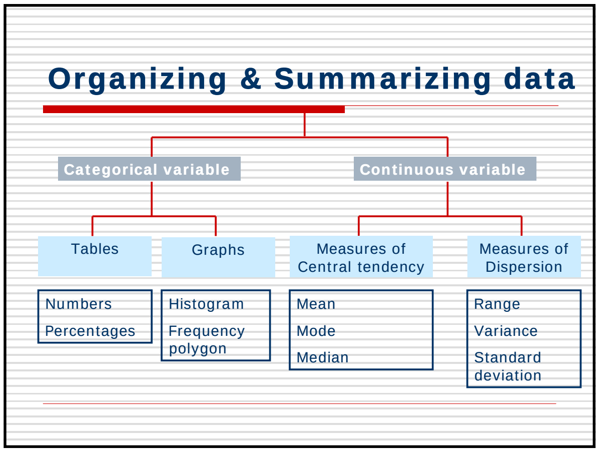
2 levels

More than 2 levels



Descriptive statistics: example

Patient ID	Gender (1=Male, 2=Female)	Age (years)	Smoking status (1=none, 2=light, 3=heavy)	DBP (diastolic Blood Pressure)
1	1	58	1	47
2	1	38	1	61
3	2	59	1	42
4	2	51	1	75
5	2	45	2	103
6	2	45	2	91
7	1	35	1	76
8	1	52	1	84
9	1	36	3	99
10	1	51	3	104
11	1	42	3	69
12	1	41	3	97
13	1	42	1	59
14	1	46	2	69



Organizing & Summarizing data

Categorical variable

Tables

Numbers

Percentages

Graphs

Histogram

Frequency polygon

Continuous variable

Measures of Central tendency

Mean

Mode

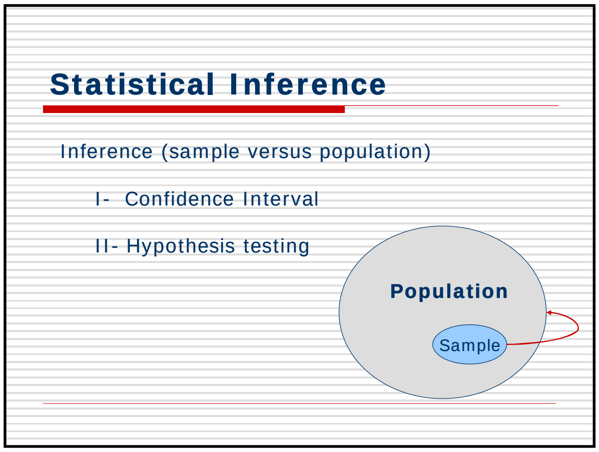
Median

Measures of Dispersion

Range

Variance

Standard deviation

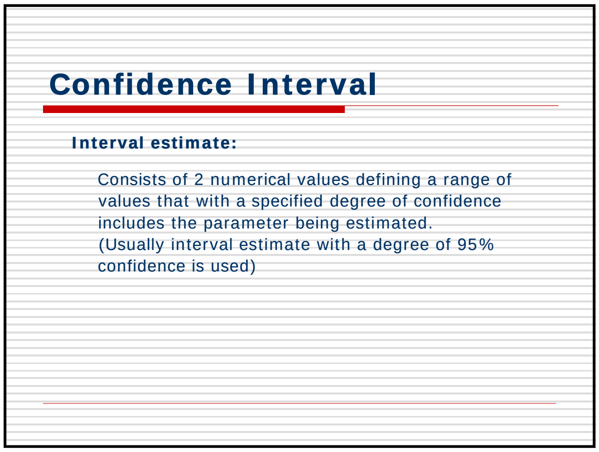


Statistical Inference

Inference (sample versus population)

I- Confidence Interval

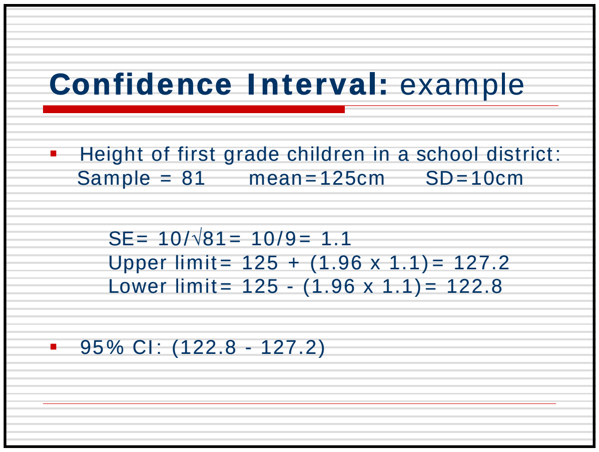
II- Hypothesis testing



Confidence Interval

Interval estimate:

Consists of 2 numerical values defining a range of values that with a specified degree of confidence includes the parameter being estimated.
(Usually interval estimate with a degree of 95% confidence is used)

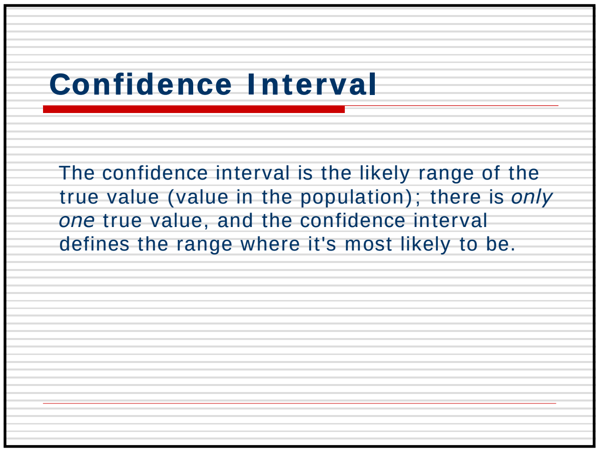


Confidence Interval: example

- Height of first grade children in a school district:
  Sample = 81 mean=125cm SD=10cm

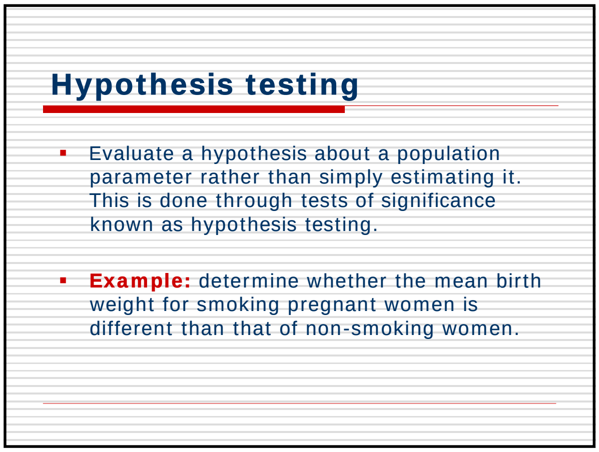
  SE= 10/√81= 10/9= 1.1
  Upper limit= 125 + (1.96 x 1.1)= 127.2
  Lower limit= 125 - (1.96 x 1.1)= 122.8

- 95% CI: (122.8 - 127.2)



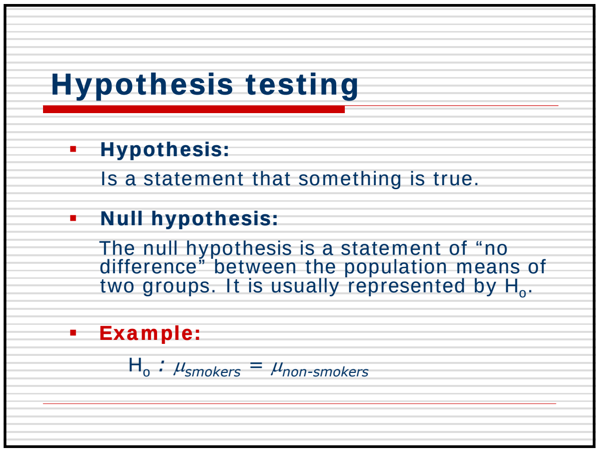
Confidence Interval

The confidence interval is the likely range of the true value (value in the population); there is *only one* true value, and the confidence interval defines the range where it's most likely to be.



Hypothesis testing

- Evaluate a hypothesis about a population parameter rather than simply estimating it. This is done through tests of significance known as hypothesis testing.
- **Example:** determine whether the mean birth weight for smoking pregnant women is different than that of non-smoking women.



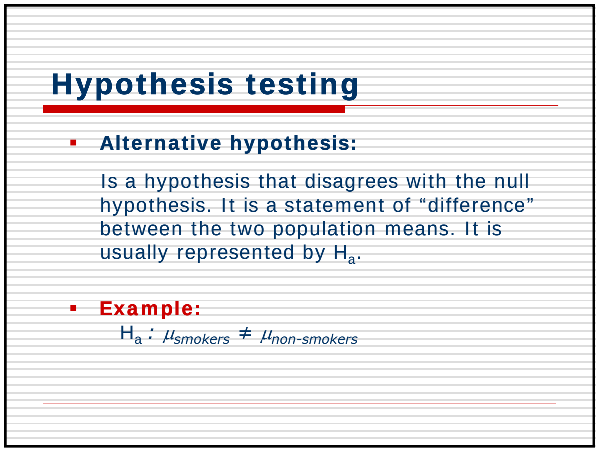
Hypothesis testing

- **Hypothesis:**

  Is a statement that something is true.
- **Null hypothesis:**

  The null hypothesis is a statement of “no difference” between the population means of two groups. It is usually represented by H_o.
- **Example:**

  $H_o : \mu_{smokers} = \mu_{non\text{-}smokers}$

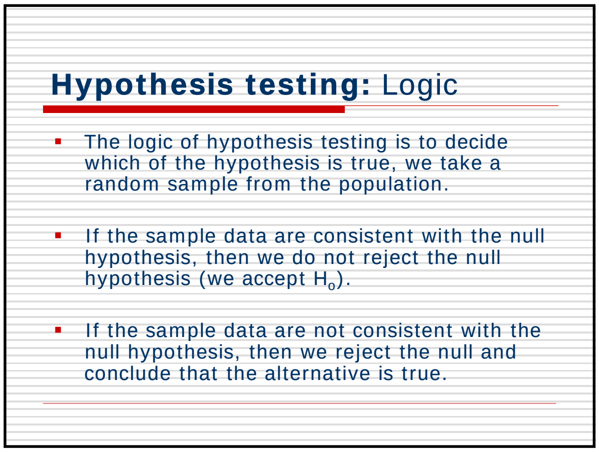


Hypothesis testing

- **Alternative hypothesis:**

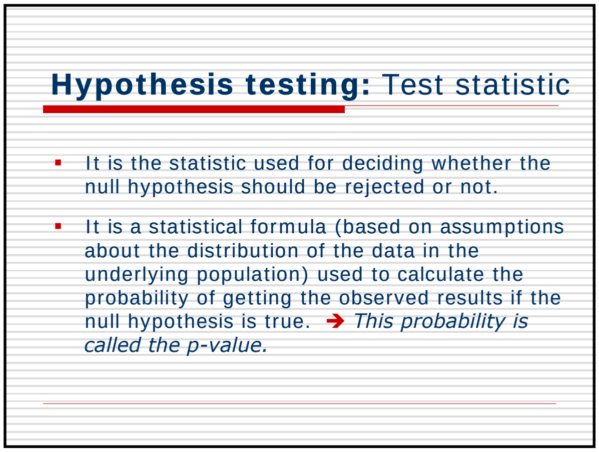
  Is a hypothesis that disagrees with the null hypothesis. It is a statement of “difference” between the two population means. It is usually represented by H_a.
- **Example:**

  $H_a : \mu_{smokers} \neq \mu_{non\text{-}smokers}$



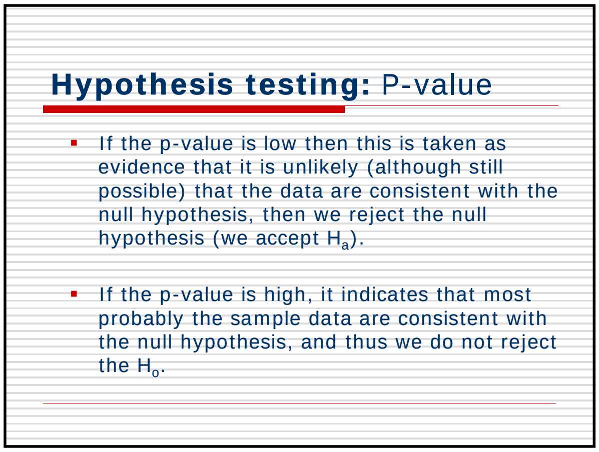
Hypothesis testing: Logic

- The logic of hypothesis testing is to decide which of the hypothesis is true, we take a random sample from the population.
- If the sample data are consistent with the null hypothesis, then we do not reject the null hypothesis (we accept H_o).
- If the sample data are not consistent with the null hypothesis, then we reject the null and conclude that the alternative is true.



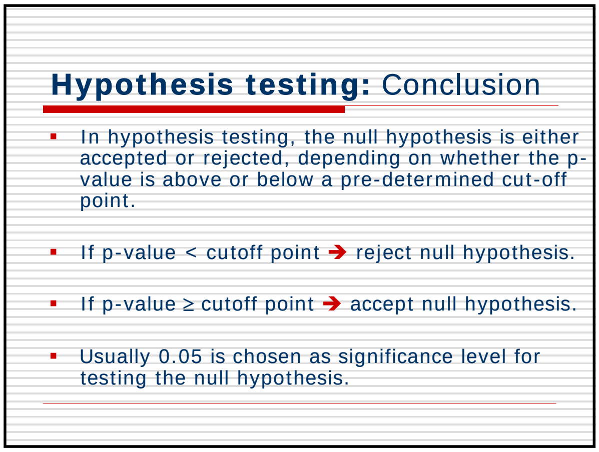
Hypothesis testing: Test statistic

- It is the statistic used for deciding whether the null hypothesis should be rejected or not.
- It is a statistical formula (based on assumptions about the distribution of the data in the underlying population) used to calculate the probability of getting the observed results if the null hypothesis is true. ➔ *This probability is called the p-value.*



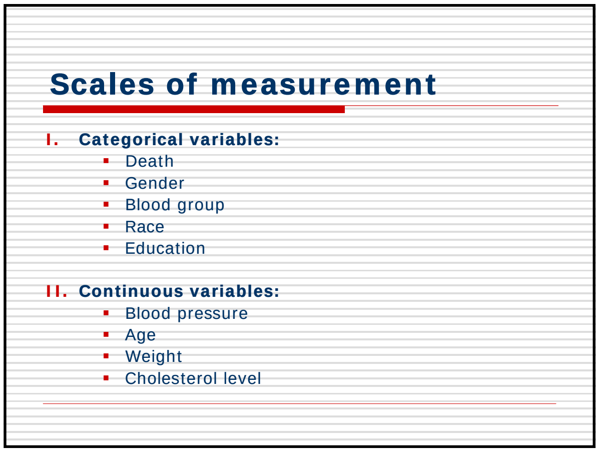
Hypothesis testing: P-value

- If the p-value is low then this is taken as evidence that it is unlikely (although still possible) that the data are consistent with the null hypothesis, then we reject the null hypothesis (we accept H_a).
- If the p-value is high, it indicates that most probably the sample data are consistent with the null hypothesis, and thus we do not reject the H_o.



Hypothesis testing: Conclusion

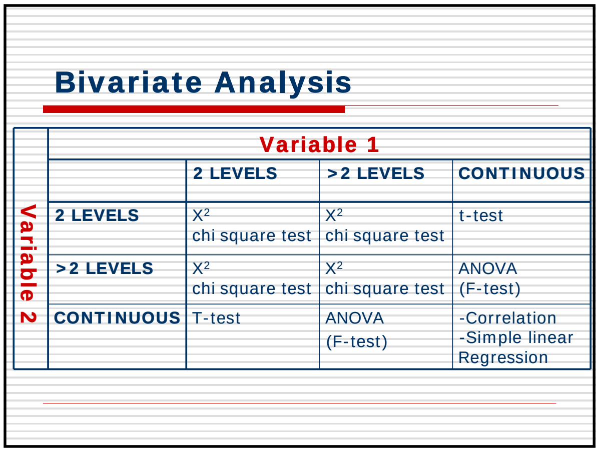
- In hypothesis testing, the null hypothesis is either accepted or rejected, depending on whether the p-value is above or below a pre-determined cut-off point.
- If p-value < cutoff point ➔ reject null hypothesis.
- If p-value ≥ cutoff point ➔ accept null hypothesis.
- Usually 0.05 is chosen as significance level for testing the null hypothesis.



Scales of measurement

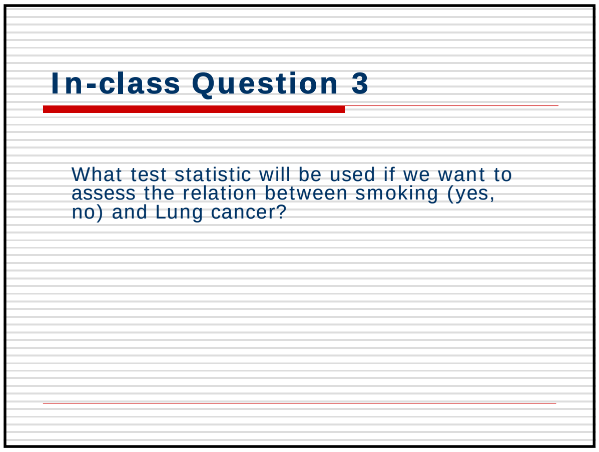
I. **Categorical variables:**
   - Death
   - Gender
   - Blood group
   - Race
   - Education

II. **Continuous variables:**
   - Blood pressure
   - Age
   - Weight
   - Cholesterol level



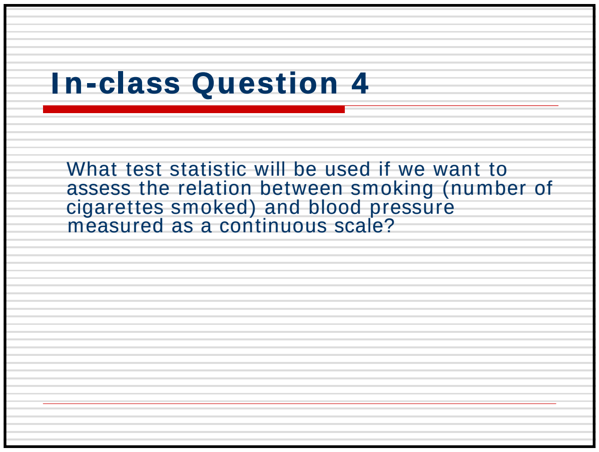
Bivariate Analysis

		Variable 1		
		2 LEVELS	**>2 LEVELS**	**CONTINUOUS**
Variable 2	**2 LEVELS**	X^2 chi square test	X^2 chi square test	t-test
	>2 LEVELS	X^2 chi square test	X^2 chi square test	ANOVA (F-test)
	CONTINUOUS	T-test	ANOVA (F-test)	-Correlation -Simple linear Regression



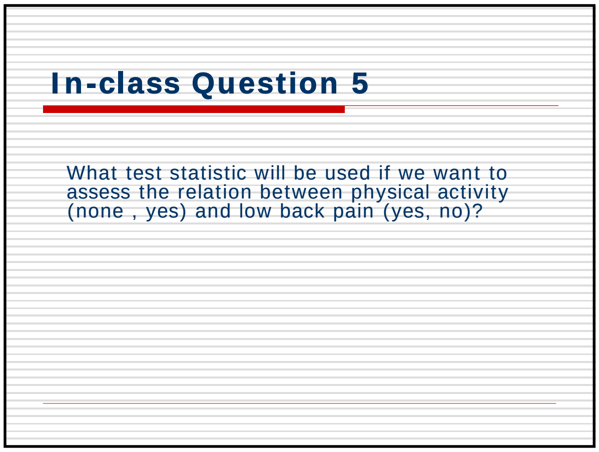
In-class Question 3

What test statistic will be used if we want to assess the relation between smoking (yes, no) and Lung cancer?



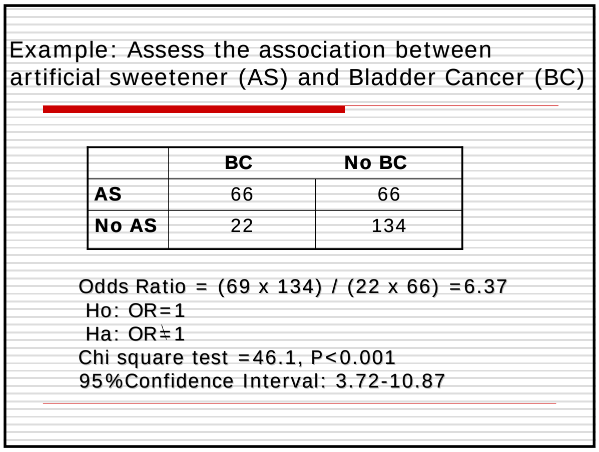
In-class Question 4

What test statistic will be used if we want to assess the relation between smoking (number of cigarettes smoked) and blood pressure measured as a continuous scale?



In-class Question 5

What test statistic will be used if we want to assess the relation between physical activity (none , yes) and low back pain (yes, no)?



Example: Assess the association between artificial sweetener (AS) and Bladder Cancer (BC)

	BC	No BC
AS	66	66
No AS	22	134

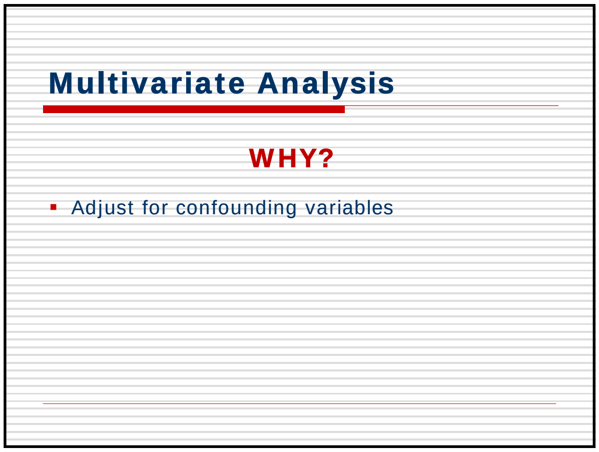
Odds Ratio = (69 x 134) / (22 x 66) =6.37

Ho: OR=1

Ha: OR≠1

Chi square test =46.1, P<0.001

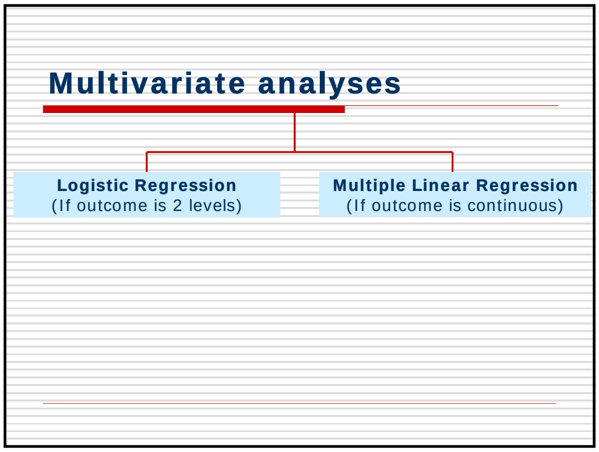
95%Confidence Interval: 3.72-10.87



Multivariate Analysis

WHY?

- Adjust for confounding variables



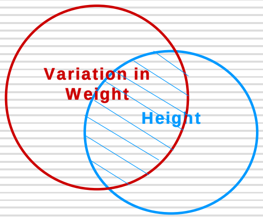
Multivariate analyses

Logistic Regression
(If outcome is 2 levels)

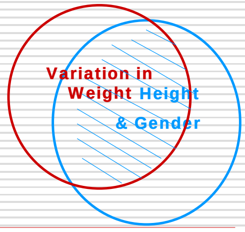
Multiple Linear Regression
(If outcome is continuous)

Multiple Linear Regression: Ex 1

Simple Linear Regression



Multiple Linear Regression


Variation in
Weight
Height
& Gender