Inferential Statistics: Comparing Groups

Dr. Aung Ye Naung Win

MBBS, M.Sc. (Tropical Medicine)

Department of Medical Research (HQ)

Outline of Presentation

-Types of data
-Types of statistical methods
-Inferential statistics
-Common statistical tests

Types of Data

QUALITATIVE

Data expressed by type

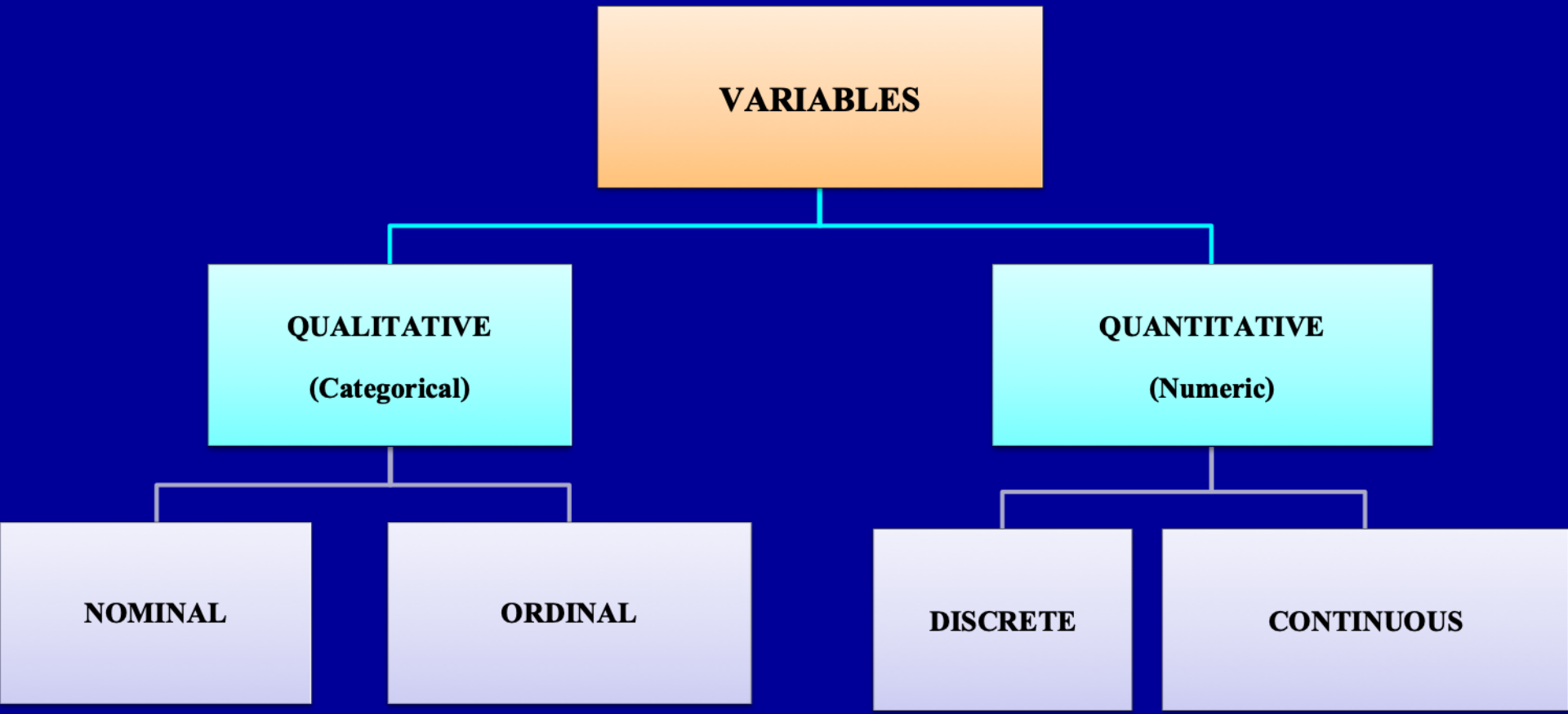
Data that has been described

QUANTITATIVE

Data classified by numeric value

Data that has been measured or counted

QUALITATIVE and QUANTITATIVE data are not mutually exclusive


VARIABLES
QUALITATIVE
(Categorical)
QUANTITATIVE
(Numeric)
NOMINAL
ORDINAL
DISCRETE
CONTINUOUS

Types of Data: Qualitative (Categorical) Data

NOMINAL DATA

- values that the data may have do not have specific order
- values act as labels with no real meaning
- Binomial: two possible values (categories, states)
- Multinomial: more than two possible values (categories, states)

e.g. Health status healthy =1 sick=2

e.g. Treatment new regimen = 1 standard regimen = 2

e.g. hair colour brown =1 blond =2 black =100

ORDINAL DATA

- values with some kind of ordering
- data that has been measured or counted

e.g. social class: upper=1 middle = 2 working = 3

e.g. glioblastoma tumor grade: 1 2 3 4 5

e.g. position in a race: 1 st 2 nd 3 rd

Adapted from: Dr. Craig Jackson, University of Central England

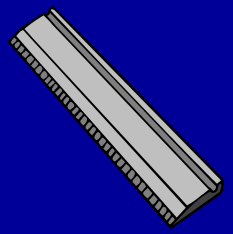
UCE Birmingham

Types of Data: Quantitative Data

DISCRETE

- distinct or separate parts, with no finite detail

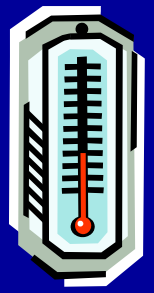
e.g children in family



CONTINUOUS

- between any two values, there would be a third

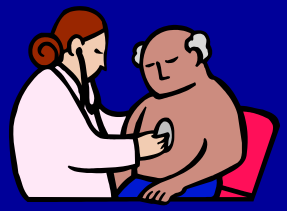
e.g between meters there are centimetres



INTERVAL

- equal intervals between values and an arbitrary zero on the scale

e.g temperature gradient



RATIO

- equal intervals between values and an absolute zero

e.g body mass index

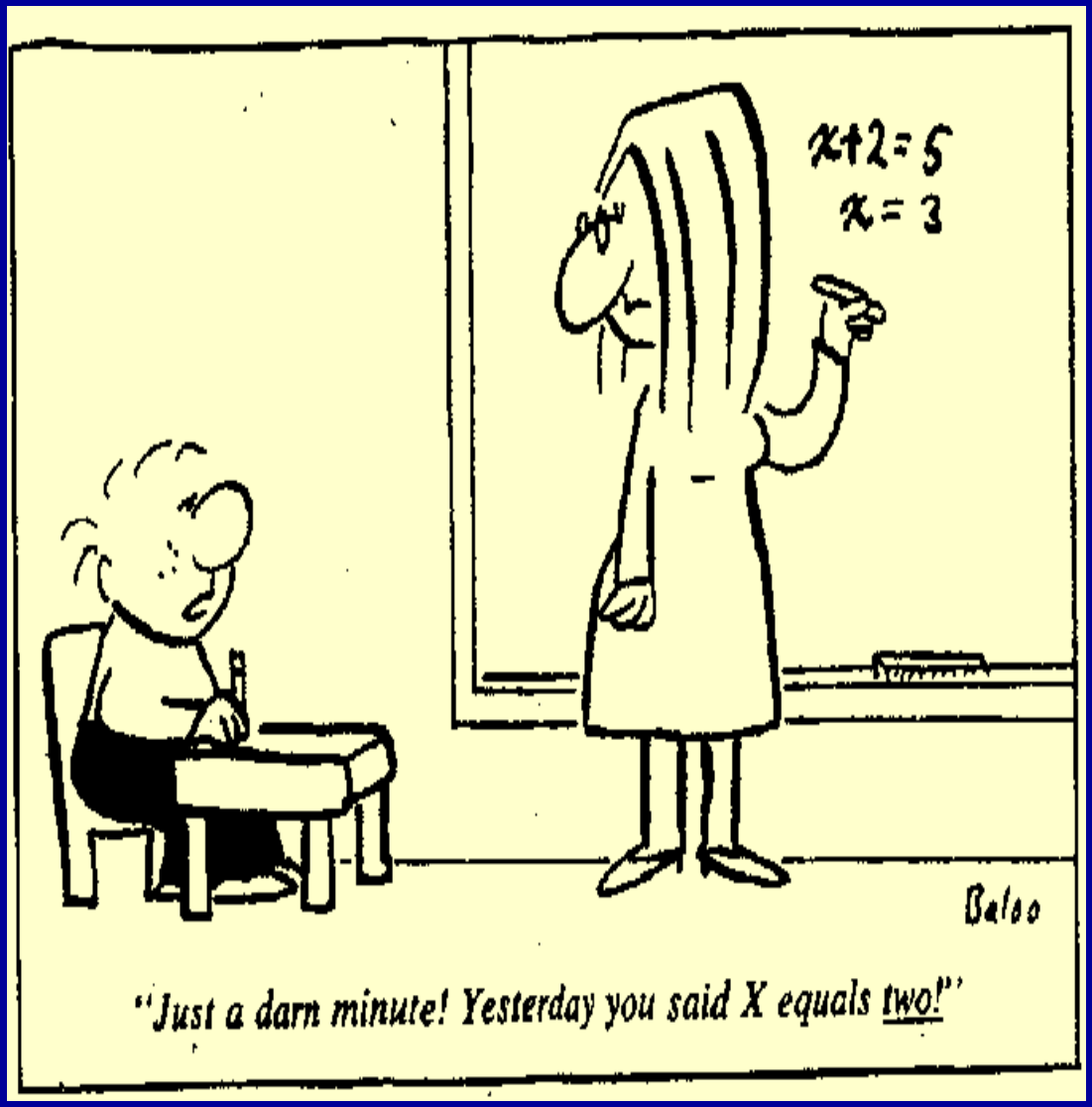
Adapted from: Dr. Craig Jackson, University of Central England

UCE Birmingham

Constant vs. Variable

Variables are the specific properties that have the ability to take different values.

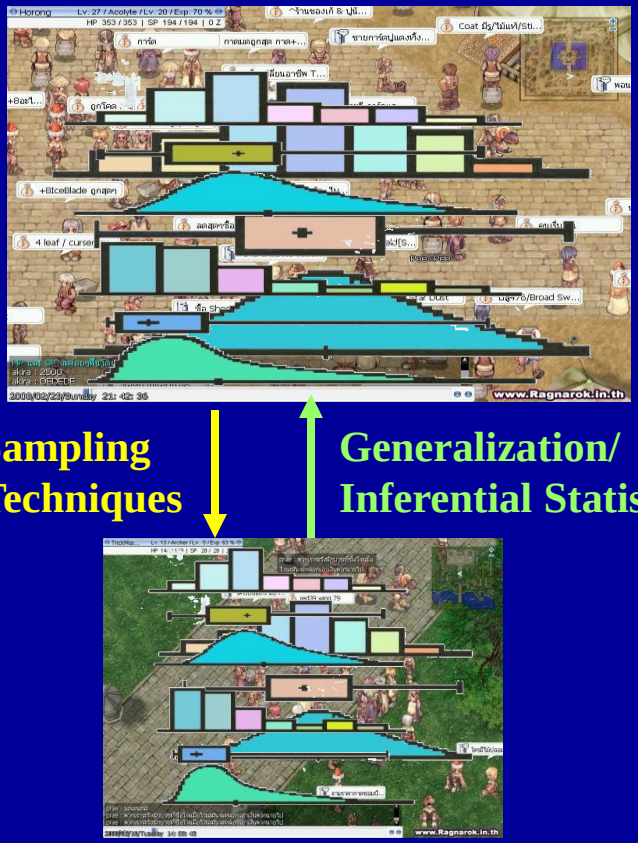
Constants are the specific properties that cannot vary or won't be made to vary.



Types of Statistical Methods

Types of Statistics

- By Level of Generalization
  - Descriptive Statistics
  - Inferential Statistics
    - Parameter Estimation
    - Hypothesis Testing
      - Comparison
      - Association
      - Multivariable data analysis



- By Level of Underlying Distribution
  - Parametric Statistics
  - Non-parametric Statistics

- **Statistics** - set of methods for **organizing, summarizing and interpreting** information

- **Inferential statistics** allow us to study samples and then make **generalizations about the population** from which they were selected.
  - Estimation (Confidence Interval) and Hypothesis testing (Statistical Tests of Significance)

Inferential statistics

- Use of statistical tests to test for significant relationships among variables
- To compare groups and test hypothesis
- To Answer research question in the study

Two approaches to statistical inference

- Estimation
  - Point estimate: disease measure (rate, mean) or measure of effect (Odds Ratio, Relative Risk, etc)
  - Confidence interval
- Hypothesis testing

Inferential Statistics (cont.)

Hypothesis testing

- Generate null and alternative hypothesis
- Test statistic
- P value
- Type I error and Type II error
- Power of a test

*A **HYPOTHESIS** is a **testable declarative statement** about an **expected relationship** between one or more independent variables and the dependent variable under study*

Hypothesis testing

Null hypothesis

- Negative declaration/ Hypothesis of no difference
- Any observed differences are entirely due to sampling errors (chance)
- Example: The hypoglycemic action of Drug A is similar to that of Drug B

Alternative hypothesis

- Positive declaration/ Hypothesis of difference
- Example: Under 3 years children in Township A are taller than those in Township B

Hypothesis testing (cont.)

Test statistic and p-value

- A test of hypothesis can give one of four results as follows:

		Your decision	
		Accept Ho	**Reject Ho**
	Accept Ho	√	**Type I error (α)**
Truth			
	Reject Ho	**Type II error (β)**	√

Hypothesis testing (cont.)

Test statistic and p-value

- A small p value (e.g., p value < 0.05) would suggest that there is less than a 5% probability of error in generalizing the results from the sample to the population (Type I error)
- So, we could inference the result of sample will be the same as the population
- So, the statement is statistically significant

Statistics – *P*-value

- Statistical significance:
  - *P* value $\leq$ 0.05 - Significant
  - *P* value < 0.01 - Very significant
  - *P* value < 0.001 - Highly significant
- Express exact *P*-values rather than “significant” or “Non-significant”.

P-value

- In the context of significance tests (eg chi-square),
- the *P* - value expresses the probability that an observed difference would occur by chance alone
- Small *P*- values indicate stronger evidence of difference (evidence to reject the null hypothesis)

‘P’ value < 0.05 – what does it really mean?

- “p” is the probability that the result is just by chance
- It is nothing but alpha error
- That is when we conclude that there is an association but in reality there is no association.
- Lower the p value more confident we are that the result is true

Inferential Statistics

- Procedure for statistical inference may involve:

Univariate Analysis

-One variable
E.g: Age, Gender, Income, etc

Bivariate Analysis

-Two variables
E.g: Parity & Fetal outcome

Multivariate Analysis

- Many variables
E.g: Age, Gender & knowledge level

Univariate Analysis

- Analysis of a single variable
- Does not involve relationship between two or more variables
- Purpose is more towards descriptive rather than explanatory

Univariate Analysis (cont.)

Type of Variable		Example	Statistic
Categorical	Dichotomous	Yes/No Male/Female	Percentage and 95% confidence interval (CI)
	Nominal	Ethnicity Cause of death	Percentage and 95% CI
	Ordinal	Stage of cancer Attitude scale	Percentage and 95% CI
Continuous		Age Serum bilirubin	Mean and 95% CI
Time to event (survival time)		Time to metastasis	5-year survival rate and 95% CI

Bivariate Analysis

- Bivariate analysis is done to assess relationship between 2 variables

- Usually one variable is designated as independent variable (or exposure) and another variable as dependent variable (or outcome)

Bivariate Statistical Methods

(A) Comparing groups on a categorical variable

- Independent variable = categorical
- Dependent variable = categorical

(B) Comparing groups on a continuous variable

- Independent variable = categorical
- Dependent variable = continuous

(C) Relating a continuous variable to another continuous variable

- Independent variable = continuous
- Dependent variable = continuous

Predictor variable	Outcome variable			
	Continuous normally dist.	Continuous not normally dist, ordinal>2category	Nominal >2 category	Dichotomo us
Continuous normally dist	Correlation Regression	Spearman rank correlation	ANOVA	Logistic regression
Continuous not normally dist, ordinal>2cate:	Spearman rank correlation	Spearman rank correlation	Kruskal-Wallis	Wilcoxon rank sum
Nominal >2 categories	ANOVA	Kruskal-Wallis	Chi-square	Chi-square
Dichotomous	t-test	Wilcoxon rank sum	Chi-square	Chi-square

Common Statistical Tests

Chi square test

- Commonly used statistical test & most flexible
- Use for categorical variables
- Based on cross-tabulation
- Each observation must be independent
- 80% of the expected cell frequencies >5
- All expected frequencies >1
- (no zero for an observed frequency in general)
- Fisher exact test is used when any cell contains expected frequency<5

Table: Prematurity by sex

Sex	Premature No. (%)#	Term No. (%)#	Total No. (%)
Male	33 (40.2)	49 (59.8)	82 (100)
Female	12 (20.3)	47 (79.7)	59 (100)
Total	45 (31.9)	96 (68.1)	141 (100)

 # Row percent, i.e., % within gender; p=0.02

'There was a significant difference in pre-maturity between males and females (40.2% vs 20.3%, P = 0.02)'

Table 1 Social situation of HIV-OVC and their neighboring children

Social characteristics	HIV-OVC (n=300)		Non-HIV (n=300)	
	Number	Percent	Number	Percent
Sibling/child displacement***				
Yes	60	20.0	8	2.7
No	240	80.0	292	97.3
Family dispersion***				
Yes	61	20.3	4	1.3
No	239	79.7	296	98.7

*** *p< 0.001, OVC-Orphans and Vulnerable Children*

How to cite this article: Mon MM, Saw S, Nu-Oo YT, San KO, Myint WW, Aye SS, *et al*. Threat of HIV/AIDS in children: social, education and health consequences among HIV orphans and vulnerable children in Myanmar. WHO South-East Asia J Public Health 2013;2:41-6.

SPSS output – Chi square test

Crosstab

				typchd type of children		
				1	2	Total
q2.4	family dispersion	yes	Count	61	4	65
			% within typchd type of children	20.3%	1.3%	10.8%
		no	Count	239	296	535
			% within typchd type of children	79.7%	98.7%	89.2%
Total			Count	300	300	600
			% within typchd type of children	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	56.058[a]	1	.000		
Continuity Correction[b]	54.108	1	.000		
Likelihood Ratio	66.140	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	55.964	1	.000		
N of Valid Cases[b]	600				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 32.50.

b. Computed only for a 2x2 table

t-test

- For categorical and continuous variables
- It uses the means of the two sets of data and their SD

SPSS output – t-test

Group Statistics

	Country of Origin	N	Mean	Std. Deviation	Std. Error Mean
Miles per Gallon	American	248	20.13	6.377	.405
	European	70	27.89	6.724	.804

Independent Samples Test

	t-test for Equality of Means						
						95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Miles per Gallon	-8.887	316	.000	-7.76	.874	-9.482	-6.045

Correlation

- For two continuous variables
- It can be visualized by scatter plot
- These need ratio, interval or ordinal data
- It cannot be used with nominal data
- It requires reasonably large data sets to work best
- Statistical test used based on data
  - **Spearman's rank correlation coefficient:** works with ordinal data
  - **Pearson's correlation coefficient:** works with interval or ratio data

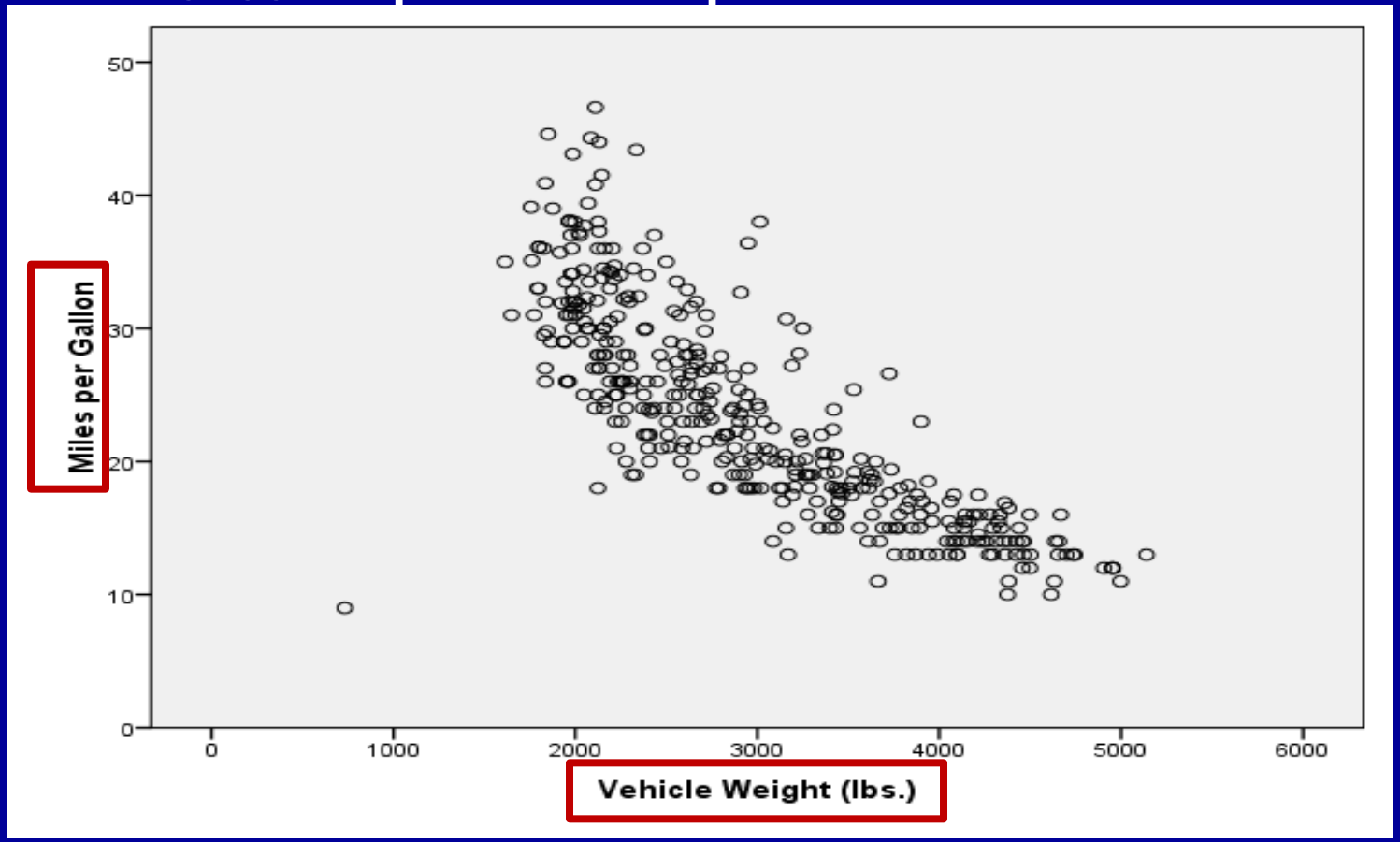
Correlation (cont.)

- To explore whether the two variables are associated with each other
- Pattern on the scatter diagram → basic nature and strength of the relationship
- To what degree are they associated?

  "strong", "moderate", or "poor"
- What direction does the relationship?

  "positive", "negative"

SPSS output: scatter plot



SPSS output: correlation

Correlations

		Miles per Gallon	Vehicle Weight (lbs.)
Miles per Gallon	Pearson Correlation	1	-.807**
	Sig. (2-tailed)	.	.000
	N	398	398
Vehicle Weight (lbs.)	Pearson Correlation	-.807**	1
	Sig. (2-tailed)	.000	.
	N	398	406

**. Correlation is significant at the 0.01 level (2-tailed).

Pattern & magnitude

significance

Regression

- Explaining the variation in an outcome (Y) by a predictor variable (X)
- Obtaining quantification of the relationship
- Making prediction

Two continuous variables

Methods for relating a continuous variable to another continuous variable		
Independent variable	**Dependent variable**	
	Continuous, Normal	***Continuous, non-Normal***
Continuous, Normal	Linear regression Pearson correlation	Spearman's rank correlation
Continuous, non-Normal	Spearman's rank correlation	Spearman's rank correlation

Multivariable Analysis

- Best prediction of outcome in a situation with many factors
- Control the effect of confounders
- A statistical tool for determining the unique contributions of various factors to a single event or outcome
- E.g: numerous associated factors with the development of coronary heart disease, including smoking, obesity, sedentary lifestyle, diabetes, elevated cholesterol level, and hypertension

Type of multivariable analysis according to type of outcome variable		
Type of outcome	Example of outcome variable	Type of multivariable analysis
Continuous	Blood pressure, weight, temperature	**Multiple linear regression**
Dichotomous	Death, cancer, intensive care unit admission	**Multiple logistic regression**
Time to event (survival time)	Time to death, time to cancer	**Proportional hazards analysis (Cox regression)**
Rare outcomes and counts	Time to leukemia, number of infections	**Poisson regression**

Choice of descriptive statistics

Type of outcome variable		Example	Statistic
Categorical	Dichotomous	Yes/No Male/Female	Frequency & percentage
	Nominal	Ethnicity Cause of death	Frequency & percentage
	Ordinal	Stage of cancer Attitude scale	Frequency & percentage
Continuous		Age Serum bilirubin	Mean ± SD (normally distributed data) Median & IQR (non-normally distributed)
Time to event (survival time)		Time to metastasis	Kaplan-Meir curve (median survival time and 5-year survival rate)

Statistics – Confidence interval Example

Study of HIV prevalence in TB patients at Site X

- Sample = 244 TB patients
- HIV prevalence = 70% (95% CI 66-76)

- Implies that we can be 95% confident that the true HIV prevalence among the TB population would lie anywhere between 66% and 76%.

Statistics – Confidence interval (95% CI)

- A Confidence Interval is a range of values within which the "true" population parameter is believed to be found with a given level of confidence

- The rationale for calculating CIs is the uncertainty which is always associated with using samples to make inferences on populations from which these samples originate

- A 95% CI means: We can be 95% confident that the true population value lies within its limits

The End of Inferential Statistics

Thank You