Research Methodology Workshop

23-25 March 2014

Principles of Quantitative Data Analysis

Dr. Ahmed Mohamed AlQasmi

Master of Science in Public Health (Biostatistics)

Doctor of Science (Biostatistics)

Director General of Planning

There are two major branches of statistics

1. Descriptive Statistics
2. Inferential Statistics

- Descriptive Statistics are used to describe data by summarizing them into more understandable way.

- We describe data in two ways
  1. Graphical description
  2. Numerical description

- In graphical description, we can use different techniques according to type of data, such as pie chart, bar chart, histograms, stem-and-leaf plots

  and others.

- In numerical description, we can use Measures of Central Tendency (such as the mode, median, and mean) and Measures of Variability (such as the range, percentiles, variance, and standard deviation)

- Inferential Statistics are used to provide inferences about populations based on information in a sample from the population.

- In inferential statistics, we mostly focus on:
    1. Estimation
    2. Hypothesis Testing

Probability Distribution

This is the listing of all possible values of a variable and their probabilities.

- A useful probability distribution for a discrete variable is the Binomial Distribution
- A useful probability distribution for a continuous variable is the Normal Distribution
- The t-distribution is another useful one for a continuous variable in the case when the sample size is less than 30 (that is, $n < 30$).

Estimation

- Estimation procedures can be divided into two types:
  1. Point Estimation
  2. Interval Estimation

- In point estimation, a single number (point estimate) from the sample data is calculated and used to estimate the population parameter of interest.

- Here, we calculate one number and infer that the parameter is that number.

Interval Estimation

- In interval estimation, two numbers from the sample data are calculated and the interval formed by the two numbers is used as an interval estimate of the population parameter of interest.

- Here, we calculate two numbers and infer that the parameter lies in the interval between them.

- We evaluate how good the interval estimation procedure by examining the fraction of times in repeated sampling that interval estimates would include the parameter. This fraction is called the confidence coefficient (1-α).

- If we set the confidence coefficient to be 0.95, we are saying that 95% of the time in repeated sampling, intervals will contain the population parameter.

Common values for the t-coefficients with 90%, and 95% confidence levels

	95%	90%
Df	t-coefficient	t-coefficient
5	2.571	2.015
9	2.262	1.833
10	2.228	1.812
12	2.179	1.782
15	2.131	1.753
20	2.086	1.725
24	2.064	1.711
27	2.052	1.703
∞	1.960	1.65

SPSS Output

Descriptives

			Statistic	Std. Error
EDUCATIONAL LEVEL	Mean		13.49	.133
	95% Confidence Interval for Mean	Lower Bound	13.23	
		Upper Bound	13.75	
	5% Trimmed Mean		13.48	
	Median		12.00	
	Variance		8.322	
	Std. Deviation		2.885	
	Minimum		8	
	Maximum		21	
	Range		13	
	Interquartile Range		3	
	Skewness		-.114	.112
	Kurtosis		-.265	.224

Hypothesis Testing of Population Parameter

- Hypothesis testing is the second type of inferential statistics. Here, we try to find whether the population parameter is equal to a specified sample statistic, or whether the differences between sample group statistics are real ones.

- Chance can play a big role to differences between sampled groups, and so every time a difference is observed the question arises as to its statistical significance; That is, whether the difference is unlikely to have occurred purely by chance only or something else.

- Every hypothesis testing situation starts with the statement of a hypothesis. There are two types of statistical hypotheses for each situation:

  (1) the Null hypothesis, H_0

  (2) the Alternative hypothesis, H_1

- The Null hypothesis proposes no difference or relationship between the variables of interest.

- If a significant difference is found, the null hypothesis is rejected. If no significant difference is found, the null hypothesis is accepted.

- The Alternative hypothesis contradicts the null hypothesis.

- The Alternative hypothesis indicates the direction of the difference or relationship that you expect. It is also known as the Research hypothesis.

One-tailed versus Two-tailed Hypothesis

- The tails refer to the ends of the probability curve.
- We use a one-tailed test of significance when a directional hypothesis is stated, and a two-tailed test in all other situations.
- For example, if we would like to test the hypothesis about a single population mean (μ_0) then we could state the hypothesis as follow:

One-tailed :

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

or

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

Two-tailed :

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

- To make a decision about a hypothesis, you need to calculate a statistical test.

- This statistical test uses the data obtained from a sample to make a decision about whether the null hypothesis should be rejected.

- The numerical value obtained from a statistical test is called the test value, given as follow :

$$\text{Test Value} = \frac{O - E}{SE}$$

where O is the observed value, E is the expected value, and SE is the standard error.

- In any decision, you could be Correct or Wrong. That is, you could make a correct decision about the null hypothesis, or you could make a wrong one.

- Errors could occur in your decision (called Inference Errors). There are two types of Inference Errors:

DECISION	NULL HYPOTHESIS (H_0)	
	TRUE	FALSE
REJECT H_0	Type I Error (α)	Correct (1- β)
ACCEPT H_0	Correct (1 - α)	Type II Error (β)

Level of Significance

- The probability of a type I error is denoted by α and is referred to as the significance level of a test.

- Statisticians generally use arbitrary significance levels: the 0.10, 0.05, and 0.01 levels. That is, if the null hypothesis is rejected, the probability of a type I error will be 10%, 5% or 1%, depending on which level is used.

- When α =0.10, there is a 10% chance of rejecting a true null hypothesis. When α =0.05, there is a 5% chance of rejecting a true null hypothesis, and so on.

Probability value ($p-value$)

- The $p-value$ is the probability of getting a sample statistic (such as the mean) or a more extreme sample statistic in the direction of the alternative hypothesis when the null hypothesis is true.

- Ranging from 0 to 1, the $p-value$ is used to judge the extent of the evidence against H_0 .

- The $p-value$ is the maximum probability of committing a type I error. It represents the probability of observing a sample outcome more contradictory to H_0 than the observed sample result.

- The smaller the value of the $p-value$, the stronger the evidence for rejecting H_0 .

Hypothesis Testing of Population Mean μ

- To test hypotheses regarding a population mean, there are two commonly statistical tests:

(1)The z-test

- The z-test can be used when $n \geq 30$ or when the population is normally distributed and σ is known.

(2)The t-test

- The t-test can be used when σ is unknown (estimated by sample standard deviation, s) and $n < 30$

Hypothesis Testing for difference between two population means

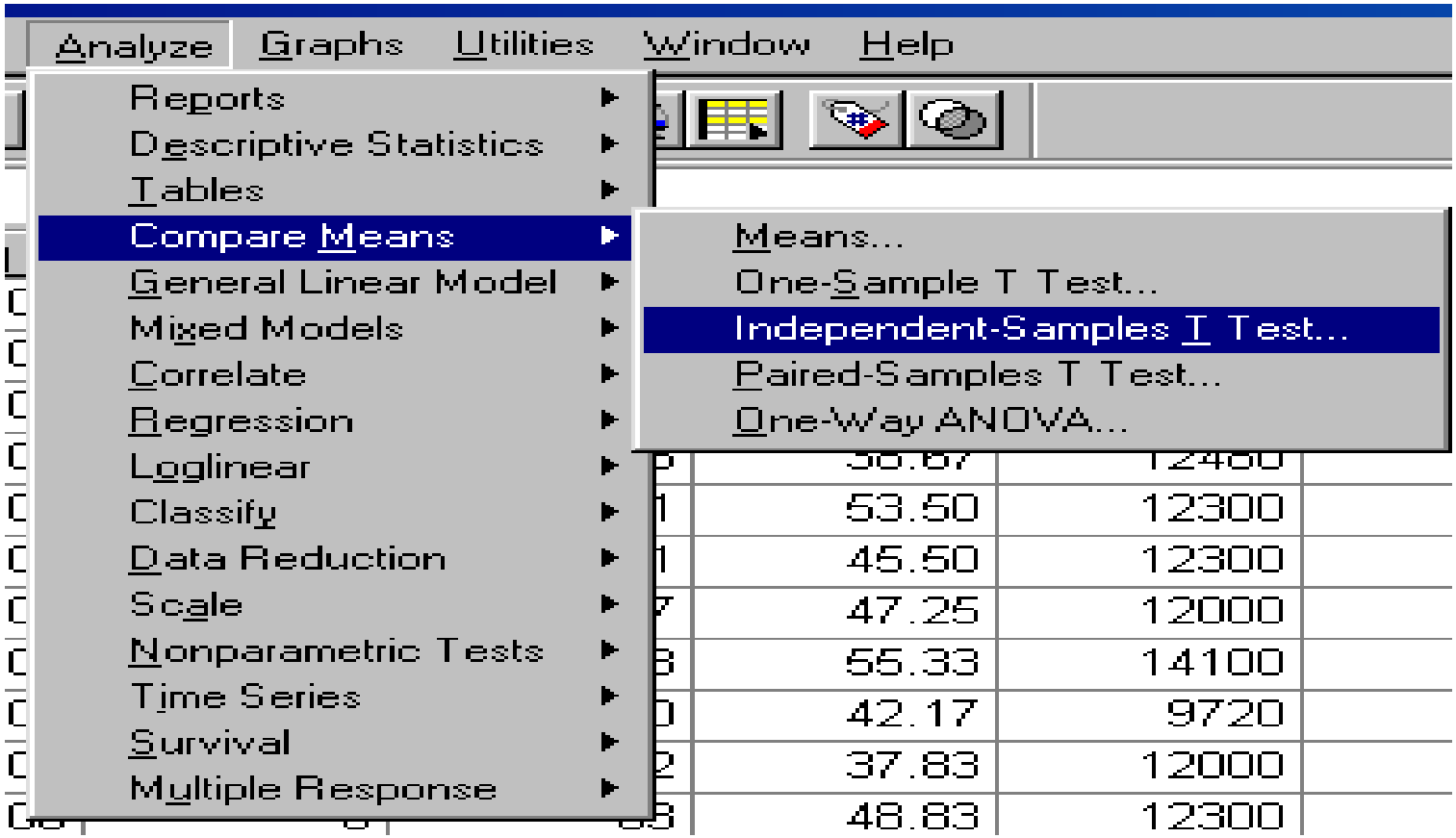
- There are many instances when researchers wish to compare more than one parameter (means) from different populations.

In this case

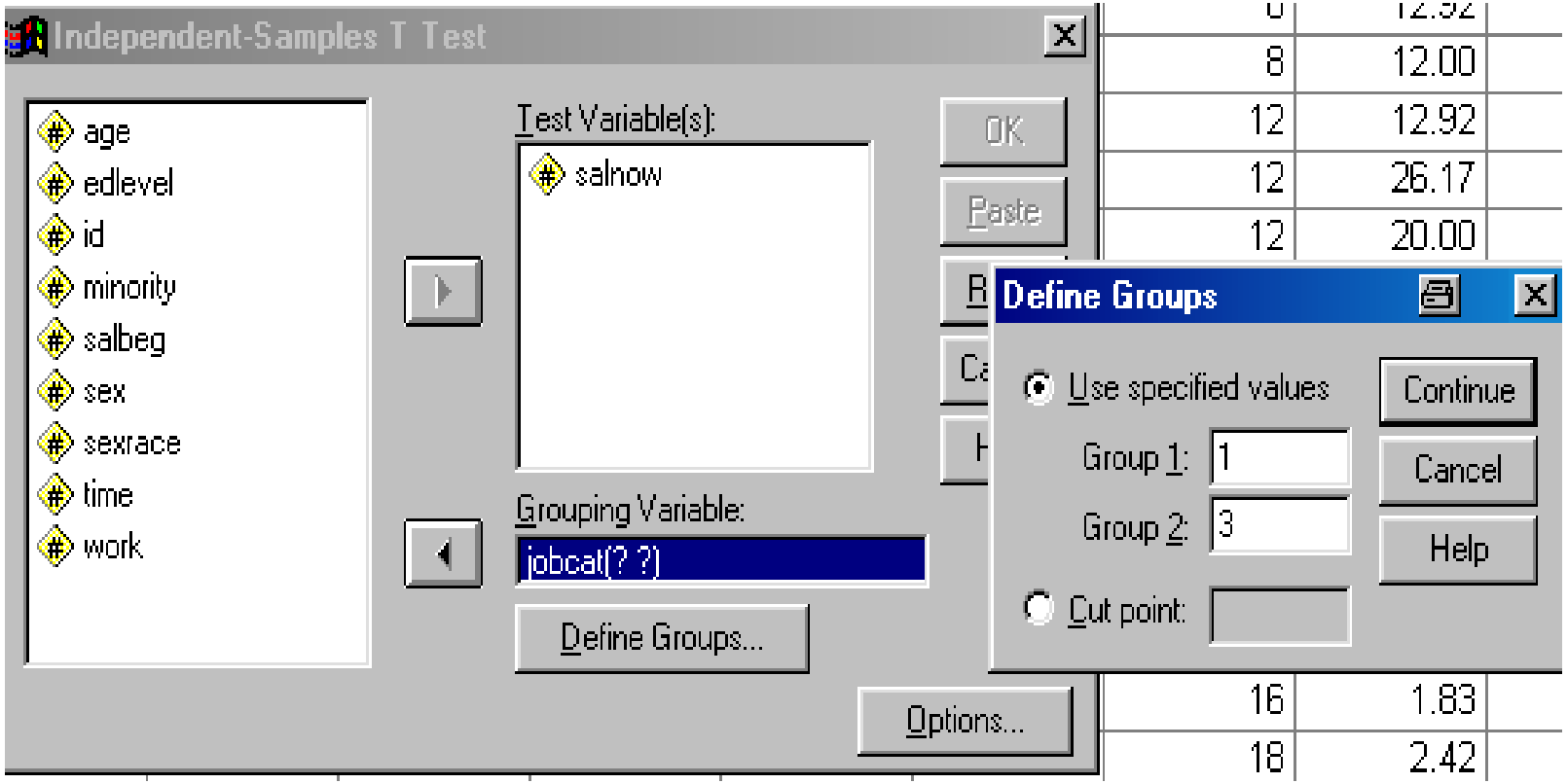
- The Z-Test will be used to test the difference between the two means when the population standard deviations are known and the variables are normally or approximately normally distributed, or both sample sizes are greater than or equal to 30.

- However, when the population standard deviations are not known, and one or both sample sizes are less than 30, the T-test is used to test the difference between the two means.

Conducting T-Test



Conducting T-Test



T-Test

Group Statistics

	EMPLOYMENT CATEGORY	N	Mean	Std. Deviation	Std. Error Mean
CURRENT SALARY	CLERICAL	227	11134.82	3196.569	212.164
	SECURITY OFFICER	27	12375.56	845.847	162.783

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Salnow	Equal variances assumed	21.530	.000	-2.005	252	.046	-1240.7	618.732	-2459.280	-22.192
	Equal variances not assumed			-4.640	142.17	.000	-1240.7	267.417	-1769.364	-712.109

Differences between more than two groups means

- We considered previously comparing two groups means using either the z-test or the t-test.

- Sometimes we need to compare more than two groups. In this case, we need to have multiple comparisons of means.

- To solve the problems of using many tests, we have a single test which considers the variation (differences) across all groups at once. This is called the Analysis of Variance (ANOVA).

- In ANOVA, we need a dependent variable and an independent variable.

- The independent variable is nominal (with categories).
- The dependent variable is continuous.

- To use ANOVA, three assumptions must be satisfied:
    1. The dependent variable must be normally distributed.
    2. The population variances must be equal for all groups.
    3. The observations must be independent.

- A one-way ANOVA means there is one independent variable.
- A two-way ANOVA means there are two independent variables.

- The variation in data (observations) can be divided into two parts:

    1. The Within-Group Variation
       This is the variation between observations within the same group.

    2. The Between-Group Variation
       This is the variation between each group.

- Total Variation (Variability) is the Sum of the Within-Group Variation and the Between-Group Variation

Multiple Comparisons

- The F-test was used to determine whether the

  between group-variation was large relative to the within-group variation.

- If the calculated $F > F_{k-1,n-k,1-\alpha}$ then the null hypothesis is rejected. This means that at least two population means are different.

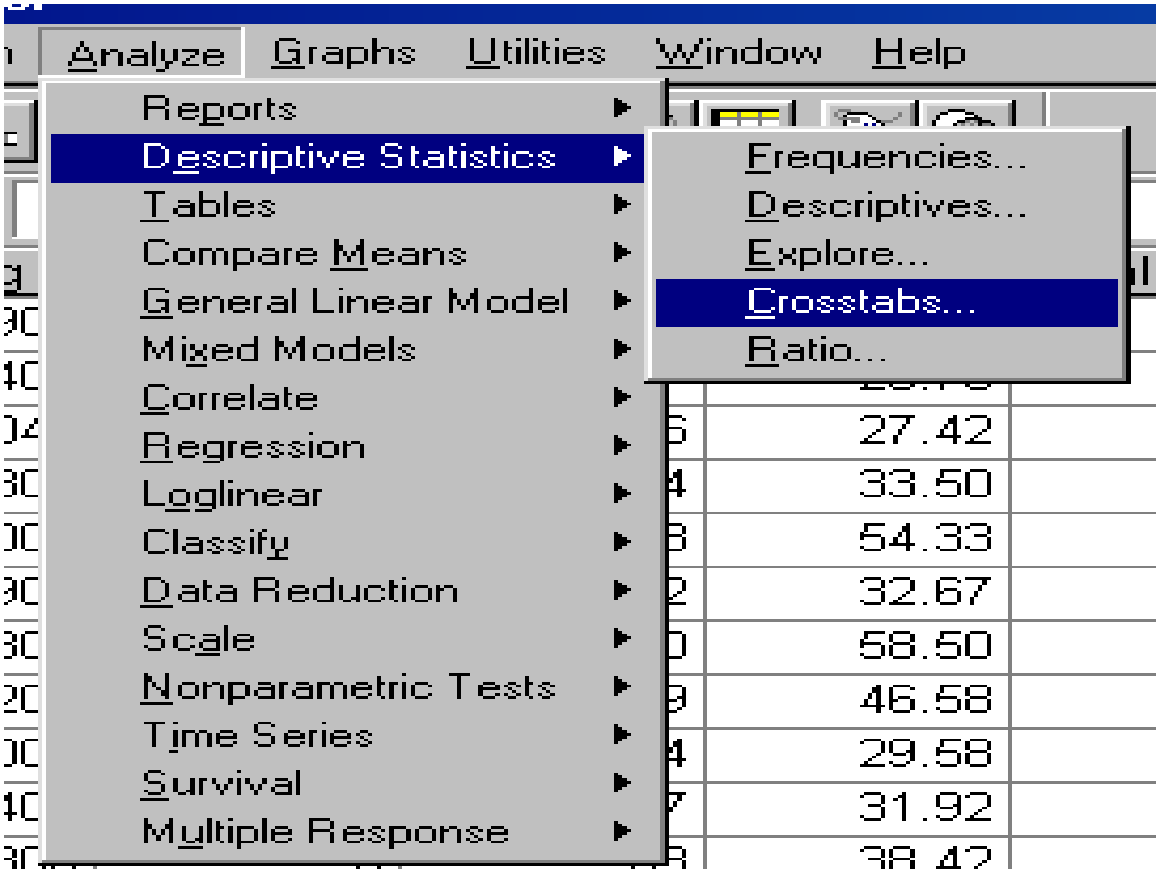
- Multiple comparisons

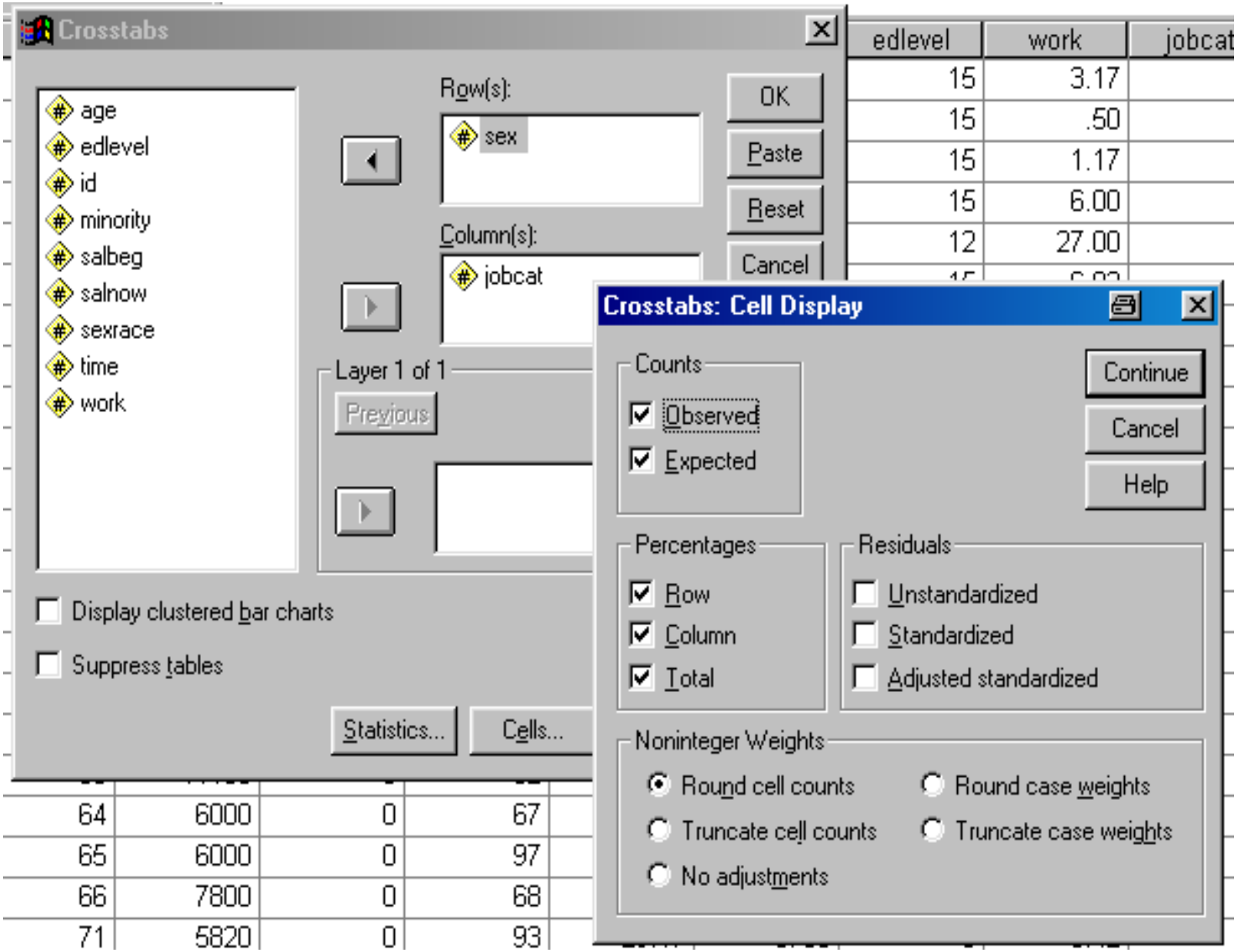
  Tukey's test, Scheffe's test, Bonferroni test and the Least Significant Difference (LSD) test.

Correlation and Regression

- Another area of inferential statistics involves determining whether a relationship between two or more variables exists.
- If both variables are categorical, we use the chi-square.
- If both variables are continuous, we use the correlation coefficient.
- For predicting the value of one variable based on the value of another, we use regression analysis.
- Regression is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear.

Steps of Conducting a Chi-Square Test

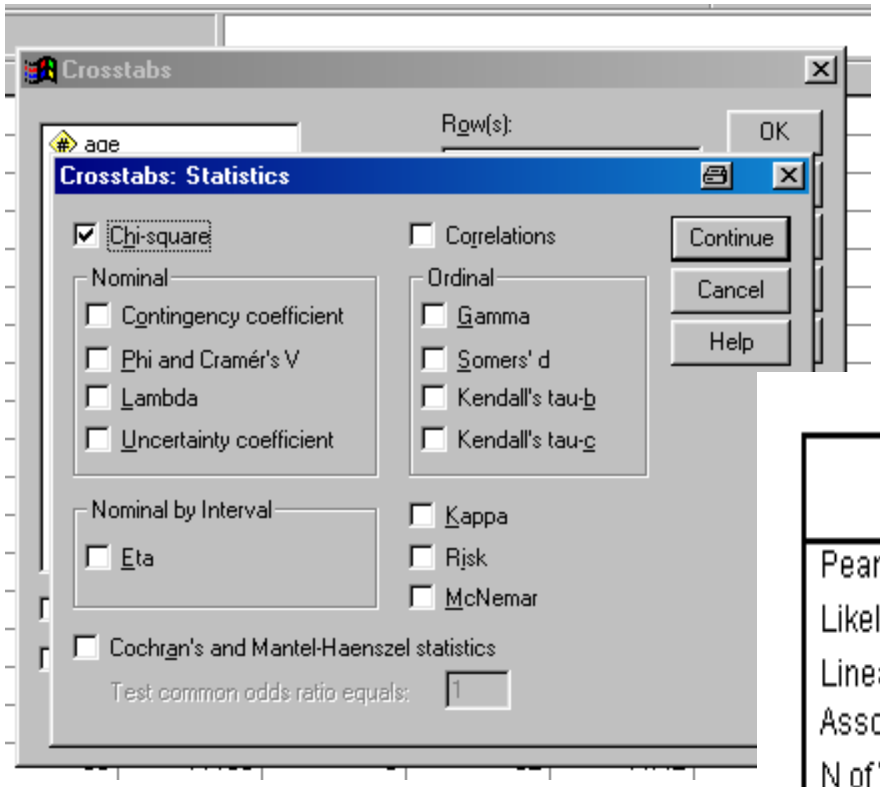



Crosstabs
age
edlevel
id
minority
salbeg
salnow
sexrace
time
work
Row(s):
sex
Column(s):
jobcat
OK
Paste
Reset
Cancel
Layer 1 of 1
Previous
Display clustered bar charts
Suppress tables
Statistics...
Cells...
Crosstabs: Cell Display
Counts
Observed
Expected
Continue
Cancel
Help
Percentages
Row
Column
Total
Residuals
Unstandardized
Standardized
Adjusted standardized
Noninteger Weights
Round cell counts
Round case weights
Truncate cell counts
Truncate case weights
No adjustments
edlevel work jobcat
15 3.17
15 .50
15 1.17
15 6.00
12 27.00
64 6000 0 67
65 6000 0 97
66 7800 0 68
71 5820 0 93

SEX OF EMPLOYEE * EMPLOYMENT CATEGORY Crosstabulation

			EMPLOYMENT CATEGORY					
			CLERICAL	OFFICE TRAINEE	SECURITY OFFICER	EXEMPT EMPLOYEE	TECHNICAL	Total
SEX	MALES	Count	110	85	27	30	6	258
		Expected Count	123.6	99.1	14.7	17.4	3.3	258.0
		% within SEX OF EMPLOYEE	42.6%	32.9%	10.5%	11.6%	2.3%	100.0%
		% within EMPLOYMENT CATEGORY	48.5%	46.7%	100.0%	93.8%	100.0%	54.4%
		% of Total	23.2%	17.9%	5.7%	6.3%	1.3%	54.4%
	FEMALES	Count	117	97	0	2	0	216
		Expected Count	103.4	82.9	12.3	14.6	2.7	216.0
		% within SEX OF EMPLOYEE	54.2%	44.9%	.0%	.9%	.0%	100.0%
		% within EMPLOYMENT CATEGORY	51.5%	53.3%	.0%	6.3%	.0%	45.6%
		% of Total	24.7%	20.5%	.0%	.4%	.0%	45.6%
Total		Count	227	182	27	32	6	474
		Expected Count	227.0	182.0	27.0	32.0	6.0	474.0
		% within SEX OF EMPLOYEE	47.9%	38.4%	5.7%	6.8%	1.3%	100.0%
		% within EMPLOYMENT CATEGORY	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	47.9%	38.4%	5.7%	6.8%	1.3%	100.0%

Crosstabs dialog box “ Statistics”



Output View

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	87.230[a]	6	.000
Likelihood Ratio	106.109	6	.000
Linear-by-Linear Association	48.161	1	.000
N of Valid Cases	474		

a. 4 cells (28.6%) have expected count less than 5. The minimum expected count is 2.28.

Prediction: Linear Regression

- Once we find that the two variables have a statistically significant relationship (significant r), we wonder if we can predict y by knowing x .

- The method used to find such prediction is regression.

- In simple linear regression, the regression equation contains one independent variable x and one dependent variable y and is written as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{\beta}_0$ is called the y-intercept of the line (ie. it is the predicted value for y when x is zero), and $\hat{\beta}_1$ is called the slope (ie. it is the change in y when x changes by one unit). We use the symbol $\hat{}$ to show that it is a “fitted or predicted" estimate.

- In multiple regression, there are several independent variables and one dependent variable, written as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$$

- The question is whether a given independent variable is linearly associated with the outcome (dependent variable) after controlling (adjusting) for a number of other factors.

Using SPSS, the output is:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-44.810	69.247		-.647	.584
	x1	87.640	15.237	.639	5.752	.029
	x2	14.533	2.914	.554	4.988	.038

a. Dependent Variable: y

Logistic Regression

This is used when the outcome (dependent) is categorical (binary)

Advantages:

Logistic function is a good model for probability

Bound by 0 and 1

Shape of the function has biologic interpretability

Other advantages:

Widely used in a variety of fields

Odds ratios easy to obtain

Lots of software available

Logistic Regression Model

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X_1 + \ldots + b_k X_k$$

Linear relationship between the log odds and the predictors, rather than the probability and the predictors.

Other Measures

Proportions

- Many outcomes can be classified as belonging to one of two possible categories: Presence or Absence, Improved or Not-improved, With disease or Without disease, Success or Failure, …and so on.

- We can re-label the two outcome categories as Success or Failure. An outcome is a success if the primary category is observed and failure if the other category is observed.

- The number of successes, x divided by the total outcomes, n (successes and failures) is what we call proportion.

Proportion =

$$\frac{x}{n}$$

Ratios

- A ratio is written as :

$$\text{Ratio} = \frac{a}{b}$$

Where a and b are similar quantities measured from different groups or under different circumstances.

EXAMPLE

In a hospital, there were 200 nurses and 80 physicians. What is the nurses-physician ratio ?

SOLUTION
Nurses-physician ratio = 2.5

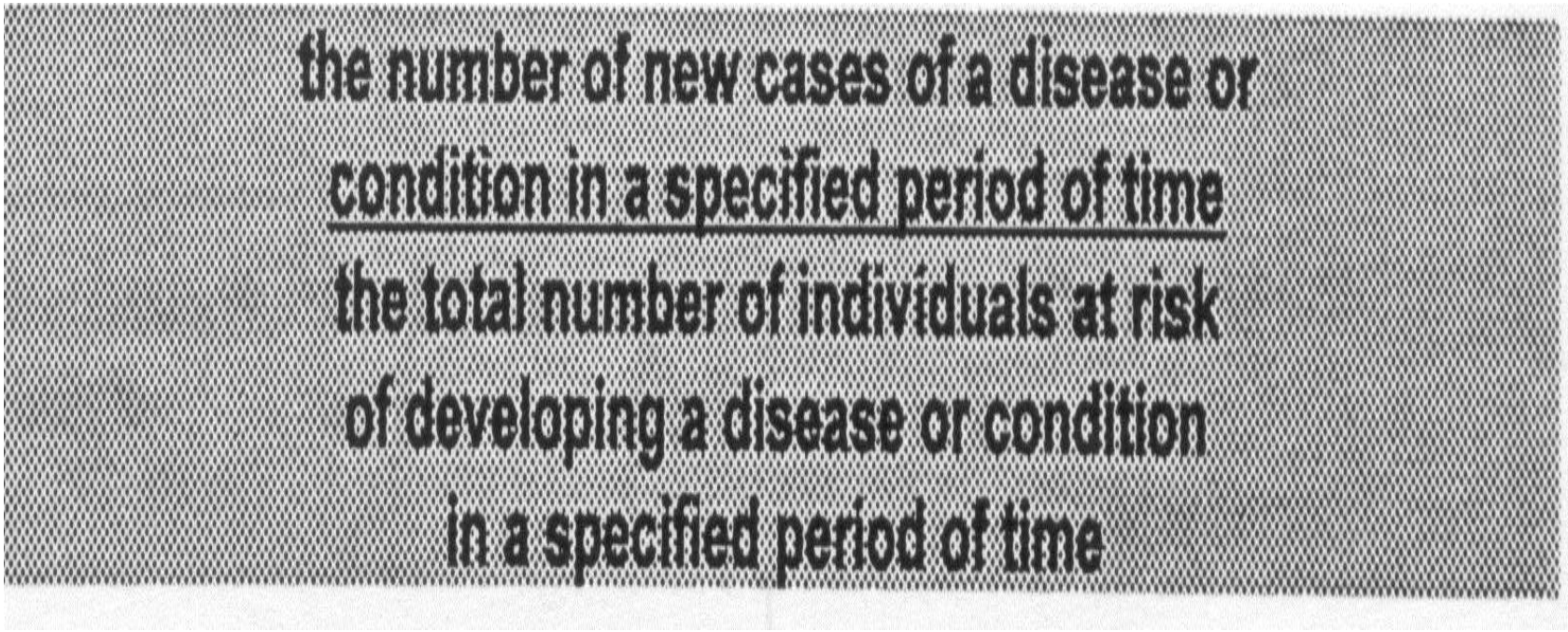
Rates

Rate is a ratio in which there is a distinct relationship between the numerator and the denominator and a measure of time is an intrinsic part of the denominator.

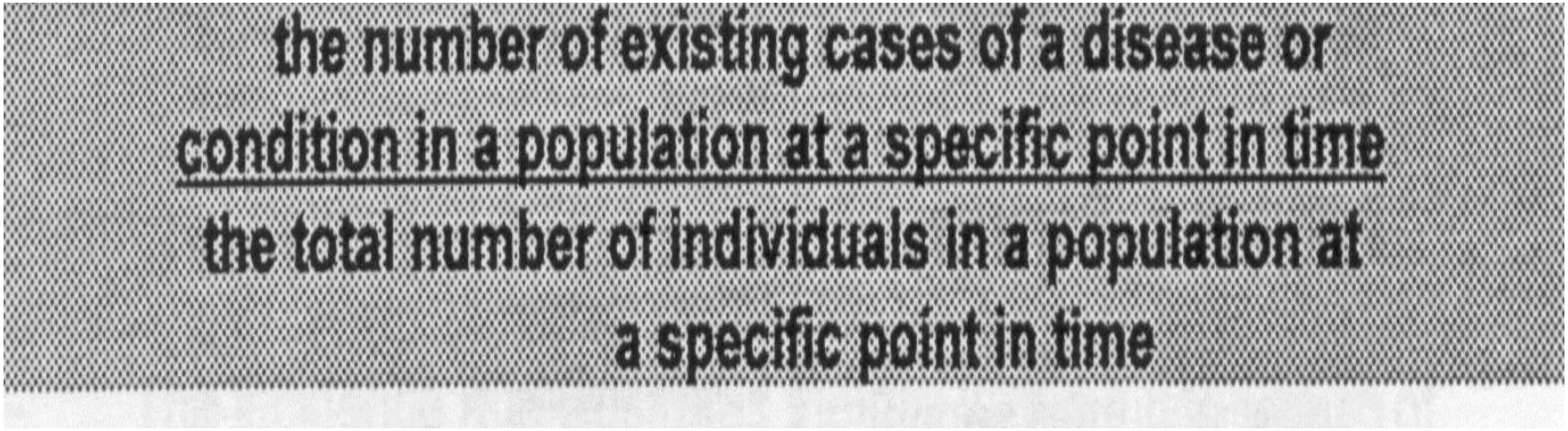
EXAMPLE

Incidence Rate , Prevalence Rate, Mortality Rates, Cause-Specific mortality Rates, Case-Fatality Rates

Incidence Rate



Prevalence Rate



Odds Ratio and Relative Risk

Other Measures

Proportions

- Many outcomes can be classified as belonging to one of two possible categories: Presence or Absence, Improved or Not-improved, With disease or Without disease, Success or Failure, …and so on.

- We can re-label the two outcome categories as Success or Failure. An outcome is a success if the primary category is observed and failure if the other category is observed.

- The number of successes, x divided by the total outcomes, n (successes and failures) is what we call proportion.

Proportion =

$$\frac{x}{n}$$

For a case-control study , the contingency table is:

exposure	cases	controls	Total
yes	a	b	a+b
no	c	d	c+d
Total	a+c	b+d	a+b+c+d

Odds of exposure among cases $= \frac{a}{c}$

Odds of exposure among controls $= \frac{b}{d}$

Odds Ratio $= \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$

In Cohort study, the contingency table is shown as:

	Disease		
Group	Present	Absent	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

Incidence in the exposed group $= \frac{a}{a+b} = I_{ex}$

Incidence in the unexposed group $= \frac{c}{c+d} = I_{un}$

Relative Risk $= \frac{I_{ex}}{I_{un}}$