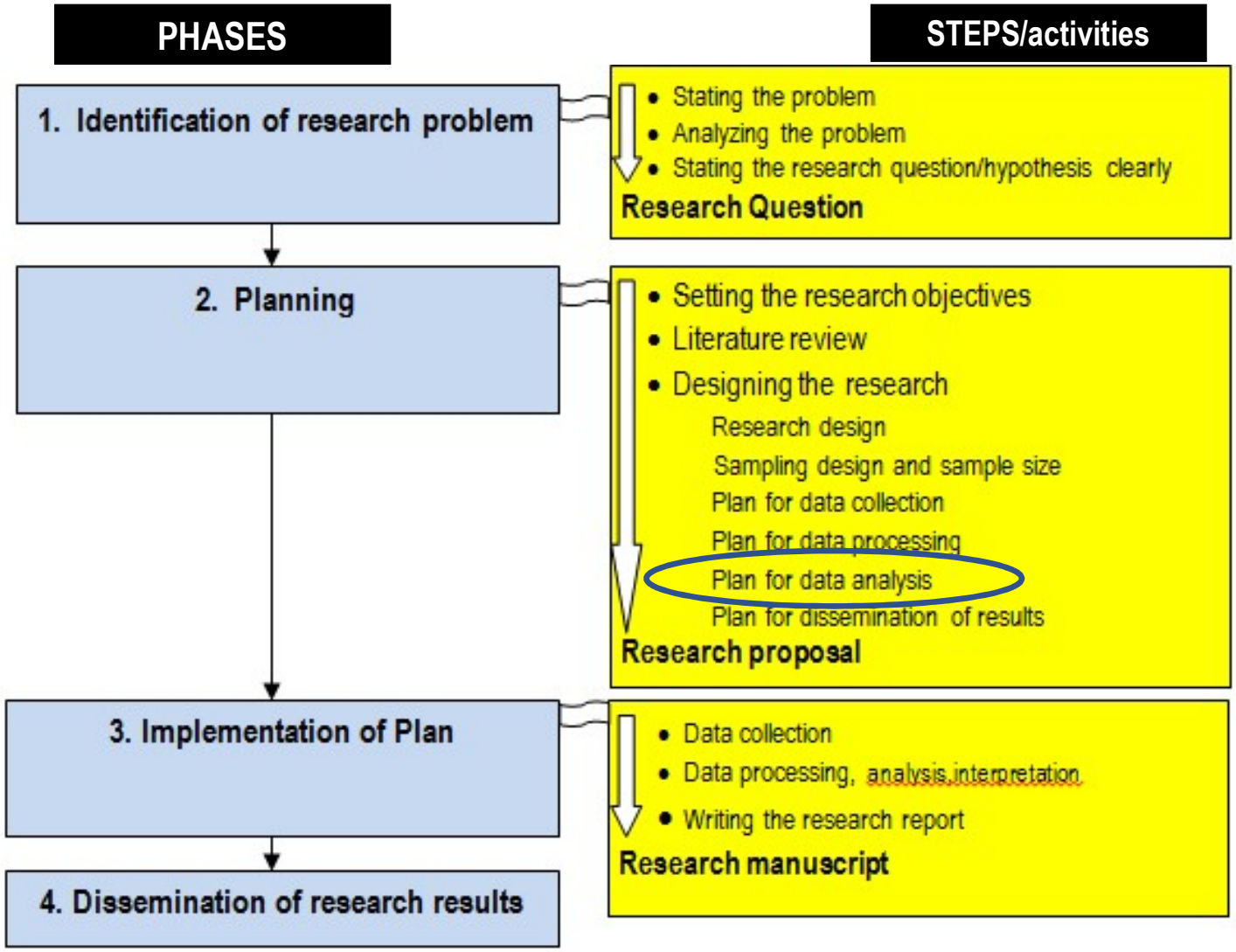
Plan for Data Analysis

WRITESHOP: Capacity Building in Technical Writing for Mental Health Research Proposal Development
05 December 2020

Derick Erl P. Sumalapao, MD, MSc, DrPH
Department of Epidemiology and Biostatistics
College of Public Health
University of the Philippines Manila



Adapted from Learning and Development Intervention for DOH Regional Health Policy and Systems Research Fellowship Program CPHFI- University of the Philippines Manila

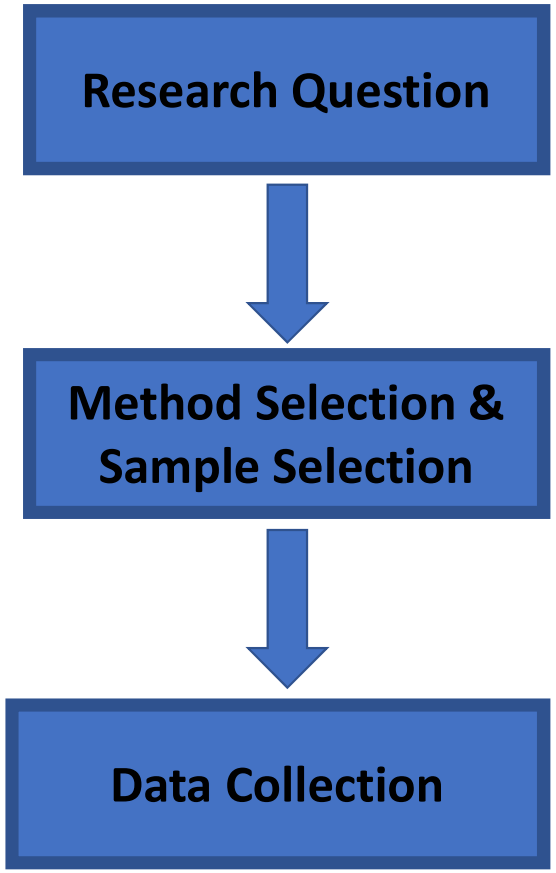
Session Objectives

DOMAIN	CONTENT
KNOWLEDGE	❑ identify appropriate numerical and graphical summaries for a specific variable type ❑ identify appropriate statistical tool for data analysis ❑ interpret confidence interval for a variety of parameters ❑ interpret *p*-values in making statistical inferences
SKILLS	❑ calculate and interpret descriptive measures ❑ apply the basic principles of estimation and hypothesis testing using a software
ATTITUDE	❑ understand how to reduce data sets into a useful descriptive measure

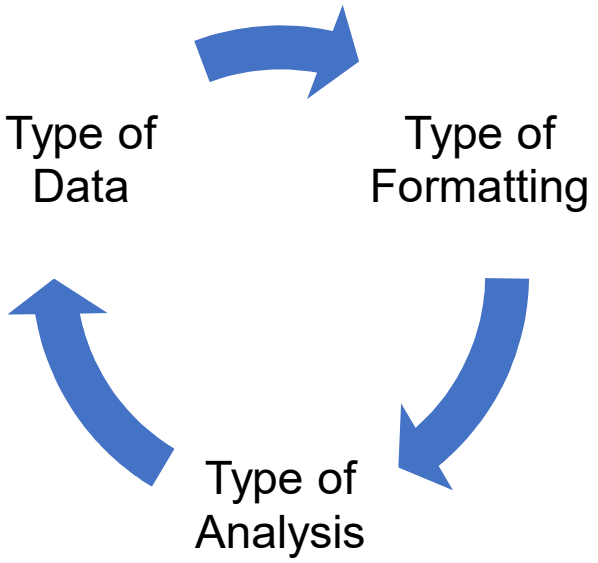
Outline

- Introduction
- Statistical Tables
- Graphical Presentation of Statistical Data
- Summary Measures and Descriptive Statistics
- Inferential Statistics: Estimation and Hypothesis Testing
  - One Sample with Continuous Outcome
  - One Sample with Dichotomous Outcome
  - Two Independent Samples with Continuous Outcome
  - Two Independent Samples with Dichotomous Outcome
- Chi-square test
- Analysis of Variance
- Linear Regression
- Summary

Research Steps



Plan for Data Analysis



Statistical Table

- a common way of organizing and presenting statistical data
- the categories of the variables and their corresponding counts, percentages, and other statistical measures are presented in rows and columns

Table 1. Distribution of residents according to sex

Sex	Frequency	Relative Frequency, %
Male	1,625	45.9
Female	1,914	54.1
Total	**3,539**	**100.0**

- One of the variables recorded was sex (nominal, qualitative)
- Frequency distribution tables are a common and useful way of summarizing discrete variables
- Relative frequencies
  - Divide the frequency in each response category by the sample size (e.g., 1,625/3,539 = 0.459).
- With dichotomous variables, the relative frequencies are often expressed as percentages (by multiplying by 100).

Frequency Distribution

- a special kind of statistical table
- data on the number or the percentage of observations falling in different categories of the variable are presented

Table 3. Distribution of participants according to marital status

Marital Status	Frequency	Relative Frequency, %
Separated	46	1.3
Single	203	5.8
Widowed	334	9.5
Divorced	367	10.4
Married	2,580	73.1
Total	3,530	100.0

- With categorical variables
  - *e.g. marital status, handedness, smoking status*
- Responses are unordered
  - the order of the responses or categories in the summary table can be changed
  - present the categories alphabetically or perhaps from the most frequent to the least frequent.

Frequency distribution table

- For discrete variables which are inherently ordinal
- Ordered categories (*e.g.*, excellent, very good, good, fair, poor)
- Continuously distributed measures, but then categorized for decision making
  - *e.g.* blood pressure, body mass index

Table 4. Frequency Distribution for Blood Pressure Category

Blood Pressure	Frequency	Relative Frequency (%)	Cumulative Frequency	Cumulative Relative Frequency, %
Normal	1,206	34.1	1,206	34.1
Pre-Hypertension	1,452	41.1	2,658	75.2
Stage I Hypertension	653	18.5	3,311	93.7
Stage II Hypertension	222	6.3	3,533	100.0
Total	**3,533**	**100.0**		

	Blood Pressure
Normal	systolic blood pressure <120 and diastolic blood pressure <80
Pre-hypertension	systolic blood pressure between 120-139 or diastolic blood pressure between 80-89
Stage I hypertension	systolic blood pressure between 140-159 or diastolic blood pressure between 90-99
Stage II hypertension	systolic blood pressure of 160 or more or diastolic blood pressure of 100 or more

Cross-tabulation

- a frequency distribution showing the simultaneous categorization of two or more variables
- to determine whether the variables being considered are related to each other

EXAMPLE.

A study identified the nutritional status and type of feeding of infants in Municipality X. The table below summarizes the distribution of infants according to their nutritional status and type of feeding.

Table 5. Distribution of infants according to nutritional status and type of feeding

Nutritional Status	Type of Feeding		Total
	Bottle/Mixed	Breastfed	
Normal	69	528	597
Malnourished	105	132	237
Total	174	660	834

Computing for Percentages

- Computed percentages are important when making comparisons, especially when the totals of the groups being compared are very different from each other

Question:
Which type of mothers are more likely to be breastfeeding their baby, working or non-working?

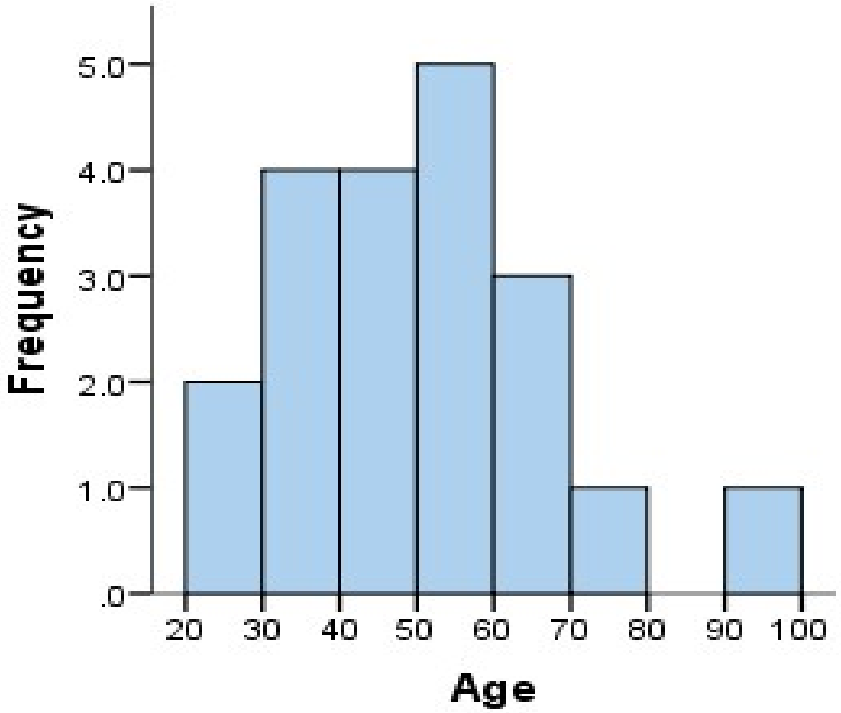
Occupational Status of Mother	Total Number Interviewed	Number breastfeeding their baby
Working	200	80
Non-working	50	45
Total	250	125

Occupational Status of Mother	Total Number Interviewed	Number breastfeeding their baby	% Breastfeeding their baby
Working	200	80	40.0
Non-working	50	45	90.0
Total	250	125	50.0

Types of Graphs Commonly Used

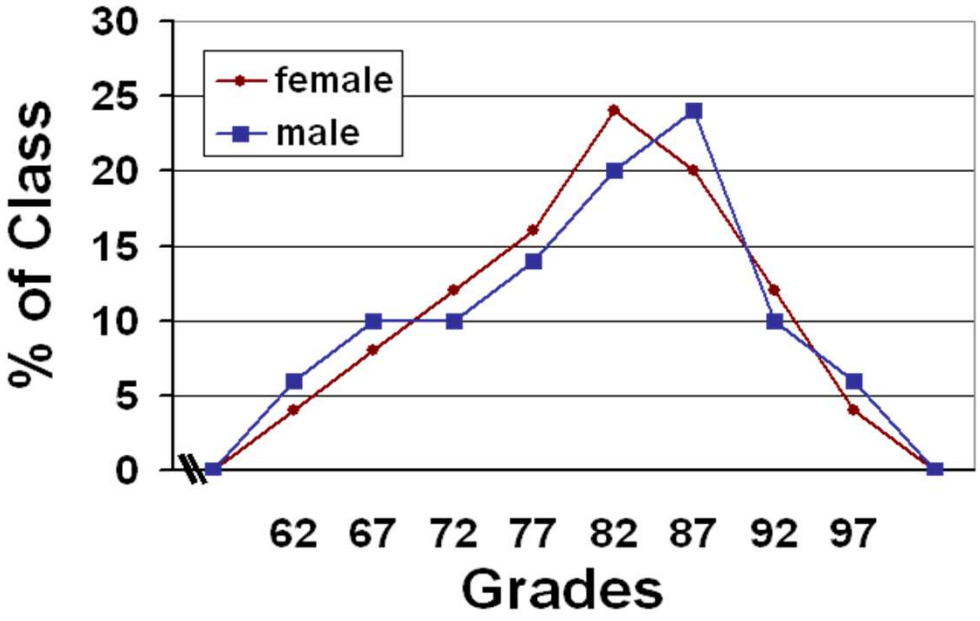
Type of graph	Type of variable or data being graphed	Purpose of presenting the graph
Histogram	Continuous quantitative	To present a frequency distribution of a quantitative continuous variable like age, height, etc.
Frequency polygon	Continuous quantitative	The same use as the histogram, but is better to use when presenting more than one frequency distribution in the same graph (ex. Comparison of the weights of male and female children).
Bar chart (horizontal or vertical)	Qualitative, or discrete quantitative	To show or compare absolute counts or relative figures (percentages, rates, etc.) of qualitative or discrete quantitative variables
Line diagram	Absolute counts as well as relative or summary figures of both quantitative and qualitative variables for which the analysis of trends is relevant	To show trends in absolute counts, rates, or means with respect to time, age, etc.
Pie chart	Qualitative, or broad categories of quantitative variables	To show how a total is divided into sub-categories (when the number of categories is not too many)
Component bar	Qualitative, or broad categories of quantitative variables	Same as the pie chart, but is better to use when presenting or comparing two or more sets of data
Scatterplot	Quantitative (discrete or continuous)	To show the nature and the strength of the relationship between two continuous quantitative variables

Histogram



- Continuous quantitative
- To present a frequency distribution of a quantitative continuous variable like age, height, etc.

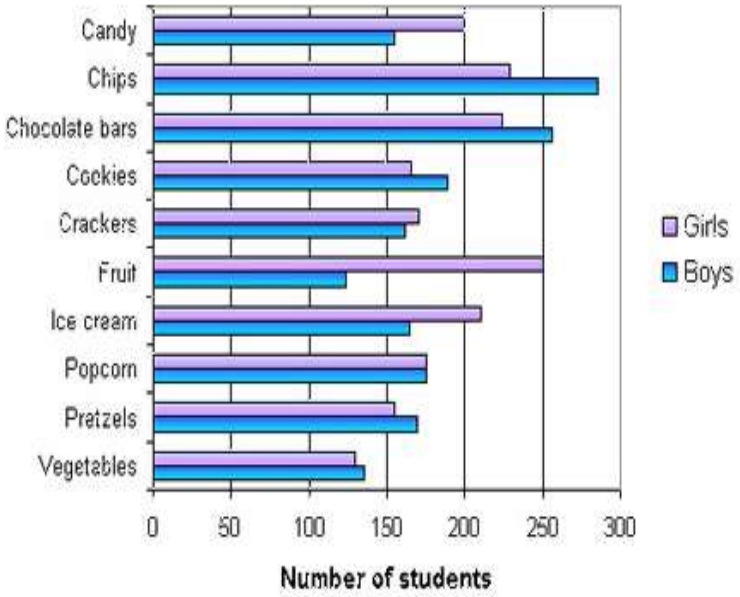
Frequency polygon

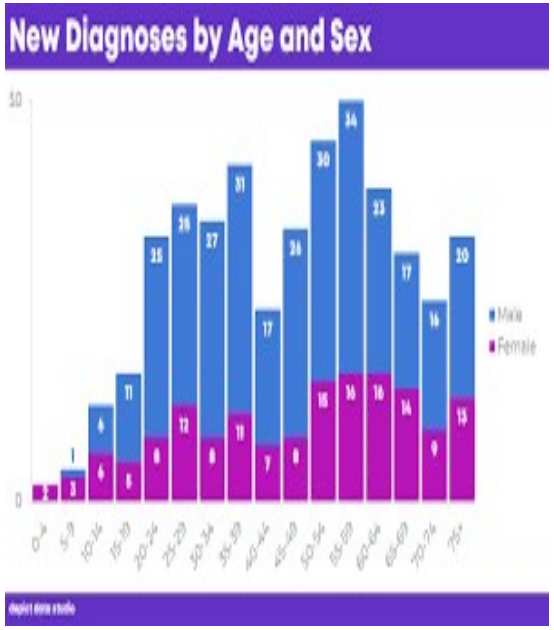


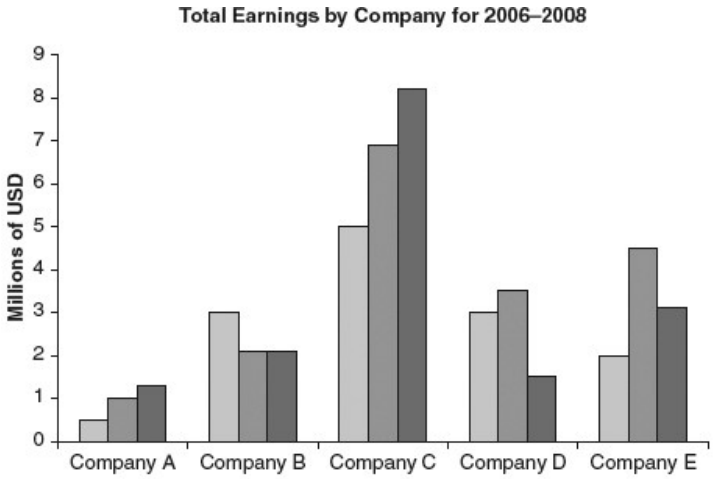
- Continuous quantitative
- The same use as the histogram
- Better when presenting more than one frequency distribution in the same graph (ex. comparison of the weights of male and female children)

Bar chart (horizontal or vertical)

- Qualitative, or discrete quantitative
- To show or compare absolute counts or relative figures (percentages, rates, etc.) of qualitative or discrete quantitative variables







	2006	*2007*	*2008*
Company A	0.5	1	1.3
Company B	3	2.1	2.1
Company C	5	6.9	8.2
Company D	3	3.5	1.5
Company E	2	4.5	3.1

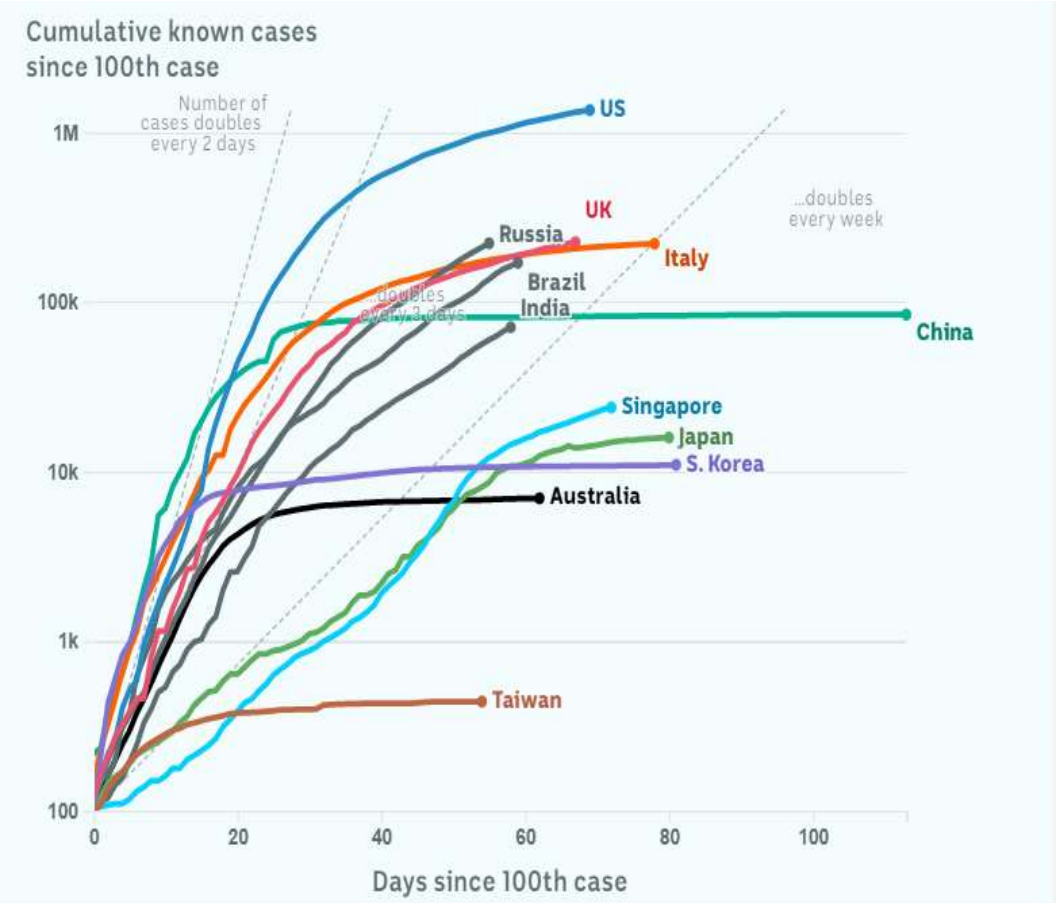
Figure 3 Clustered Bar Chart and Associated Data

Note: USD = U.S. dollars.

Line diagram

- Absolute counts as well as relative or summary figures of both quantitative and qualitative variables for which the analysis of trends is relevant

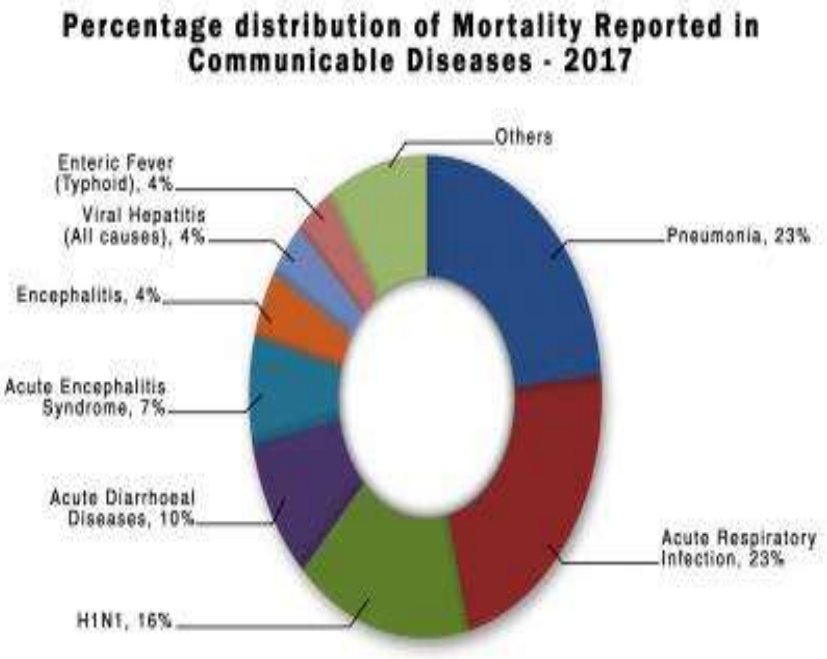
	Year 1990	Year 1999	Year 2003	Year 2008
Early	31.4	40.6	32.4	39.2
Delayed	68.6	59.4	67.6	60.8

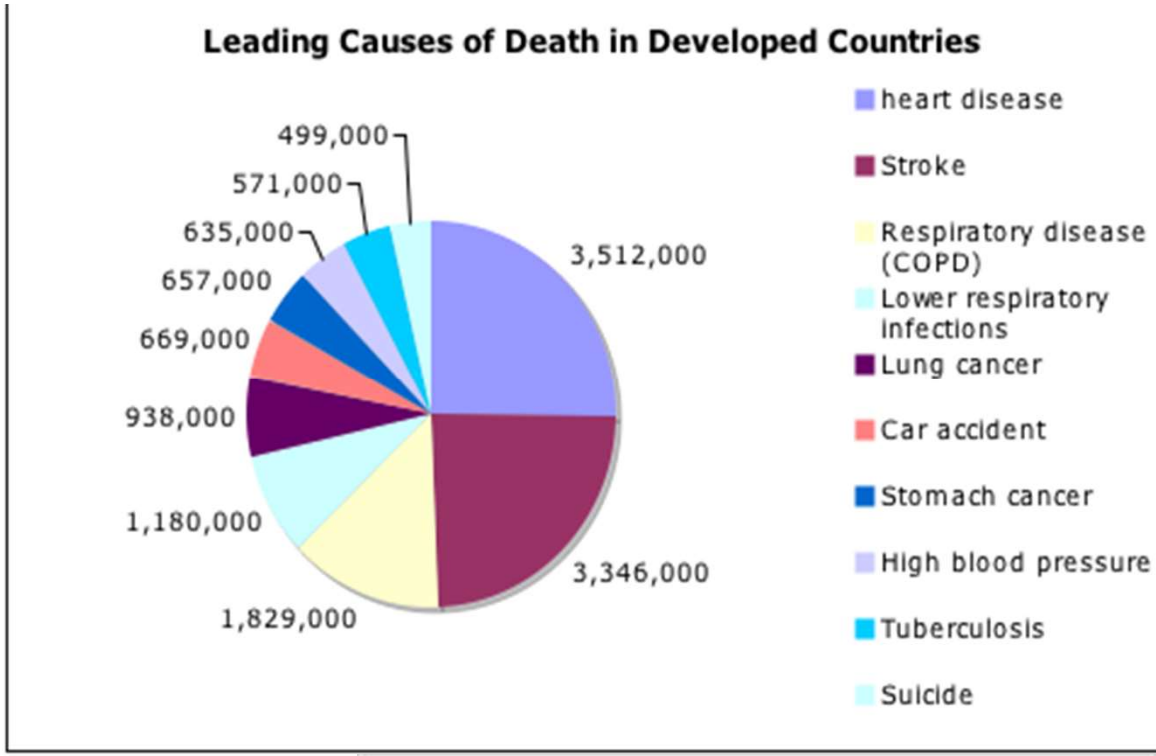


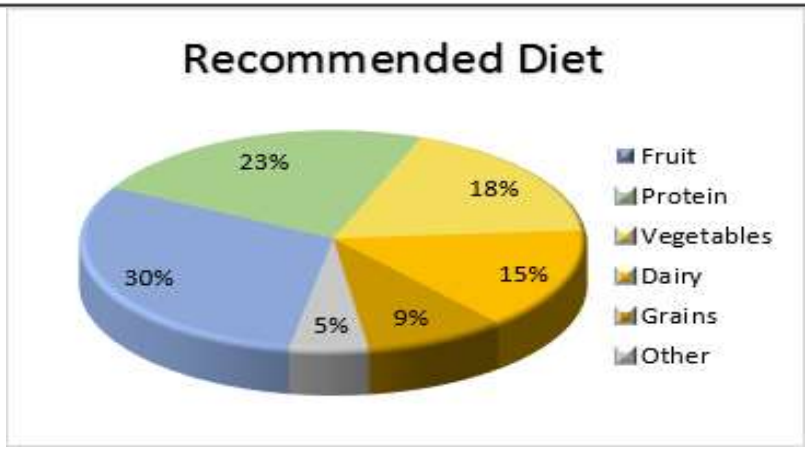
- To show trends in absolute counts, rates, or means with respect to time, age, etc.

Pie chart

- Qualitative, or broad categories of quantitative variables
- To show how a total is divided into sub-categories (when the number of categories is not too many)



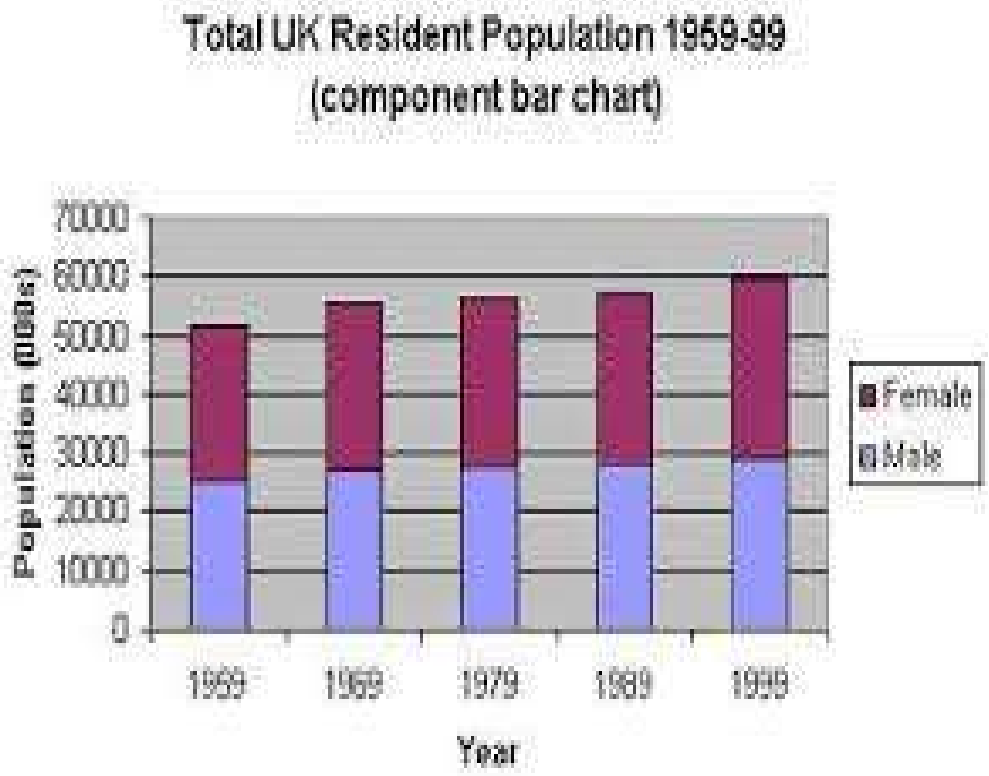




Component Bar

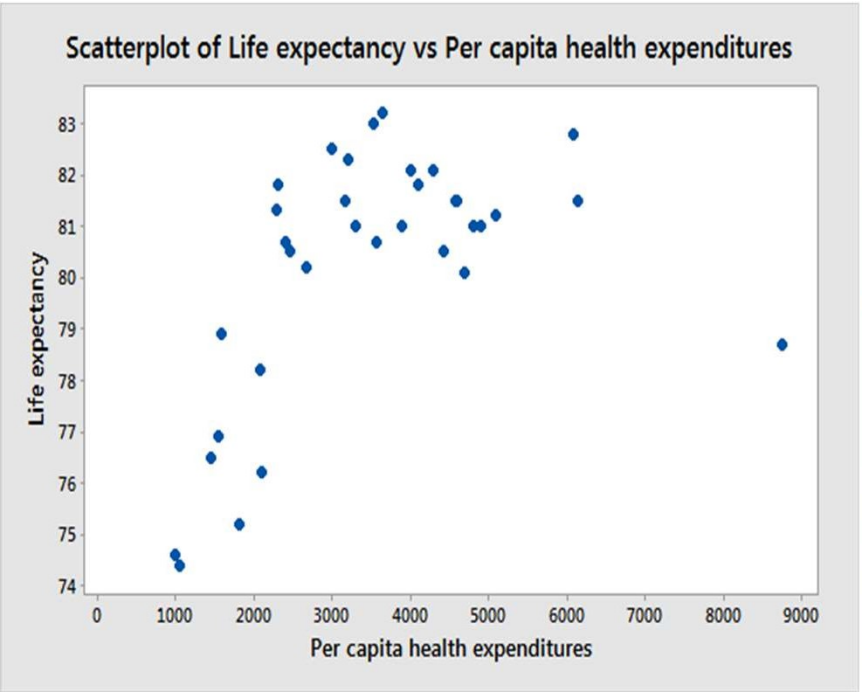
- Qualitative, or broad categories of quantitative variables

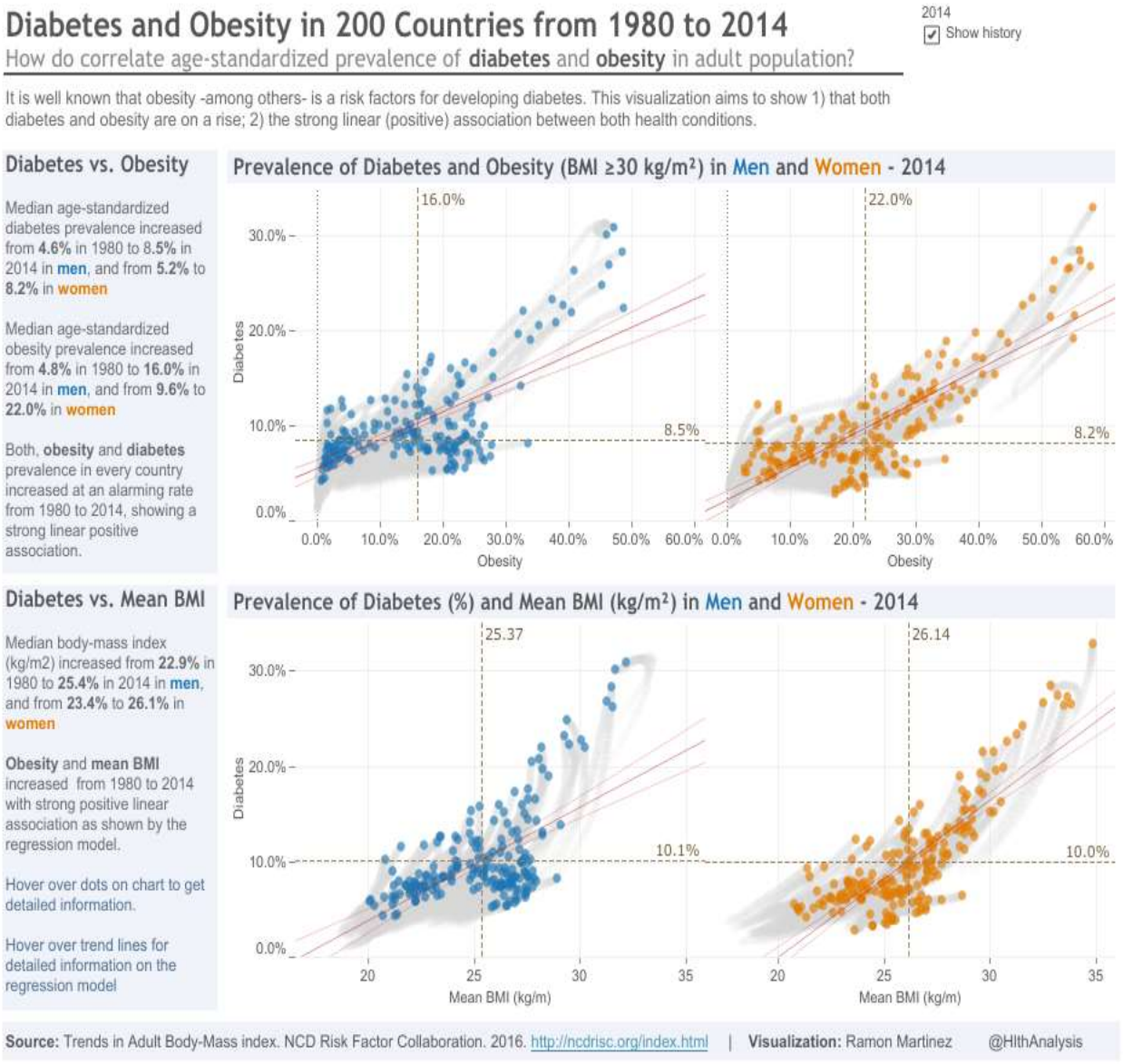
- Same as the pie chart, but is better to use when presenting or comparing two or more sets of data



Scatterplot

- ❑ Quantitative (discrete or continuous)
- ❑ To show the nature and the strength of the relationship between two continuous quantitative variables





Descriptive Statistics for Continuous Variables

- ❑ Sample Mean
- ❑ Median
- ❑ Mode
- ❑ Range
- ❑ Variance and Standard Deviation
- ❑ Inter-quartile Range

Measures of Central Tendency

MEAN

- ❑ statistical measure derived by dividing the sum of all observations by the total number of observations included in the computations
- ❑ is easily affected by outliers with extremely high or low values – not recommended in the presence of extreme values

MEDIAN

- ❑ middlemost value in a set of observations
- ❑ not affected by outliers
- ❑ choice for the measure of central tendency when the distribution is skewed

MODE

- ❑ most frequently occurring observation in the data set
- ❑ possible for an absence of mode or presence of several modes

Measures of Dispersion

Range

- A relatively crude, yet important, measure of variability
- Difference between the highest and the lowest value in the data set
  - **Sample Range = Maximum – Minimum**
- Easily affected by outliers with extremely low or high values

Coefficient of Variation (CV)

- Computed by expressing the standard deviation as a percentage of the mean
- Comparing the variability of two variables with different units of measurement

Variance and Standard Deviation

- Average of the squared deviations from the mean of a given data set
- Units of the computed value of the variance are in squared units
- Standard deviation is the square root of the variance

Variable	Mean	Standard deviation	Coefficient of Variation
Weight (kg)	50.0	3.9	(3.9/50.0) x 100 = 7.8%
Height (cm)	160.0	7.5	(7.5/160.0) x 100 = 4.7%

If I will be asked to describe the data by using just one value, what will that value be?

- It is not recommended to describe the data using the mean if the data set has **extreme values**.
- It is more appropriate to use the **median** when the distribution is **skewed** or there are **outliers** which are extremely low or high values.
- If it is to identify the value which occurs most frequently in the data, use the **mode**.

If I will be asked to describe how different are the data from one another by using just one value, what will that value be?

- In the absence of outliers, **standard deviation** or **variance** is the most appropriate measure of dispersion
  - Usually employed in more advanced statistical analysis like hypothesis testing
- **Interquartile range** replaces standard deviation or variance in the event the data set contains **outliers**

Descriptive vs Inferential Statistics

- **Descriptive Statistics**
  - Methods for organizing and summarizing information

- **Inferential Statistics**
  - Methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population
  - Concerns with the techniques applied when making conclusions about a target population

Examples of Research Objectives Involving the Application of Inferential Statistics

- To compare the psychological distress and anxiety levels of COVID-19 patients, medical staff, and medical students

- To determine if there is an association between smoking and substance abuse among college students

- To determine if the proportion of suicide cases differs between geographic residence (urban vs rural) and between sex group (male vs female)

Inferential Statistics:
Estimation and Hypothesis Testing

- ❑ One Sample with Continuous Outcome
- ❑ One Sample with Dichotomous Outcome
- ❑ Two Independent Samples with Continuous Outcome
- ❑ Two Independent Samples with Dichotomous Outcome

One Sample with Continuous Outcome

- ❑ According to the latest WHO* data published in 2018 life expectancy in Philippines is: Male 66.2, female 72.6, and total life expectancy is 69.3.
- ❑ Suppose in a 2020 study involving life expectancy of individuals in the Philippines, a **sample of 4,532** individuals had an **average life span of 70.5 years** with a **standard deviation of 12.1 years**.
- ❑ Does this data provide evidence that the life expectancy of Filipinos has increased?
- ❑ Obtain the 95% confidence interval for the mean life span.

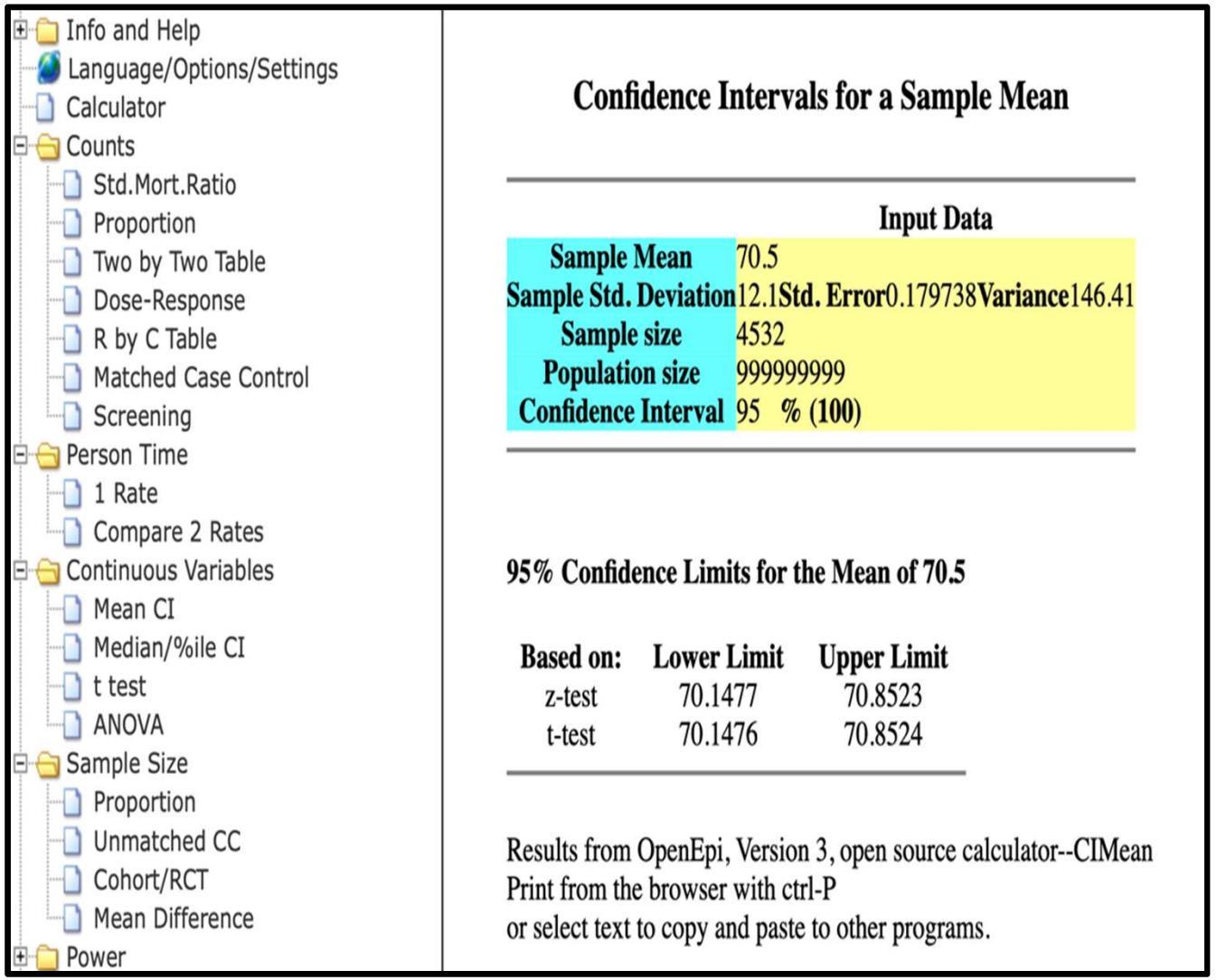
*https://www.worldlifeexpectancy.com/philippines-life-expectancy

Openepi*

- ❑ a free online statistical software
- ❑ offers an array of inferential statistics tools

*www.openepi.com

- Given:
  - Total life expectancy is 69.3.
  - sample of 4,532
  - average life span of 70.5 years
  - standard deviation of 12.1 years.
- Does this data provide evidence that the life expectancy of Filipinos has increased?
- Obtain the 95% confidence interval for the mean life span.



Interpretation

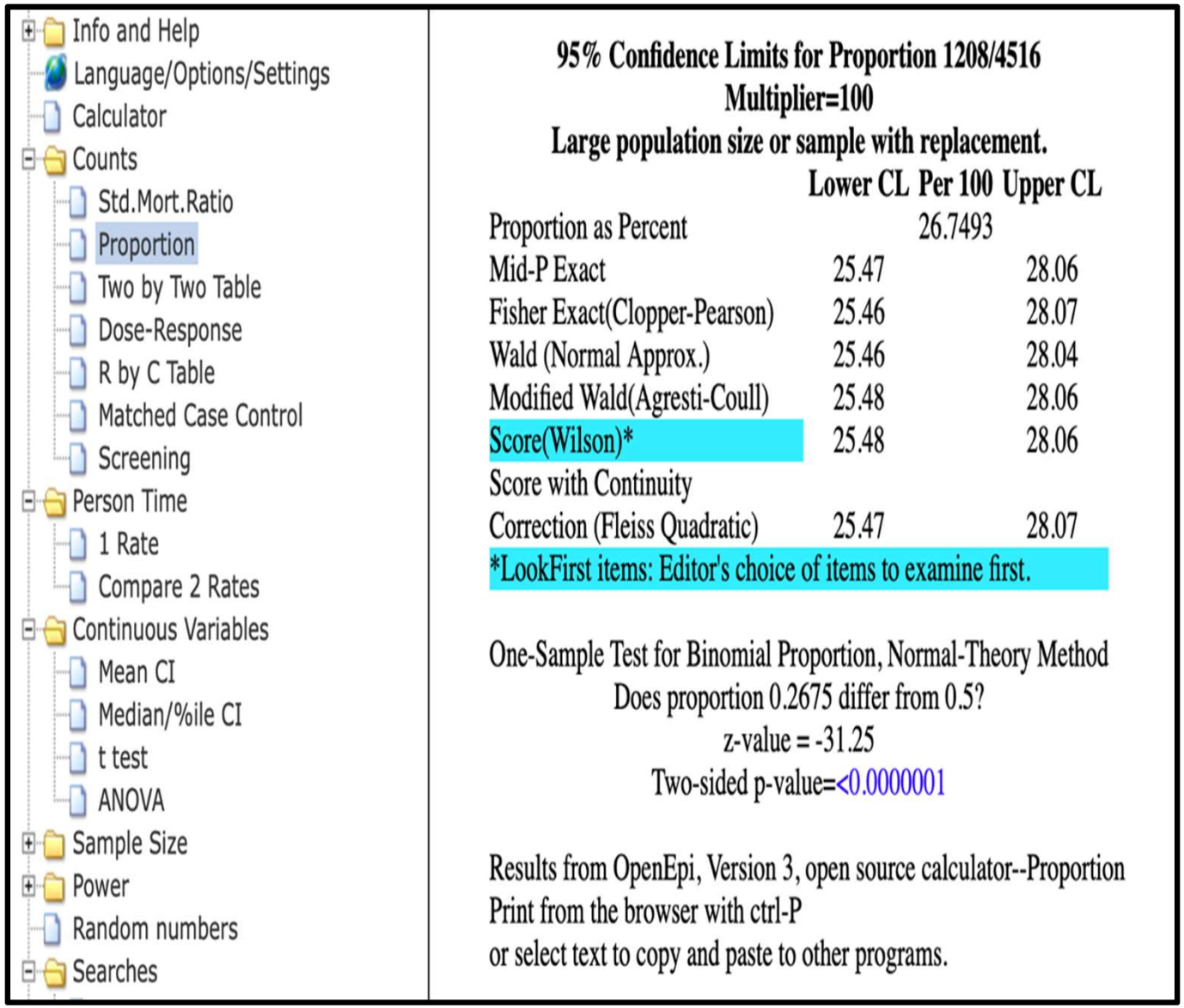
- The point estimate for the true mean life expectancy in the population is 70.5.
- The 95% CI is (70.15,70.85).
- We are 95% confident that the true mean is between 70.15 and 70.85 years.
- Given that the total life expectancy is 69.3, which is outside the interval and is lower than the left endpoint of the interval, this suggests that the mean life expectancy of Filipinos had increased compared with the 2018 data.

One Sample with Dichotomous Outcome

- In 2006, there were approximately 1.4 million hospitalizations specifically for MH conditions*. In total, **21.3%** of hospital stays included some mention of a MH condition as either a principal or secondary diagnosis.
- Does the data in 2018 provide evidence that there is a change in this proportion when **1,208** of the **4,516** of hospital stays included some mention of a MH condition as either a principal or secondary diagnosis?
- Obtain the 95% CI for this proportion.

https://www.hcup-us.ahrq.gov/reports/statbriefs/sb62.pdf

- ❑ The sample proportion is *p*=26.75%.
- ❑ Point estimate, *i.e.*, our best estimate of the proportion of the population.
- ❑ We are 95% confident that the true proportion lies between 25.47% and 28.06%.
- ❑ The proportion has significantly increased vs 21.3% which lies outside the confidence interval.



Two Independent Samples with Continuous Outcome

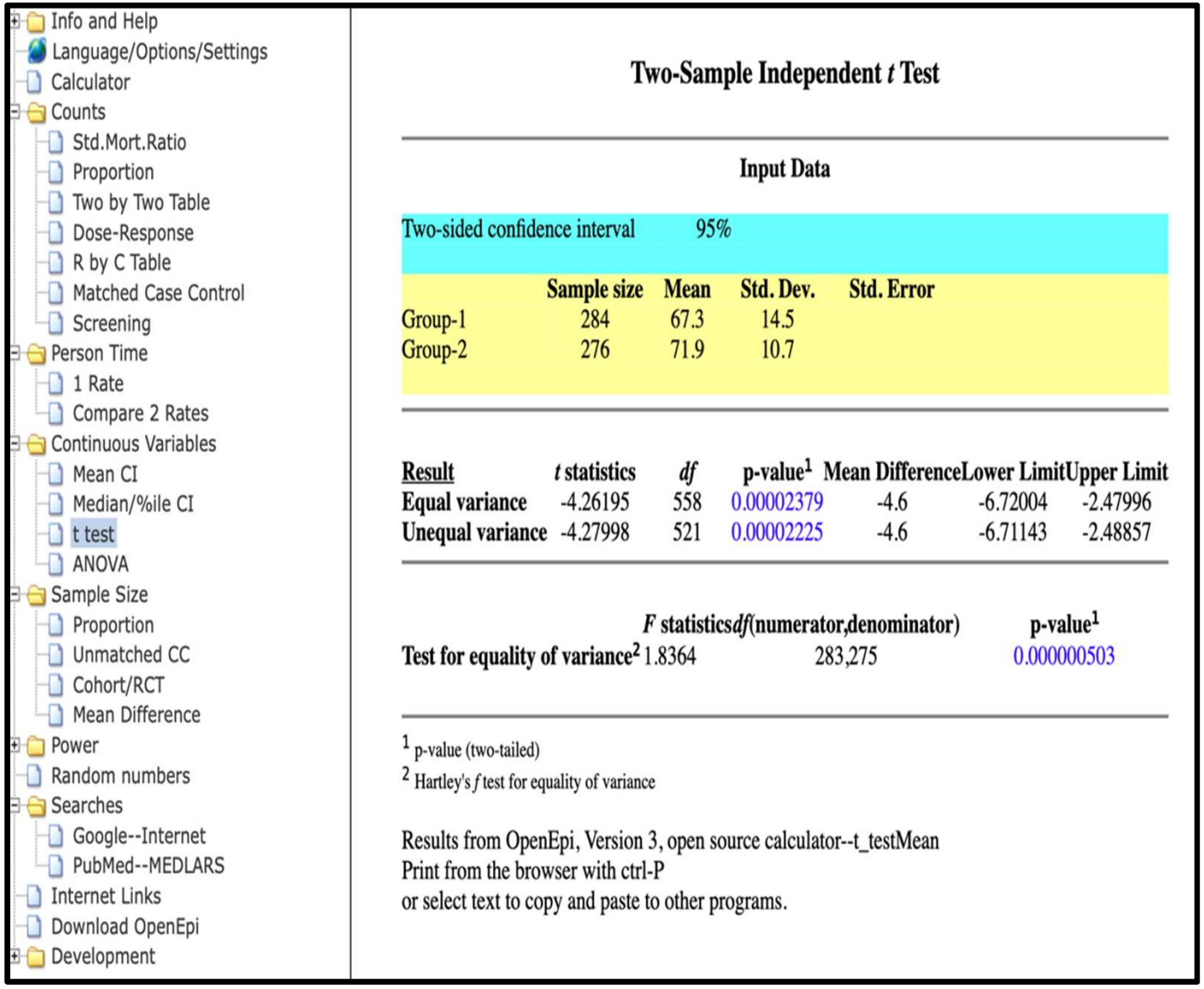
- According to the latest WHO* data published in 2018 life expectancy in Philippines is: Male 66.2, female 72.6, and total life expectancy is 69.3.
- Suppose we want to calculate the difference in the mean life expectancy between men and women, and we also want the 95% confidence interval for the difference in means.
- The descriptive statistics are given below.

Men			**Women**		
n	$\bar{x}$	s	n	$\bar{x}$	s
284	67.3	14.5	276	71.9	10.7

*https://www.worldlifeexpectancy.com/philippines-life-expectancy

Men			Women		
n	$\bar{x}$	s	n	$\bar{x}$	s
284	67.3	14.5	276	71.9	10.7

- With 95% confidence, the difference in mean life expectancy between men and women is between -6.71 and -2.49.
- Our best estimate of the difference, the point estimate, is -4.6 years.



Two-Sample Independent *t* Test

Input Data

Two-sided confidence interval 95%

	Sample size	Mean	Std. Dev.	Std. Error
Group-1	284	67.3	14.5	
Group-2	276	71.9	10.7	

Result	*t* statistics	*df*	p-value[1]	Mean Difference	Lower Limit	Upper Limit
Equal variance	-4.26195	558	0.00002379	-4.6	-6.72004	-2.47996
Unequal variance	-4.27998	521	0.00002225	-4.6	-6.71143	-2.48857

	F statistics	*df*(numerator,denominator)	p-value[1]
Test for equality of variance[2]	1.8364	283,275	0.000000503

[1] p-value (two-tailed)
[2] Hartley's *f* test for equality of variance

Results from OpenEpi, Version 3, open source calculator--t_testMean
Print from the browser with ctrl-P
or select text to copy and paste to other programs.

Two Independent Samples with Dichotomous Outcome

- The following table contains data on the prevalent substance abuse among participants who were non-smokers and those who were smokers at the time of the study.

	Substance Abuse		
	No	**Yes**	**Total**
Non-Smoker	1,738	126	1,864
Smoker	851	235	1,086
Total	2,589	361	2,950

- The point estimate of prevalent substance abuse among non-smokers is 126/1,864 = 0.0676 = 6.76%.
- The point estimate of prevalent substance abuse among smokers is 235/1,086 = 0.2164 = 21.64%.
- Does the proportion of prevalent substance abuse in smokers differ as compared to non-smokers?

- Info and Help
- Language/Options/Settings
- Calculator
- Counts
  - Std.Mort.Ratio
  - Proportion
  - Two by Two Table
  - Dose-Response
  - R by C Table
  - Matched Case Control
  - Screening
- Person Time
  - 1 Rate
  - Compare 2 Rates
- Continuous Variables
  - Mean CI
  - Median/%ile CI
  - t test
  - ANOVA
- Sample Size
- Power
- Random numbers
- Searches
  - Google--Internet
  - PubMed--MEDLARS
- Internet Links
- Download OpenEpi
- Development

2 x 2 Table Statistics

Single Table Analysis

		Substance Abuse (+)	Substance Abuse (-)	
Smoking	(+)	235	851	1086
	(-)	126	17381	864
		361	25892	950

Chi Square and Exact Measures of Association

Test	Value	p-value(1-tail)	p-value(2-tail)
Uncorrected chi square	141.5	<0.0000001	<0.0000001
Yates corrected chi square	140.1	<0.0000001	<0.0000001
Mantel-Haenszel chi square	141.4	<0.0000001	<0.0000001
Fisher exact		<0.0000001	<0.0000001
Mid-P exact		<0.0000001	<0.0000001

All expected values (row total*column total/grand total) are >=5
OK to use chi square.

Risk-Based* Estimates and 95% Confidence Intervals
(Not valid for Case-Control studies)

Point Estimates		Confidence Limits	
Type	**Value**	**Lower, Upper**	**Type**
Risk in Exposed	21.64%	19.29, 24.19	Taylor series
Risk in Unexposed	6.76%	5.703, 7.994	Taylor series
Overall Risk	12.24%	11.1, 13.47	Taylor series
Risk Ratio	3.201	2.613, 3.922[1]	Taylor series
Risk Difference	14.88%	12.18, 17.58°	Taylor series
Etiologic fraction in pop.(EFp)	44.76%	37.37, 52.16	
Etiologic fraction in exposed(EFe)	68.76%	61.73, 74.5	

Odds-Based Estimates and Confidence Limits

Point Estimates		Confidence Limits	
Type	**Value**	**Lower, Upper**	**Type**
CMLE Odds Ratio*	3.807	3.024, 4.808[1]	Mid-P Exact
		3.005, 4.84[1]	Fisher Exact
Odds Ratio	3.809	3.022, 4.801[1]	Taylor series
Etiologic fraction in pop.(EFpIOR)	48.01%	40.55, 55.46	
Etiologic fraction in exposed(EFeIOR)	73.75%	66.91, 79.17	

*Conditional maximum likelihood estimate of Odds Ratio
(P)indicates a one-tail P-value for Protective or negative association; otherwise one-tailed exact P-values are for a positive association.
Martin,D; Austin,H (1991) An efficient program for computing conditional maximum likelihood estimates and exact confidence limits for a common odds ratio. Epidemiology 2, 359-362.
° [1] 95% confidence limits testing exclusion of 0 or 1, as indicated
P-values < 0.05 and confidence limits excluding null values (0,1, or [n]) are highlighted.
LookFirst items: Editor's choice of items to examine first.

- ❑ The 95% confidence interval on the difference **does not** include zero.
- ❑ The difference in prevalent substance abuse between smokers and non-smokers is **statistically significant.**

p-value

- probability of obtaining a result as extreme or more extreme than the actual sample obtained, given that the null hypothesis is true

***p*-value**	**Interpretation**
$p \leq 0.05$	The null hypothesis is rejected unless the level of significance was set at 0.01, and the results are said to be statistically significant
$0.01 \leq p \leq 0.05$	The results are significant, and is usually denoted by one asterisk (*)
$0.001 < p < 0.01$	The results are highly significant, and is usually denoted by two asterisks (**)
$p > 0.05$	The results are not statistically significant, unless level of significance was set at 0.10
$0.05 < p \leq 0.10$	There is a trend towards statistical significance. Borderline significance, usually due to sample size problems

Chi-square test for independence

In an experiment to study the dependence of substance abuse on smoking habits, the following data were taken on 180 individuals:

	Nonsmokers	Moderate smokers	Heavy smokers
Substance abuse	21	36	30
No substance abuse	48	26	19

Test the hypothesis that the presence or absence of substance abuse is independent of smoking habits at 5% level of significance.

Chi-square test for independence

	Nonsmokers	Moderate smokers	Heavy smokers
Substance abuse	21	36	30
No substance abuse	48	26	19

```
. tabi 21 36 30\48 26 19, chi2 exp

| Key                    |
|------------------------|
|      frequency         |
| expected frequency     |

           |        col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        21         36         30 |        87
           |      33.4       30.0       23.7 |      87.0
-----------+---------------------------------+----------
         2 |        48         26         19 |        93
           |      35.6       32.0       25.3 |      93.0
-----------+---------------------------------+----------
     Total |        69         62         49 |       180
           |      69.0       62.0       49.0 |     180.0

          Pearson chi2(2) =  14.4636   Pr = 0.001
```

- **Statistical decision:**

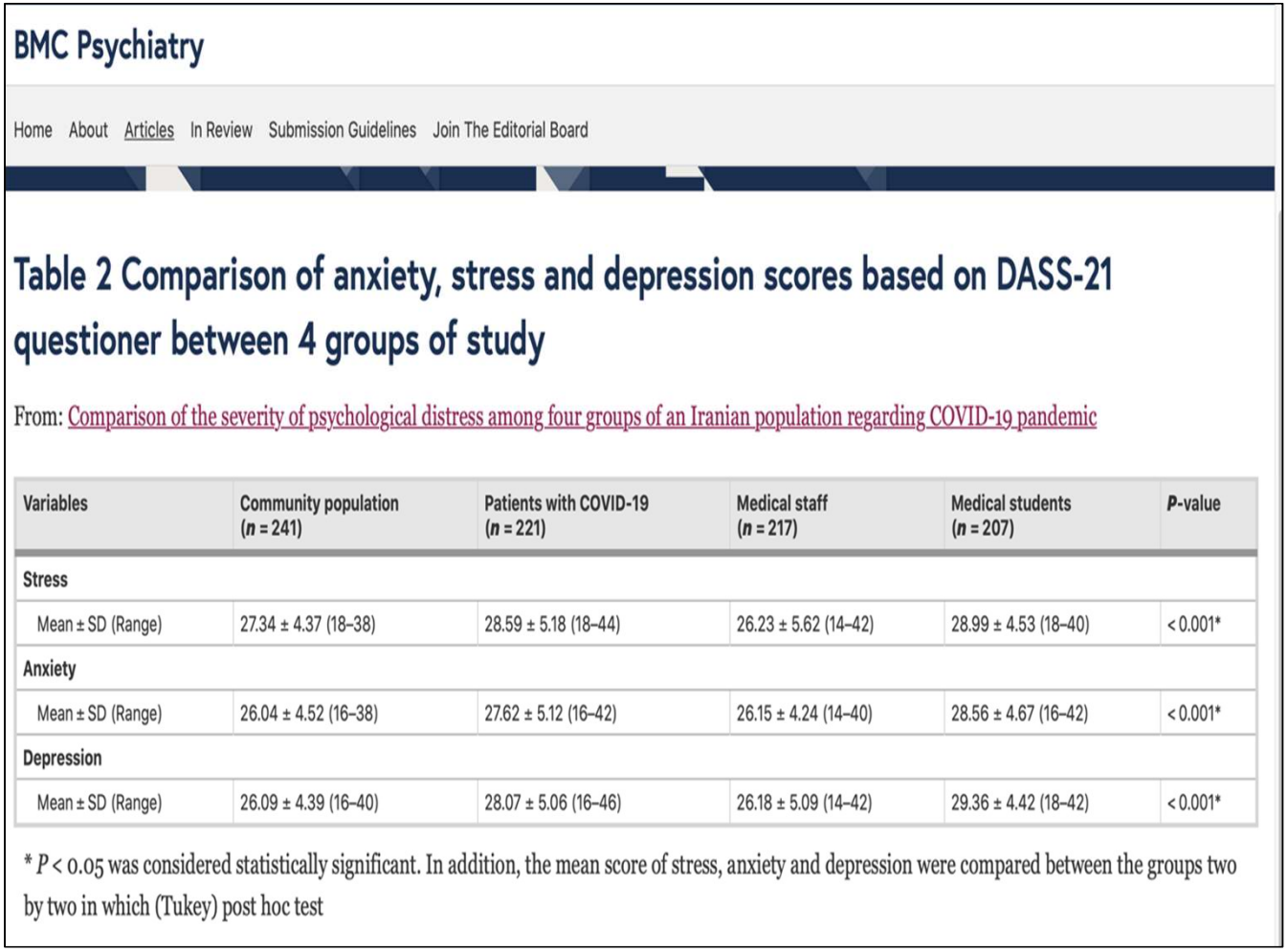
The *p*-value (0.001) is less than α=0.05, reject Ho.

- **Conclusion:**

The presence or absence of substance abuse is not independent of smoking habits.

Analysis of Variance

- Objective: To measure and compare the severity of psychological distress which includes anxiety, stress, and depression among four groups of an Iranian population.
- ANOVA test was used to compare the severity of stress, anxiety, and depression among the four study groups.
- Is there a statistically significant difference in the mean psychological stress score among the four population groups?



BMC Psychiatry

Home About Articles In Review Submission Guidelines Join The Editorial Board

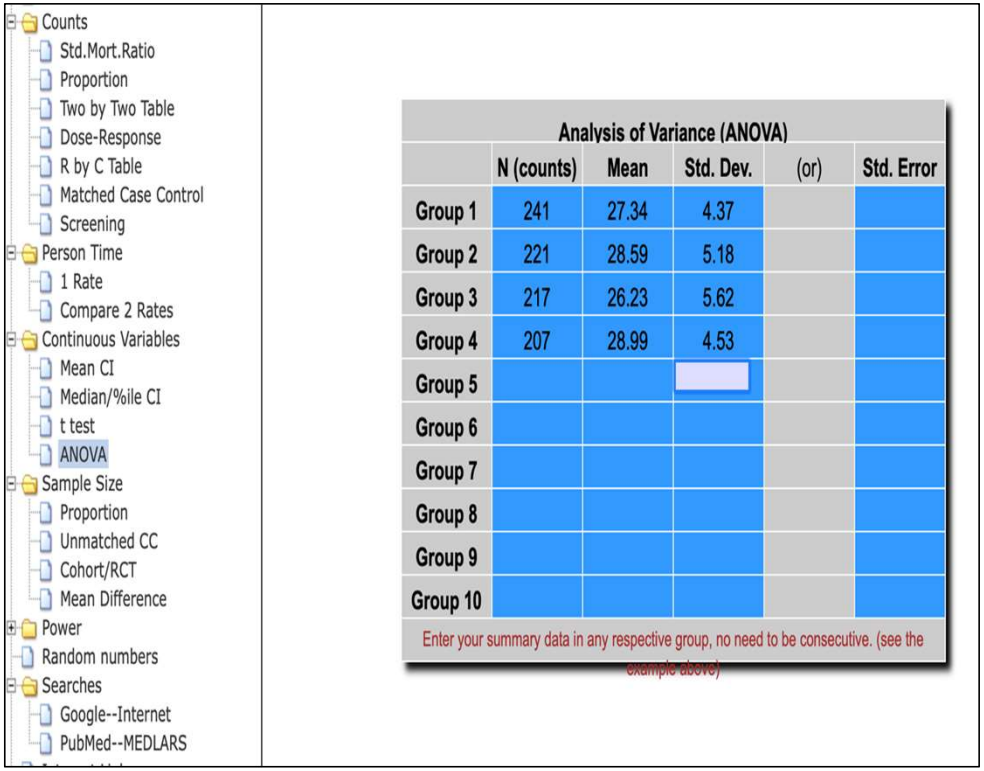
Table 2 Comparison of anxiety, stress and depression scores based on DASS-21 questioner between 4 groups of study

From: Comparison of the severity of psychological distress among four groups of an Iranian population regarding COVID-19 pandemic

Variables	Community population (*n* = 241)	Patients with COVID-19 (*n* = 221)	Medical staff (*n* = 217)	Medical students (*n* = 207)	*P*-value
Stress					
Mean ± SD (Range)	27.34 ± 4.37 (18–38)	28.59 ± 5.18 (18–44)	26.23 ± 5.62 (14–42)	28.99 ± 4.53 (18–40)	< 0.001*
Anxiety					
Mean ± SD (Range)	26.04 ± 4.52 (16–38)	27.62 ± 5.12 (16–42)	26.15 ± 4.24 (14–40)	28.56 ± 4.67 (16–42)	< 0.001*
Depression					
Mean ± SD (Range)	26.09 ± 4.39 (16–40)	28.07 ± 5.06 (16–46)	26.18 ± 5.09 (14–42)	29.36 ± 4.42 (18–42)	< 0.001*

* $P < 0.05$ was considered statistically significant. In addition, the mean score of stress, anxiety and depression were compared between the groups two by two in which (Tukey) post hoc test

Vahedian-Azimi, A., Moayed, M.S., Rahimibashar, F. *et al.* Comparison of the severity of psychological distress among four groups of an Iranian population regarding COVID-19 pandemic. *BMC Psychiatry* **20,** 402 (2020). https://doi.org/10.1186/s12888-020-02804-9



Analysis of Variance (ANOVA)

Input Data

Group	N (count)	Mean	Std. Dev.	Std. error
1	241	27.34	4.37	
2	221	28.59	5.18	
3	217	26.23	5.62	
4	207	28.99	4.53	
5				
6				
7				
8				
9				
10				

ANOVA Table

Source of variation	Sum of squares	d.f	Mean square	*F* statistics	p-value[1]
Between Groups	1015.88	3	338.626	13.8684	0.00000000765784
Within Groups	21535.9	882	24.4171		
Total	22551.8	885			

	Chi square	d.f	p-value[1]
Test for equality of variance	18.4137	3	0.000361361

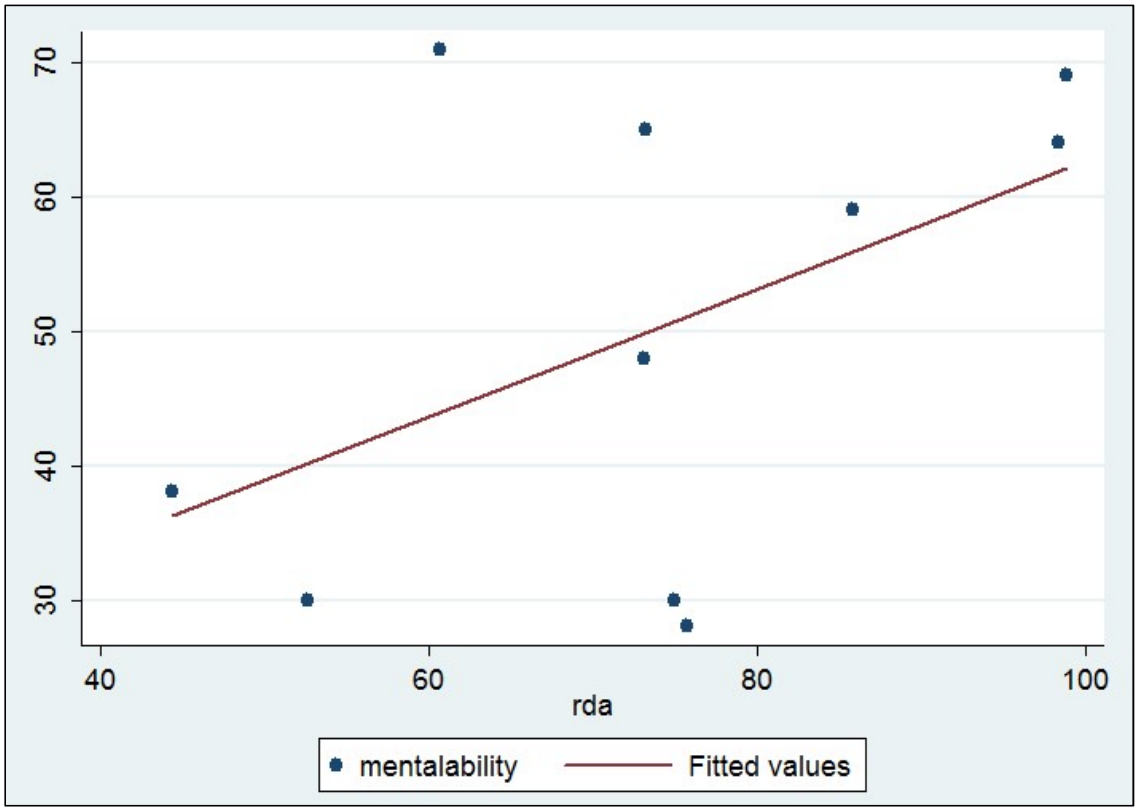
		95% CI of individual sample mean		95% CI assuming equal variance	
Group	Mean	Lower Limit	Upper Limit	Lower Limit	Upper Limit
1	27.34	26.7855	27.8945	26.713	27.967
2	28.59	27.9033	29.2767	27.9349	29.2451
3	26.23	25.478	26.982	25.5688	26.8912
4	28.99	28.3693	29.6107	28.3129	29.6671

Linear Regression

- A study determined the relationship between nutrient intake and the academic performance of grade school pupils.
- The specific variables considered are the mental ability scores and caloric intake, measured in terms of the percentage of the required daily allowance (%RDA) for their age and sex.
- The researchers wanted to derive a formula which will enable them to predict the mental ability score of a pupil given his %RDA for calorie.

Mental ability	%RDA for calories
65	73.2
30	74.9
28	75.7
64	98.3
38	44.4
71	60.7
48	73.1
69	98.8
30	52.6
59	85.8

Scatterplot with the best-fit line



Mental ability	%RDA for calories
65	73.2
30	74.9
28	75.7
64	98.3
38	44.4
71	60.7
48	73.1
69	98.8
30	52.6
59	85.8

Regression Equation

```
. regress  mentalability rda
```

Source	SS	df	MS
Model	643.225514	1	643.225514
Residual	2092.37449	8	261.546811
Total	2735.6	9	303.955556

Number of obs	= 10
F(1, 8)	= 2.46
Prob > F	= 0.1555
R-squared	= 0.2351
Adj R-squared	= 0.1395
Root MSE	= 16.172

mentalabil~y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rda	.4742152	.3023908	1.57	0.155	-.2230992	1.17153
_cons	15.22663	22.8802	0.67	0.524	-37.5352	67.98846

Mental ability = 15.227 + 0.474 (%RDA)

Appropriate Types of Analysis

		Independent Variable		
		Nominal	Ordinal	Interval/Ratio
Dependent Variable	Nominal	*Crosstabs* Chi-square Lambda	*Crosstabs* Chi-square Lambda	
	Ordinal	*Crosstabs* Chi-square Lambda	*Crosstabs* Chi-square Lambda Gamma Kendall's tau Sommers' *d*	
	Interval/Ratio	*Means* *t*-test ANOVA	*Means* *t*-test ANOVA	*Correlate* Pearson *r* Regression (*R*)

Summary

- Computer and statistical software as tools for calculating descriptive measures and constructing various distributions from large data sets
- Confidence intervals are useful for comparing means or proportions and can be used to assess whether there is a statistically meaningful difference
- Check if the confidence interval includes the null value (*e.g.*, 0 for the difference in means, mean difference and risk difference or 1 for the relative risk and odds ratio).
- When tests of hypothesis are conducted using statistical computing packages, exact *p*-values are computed

Thank you!