Workbook for Biostatistics

On

Concepts of Statistical Inference

Prepared by:
Bandit Thinkhamrop, Ph.D. (Statistics)
Associate Professor of Biostatistics
Department of Biostatistics and Demography, Faculty of Public Health
Khon Kaen University, THAILAND

E-mail: bandit@.kku.ac.th

INTRODUCTION:

A good research needs to have involved an important issue, had useful objectives, used sensible design, used adequate sample size, performed appropriate analysis, and drawn reasonable inferences from the findings. Statistics plays a very important role not only on analysis but also on the other components.

> *Statistics is a curious amalgam of mathematics, logic and judgement. It is logical process that cause more difficult than mathematics- the principles of good design, and the concepts underlying data analysis and interpretation* (Altman, 1991; page 9).

This workbook is designed for participants who have a limited background in mathematics and statistics. It is believed that if participants can be convinced that statistics is a very important tool and useful for their professional, they will then open themselves for this difficult subject. Once they understood the key concept behind basic statistics, they will find their own way in trying to understand a more complicated statistical methods. An example of the evidence of success, reported by a participant of a previous a course in which this workbook had been used, is as quoted - "*The class has been finished long time ago but I found many of my classmates keep on reading statistics books. I bought several books of this kind for myself in which I have never even taken a glance on it prior to the course."* However, participants who gained a little from this are not abnormal. But this little gain is hopefully a great starting point of the participants.

OBJECTIVES:

Upon completion of this workbook, participants should be able to:

1. describe the underlying "core" concepts of statistical methods,
2. outline the principle of choosing appropriate statistical methods,
3. outline important steps in data analysis including examining the data, describing the study sample, and answering the research questions,
4. describe the "core" concepts underlying statistical inference,
5. interpret correctly the confidence intervals and the p-value, and
6. describe how to apply probability theory in research practice

CONTENTS:

PART 1: PRETEST

A controlled trial of a new treatment led to the conclusion that it is significantly better than placebo (p-value < 0.05).

Which of the following statement do you prefer?

[ ]1. It has been proved that treatment is better than placebo.

[ ]2. If the treatment is not effective, there is less than a 5 percent chance of obtaining such results.

[ ]3. The observed effect of the treatment is so large that there is less than a 5 percent chance that the treatment is no better than placebo.

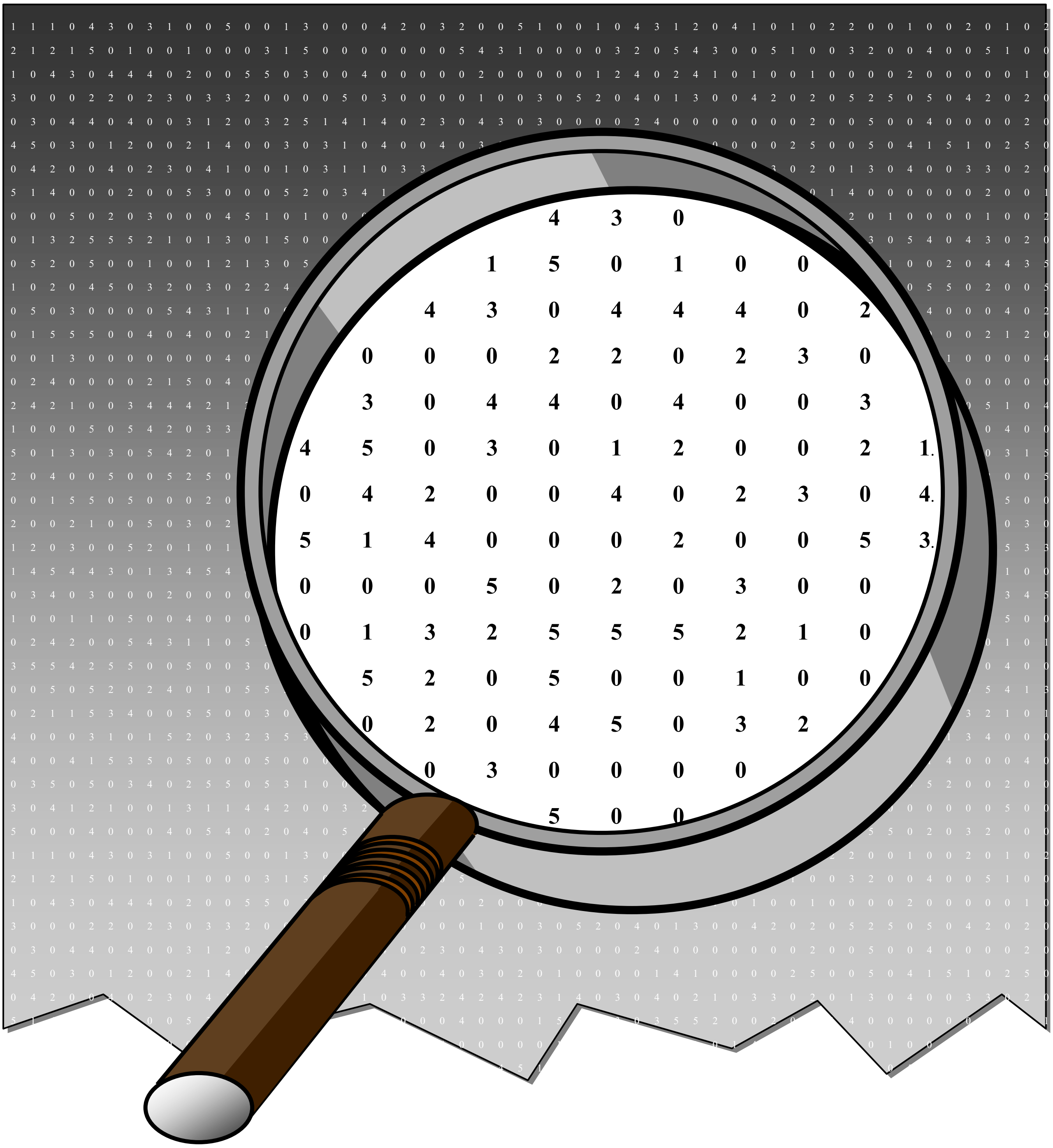
[ ]4. I do not really know what a p-value is and do not want to guess.

Source: *Wulff, H.R., Andersen, B., Brandenhoff, P., and Guttler, F. (1987) What do doctors know about statistics?. Statistics in Medicine,* ***6****, 3-10.*

Justifications of the answer:

PART 2: STATISTICAL INFERENCE

1. Followings are parts of unlimited item of data. Let we call these the ***target population***. The data represent the number of cigarettes smoked per day reported by each member of a community. Assuming these data are located at random in this long piece of paper, lists of items within the magnifying glass are what we can clearly see, i.e. what we only have at hand. Let we call these the ***sampling frame***. Now please randomly selected your study sample from the sampling frame with the sample size of 10.



2. List all of your samples (n = 10)

3. Construct frequency table for weight of the student of n = 10

Number of cigarettes per day	**Frequency**	**Percent**	**Cumulative percentage**
Total			

Question: For how many percents of the sample smoke 3 or more cigarettes per day?……….%

4. Construct a histogram and then superimpose by a curve covering the same size of area, i.e., n blocks or 10 blocks for this example.

Question: For how many percents of the area under the curve corresponds to smoking of 3 or more extreme?……….%

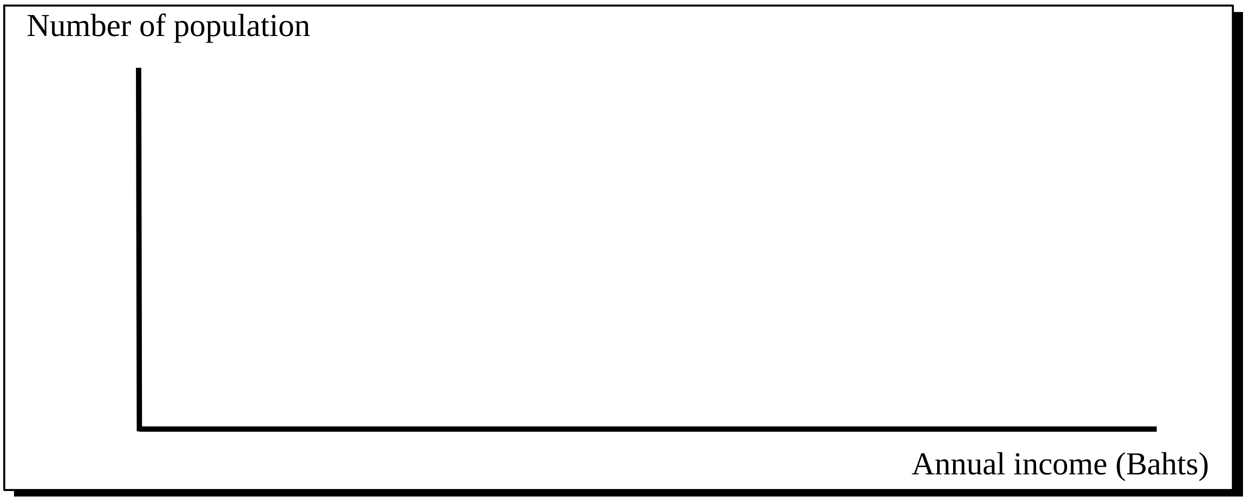
If we can validly assume that the distribution in anywhere in Thailand is similar to what was found by this study, how many people smoking 3 or more cigarette per day within a community of 400,000 population? ………………

5. Describe type of the distribution of weight

[ ] Symmetry
[ ] Left skew (negative skew)
[ ] Right skew (positive skew)
[ ] Others ……………………………………………………………………………

6. Followings are hypothetical data ***"A study of annual income per capita among Thai people found that 80% were poor (poor is classified as those whose income of less than 20000 bahts per year). The range is as small as 500 bahts to several thousand million bahts "***

6.1 Construct the approximate the frequency curve to reflect the findings



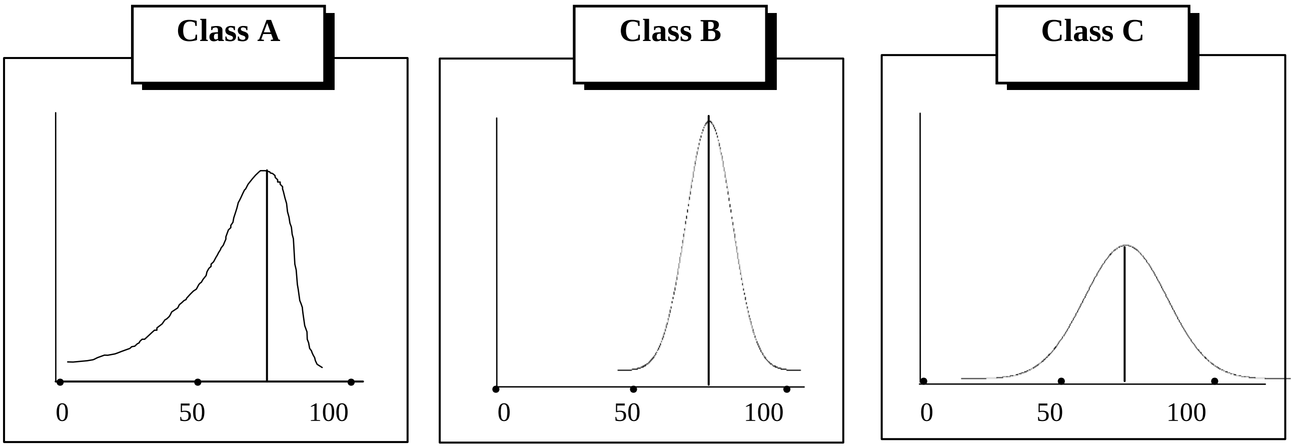
6.2 What are the income which make the area under the curve covering 80% of total Thai population. ....
How did you get that: ....

(Note that this is similar to the question for parameter estimation, i.e. it is to obtain the lower and upper limit of a quantity that cover a specified magnitude of proportion or probability)

6.3 What is the probability of having found people who have the annual income of 20,000 Bahts or higher: ....
How did you get that: .... …

(Note that this is similar to the question for testing hypothesis, i.e., it is to obtain the probability from a quantity which is calculated from comparing the observed estimate with the hypothesized value. Such quantity is the statistical test and the corresponding probability is the p-value which will then be compared with the cut point such as 0.05 and interpret as significant or non-significant.)

7. Examine the frequency curve reflecting the marks of biostatistics examination for three classrooms bellow and then answer the questions:



7.1 Describe the types of the distribution and their implications
Class A ....
Class B ....
Class C ....

7.2 What is the main different of Class B and C
....

7.3 What is the most appropriate summary statistics for each class
Class A ....
Class B ....
Class C ....

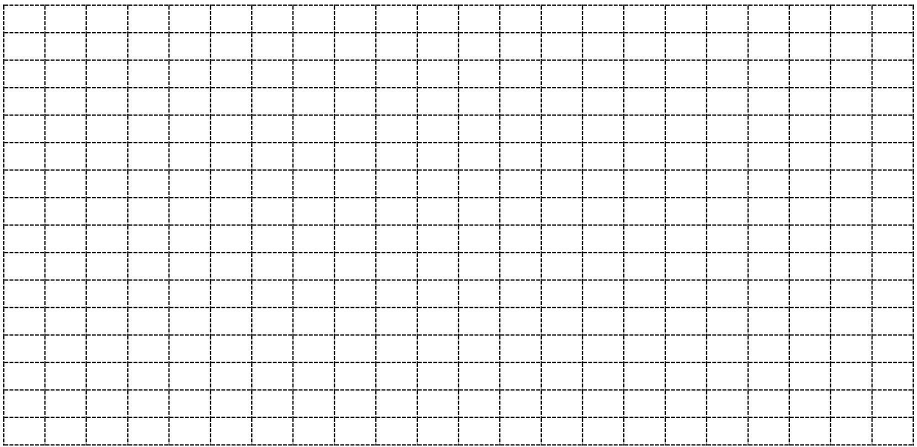
8. To summarize about a continuous data, the **two** important numbers to be reported are
8.1 ………………………………… 8.2 ……………………………………

9. Write on your own word regarding the important of examining the distribution of data
....
....

10. Sampling variation
10.1 Write down the mean number of cigarettes per day you have obtained from your smoking survey. Then take that figure from all of your class mate.

10.2 If someone conclude their findings based solely on the statistic, what kind of comments you will make. …………………………………………………………
…………………………………………………………………………………………
...………………………………………………………………………………………

10.3 Construct a histogram of the mean number of cigarettes per day superimposed by a frequency curve



10.4 Estimate the mean of the sample mean number of cigarettes per day from the graph shown above = ...
Describe how did you guess that amount ....

10.5 The mean number of cigarettes per day obtained from all 115 subjects was ....

10.6 Let's summarize:

10.6.1 From the data of "n" of 10, we get the sample mean. The mean here is called the "statistics" denoted by $\overline{X}$, pronounced x-bar.

10.6.2 From the data of "N" of 115, we get the population mean. The mean here is called the "parameter" denoted by μ, pronounced mu.

10.6.3 From the repeated sample, we get the mean of the sample mean which is equivalent to the population mean.

10.6.4 What we aim to estimate is the ***parameter***, i.e., the population mean in this example. In reality, we cannot do like #10.6.2 for several reasons and thus, almost all situations, the parameter is unknown. The repeated sample (in #10.6.3) told us that there is the way to get the parameter without having been collected the data from all members of the population. But how? No body did like #10.6.3 too. We did a study only once. What we have at hand now is only the ***statistic***, i.e., the sample mean (in #10.6.1) which we are ***not sure*** at all whether or not it closes to the parameter. Thus we need to quantify the "***not sure***" so that we can tell about the parameter using the statistic in terms of "***how many percent sure***". The *percentage* and *probability* in this sense are closely related.

11. Try to understand the Central Limit Theorem

11.1 Compare the distribution of the three types of data from the smoking survey.

Fig. 1. Distribution of the data from all members of the population (N = 115).

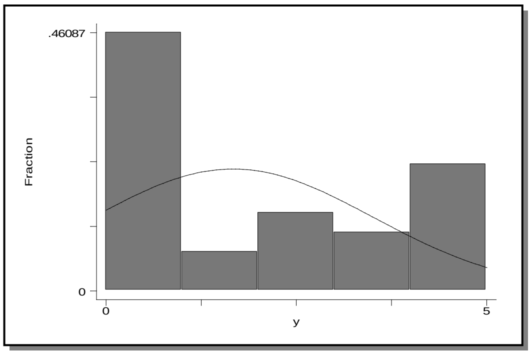
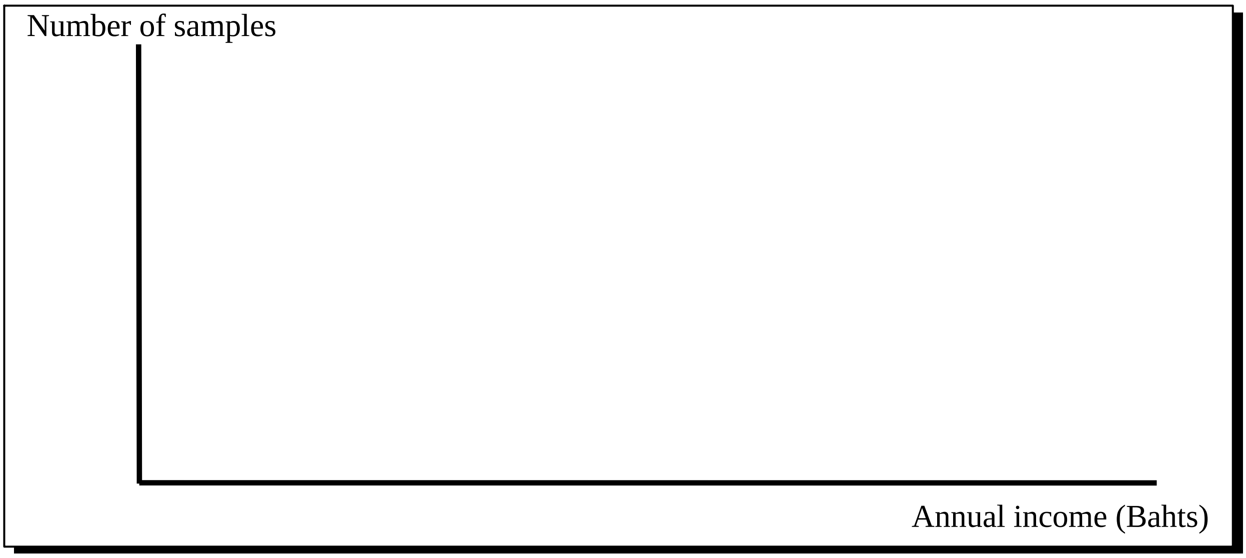


Fig. 2. Distribution of the data from a sample of size n = 10

Fig. 3. Distribution of the data from repeated sample of size n = 10

Comments: ………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………
………………………………………………………………………………

11.2 From the information in Part 3, Number 4.1, regarding the hypothetical data that ***"A study of annual income per capita among Thai people found that 80% were poor (poor is classified as those whose income of less than 20000 bahts per year). The range is as small as 500 bahts to several thousand million bahts ".*** If several researchers doing the same study. The mean income they obtained from their study can be used to construct a frequency curve as the bellow figure.



Describe in the probability words about a guy who earn 50,000 Bahts annually:
……………………………………………………………………………………………
……………………………………………………………………………………………

11.3 Back to the smoking survey, what shape of the distribution would be if all researchers increase the sample size to 50?

The middle of the curve, i.e. the mean, is ………… (the same / smaller / larger).

The base of the curve is …………………………… (the same / narrower / wider).

11.4 We have seen from #11.3 that the width of the base of the curve is affected by sample size. Such width quantify the spread or the deviation of the data. Let's calculate the standard deviation of your smoking survey data. Firstly calculate the mean. Secondly subtract each data from the mean, then square this amount. Thirdly sum all the square of the difference between the data and its mean. Fourthly divided the sum by the sample size. This amount is called "variance". Square root of this amount is the standard deviation. We can write in formula as:

$$SD = \sqrt{\frac{\sum\left(X-\bar{X}\right)^2}{(n-1)}}$$

Therefore, your SD is ……………………………………………………………………………

This is the SD of your sample size of 10.

11.5 Do the same for SD of sample mean

Therefore, the SD of the sample mean is ………………………………..………………

This is called the standard error, denoting SE.

11.6 Let's divided SD obtained in #11.4 by square root (n), you get ……………………

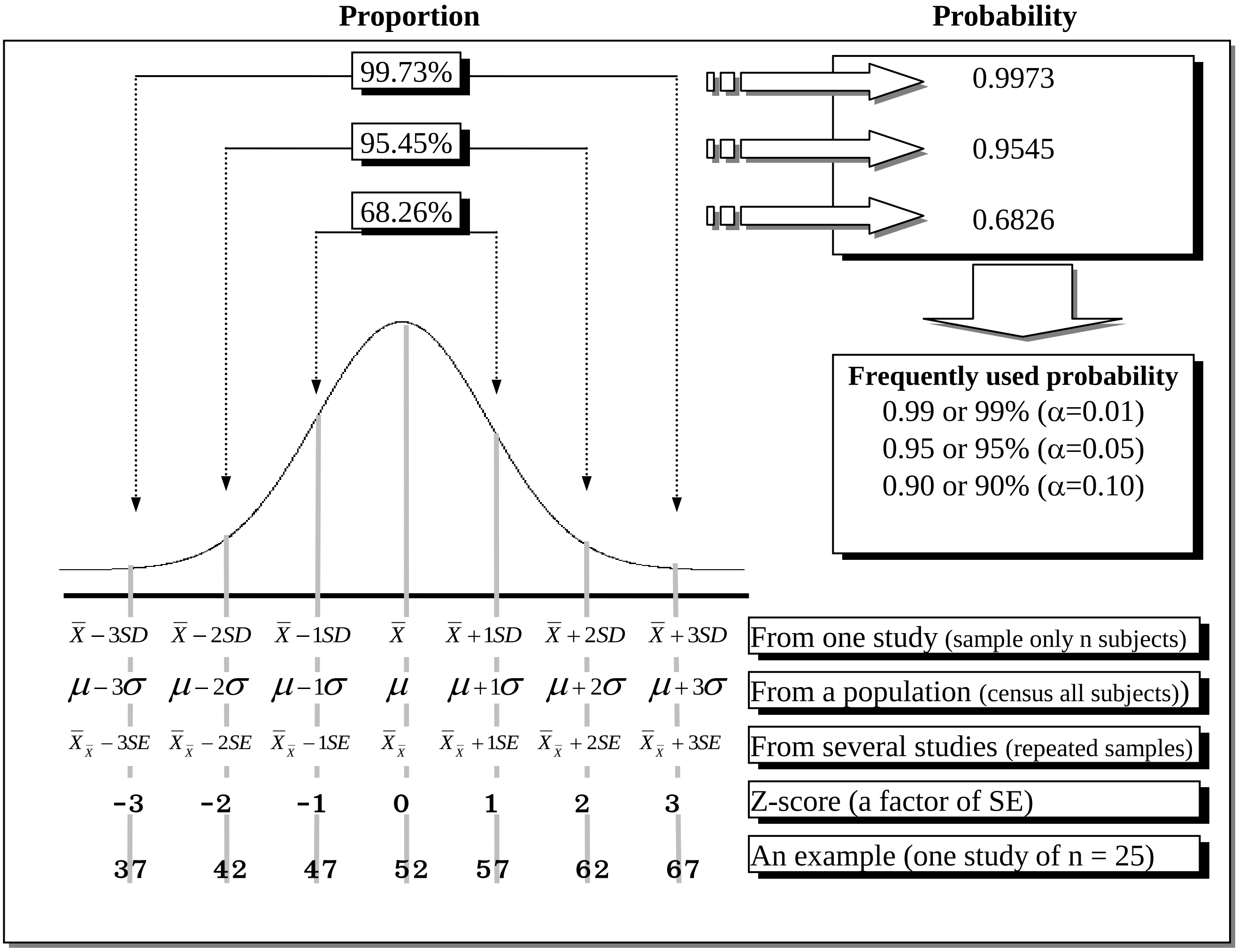
11.7 Thus SE is approximately equal to SD divided by square root (n).

11.9 Let's summarize:

We have known that we can estimate the population mean without having been collected the data from all members of the population, by estimating it from the sampling distribution obtained from the repeated samples. But we still no hope since nobody does the repeated sample. Then the central limit theorem told us that whatever the distribution of the population data, the sampling distribution always become symmetry. It became the "***normal distribution***" if the sample size is sufficiently large. (*This is always true for the infinite population or a very large population, but for the definite or small population such as the present example it is not so. That is why we need to adjust for the definite population in making an inference.*) Knowing the distribution is normal, we can make use of its properties regarding the area under the curve and the score as we have done previously. We have known that the curve can be fully describe if we know its mean and SD. Since the SD of the repeated samples, i.e. the SE can be estimated from the SD that we only have at hand as if we did the repeated sample. Playing around the distribution of the sampling distribution can lead to knowing the population parameter.

12. Make use of the property of the normal distribution.

Area under the Normal Distribution Curve and other related parameters



Number 12.1 to 12.5 uses the example at the last line of the diagram

12.1 A study yields a mean score of 52. The SD of this data is …………………..

12.2 The SE for this study is …………………

12.3 If the distribution of 'raw' data from some observations of a population, i.e., the sample data, is found to be normal, the scores that ***most likely*** to include 95.45% of the observations in this sample is between …… and ……

Note that this has nothing to do with the statistical inference, just illustrate its use.

12.4 If the distribution of the means from repeated samples is found to be normal, the mean scores that most likely to include 95.45% of the repeated samples is between ………… and ……..….

Note that this is a 95.45% confidence interval. It is the magnitude of the parameter of interest. In comparative studies, we called it the magnitude of effect.

12.5 If the true mean score is in fact 55, what is the probability that a study will yield a mean of 52 or more extreme? It is ………………… .

*Note that this is the p-value. Comparing this with a pre-specified cut point α, such as 0.05, leads to making decision of reject or not reject that the true mean is 55. If the p-value is less than the α, then we reject that the true mean is 55. That is, it is unlikely that a study yields a mean of 52 or more extreme if the true mean score is 55. We say that 52 is statistically **significantly** different from 55. On the other hand, if the p-value is greater than the α, then we do not reject that the true mean is 55. That is, the chance to obtain a mean of 52 or more extreme even if the true mean score is 55 is still high. In other words, the difference between 52 and 55 is due to the role of chance. We say that 52 is **not** statistically **significantly** different from 55. But we can neither conclude that they are equal or they are from the same population, nor that they are from the different population. The finding stills inconclusive. We have insufficient evident to conclude that they are different. Thus, the p-value provides only the strength of the evidence.*

12.6 If the sample size is small, the distribution is known not the normal distribution. The base of the curve must be ……………. (wider / narrower). It is called the "t - distribution". It will approximate the normal distribution if the sample size is sufficiently large. Thus the shape of the t-distribution depend on the sample size, denoting the degree of freedom (df). For the study that involve one group, the df = n-1. Another popular distribution is the Chi-square distribution. Like the t-distribution, it depend on df and approximate normal distribution for large sample.

12.7 Appendix 1 presents the three statistical tables for the normal distribution, the t-distribution, and the Chi-square distribution, respectively. The effective way to understand and be able to use them is by noticing their values corresponding a fixed probability such as 0.05. In the tables, they are highlighted by the ovals.

12.8 Relationship between Z-distribution, t-distribution, and χ^2 -distribution

Example: An example showing the relationship using two-tailed probability from the tables in Appendix 1 can be shown below.

$$\chi^2_{\alpha=0.05;df=1} = Z^2_{\alpha=0.05/2} = t_{\alpha=0.05/2;df=\infty} = 1.96^2 = 3.84$$

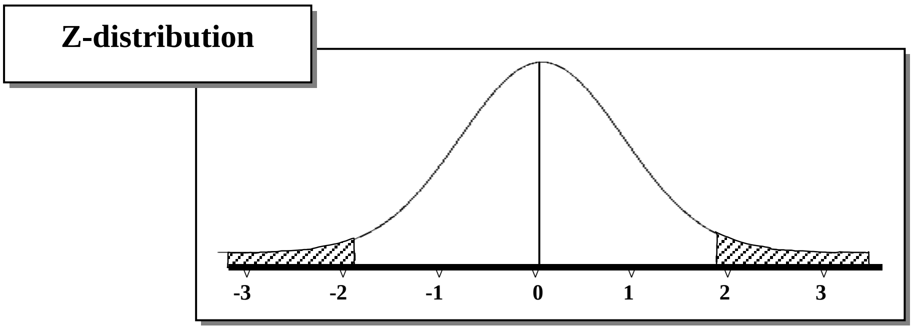
Question : The three graph shown below are the distribution of the three statistics. The shaded area correspond the probability of 0.05.

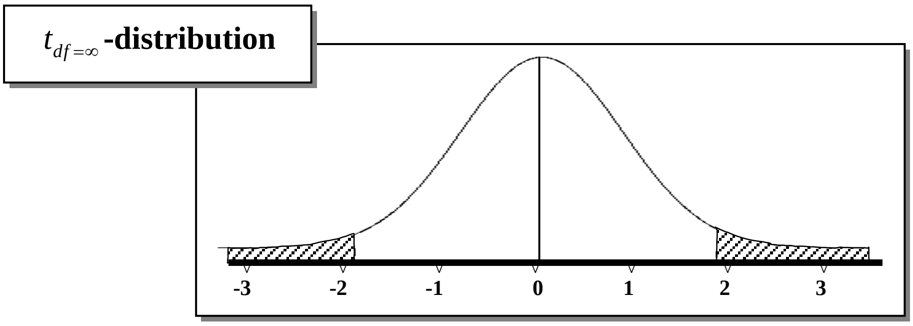
The vertical axis represent ………………………………………………………..…

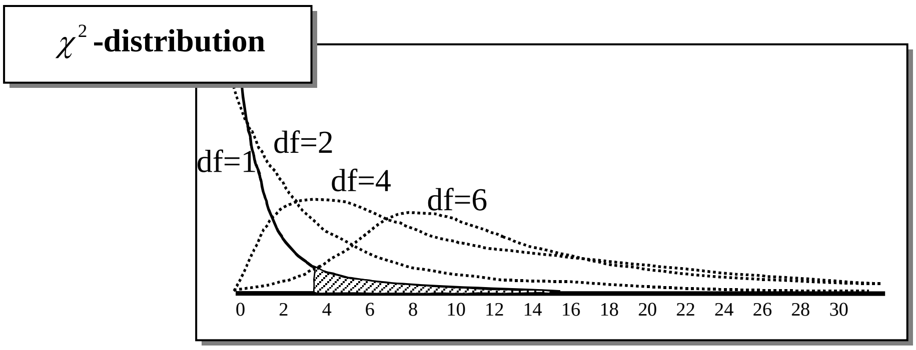
The horizontal axis represent …………………………………………………

Write down a curve of t-distribution given df = 29 (Write on the graph below.)

Write down a curve of χ^2 -distribution given df = 100 (Write on the graph below.)





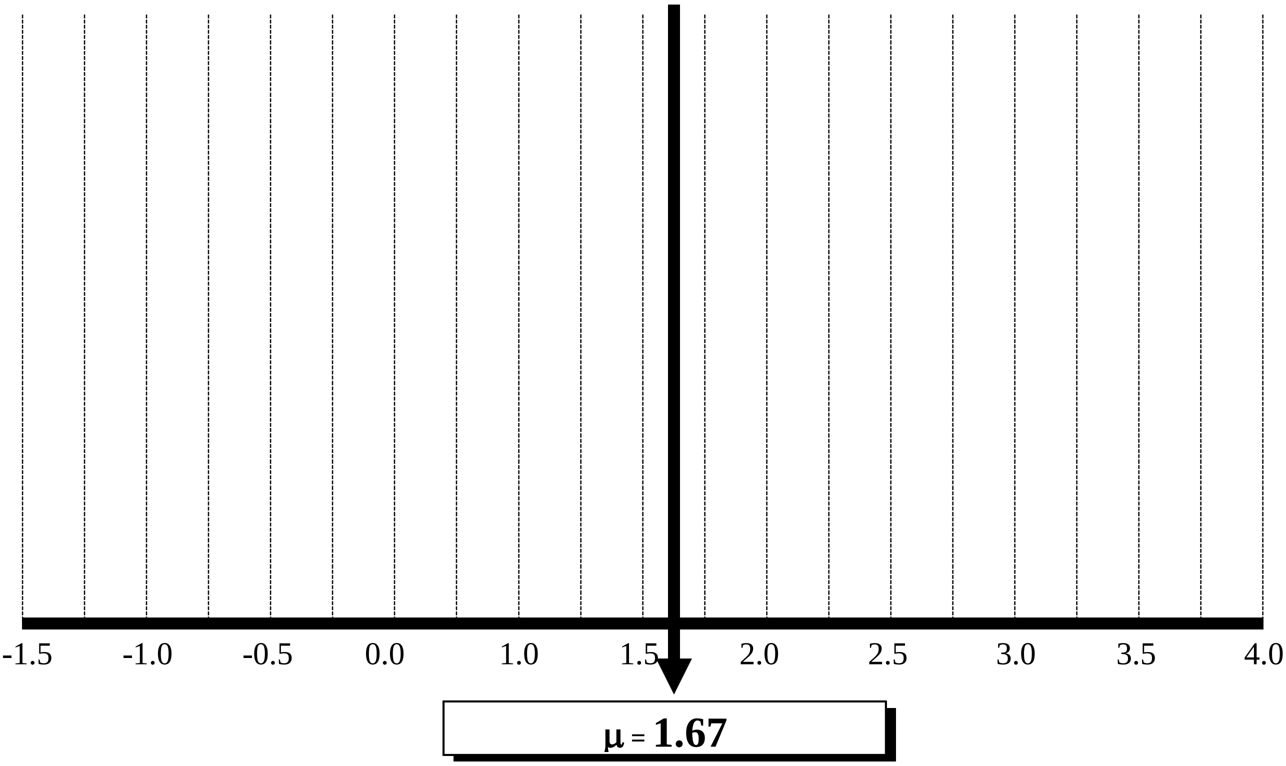


13. Understand statistical inference: the estimation

13. 1 Estimate the 95% Confidence interval of the mean number of cigarette per day in the population.

Formula : $\overline{X} \pm t_{0.025}\left(S / \sqrt{n-1}\right)$

13.2 Draw a horizontal line showing the 95% confidence intervals for mean weight from your study as well as your class mates'.



There are a total of ………….. intervals in the graph, …………. intervals did not cover the population mean. Thus, there are ……….. % of the intervals that cover the population parameter.

Interpret the confidence intervals in your own words:

13.3. The Ministry of Public Health reported that the average number of cigarette per day is 3. Perform hypothesis testing to provide evidence against that claim.

i) State the hypothesis:

ii) Set the significant level:

iii) Calculate the statistical test: Formula: $t = \dfrac{\overline{x} - k}{\dfrac{s}{\sqrt{n-1}}}$

iv) Find the p-value

v) Make decision regarding the hypothesis and conclude the findings:

13.4 Compare the 95% confidence intervals obtained in #4.1 with your conclusion from hypothesis testing obtained in #4.3 and describe your implications.

13.5 Report results of the smoking survey in your own word as complete as possible so that this paragraph could be put in the section of Results in the research report.

----

14. Summary of general formula for statistical inference

Confidence interval = Statistic calculated from the study sample ± [(Coefficient × (SE of the statistic]

$$\text{Statistical test} = \frac{\text{Statistic calculated from the study sample - Null value specified under } H_O}{\text{SE of the statistic}}$$

15. The main difference between the two components of statistical inference

15.1 The confidence interval provides:
[ ]the magnitude of the effect / [ ]the strength of evidence.

15.2 The p-value provides:
[ ]the magnitude of the effect / [ ]the strength of evidence.

16. The value of α and the coefficient frequently used for Normal distribution

Values that frequently used

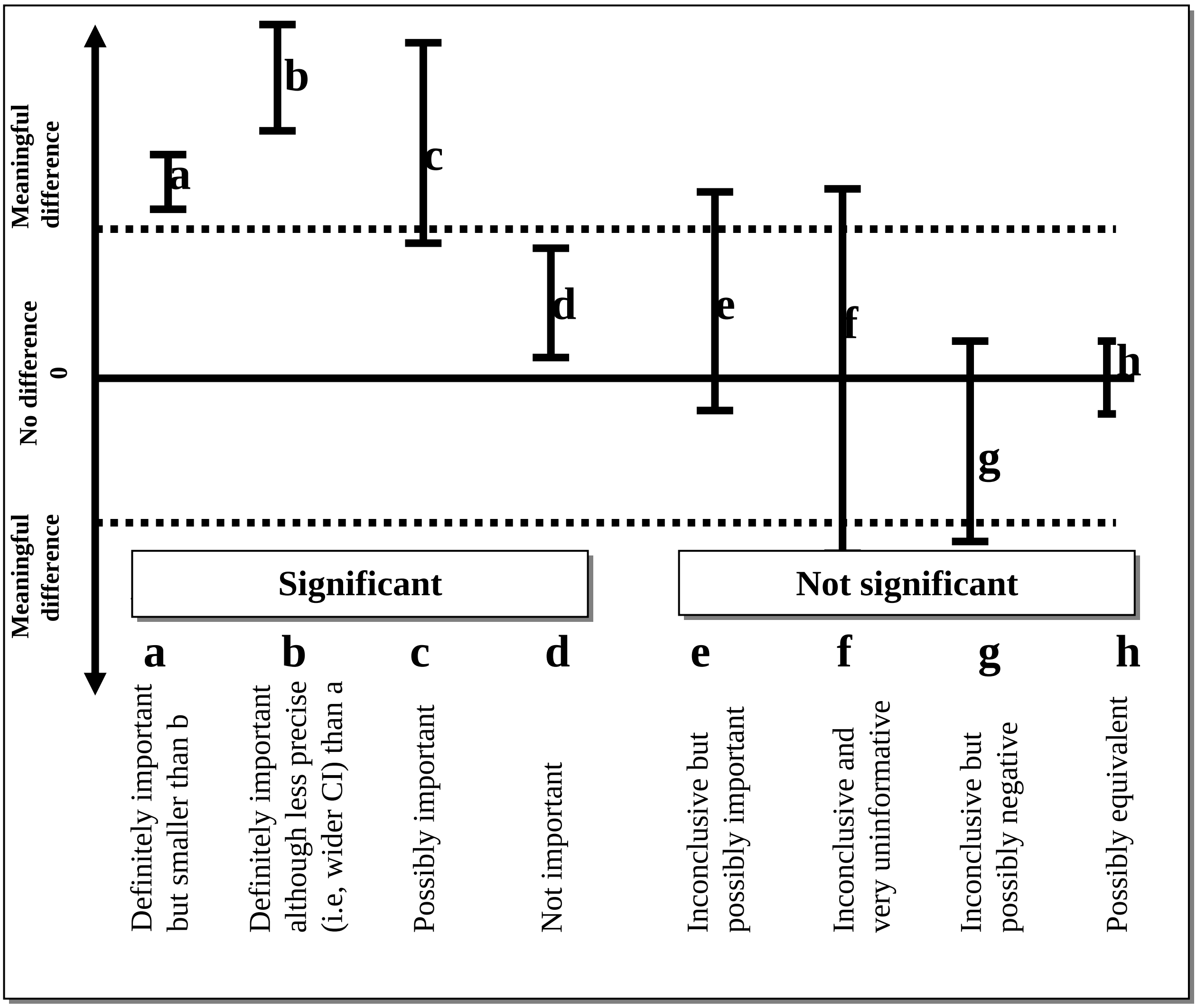
Level of significance (α)	**Level of confidence $(1-\alpha)$**	**Coefficient (Z)**
0.10	**0.90**	**1.64**
0.05	**0.95**	**1.96**
0.01	**0.99**	**2.58**

See Appendix 1 for other values. Those in ovals correspond to the values in the box shown above. Such values from other distributions, t and chi-square, are also provided.

PART 3: CONFIDENCE INTERVAL VS. P-VALUE

Confidence intervals showing eight possible interpretations in terms of statistical significance and practical importance.

(Adapted from: Armitage, P. and Berry, G. Statistical methods in medical research. 3rd edition. Blackwell Scientific Publications, Oxford. 1994.page 99)



Discuss:

PART 4: POST-TEST

A controlled trial of a new treatment led to the conclusion that it is significantly better than placebo (p-value < 0.05).

Which of the following statement do you prefer?

[ ]1. It has been proved that treatment is better than placebo.

[ ]2. If the treatment is not effective, there is less than a 5 percent chance of obtaining such results.

[ ]3. The observed effect of the treatment is so large that there is less than a 5 percent chance that the treatment is no better than placebo.

[ ]4. I do not really know what a p-value is and do not want to guess.

Source: *Wulff, H.R., Andersen, B., Brandenhoff, P., and Guttler, F. (1987) What do doctors know about statistics?. Statistics in Medicine, **6**, 3-10.*

Justifications of the answer:

Report the findings of such controlled trial in your own word (*Hints: make up any numbers you need for the report*):

ANNOTATED BLIBLIOGRAPHIES

Armitage, P., and Berry, G. (1994). *Statistical methods in medical research*. Oxford: Blackwell Scientific Publications.

Bland, M. (1998). *An introduction of medical statistics*. Oxford: Oxford University Press.

Daniel, W.W. (1991). *Biostatistics: A foundation for analysis in the health sciences*. 5th ed. New York: John Wiley & Sons.

Dawson-Sounder, B., and Trapp, R.G. (1990). *Basic and clinical biostatistics*. Norwalk, Connecticut: Appleton & Lange.

Diggle, P.J., Liang, K-Y, and Zeger, S.L. (1994). *Analysis of longitudinal data*. New York: Oxford University Press.

Everitt. B.S. (1994). *Statistical methods for medical investigations*. 3rd edition. London: Edward Arnold.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. 2nd edition. New York: John Willey & Sons.

Guyatt, G., Jaeschke, R., Heddle, N., Cook, D., Shannon, H., and Walter S. (1995). Interpreting study results: confidence intervals. *Canadian Medical Association Journal*. 152:169-173.

Hosme, D.W. Lemeshow, S. (1989). *Applied Logistic Regression*. New York: John Wiley & Sons.

Kleinbaum, D.G. (1994). *Logistic Regression: A self-learning text*. New York Springer-Verlag.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982). *Epidemiologic research: principles and qualitative methods*. London: Lifetime Learning Publications.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E, and Nizam, A. (1998). *Applied regression analysis and other multivariable methods*. Pacific Grove: Duxbury Press.

Lang, TA., Secic, M. (1997). *How to report statistics in medicine: annotated guidelines for authors*, editors, and reviewers. Philadelphia: American College of Physician.

Liang, K-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73(l):13-22.

Meinert, C.L. and Tonascia, S. (1986). *Clinical trials, design, conduct, and analysis*. New York: Oxford University Press.

Rosner, Bernard. (1995). *Fundamentals of Biostatistics*. 4th ed. Belmont, California: Duxbury Press.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

Wonnacott TH., Wonnacott RJ. (1990). *Introductory statistics.* 5th Edition. New York: John Willey & Sons .

Wulff, H.R., Andersen, B., Brandenhoff, P., and Guttler, F. (1987). What do doctors know about statistics? Statistics in Medicine, 6, 3-10.

APPENDIX 1
Statistical Tables

Normal distribution

(A) Z	(B) AREA BETWEEN MEAN AND Z	(C) AREA BEYOND Z	(A) Z	(B) AREA BETWEEN MEAN AND Z	(C) AREA BEYOND Z	(A) Z	(B) AREA BETWEEN MEAN AND Z	(C) ARED BEYOND Z
0.00	0.0000	0.5000	.035	0.1368	0.3632	0.70	0.2580	0.2420
0.01	0.0040	0.4960	0.36	0.1406	0.3594	0.71	0.2611	0.2389
0.02	0.0080	0.4920	0.37	0.1443	0.3557	0.72	0.2642	0.2358
0.03	0.0120	0.4880	0.38	0.1480	0.3520	0.73	0.2673	0.2327
0.04	0.0160	0.4840	0.39	0.1517	0.3483	0.74	0.2704	0.2296
0.05	0.0199	0.4801	0.40	0.1554	0.3446	0.75	0.2734	0.2266
0.06	0.0239	0.4761	0.41	0.1591	0.3409	0.76	0.2764	0.2236
0.07	0.0279	0.4721	0.42	0.1628	0.3372	0.77	0.2794	0.2206
0.08	0.0319	0.4681	0.43	0.1664	0.3336	0.78	0.2823	0.2177
0.09	0.0359	0.4641	0.44	0.1700	0.3300	0.79	0.2852	0.2148
0.10	0.0398	0.4602	0.45	0.1736	0.3264	0.80	0.2881	0.2119
0.11	0.0438	0.4562	0.46	0.1772	0.3228	0.81	0.2910	0.2090
0.12	0.0478	0.4522	0.47	0.1808	0.3192	0.82	0.2939	0.2061
0.13	0.0517	0.4483	0.48	0.1844	0.3156	0.83	0.2967	0.2033
0.14	0.0557	0.4443	0.49	0.1879	0.3121	0.84	0.2995	0.2005
0.15	0.0596	0.4404	0.50	0.1915	0.3085	0.85	0.3023	0.1977
0.16	0.0636	0.4364	0.51	0.1950	0.3050	0.86	0.3051	0.1949
0.17	0.0675	0.4325	0.52	0.1985	0.3015	0.87	0.3078	0.1922
0.18	0.0714	0.4286	0.53	0.2019	0.2981	0.88	0.3106	0.1894
0.19	0.0753	0.4247	0.54	0.2054	0.2946	0.89	0.3133	0.1867
0.20	0.0793	0.4207	0.55	0.2088	0.2912	0.90	0.3159	0.1841
0.21	0.0832	0.4168	0.56	0.2123	0.2877	0.91	0.3186	0.1814
0.22	0.0871	0.4129	0.57	0.2157	0.2843	0.92	0.3212	0.1788
0.23	0.0910	0.4090	0.58	0.2190	0.2810	0.93	0.3238	0.1762
0.24	0.0948	0.4052	0.59	0.2224	0.2776	0.94	0.3264	0.1736
0.25	0.0987	0.4013	0.60	0.2257	0.2743	0.95	0.3289	0.1711
0.26	0.1026	0.3974	0.61	0.2291	0.2709	0.96	0.3315	0.1685
0.27	0.1064	0.3936	0.62	0.2324	0.2676	0.97	0.3340	0.1660
0.28	0.1103	0.3897	0.63	0.2357	0.2643	0.98	0.3365	0.1635
0.29	0.1141	0.3859	0.64	0.2389	0.2611	0.99	0.3389	0.1611
0.30	0.1179	0.3821	0.65	0.2422	0.2578	1.00	0.3413	0.1587
0.31	0.1217	0.3783	0.66	0.2454	0.2546	1.01	0.3438	0.1562
0.32	0.1255	0.3745	0.67	0.2486	0.2514	1.02	0.3461	0.1539
0.33	0.1293	0.3707	0.68	0.2517	0.2483	1.03	0.3485	0.1515
0.34	0.1331	0.3669	0.69	0.2549	0.2451	1.04	0.3508	0.1492
0.35	0.1368	0.3632	0.70	0.2580	0.2420	1.05	0.3531	0.1469
1.05	0.3531	0.1469	1.40	0.4192	0.0808	1.75	0.4599	0.0401
1.06	0.3554	0.1446	1.41	0.4207	0.0793	1.76	0.4608	0.0392

(A) Z	(B) AREA BETWEEN MEAN AND Z	(C) AREA BEYOND Z	(A) Z	(B) AREA BETWEEN MEAN AND Z	(C) AREA BEYOND Z	(A) Z	(B) AREA BETWEEN MEAN AND Z	(C) ARED BEYOND Z
1.07	0.3577	0.1423	1.42	0.4222	0.0778	1.77	0.4616	0.0384
1.08	0.3599	0.1401	1.43	0.4236	0.0764	1.78	0.4625	0.0375
1.09	0.3621	0.1379	1.44	0.4251	0.0749	1.79	0.4633	0.0367
1.10	0.3643	0.1357	1.45	0.4265	0.0735	1.80	0.4641	0.0359
1.11	0.3665	0.1335	1.46	0.4279	0.0721	1.81	0.4649	0.0351
1.12	0.3686	0.1314	1.47	0.4292	0.0708	1.82	0.4656	0.0344
1.13	0.3708	0.1292	1.48	0.4306	0.0694	1.83	0.4664	0.0336
1.14	0.3729	0.1271	1.49	0.4319	0.0681	1.84	0.4671	0.0329
1.15	0.3749	0.1251	1.50	0.4332	0.0668	1.85	0.4678	0.0322
1.16	0.3770	0.1230	1.51	0.4345	0.0655	1.86	0.4686	0.0314
1.17	0.3790	0.1210	1.52	0.4357	0.0643	1.87	0.4693	0.0307
1.18	0.3810	0.1190	1.53	0.4370	0.0630	1.88	0.4699	0.0301
1.19	0.3830	0.1170	1.54	0.4382	0.0618	1.89	0.4706	0.0294
1.20	0.3849	0.1151	1.55	0.4394	0.0606	1.90	0.4713	0.0287
1.21	0.3869	0.1131	1.56	0.4406	0.0594	1.91	0.4719	0.0281
1.22	0.3888	0.1112	1.57	0.4418	0.0582	1.92	0.4726	0.0274
1.23	0.3907	0.1093	1.58	0.4429	0.0571	1.93	0.4732	0.0268
1.24	0.3925	0.1075	1.59	0.4441	0.0559	1.94	0.4738	0.0262
1.25	0.3944	0.1056	1.60	0.4452	0.0548	1.95	0.4744	0.0256
1.26	0.3962	0.1038	1.61	0.4463	0.0537	1.96	0.4750	0.0250
1.27	0.3980	0.1020	1.62	0.4474	0.0526	1.97	0.4756	0.0244
1.28	0.3997	0.1003	1.63	0.4484	0.0516	1.98	0.4761	0.0239
1.29	0.4015	0.0985	1.64	0.4495	0.0505	1.99	0.4767	0.0233
1.30	0.4032	0.0968	1.65	0.4505	0.0495	2.00	0.4772	0.0228
1.31	0.4049	0.0951	1.66	0.4515	0.0485	2.01	0.4778	0.0222
1.32	0.4066	0.0934	1.67	0.4525	0.0475	2.02	0.4783	0.0217
1.33	0.4082	0.0918	1.68	0.4535	0.0465	2.03	0.4788	0.0212
1.34	0.4099	0.0901	1.69	0.4545	0.0455	2.04	0.4793	0.0207
1.35	0.4115	0.0885	1.70	0.4554	0.0446	2.05	0.4798	0.0202
1.36	0.4131	0.0869	1.71	0.4564	0.0436	2.06	0.4803	0.0197
1.37	0.4147	0.0853	1.72	0.4573	0.0427	2.07	0.4808	0.0192
1.38	0.4162	0.0838	1.73	0.4582	0.0418	2.08	0.4812	0.0188
1.39	0.4177	0.0823	1.74	0.4591	0.0409	2.09	0.4817	0.0183
1.40	0.4192	0.0808	1.75	0.4599	0.0401	2.10	0.4821	0.0179
2.10	0.4821	0.0179	2.40	0.4918	0.0082	2.90	0.4981	0.0019
2.11	0.4826	0.0174	2.41	0.4920	0.0080	2.92	0.4982	0.0018
2.12	0.4830	0.0170	2.42	0.4922	0.0078	2.94	0.4984	0.0016
2.13	0.4834	0.0166	2.43	0.4925	0.0075	2.96	0.4985	0.0015
2.14	0.4838	0.0162	2.44	0.4927	0.0073	2.98	0.4986	0.0014
2.15	0.4842	0.0158	2.45	0.4929	0.0071	3.00	0.4987	0.0013
2.16	0.4846	0.0154	2.46	0.4931	0.0069	3.02	0.4987	0.0013
2.17	0.4850	0.0150	2.47	0.4932	0.0068	3.04	0.4988	0.0012
2.18	0.4854	0.0146	2.48	0.4934	0.0066	3.06	0.4989	0.0011
2.19	0.4857	0.0143	2.49	0.4936	0.0064	3.08	0.4990	0.0010
2.20	0.4861	0.0139	2.50	0.4938	0.0062	3.10	0.4990	0.0010
2.21	0.4864	0.0136	2.52	0.4941	0.0059	3.12	0.4991	0.0009
2.22	0.4868	0.0132	2.54	0.4945	0.0055	3.14	0.4992	0.0008

(A) Z	(B) AREA BETWEEN MEAN AND Z	(C) AREA BEYOND Z	(A) Z	(B) AREA BETWEEN MEAN AND Z	(C) AREA BEYOND Z	(A) Z	(B) AREA BETWEEN MEAN AND Z	(C) ARED BEYOND Z
2.23	0.4871	0.0129	2.56	0.4948	0.0052	3.16	0.4992	0.0008
2.24	0.4875	0.0125	2.58	0.4951	0.0049	3.20	0.4993	0.0007
2.25	0.4878	0.0122	2.60	0.4953	0.0047	3.25	0.4994	0.0006
2.26	0.4881	0.0119	2.62	0.4956	0.0044	3.30	0.4995	0.0005
2.27	0.4884	0.0116	2.64	0.4959	0.0041	3.35	0.4996	0.0004
2.28	0.4887	0.0113	2.66	0.4961	0.0039	3.40	0.4997	0.0003
2.29	0.4890	0.0110	2.68	0.4963	0.0037	3.45	0.4997	0.0003
2.30	0.4893	0.0107	2.70	0.4965	0.0035	3.50	0.4998	0.0002
2.31	0.4896	0.0104	2.72	0.4967	0.0033	3.55	0.4998	0.0002
2.32	0.4898	0.0102	2.74	0.4969	0.0031	3.60	0.4998	0.0002
2.33	0.4901	0.0099	2.76	0.4971	0.0029	3.65	0.4999	0.0001
2.34	0.4904	0.0096	2.78	0.4973	0.0027	3.70	0.4999	0.0001
2.35	0.4906	0.0094	2.80	0.4974	0.0026	3.75	0.4999	0.0001
2.36	0.4909	0.0091	2.82	0.4976	0.0024	3.80	0.4999	0.0001
2.37	0.4911	0.0089	2.84	0.4977	0.0023	3.85	0.4999	0.0001
2.38	0.4913	0.0087	2.86	0.4979	0.0021	3.90	0.4999	0.00005
2.39	0.4916	0.0084	2.88	0.4980	0.0020	3.95	0.4999	0.00004
2.40	0.4918	0.0082	2.90	0.4981	0.0019	4.00	0.4999	0.00003

t-Distribution

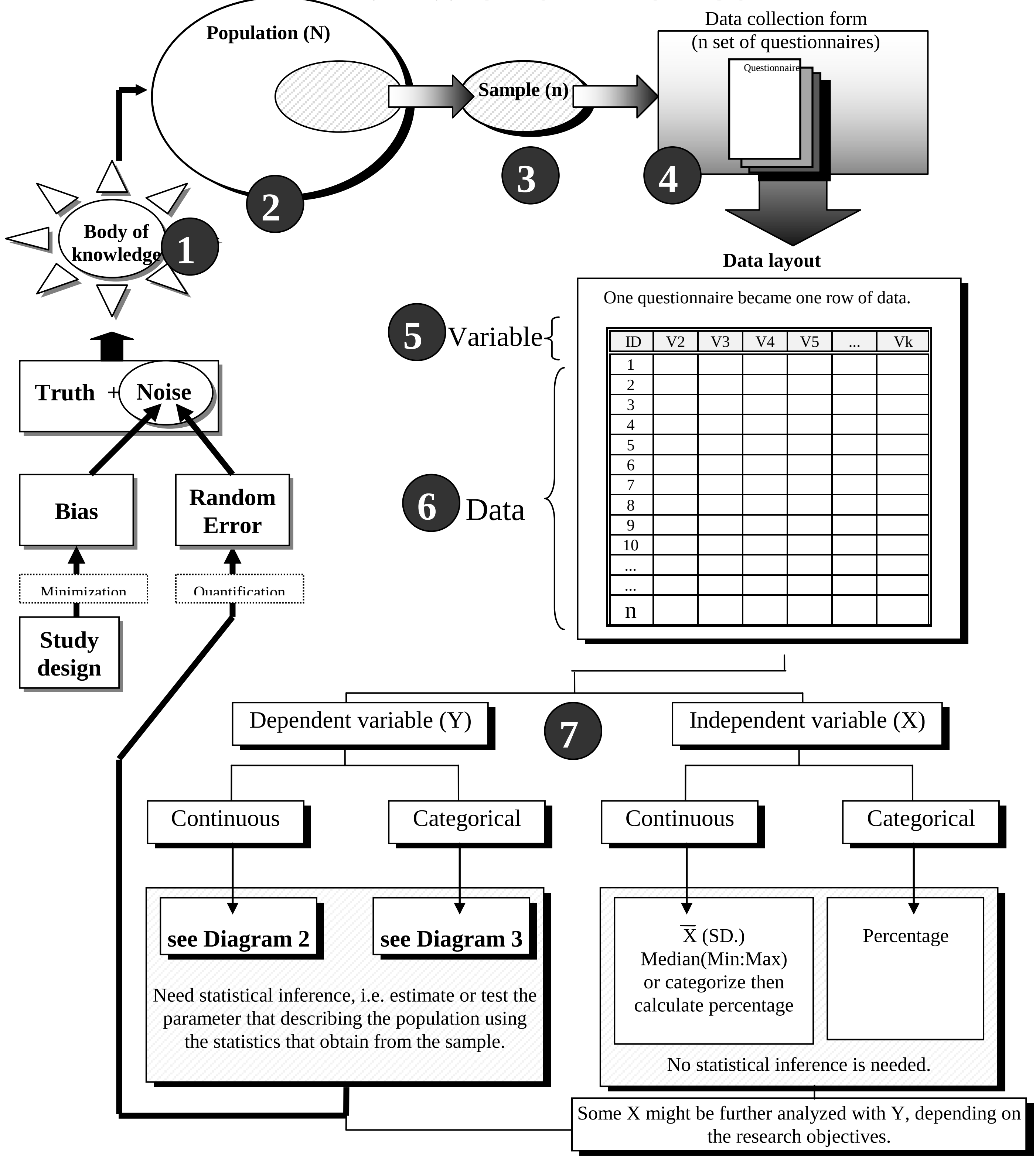
	Level of Significance					
DF	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
1	3.078	6.314	12.706	31.821	63.657	63.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.992
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674

DF	Level of Significance 0.10	0.05	0.025	0.01	0.005	0.0005
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Chi-square(χ^2) Distribution

DF	Probability (α) 0.995	0.990	0.975	0.950	0.900	0.500	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	0.45	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	1.39	4.61	5.99	7.38	19.21	10.60
3	0.07	0.11	0.22	0.35	0.58	2.37	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.38	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.50	13.24	20.38	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.80	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

APPENDIX 1
Summary
Diagram 1
OVERVIEW OF STATISTICS



1 is the body of knowledge which is the goal of all research. We can take the Pre-test as an example. What you have been trying to answer the pre-test question is really that you are trying to draw a conclusion from the study. It's the knowledge gain from reading the paper.

A controlled trial of a new treatment led to the conclusion that it is significantly better than placebo (p-value < 0.05).

Which of the following statement do you prefer?

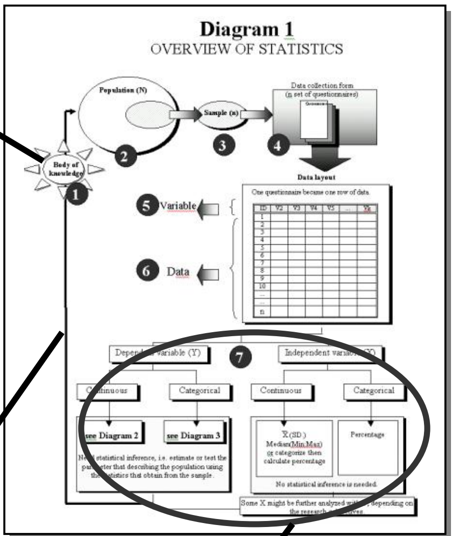
[....]1. It has been proved that treatment is better than placebo.

[....]2. If the treatment is not effective, there is less than a 5 percent chance of obtaining such results.

[....]3. The observed effect of the treatment is so large that there is less than a 5 percent chance that the treatment is no better than placebo.

[....]4. I do not really know what a p-value is and do not want to guess.

Source: *Wulff, H.R., Andersen, B., Brandenhoff, P., and Guttler, F. (1987) What do doctors know about statistics?. Statistics in Medicine,* ***6****, 3-10.*



Statistical inference:
Describe characteristics of the population
(i) Estimation
Commonly reported as 95% Confidence intervals
(ii) Hypothesis testing
Commonly reported as p-value
(or significant / not significant)

Descriptive statistics:
Describe characteristics of the sample.

2 and **3** are about the population to which the findings from the study will be applied.

4 to **7** are about data collection. Then data processing which (always) use the computer followed by data analysis.

Diagram 2

ANALYSIS OF CONTINUOUS OUTCOME

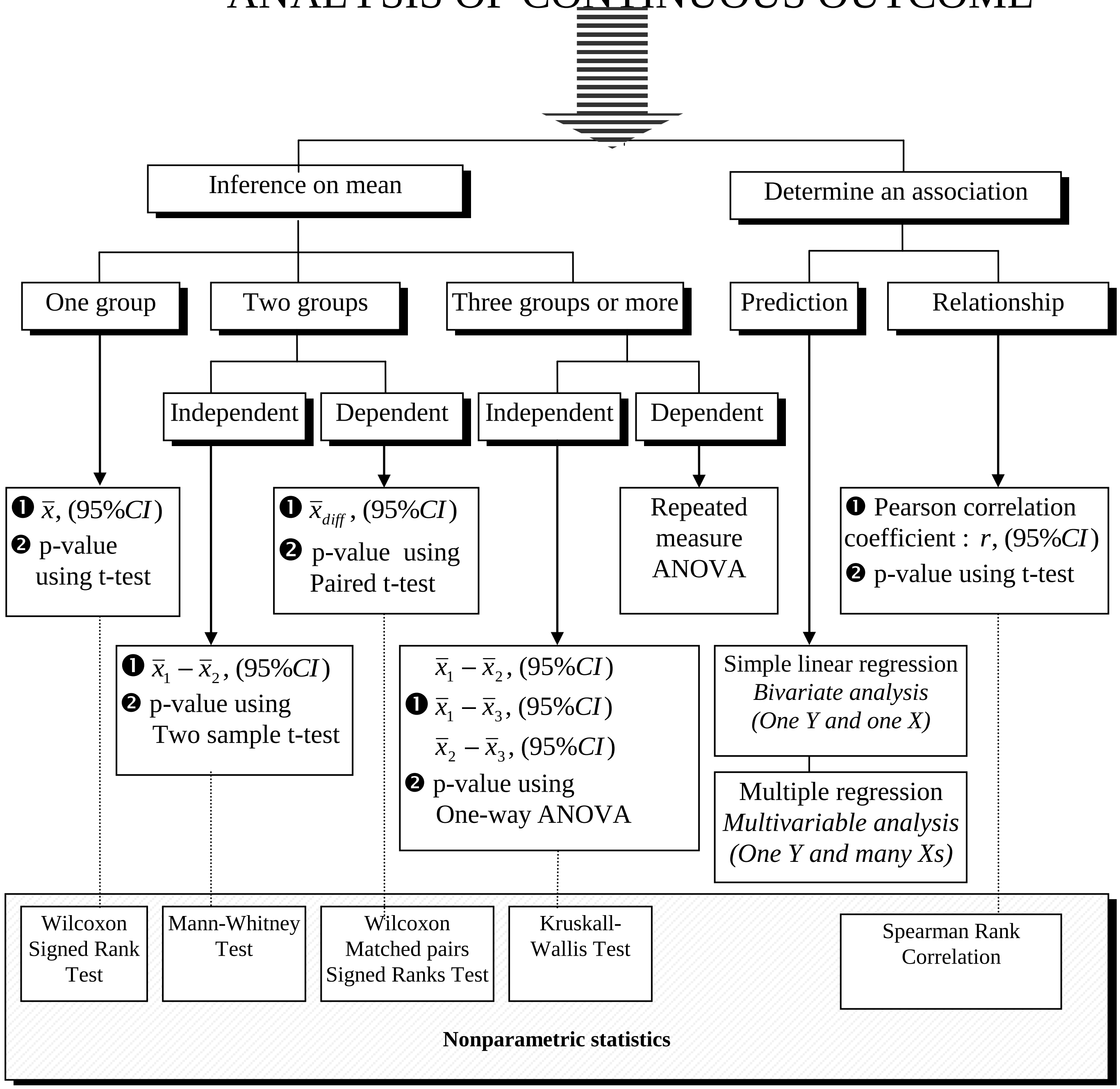


Diagram 3

ANALYSIS OF CATEGORICAL OUTCOME