INTRODUCTION TO STATISTICS IN RESEARCH

Prof Kehinde Oluwadiya

www.Oluwadiya.com

Objectives

Participants gets the basic requirement to adequately determine the statistics to use for their studies

- Raise our awareness of the need to raise the quality of statistics used in researches
- Functions of statistics: Why you need it!
- Data collection methods
- Sample size determination
- Basics of randomization
- Randomization techniques
- Appropriate use of statistical methods
- Useful packages
- Other useful tools in research

Data Collection

Fundamentally--2 types of data

- **Quantitative** – Numbers, tests, counting, measuring

- **Qualitative** – Words, images, observations, conversations, photographs

Data Collection Techniques

- Observations
- Tests
- Surveys
- Document analysis

Data collection choice

- What you must ask yourself:
  - Will the data answer my research question?

Data collection choice

- To answer that
  - You much first decide what your research question is
  - Then you need to decide what data/variables are needed to scientifically answer the question

Data collection QUESTIONS

- Where then does ***data*** come from?
- How is it gathered?
- How do we ensure it is accurate?
- Is the data reliable?
- Is it representative of the population from which it was drawn?

Methods of Collecting Data

- There are many methods used to collect or obtain data for statistical analysis.
- Three of the most popular methods are:
  - Direct Observation
  - Experiments
  - Surveys

Surveys

- A ***survey*** solicits information from people; e.g. usage of health facilities, eating habits etc. using an array of instruments.
- The ***Response Rate*** (i.e. the proportion of all people selected who complete the survey) is a key survey parameter.
- Surveys may be administered in a variety of ways, e.g.
  - Personal Interview
  - Telephone Interview
  - Self Administered Questionnaire
  - Schedules and;
  - Internet

Sources of questionnaire

- **Pre-existing questionnaire**:
  - Must be adapted to the study population (**Validation**)
- Newly Developed:
  - Must be
    - Valid,
    - Reliable and
    - Appropriate.

Questionnaire Design…

Over the years, a lot of thought has been put into the science of the design of survey questions. Key design principles:

1. Keep the questionnaire as short as possible.
2. Ask short, simple, and clearly worded questions.
3. Start with demographic questions to help respondents get started comfortably.
4. Use dichotomous (yes/no) and multiple choice questions.
5. Use open-ended questions cautiously.
6. Avoid using leading-questions.
7. Pretest a questionnaire on a small number of people.
8. Think about the way you intend to use the collected data when preparing the questionnaire.

Merits of using questionnaire

- Relatively low cost even when the population is large and is widely spread geographically.
- *It is free from the bias of the interviewer; answers are in respondents' own words.
- *Respondents have adequate time to give well thought out answers.
- *Respondents, who are not easily approachable, can also be reached conveniently.
- Large samples can be made use of and thus the results can be made more dependable and reliable.

***Unless when interviewer administered**

Demerits of using questionnaire

- Low rate of return of the duly filled in questionnaires; bias due to no-response is often indeterminate.
- *It can be used only when respondents are educated and cooperating
- *The control over questionnaire may be lost once it is sent.
- *There is inbuilt inflexibility because of the difficulty of amending the approach once questionnaires have been dispatched
- *There is also the possibility of ambiguous replies or omission of replies altogether to certain questions; interpretation of omissions is difficult.
- It is difficult to know whether willing respondents are truly representative.
- This method can be very slow.

***Unless when interviewer administered**

Electronic data collection using smartphones & Tablets

- They’re portable
- Come with an on-board GPS receiver (helps to guide against fudging)
- Have on board cameras
- Automatically record time taken to enter data ((helps to guide against fudging)
- No need to enter data separately after collection
- Can connect to wireless networks
- Access to the internet
- Email is available
- There’s an app for that!

Electronic data collection using smartphones

Different (Free) Apps:

- Open Data Kit (ODK) (https://www.google.com.au/intl/en/earth/outreach/tutorials/odk_gettingstarted.html)
- Epicollect (http://www.epicollect.net/instructions/)
- Epicollect+ (http://www.epicollect.net/plus_Instructions/default.html)

What about online survey tools

- All the advantages of questionnaire plus the convenience of online tools
- Free versions available e.g. Google Forms, Survey Monkey
- Reach and scalability is much more than traditional questionnaire
- Cheap
- Less time consuming for the researcher
- More accurate data entering
- Quicker
- Ensures better anonymity and therefore improves confidentiality

Lies, damn lies and statistics

Beware, statistics can be unreliable!

Errors galore

There is an increasing number of publications on the flaws and errors in much of published medical literature:

- The scandal of poor medical research- **DG Altman, 1994**
- Statistical errors in medical research, a chronic disease?- **J Young, 2007**
- Improved reporting of statistical design and analysis: guidelines, education, and editorial policies. **Mazumdar M et al 2010**
- Bay Area Research symposium 2010- “Why most published research findings are false”- **John Iaonnidis**

And many more…

Medical Lies?

PLoS Medicine | www.plosmedicine.org 0696 August 2005 | Volume 2 | Issue 8 | e124

Open access, freely available online

Essay

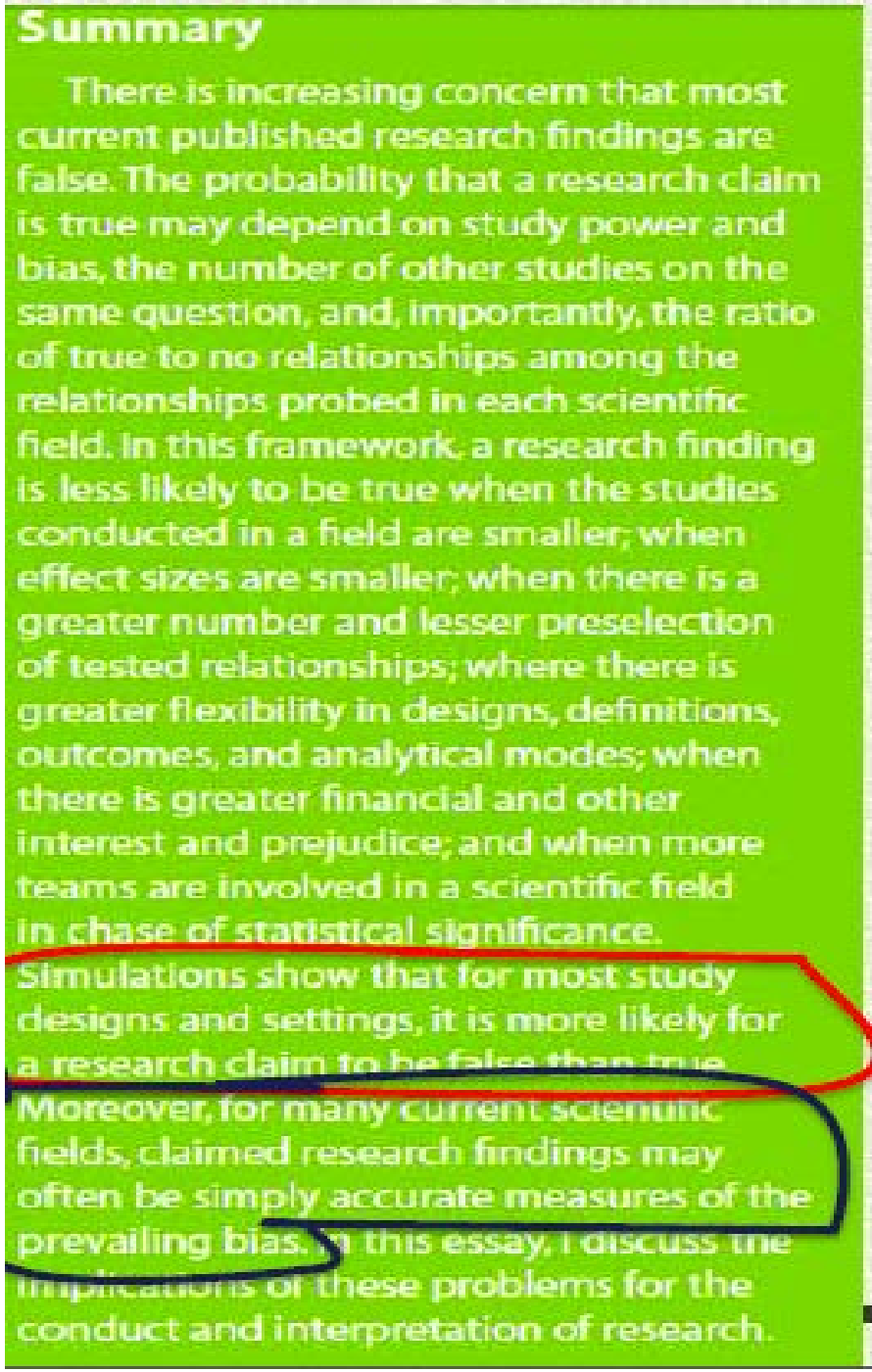
Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

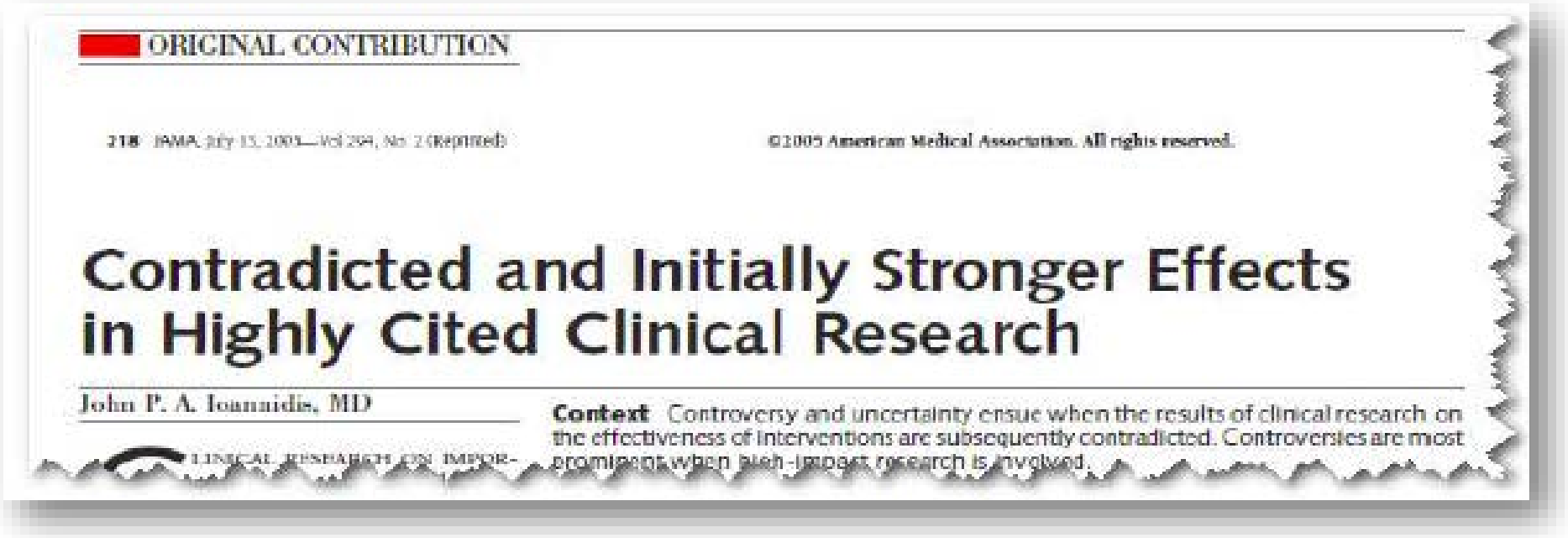
factors that influence this problem and is characteristic of the field and can

Why most published research findings are false



Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.



ORIGINAL CONTRIBUTION

218 JAMA, July 13, 2005—Vol 294, No. 2 (Reprinted)

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

CLINICAL RESEARCH ON IMPOR-

Hall of shame: % Contradicted by later studies

- 80% of non-randomized studies were wrong
- 25% of supposedly gold-standard randomized trials
- 10% of large randomized trials

Excerpts from DG Altman……

- *"We need less research, better research, and research done for the right reasons"*
- *"We need not be experts in statistics, but we should understand the principles of sound methods of research. If we can also analyze their own data, so much the better.* ***Amazingly, it is widely considered acceptable for (medical) researchers to be ignorant of statistics. Many are not ashamed (and some seem proud) to admit that they don't know anything about statistics"***

The scandal of poor medical research- DG Altman, 1994

As scientists and medical researchers, it is our sacred duty to uphold the standards of research in our specialties

To do this, we should empower ourselves, understand the underlying principles, and be ready to stand by them…

Functions of statistics…..1

1. To reduce data. This is done:

I. Graphically by compiling charts, tables, graphs, histograms, frequency polygons etc.,

II. Univariate analysis (mean, median, standard deviations, range etc)

- Aim is to determine trends, or an average of variables.

Statistics is a tool for converting ***DATA*** into ***INFORMATION***

Functions of statistics….2

2. To provide methods of applying tests of significance.

- Tests of significance are used to separate real differences from those due to chance. In general, the level of significance is arbitrarily set at 5% (p = probability = 0.05).

Functions of statistics…..3

3. To provide a sound basis for experimental design

- Experiments must be carefully designed because a good design may mean the difference between a sound, scientific research and worthless data which yield little or no information.
- In many instances, more information are obtained with the same amount of work if the researcher has a knowledge of statistical methods, and plans his experiment accordingly.

Determining the statistical method to use

How to chose the right statistic

1 . Sampling methodology

- It is important that the characteristics of the sample be a true indication of the characteristics of the population to which he wants to generalized
- For the sample to be representative of the population it must be obtained in such a way that no bias enters into its selection
- For greatest likelihood of this being the case every individual in the population must have the same chance of being selected for the sample.

Most hospital based studies are blighted by selection bias

Assessing the sample size of the proposed study

Why is it important to determine the sample size in a study?

- To determine the number of participants needed to detect a clinically relevant treatment effect.
- Too small a sample size may lead to failure to detect an important effect.
- Too large a sample size will waste time and resources. It is also likely to identify trivial effects.
- It is important to use the appropriate sample size estimation technique for your study design.

Assessing the sample size of the proposed study

THERE IS A QUESTION OF ETHICS TOO…..

In an experiment involving human or animal subjects, sample size is an important ethical consideration:

1. An undersized experiment exposes the subjects to potentially harmful treatments without advancing knowledge.
2. In an oversized experiment, an unnecessary number of subjects are exposed to a potentially harmful treatment.

Assessing the sample size of the proposed study

- What are the components of sample size estimation:
  - Type I error (alpha)
  - Power (1- beta): i.e. the probability of not making a type II error
  - The smallest effect of interest (effect size)
  - Inherent variability of population (usually estimated from preliminary data or previous study)
- The parameters of the sample size estimation must be stated in the paper.
- If necessary, consult a statistician

Assessing the sample size of the proposed study

An Hypothetical Case

Here are the objectives of a proposed study:

1. To estimate the proportion of patients with fracture neck of the femur who will develop AVN after internal fixation
2. To estimate the mean time to fracture union in the three types of fracture neck of femur and whether these were significantly different from each other
3. To determine the risk factors to AVN in the patients.

What sample size should be used?

Assessing the sample size of the proposed study

An Hypothetical Case

It is important for us to know that each objective requires different sampling technique.

1. To estimate the proportion of patients with fracture neck of the femur who will develop AVN after internal fixation → SS formula for proportion is used. Estimated sample size is 72
2. To estimate the mean time to fracture union in the three types of fracture neck of femur and whether this were significantly different from each other → SS formula for mean is used. Estimated sample size is 35
3. To determine the risk factors to AVN in the patients → SS formula for regression is used. Estimated sample size is 105

The largest estimated sample size should be used.
In this case, this is 105

Assessing the sample size of the proposed study

In summary:

- Sample sizes are calculated for the various statistics used in the study.
- The largest calculated sample size should be used for the study
- Allowances should be made for attrition.

2. Randomization

Randomization gives each participant an equal chance of being assigned to any treatment group (Intervention versus Control).

Successful randomization requires that assignment to groups cannot be predicted in advance.

Why Randomized?

- If, at the end of a clinical trial, a difference in outcomes occurs between two treatment groups, possible explanations for this difference would include:
  i. the treatment exhibits a real effect;
  ii. the difference in outcome between the two groups is solely due to chance
  iii. there is a systematic difference (or bias) between the groups due to factors other than the intervention.

Randomization aims to remove the third possibility.

Types of randomization

- Simple Randomization
- Permuted Block Randomization
- Stratified Block Randomization
- Dynamic (adaptive) random allocation

Inappropriate randomization methods

- Assigning patients alternately to treatment group
- Assigning the first half of the population to one group
- Assignments by methods based on patient characteristics such as date of birth, order of entry into the clinic, day of clinic attendance, hospital numbers etc are not reliably random

What about when randomization is not possible?

1. Controlled Pre & Post-test Quasi-experimental design (Groups are similar on all known and unknown factors except for exposure to the event)
2. Regression Techniques (There several regression that can adjust for differences in baseline patient characteristics included in different groups

Sampling Vs Randomization

- Random Samples and Randomization are two different things, but they have something in common as the presence of random in both names suggests — both involve the use of a probability device.
- With *random samples*, chance determines who will be in the sample.
- With *randomization (also called* random assignment] chance determines the assignment of into treatment groups.

3 Appropriate use of statistical methods:

An Example from an actual research proposal:

Data will be analyzed using SPSS 15. Univariate analyses will be conducted for the association of outcome variables with other characteristics. Chi-square test for significance will be used for categorical variables, t-test for parametric and Wilcoxon rank-sum test for non-parametric variables. Multivariate analyses using either linear or logistic regression or correlations will be conducted for characteristics like age, experience and alcohol/herbal mixture use.

The level of significance will be 0.05 while 95% confidence intervals of estimates of the outcome variables will be calculated.

Tests of association are not univariate analysis

Wicoxon Rank Test?
Yes, it is a parametric technique, but it is used for paired samples!

???????

Regression are not conducted for i.ndividual variables. They are used for building models to determine causes and effects

Appropriate use of statistical methods

Another Example from an actual research proposal:

- Data that will be generated from this study will be analyzed using SPSS. Statistical technique that will be employed will include descriptive statistics such as frequency, percentages, mean, mode etc., inferential statistics such as chi square and correlation coefficient will be used to test the hypothesis.

How to choose the appropriate test?

This is based on several factors

Determining the Test (I)

What kind of variables are they?

1. Numerical variable: Ratio or Interval
2. Categorical: Ordinal or Nominal

DATA IS:	Numeric	Categorical
METHOD TO USE:	Parametric	Nonparametric

Determining the Test (II)

What is the distribution of the data?

I. For normally distributed data, use parametric techniques
II. For non-normally distributed data, use nonparametric techniques

Parametric	Nonparametric
T-test	Mann-Whitney U test
Paired t-test	Wilcoxon Rank test
ANOVA	Kruskal-Wallis test
Pearson Correlation	Spearman's correlation
Linear/Multiple Regression	Logistic Regression

Generally Speaking….

Parametric tests are used for…

- Continuous data; normally distributed
- Other assumptions may need to be met

Nonparametric tests are used for….

- Continuous data; not normally distributed
- Categorical or Ordinal data

Determining the Test (Iii)

How many groups are in the dependent (outcome variable)?

I. Two groups or categories
II. More than two groups or categories
III. Continuous data

Two Groups	More than two groups	Continuous
T-Test	ANOVA	NA
Mann Whitney U	Kruskal Wallis Test	NA
Logistic Regression	Multinomial Regression	NA
NA	NA	Multiple/Linear Regression/Correlation

Remember also that if **ANOVA** is done, then a **POSTHOC** test must also be done to determine the groups that actually differ from each other.

Determining the Test (Iv)

Are measurements taken from the same patient more than one time (before and after treatment)?

- In such situations you should use
  - Paired *t*-test
  - Repeated-measures ANOVA
  - Wilcoxon Rank Test
  - Kruskas Wallis Test

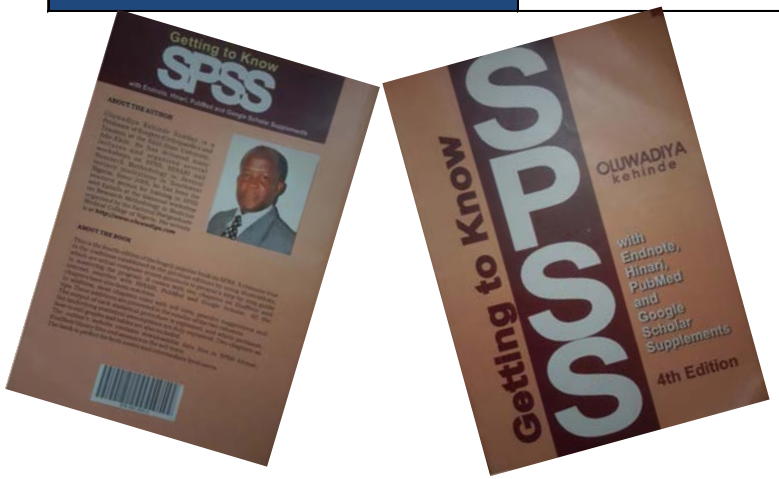
Determining the Test (IV)

When time to an event is the item of interest, then you must consider individuals who will not experience the event during the study (Censored cases).

- For these data; do survival analysis

Selecting the appropriate procedure among the common statistical procedures: A summary

Independent Variable		Dependent Variable	
		Categorical	**Continuous**
	Categorical	**Chi Square** **Logistic regression**	**t-tests** **One-way ANOVA**
	Continuous	**Logistic regression**	**Correlation** **Linear regression**



For a more complete table, please see the book:
Getting to know SPSS. Fourth Edition
by **Oluwadiya Kehinde**

Some common statistic Software

- **SPSS**……Most commonly used statistic software in reported literature. Easy to use
- **EPIINFO**….Free, open source and GUI interphase
- **STATA**….Popular among statistician. GUI but programming language required. Steep learning curve.
- **R**….. Free, open source. Programming language required. Steep learning curve.
- **SAS**…. Popular among statistician. Programming language required. Steep learning curve.
- **MedCalc**…Easy to use. Especially useful for ROC curve determination
- **NISS**…..Useful for sample size determination
- **GRAPHPAD**….Has an easy to use tutorial section. Also has web-based components
- …….and many many more

Some other useful software for researchers

- **Reference Managers** e.g. EndNote, Mendeley, Zoltero etc.
  - If you have not been using these, then you have not been fair to yourself
- **Cloud Storage** e.g. DropBox, Microsoft One Drive, Google Drive etc.
  - Use any of these and you'll never lose your data again.
- **Word processors** e.g. Microsoft Words, Google Docs, Open Office etc.
  - If you've never used any of this then, you are a dinosaur!
- **Archiving and note taking software** e.g. Microsoft OneNote, Evernote etc.
  - These are the catch-all software. They store anything you put in them.
- **Literature Search** e.g. PubMed, HINARI, Google Scholar
  - Is there anyone here who has not use any of this? HINARI is a gem: every medical research should be an expert at using it.

Final thoughts…….

- When reading a research proposal/dissertation …….

Keep In Mind That….

- No study is perfect
- All data is dirty in some ways or the other; research is what you do with that dirty data
- Measurement involves making choices

Be Critical About Numbers

- Every statistic is a way of summarizing complex information into relatively simple numbers.
- How did the researchers arrive at these numbers?
- Who produced the numbers and what is their bias?
- How were key terms be defined & in how many different ways?

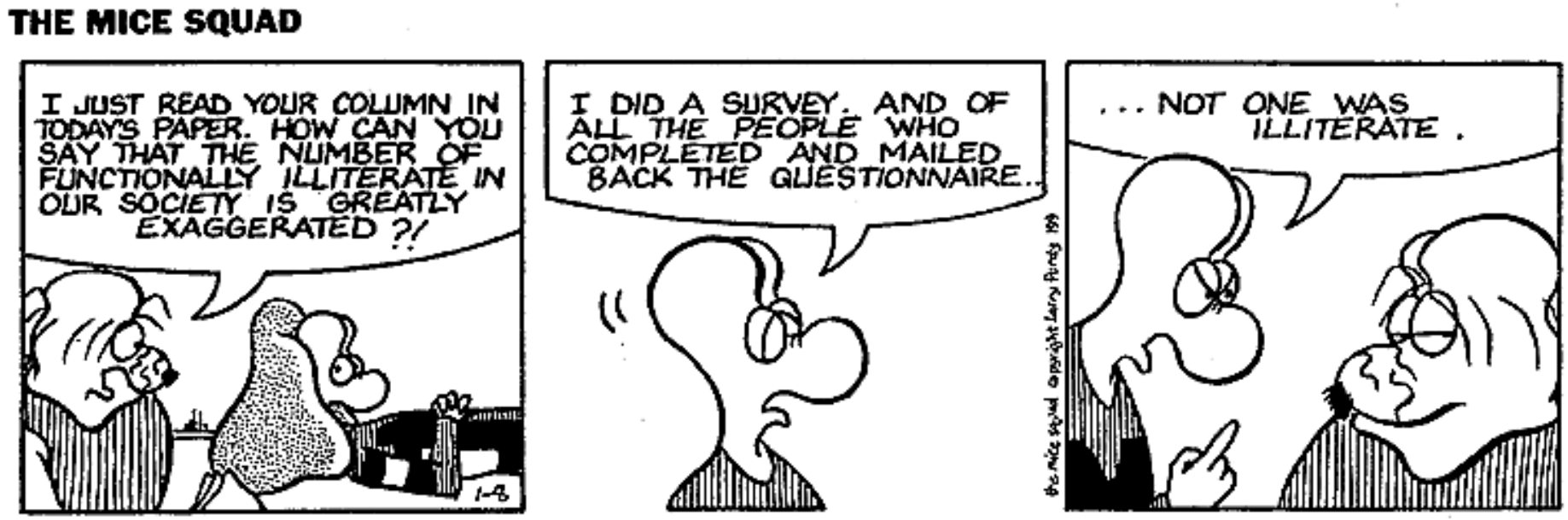
Be Critical About Numbers

- How was the choice for the measurement made?
- What type of sample was gathered & how does that affect result?
- Is the statistical result interpreted correctly?
- If comparisons are made, were they appropriate?
- Are there competing statistics?

Be Critical About Numbers

With one foot in a bucket of ice water, and one foot in a bucket of boiling water, you are, on the average, comfortable.

Be critical about numbers: Bias and Error



Thank you

To ask questions, please join the forum at www.oluwadiya.com