

## Post Doc Development Hub Workshop

### Statistics for Researchers

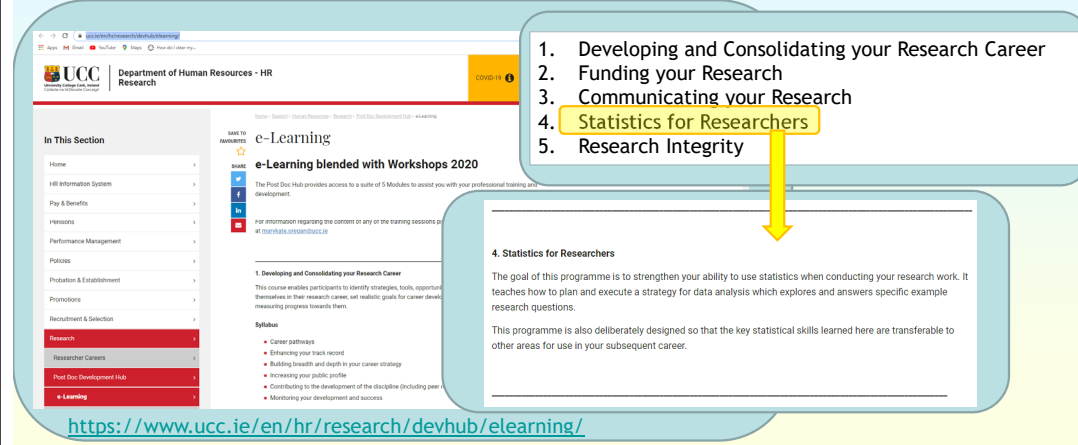
Kathleen O'Sullivan

Department of Statistics

School of Mathematical Sciences

Friday September 10, 2021 11:00-12:00

## Statistics for Researchers



1. Developing and Consolidating your Research Career
2. Funding your Research
3. Communicating your Research
4. **Statistics for Researchers**
5. Research Integrity

**4. Statistics for Researchers**

The goal of this programme is to strengthen your ability to use statistics when conducting your research work. It teaches how to plan and execute a strategy for data analysis which explores and answers specific example research questions.

This programme is also deliberately designed so that the key statistical skills learned here are transferable to other areas for use in your subsequent career.

<https://www.ucc.ie/en/hr/research/devhub/elearning/>

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

## Epigeum On-Line Training with Blended Workshops

Suit of **7** Modules

1. Getting Started
2. Thinking Statistically Describing Data Well
3. Thinking Statistically Making Good Generalisations
4. Which Hypothesis Test Should I Use
5. Statistical Modelling
6. Analysis of Categorical Data
7. Conclusion Putting Your Skills into Practice

### Statistics for Researchers Workshop

**4. Statistics for Researchers**

The goal of this programme is to strengthen your ability to use statistics when conducting your research work. It teaches how to plan and execute a strategy for data analysis which explores and answers specific example research questions.

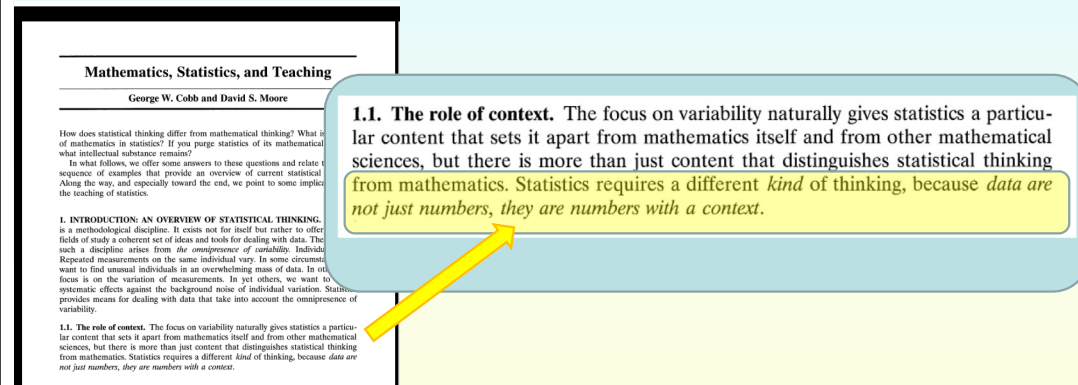
This programme is also deliberately designed so that the key statistical skills learned here are transferable to other areas for use in your subsequent career.

**Strengthen** your ability to use statistics  
Teaches how to **plan** and **execute** a strategy for data analysis  
Programme is designed so that **statistical skills** learned are **transferable** to other areas

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

## Getting Started

*The American Mathematical Monthly*, Vol. 104, No. 9. (Nov., 1997), pp. 801-823.



**Mathematics, Statistics, and Teaching**

George W. Cobb and David S. Moore

How does statistical thinking differ from mathematical thinking? What is mathematics in statistics? If you purge statistics of its mathematical what intellectual substance remains?

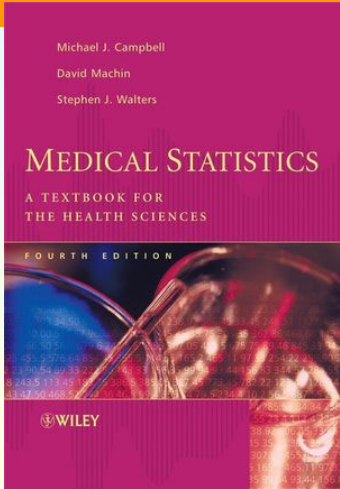
In what follows, we offer some answers to these questions and relate a sequence of examples that provide an overview of current statistical thinking. Along the way, and especially toward the end, we point to some implicit teaching of statistics.

**1. INTRODUCTION: AN OVERVIEW OF STATISTICAL THINKING.** Statistics is a methodological discipline. It exists not for itself but rather to offer fields of study a coherent set of ideas and tools for dealing with data. The such a discipline arises from the omnipresence of variability. Individuals. Repeated measurements on the same individual vary. In some circumstances, want to find unusual individuals in an overwhelming mass of data. In other, focus is on the variation of measurements. In yet others, we want to systematic effects against the background noise of individual variation. Statistics provides means for dealing with data that take into account the omnipresence of variability.

**1.1. The role of context.** The focus on variability naturally gives statistics a particular content that sets it apart from mathematics itself and from other mathematical sciences, but there is more than just content that distinguishes statistical thinking from mathematics. Statistics requires a different kind of thinking, because data are not just numbers, they are numbers with a context.

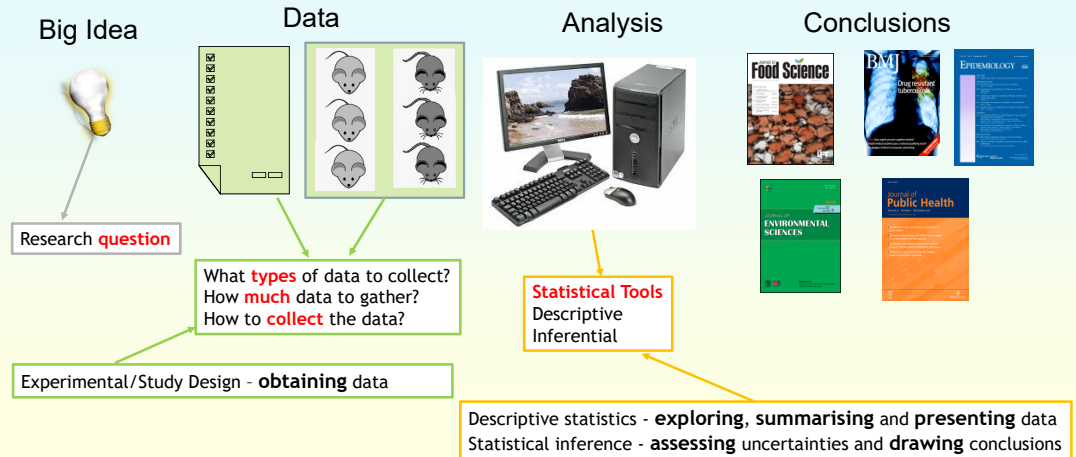
Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

# Getting Started



**1.3 Statistics is about common sense and good design**  
 A well-designed study, poorly analysed, can be rescued by a reanalysis but a poorly designed study is beyond the redemption of even sophisticated statistical manipulation. Many experimenters consult the medical statistician only at the end of the study when the data have been collected. They believe that the job of the statistician is simply to analyse the data, and with powerful computers available, even complex studies with many variables can be easily processed. However, analysis is only part of a statistician's job, and calculation of the final 'p-value' a minor one at that!  
 A far more important task for the medical statistician is to ensure that results are comparable and generalisable.

# Components of a Statistical Analysis



# Research Questions

## Big Idea



- ◆ **Descriptive questions**
  - ❖ Study designed to describe what is going on or what exists
  - ❖ Do not usually involve experimental manipulation
- ◆ **Difference questions**
  - ❖ Is there a difference?
  - ❖ Comparison between groups on a dependent variable
- ◆ **Relationship questions**
  - ❖ Study is designed to look at the relationship between two or more variables

Ask a question-clearly stated  
 Relationship between two or more variables - difference and relationship questions  
 Measurable - capable of empirical testing

**FEATURE / MANCHETTE**

**Formulating Answerable Questions: Question Negotiation in Evidence-based Practice<sup>1,2</sup>**

Lorie A. Kloda and Joan C. Bartlett

**Abstract Objectives:** This review explores the different question formulation structures proposed in the literature that aim to identify the literature for conducting the reference interview and for teaching students and clinicians. **Method:** The present study compares several known question formulation structures identified in the health and social sciences literature. **Relevance:** Health and social care professionals should be made aware of the plurality of question formulation structures and their applicability to different fields of practice, as well as their ability to address types of questions within a field of practice.

**Introduction**

Librarians have important roles in assisting healthcare professionals and students in their interactive working. One of these roles is to help users identify and express their information needs clearly. This paper discusses the literature on formulating clinical questions in the context of health care. Linking question formulation to the literature of the reference interview, this paper first introduces Taylor's model of question negotiation to explain the first step of evidence-based practice. Various question formulation structures are reviewed, with examples provided as well as research on their usefulness. Finally, a discussion of the utility and applicability of question formulation structures for librarians, clinicians and healthcare professionals is provided with suggestions for future research.

**Asking a question**

The basis of much of the interaction between librarians and information users (health professionals, researchers, students) and others begins with questions that require answers. The question, a formal expression of an individual's information need, is the precursor and, in fact, the prerequisite for preparative information seeking to take place [1]. Librarians may expect users to approach them with fully formed and well-articulated questions as requests of information needs that have been given much thought. In other words, librarians are prepared to assist users in answering precise, identifiable questions.

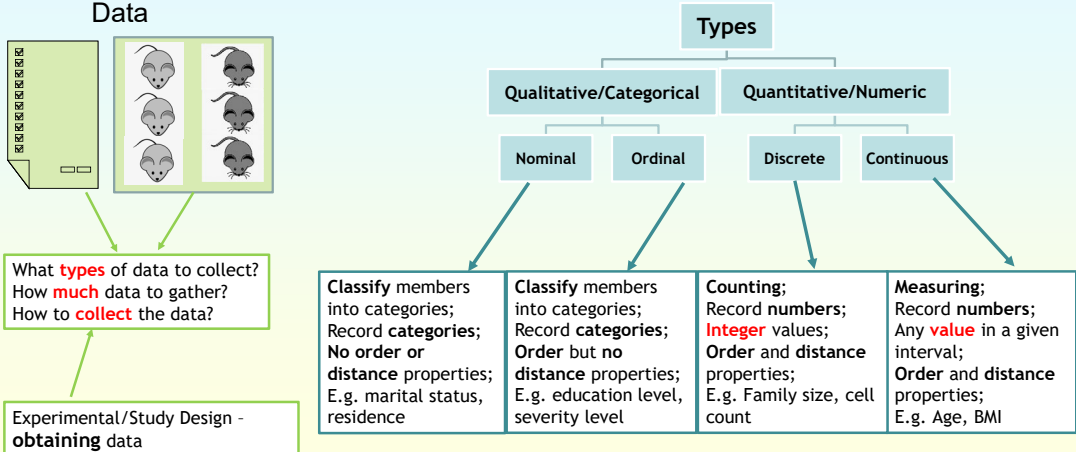
**Types of questions**

There are three components to Taylor's model: the first types (or levels) of expression of information needs, the process model for preparative decisions, and the final types that questions go through during negotiation [2]. The first component is the most relevant to question formulation. The four levels of information needs or questions are: visceral need (QN), conscious need (QCN), formulated need (QF), and compressed need (QD).

In Taylor's typology of needs, at the first level of question (QN), the user is not conscious of the need, and it remains vague and unarticulated. As it moves through the second level of question (QCN) to an articulated information need in the user's mind, but still unhelpful to the user, Taylor suggests that the user may speak to

1. Lorie A. Kloda, 'Medical University Librarian, McMaster University, 1600 McMaster Street, Hamilton, Ontario, L8S 4L7, Canada, 2019.  
2. Lorie A. Kloda, 'Medical University Librarian, McMaster University, 1600 McMaster Street, Hamilton, Ontario, L8S 4L7, Canada, 2019.  
Corresponding author (email: lorie.kloda@mcmcgill.ca).

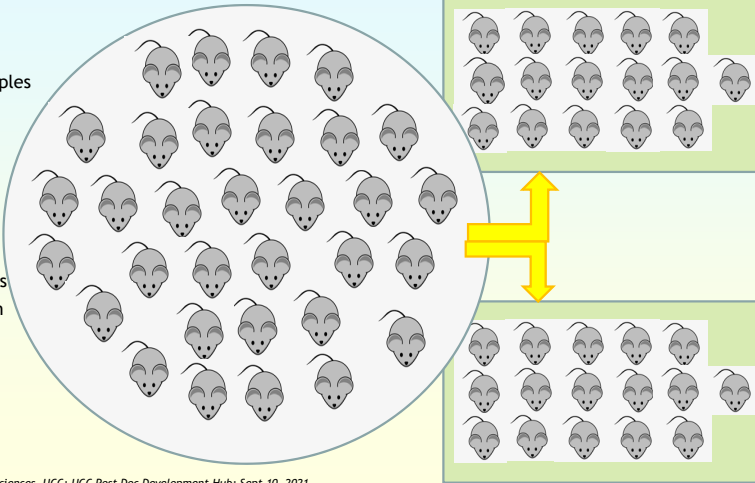
# Data: Variable Types



## Data: Sample Types

### Independent samples

- ◆ Uncorrelated samples
  - ❖ Two or more **unrelated** samples of items
- ◆ Two or more distinct groups
- ◆ No matching of cases
- ◆ “**Between-subjects**” factor
- ◆ Examples
  - ❖ First years versus third years
  - ❖ Two groups of animals given different diets



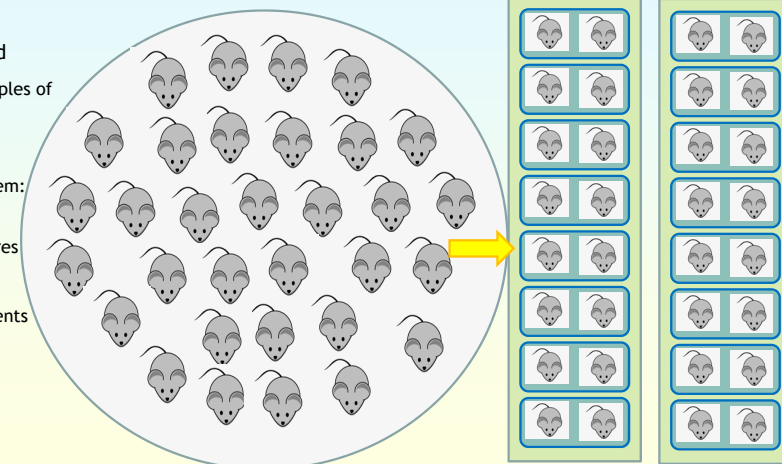
Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

8

## Data: Sample Types

### Dependent samples

- ◆ Observations are correlated
  - ❖ Two or more **related** samples of items
- ◆ “**Within-subjects**” factor
  - ❖ Measurements for each item: Two: paired data; Multiple: repeated measures
- ◆ Examples
  - ❖ Before and after experiments
  - ❖ Matched pairs
  - ❖ Twins



Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

9

## Analysis

### Analysis



**Statistical Tools**  
Descriptive  
Inferential

- ◆ What is your **hypothesis** or research question?
- ◆ How many **variables**?
- ◆ Is the **data** nominal, ordinal or numeric ?
- ◆ How many **groups**?
- ◆ Are the groups **independent**?
- ◆ Is the data **distribution** Normal?

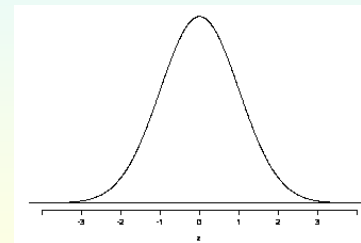
Descriptive statistics - **exploring, summarising** and **presenting** data  
Statistical inference - **assessing** uncertainties and **drawing** conclusions

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

10

## Analysis: Normal Distribution

Data distribution  
Normal distribution?



- ◆ Bell-shaped
- ◆ Uni-modal
- ◆ Symmetric
- ◆ Asymptotic to the x-axis
- ◆ Defined by two parameters  $\mu$  and  $\sigma$
- ◆  $\mu$ =centre and  $\sigma$ =spread

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

11

## Descriptive statistic, graphical display or statistical (parametric) test

- ◆ What is its name?
- ◆ What will it tell you?
- ◆ What are its requirements? - data type, number of groups

## Statistical test

- ◆ What are its underlying assumptions?
- ◆ What is the alternative nonparametric test?

## ◆ Parametric tests

- ❖ Assumption that data are Normally distributed
- ❖ Numeric data - compute means and standard deviations
- ❖ No extreme scores
- ❖ Assumption of homogeneity of variance (ANOVA)
- ❖ Robust for large samples
- ❖ Check associated assumptions
- ❖ More powerful than non-parametric procedures

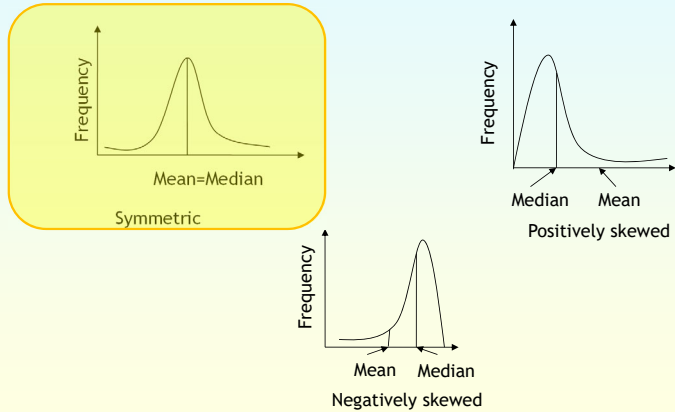
## ◆ Nonparametric tests

- ❖ Distribution-free
- ❖ Ranks
- ❖ Assumptions behind parametric tests are invalid
- ❖ Ordinal variable
- ❖ Values are "off the scale" that is, too high or low
- ❖ Usually used with small samples
- ❖ Statistical significance is more difficult to reach

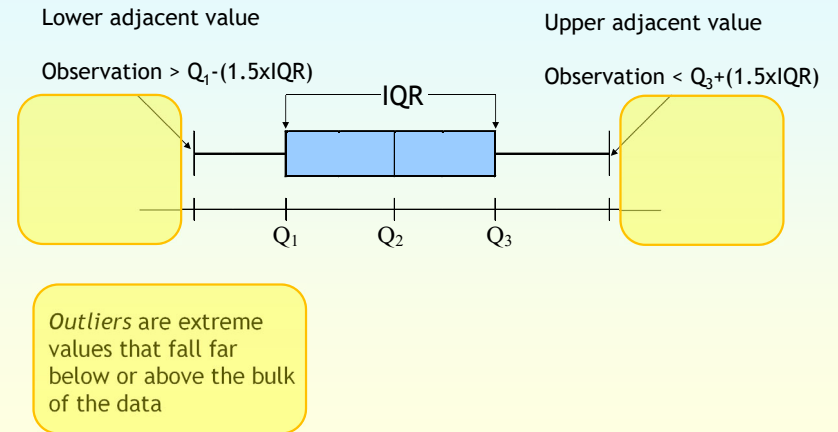
Descriptor	What will it tell me?	Requirements
<b>Numerical Summaries</b>		
Mean and standard deviation or standard error	Provides a measure of centrality and spread	Numeric variable; Symmetric distribution
Median and interquartile range	Provides a measure of centrality and spread	Numeric variable; Skewed distributions
Range	Provides a measure of spread	Numeric variable
Proportion	Provides the rate for a specific characteristic	Categorical variable; Grouped numeric variable
Frequency table	Details the number and % in each category (group)	Categorical variable; Grouped numeric variable

Descriptor	What will it tell me?	Requirements
<b>Graphical Display</b>		
Histogram	Describes the distribution of the data	Numeric variable
Box plot	Describes centrality and spread of data; Identifies potential outliers (modified box plot)	Numeric variable
Bar chart	Graphical display of frequency table	Categorical variable; Grouped numeric variable
Scatter plot	Gives a 2-D plot	Numeric variables

## Analysis: Distribution Shapes



## Analysis: Box Plots



## Analysis: Comparing Two Means

### Comparing Means

Parametric test	What will it tell me?	Requirements	Assumptions	Non-parametric equivalent
Independent two-sample t test	Is there a difference between two means?	Numeric variable; Two groups to be compared; Groups are independent; Independent observations within groups	Normality; Data within each group are Normally distributed	Mann-Whitney U test/ Wilcoxon rank sum test
Paired t test	Is the mean difference zero?	Numeric variable; Two matched samples or paired samples (dependent groups); Independent observations within group	Normality; Paired differences are Normally distributed	Sign test/ Wilcoxon matched-pairs signed rank test

## Analysis: Comparing Three or More Means

### Comparing Means

Parametric test	What will it tell me?	Requirements	Assumptions	Non-parametric equivalent
One-way analysis of variance (ANOVA)	Is there a difference among three or more means?	Numeric variable; Three or more groups to be compared; Independent groups; Independent observations within groups	Checks focus on residual diagnostics: Normality; Homogeneity of variance	Kruskal-Wallis test

# Analysis: Comparing Three or More Means

## Comparing Means

Post-hoc test	What will it tell me?	Requirements
<b>Bonferroni test</b>	Where did differences occur among means?	Following ANOVA where the F test was significant
<b>Tukey's test</b>	Where did differences occur among means?	Following ANOVA where the F test was significant
<b>Dunnnett's test</b>	Do the other groups differ to the control group?	Following ANOVA where the F test was significant
<b>Scheffé's test</b>	Where did differences occur among means?	Following ANOVA where the F test was significant
<b>Fisher's Least Significant Difference</b>	Where did differences occur among means?	Following ANOVA where the F test was significant

# Analysis: Comparing Proportions

## Comparing Proportions

Non-parametric test	What will it tell me?	Requirements	Assumptions
<b>Chi-square test (2 x 2 table)</b>	Is there a difference between two proportions?	Categorical variables; Independent groups	Expected cell frequencies $\geq 5$
<b>Yates's continuity corrected Chi-square test (2 x 2 table)</b>	Is there a difference between two proportions?	Categorical variables; Independent groups	Improves the approximation provided by the Chi-square test
<b>Fisher's Exact test (2 x 2 table)</b>	Is there a difference between two proportions?	Categorical variables; Independent groups	Expected cell frequencies $< 5$
<b>Chi-square test (2 x r table)</b>	Is there a difference between three or more proportions?	Categorical variables; Independent groups	Expected cell frequencies $\geq 1$
<b>McNemar's test</b>	Are the proportions from two matched samples (a paired sample) the same?	Categorical variables; Two matched samples or paired samples	
<b>Cochran's Q test</b>	Are the proportions from three or more matched samples the same?	Categorical variables; Three or more matched samples	

# Analysis: Tests of Associations

## Measures of Associations

Non-parametric test	What will it tell me?	Requirements	Assumptions
<b>Chi-square test of independence</b>	Is there an association between two categorical variables?	Categorical variables	Expected cell frequencies must be approx. 1
Parametric measure	What will it tell me?	Requirements	Non-parametric measure
<b>Pearson's correlation coefficient</b>	What is the strength of the linear relationship between two variables?	Numeric variables; Linearity	Spearman's rank correlation coefficient

# Analysis: Regression Analysis

## Linear Associations

Parametric	What will it tell me?	Requirements	Assumptions
<b>Simple linear regression analysis</b>	Can I predict one variable based on another?	Continuous dependent variable	Independent observations; Linearity; Checks focus on residual diagnostics: Normality; Homoscedasticity
<b>Multiple (multivariable) linear regression</b>	Can I predict one variable based on others?	Continuous dependent variable; Categorical/numerical independent variables	Independent observations; Linearity; Checks focus on residual diagnostics: Normality; Homoscedasticity



**STATISTICALLY Speaking**

**Multivariate or Multivariable Regression?**

The terms multivariate and multivariable are often used interchangeably in the public health literature. However, these terms actually represent 2 very distinct types of analyses. We define the 2 types of analysis and assess the prevalence of use of the statistical term multivariate in a series of articles published in the American Journal of Public Health. Our goal is to make a clear distinction and to identify the nuances that make these types of analyses so distinct from one another.

Most regression models are described in terms of the regression equation, which is a statistical model in which the dependent variable (the outcome) is predicted by one or more independent variables (the predictors). As in the case with linear models, logistic and proportional hazards regression models can be simple or multivariable. Each of these model structures has a single outcome variable and 1 or more independent or predictor variables.

Multivariate, by contrast, refers to the analysis of data that are obtained from longitudinal studies, whereas an outcome is measured for each individual at multiple time points (repeated measures), or the matching of individual-level data, which often are multiple individuals in each cluster. A multivariate linear regression model would have the form:

$$Y_{ij} = \mu + \alpha_i + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \dots + \epsilon_{ij}$$

where the relationship between multiple independent variables (i.e., the analysis of multiple outcomes) and a single dependent variable (i.e., the use of multiple outcomes) is of interest. This is not the case for multivariate regression, which is used to assess the prevalence of use of the statistical term multivariate. That is, we used PubMed and the keyword "multivariate" to review articles published in the American Journal of Public Health over a 1-year span (December 2010–November 2011). We identified 30 articles in which the term multivariate or multivariable was used in the title.

Of 117% of the 30 articles, multivariate models were used in 16 (50%) articles, whereas 4 (13%) of these models were derived from longitudinal data and 1 from cross-sectional data. The remaining 29 (93%) articles included multivariate models, which included 24 (83%) or more independent or predictor variables (1 of 30, or 3%), binomially, in 2 of the 29 articles (7%). The terms multivariate and multivariable were used interchangeably. This feature indicates the need for a clear distinction in use of the 2 statistical terms.

Although many may argue that the interchangeability of multivariate and multivariable in single settings, we believe that it is important to make a clear distinction in the field of public health. In general, each term is used in public health research to indicate the number of predictors, such as binary, logistic, multivariate, or proportional hazards, to indicate the type of outcome (i.e., continuous, dichotomous, repeated measures, time to event).

Our review revealed that there is a need for more accurate application and reporting of multivariate methods. This issue is not unique to public health research and has been

## Analysis: Residual Diagnostics

Term	What will it tell me?
<b>Diagnostic</b>	"Diagnostic" refers to studying whether a particular model is appropriate for the data.
<b>Residual</b>	"Residuals" of the model represent the difference between what we observed and what we would predict based on the model. Let a particular observation be denoted by $y_i$ and our predicted value of this observation be denoted by $\hat{y}_i$ (the "hat" symbol will always be used to denote an estimated value). Then the $i^{th}$ residual is given by $y_i - \hat{y}_i$ .
<b>Outliers</b>	The occurrence of a particular large residual indicates the presence of an "outlier". That is, an extreme data point for which the model does not fit the data well. The possibility that this point represents a mistake (e.g. in either the measurement, recording, or retrieving of the data) should be considered. If there is a strong possibility of an error, the observation should be discarded. In regression, the difficulty with extreme observations is that they tend to pull the regression line disproportionately off course, leading to poor estimation for the bulk of the data.

## Analysis: Assumptions

Assumption	Method	What will it tell me?	Explanation
<b>Linearity</b>	Scatter plot	Is the relationship between the two numeric variables linear?	Plot of the response variable (DV) against the predictor variable (IV) would indicate if a linear model is appropriate for describing the relationship. A transformation may be useful if the scatter plot indicates that a non-linear relationship exists.

## Analysis: Assumptions

Assumption	Method	What will it tell me?	Explanation
<b>Normality</b>	Histogram	Is the Normality assumption valid?	The histogram should be fairly symmetric and bell shaped. If the Normality assumption is violated a transformation of the data may be used to remedy the problem. For example, replacing the outcome $y$ with the logarithm of $y$ may be helpful.
<b>Normality</b>	Normal probability plot	Is the Normality assumption valid?	It plots the sample versus the values we would get, on the average, if the sample came from a Normal distribution. This plot is approximately a straight line if the sample is from a Normal distribution.
<b>Normality</b>	Kolmogorov-Smirnov test; Shapiro-Wilks test	Is the Normality assumption valid?	Formal tests for testing if the data are Normally distributed; These tests should provide nonsignificant results ( $P > 0.05$ )

## Analysis: Assumptions

Assumption	Method	What will it tell me?	Explanation
<b>Homogeneity of variance (HOV)</b>	Box plots	Is the HOV assumption valid? (ANOVA)	Boxes should be roughly similar in size
<b>Homogeneity of variance (HOV)</b>	Levene's test	Is the HOV assumption valid? (ANOVA)	Formal test for testing if variances differ; Test should provide nonsignificant result ( $P > 0.05$ )
<b>Homogeneity of variance (HOV)/ Homoscedasticity</b>	Plot of residual against predicted values or independent variable	Is the HOV assumption valid? (ANOVA) Is the homoscedasticity assumption valid? (Regression model)	Points should be scattered around the x-axis with no trends and no variation in the extent of the scatter

# Analysis: Some Modelling Considerations



Focus on: Contemporary Methods in Biostatistics (1)  
Regression Modelling Strategies

Eduardo Núñez,<sup>a,b,\*</sup> Ewout W. Steyerberg,<sup>c</sup> and Julio Núñez<sup>a</sup>

<sup>a</sup> Servicio de Cardiología, Hospital Clínico Universitario, INCLIVA, Universidad de Valencia, Spain  
<sup>b</sup> Centre International, Faculty of Medicine, University of Valencia, Spain  
<sup>c</sup> Department of Public Health, Erasmus MC, Rotterdam, The Netherlands

Article history:  
Available online 6 May 2011

Keywords:  
Overfitting  
Number of events per variable  
Calibration  
Discrimination

## ABSTRACT

Multivariable regression models are widely used in health science research, mainly for two purposes: prediction and effect estimation. Various strategies have been recommended when building a regression model: a) use the right statistical method that matches the structure of the data; b) ensure an appropriate sample size by limiting the number of variables according to the number of events; c) prevent or correct for model overfitting; d) be aware of the problems associated with automatic variable selection procedures (such as stepwise), and e) always assess the performance of the final model in regard to calibration and discrimination measures. If resources allow, validate the prediction model on external data.

© 2011 Sociedad Española de Cardiología. Published by Elsevier España, S.L. All rights reserved.

## Statistically Speaking

### Multivariate Regression: The Pitfalls of Automated Variable Selection

Kristin L. Scainini, PhD

Medical researchers commonly use automatic model selection procedures (including forward, backward, and stepwise selection) to build multivariate regression models. These algorithms rarely sit through a pile of potential predictor variables to come up with a single, compact model. Their popularity is understandable. They are readily available in statistical analysis packages, trivial to implement, reduce the work of model building to the click of a button, have the allure of objectivity, and often yield seemingly exciting results with small P values. However, their use is inadvisable due to numerous, widely recognized statistical problems. This article reviews these problems, provides a simple example that illustrates how badly these methods can mislead, and suggests alternative approaches to model building.

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

Journal of Big Data

South Africa Date: 2018-1-12  
https://doi.org/10.1186/s13065-018-0164-6

## SHORT REPORT

Open Access

### Step away from stepwise

Gary Smith

Correspondence:  
gsmith@uconn.edu  
Department of Economics,  
University of Connecticut,  
College Avenue, Storrs,  
CT 06269, USA

## Abstract

**Background:** Stepwise regression is a popular data mining tool that uses statistical significance to select the explanatory variables to be used in a multiple regression model.

**Findings:** A fundamental problem with stepwise regression is that some real explanatory variables that have causal effects on the dependent variable may happen to not be statistically significant, while nuisance variables may be considered significant. As a result, the model may fit the data well in-sample, but do poorly out-of-sample.

**Conclusions:** Many Big Data researchers believe that by using the number of possible explanatory variables, the more useful a stepwise regression for selecting explanatory variables. The reality is that stepwise regression is less effective the larger the number of potential explanatory variables. Stepwise regression does not solve the in-sample problem of too many explanatory variables. Big Data exacerbates the failings of stepwise regression.

ds: Stepwise regression, Data mining, Big Data

# Analysis: Some Modelling Considerations



## Influential Observations, High Leverage Points, and Outliers in Linear Regression

Samprit Chatterjee and Ali S. Hadi

Abstract: A bewilderingly large number of statistical quantities have been proposed to study outliers and influence of individual observations in regression analysis. In this article we describe the inter-relationships which exist among the proposed measures. An examination of these relationships leads us to conclude that only three of these measures along with some graphical displays can provide an analysis a complete picture of outliers (major discrepant points) and points which excessively influence the fitted regression equation. Illustrative examples based on real data are presented.

Key words and phrases: Influence, leverage, outliers, regression diagnostics, residuals.

J. Stat. Theory and Practice  
ISSN: 2458-8668 (p); 2457-8240 (e)

© JOST, Tribhuvan University  
Research Article

## DEALING WITH OUTLIERS AND INFLUENTIAL POINTS WHILE FITTING REGRESSION

Chandra Prasad Dhatal

Tribhuvan University, Institute of Agriculture and Animal Sciences (IAAS), Rampur, Chitwan, Nepal  
Corresponding E-mail: chandra.studied@gmail.com

**ABSTRACT**  
Dealing with outliers and influential points while fitting regression is recognizing them, identifying the reasons for their existence in the process and employing the best alternatives to lessen their effect to the fitted regression model. In this paper, before considering elimination of outliers and the influential points while fitting a regression, as they contain important information, issues why unusual observations (possible outliers) appear in the process and how to analyze them to detect if they were real outliers, have been discussed thoroughly. And, when detected as outliers and influential points, to investigate and eliminate their effect in the fitted model, analytic procedures, leverage value, standardized residuals and Cook's distance were carefully employed to optimize a multiple regression model for rice production forecasting in Nepal. This model was fitted with 35 years (1981-1995) time series data, collected from Ministry of Agriculture and Cooperatives, Food and Agriculture Organization Statistics Database, International Rice Research Institute and Department of Hydrology and Meteorology which as its end was correlated to the three predictors, price at harvest, rural population and area harvested.

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

# Analysis: Some Modelling Considerations



## What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models

MICHAEL A. BAYAR, PhD

**Objective:** Statistical models, such as linear or logistic regression or survival analysis, are frequently used as a means to answer scientific questions in psychosocial research. Many who use these techniques, however, apparently fail to appreciate fully the problem of overfitting, so, capitalizing on the idiosyncrasies of the sample at hand, Overfitted models will fail to replicate in future samples, thus creating considerable uncertainty about the scientific merit of the finding. The present article is a nontechnical discussion of the concept of overfitting in terms of asking too much from the available data. Given a certain number of observations in a data set, there is an upper limit to the complexity of the model that can be derived with any acceptable degree of accuracy. Complexity arises as a function of the number of degrees of freedom expended (the number of predictors including complex terms such as interactions and nonlinear terms) against the same data set during any stage of the data analysis. Theoretical and empirical evidence—with a special focus on the results of computer simulation studies—is presented to demonstrate the practical consequences of overfitting with respect to scientific inference. Three common practices—automated variable selection, plotting of candidate predictors, and deconvolution of continuous variables—are shown to pose a considerable risk for spurious findings in models. The dilemma between overfitting and exploring candidate confounders is also discussed. Alternative means of guarding against overfitting are discussed, including variable suppression and the fixing of coefficients a priori. Techniques to account and correct the complexity, including shrinkage and penalization, also are introduced. **Key words:** statistical models, regression.

## REVIEW ARTICLE

### Linear Regression Analysis

Part 14 of a Series on Evaluation of Scientific Publications

by Astrid Schnöcker, Gerhard Hommel, and Maria Elblöcher

length in centimeters and Y is weight in kilograms. The y-intercept  $a = -133.18$  is the value of the dependent variable when  $X = 0$ , but X cannot possibly take on the value 0 in this study (one obviously cannot expect a person of length 0 centimeters to weigh negative 133.18 kilograms). Therefore, interpretation of the constant is often not useful. In general, only values within the range of observations of the independent variables should be used in a linear regression model.

prediction of the value of the dependent variable becomes increasingly inaccurate the further one goes outside that range.  
The regression coefficient of 1.16 means that in this model, a person's weight increases by 1.16 kg

Deutscher Akademischer Austauschdienst | Deutscher Akademischer Austauschdienst | Deutscher Akademischer Austauschdienst

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

## Statistical Primer for Cardiovascular Research

### Correlation and Regression

Sybil L. Crawford, PhD

#### Alternatives to Least-Squares Estimation

Ordinary least-squares regression is widely used. In part because of its ease of computation and also because it has desirable properties when the assumptions are met. Because the regression line is estimated by minimizing the squared residuals, however, outlying values can exert a relatively large impact on the estimated line. With the advent of computers, alternative methods have been developed that are computationally more demanding but are more robust to outliers. Some techniques reduce the influence of outliers by replacing squared residuals with other functions of the residuals or minimizing the median of the squared residuals rather than the sum (see Rousseeuw and Leroy). Other approaches are nonparametric, such as Tukey's resistant lines or Theil's method. It is difficult to generalize some of these approaches to the setting with multiple predictor variables, however.

#### Additional Considerations and Cautions

**Extrapolation**  
Even when an estimated regression line provides a good fit to the observed data, it is important not to extrapolate beyond the range of the sample, because the estimated line may not be appropriate. For example, as seen in Figure 1A, estimates of Y from the regression line may be invalid for extreme X values. Alternatively, the relation between X and Y may become nonlinear outside the range of the sample.

**Study Design and Interpretation of Estimates**  
Estimates of correlation and R<sup>2</sup> depend not only on the magnitude of the underlying true association but also on the variability of the data included in the sample (see Weisberg). In the preceding ha-CRP and BMI example, the estimated Pearson correlation of log ha-CRP and log BMI in the full sample is 0.62. If we restrict the sample to the middle 2 quartiles of log BMI, thereby artificially decreasing the SD of log BMI from 0.25 to 0.08, the corresponding estimated correlation is 0.31, an underestimate. Conversely, if we include only women in the top and bottom log BMI quartiles (which yields an SD of log BMI of 0.31), the estimated

**Categorical Versus Continuous Variables**  
When a variable is continuous, treating it as a continuous variable typically retains more information than collapsing it to an ordinal categorical variable. In some cases, however, the latter version may be preferable. Consider the example of labor consumption in a some population; there may be a large percentage with no consumption, which leads to a large "spike" at the value 0; hence, there may be no straightforward transformation that satisfies the assumptions of correlation or linear regression. Here, it may be more useful to categorize alcohol consumption as an ordinal variable, eg, zero consumption and quantities of ounces consumed, and to use ANOVA rather than linear regression. As another example, consider years of education. A difference of 1 year often has a different impact depending on whether the reference point is, say, 11 years compared with 13 years. In this case, a categorized ordinal variable may provide a better fit to the data. Moreover, categorized variables may be more interpretable in clinical settings.<sup>10</sup>

**Confounding**  
The above discussion assumes there is only a single predictor variable of interest. The association between X and Y, however, may be due in part to the combination of additional variables that are related to both X and Y, ie, confounding variables. For example, the estimated association between BMI and ha-CRP may be due in part to age, because both BMI and ha-CRP are themselves positively related to age. The methods summarized above can be expanded to include multiple predictors, and associations between X and Y that adjust for these confounding factors do not estimate. Returning to the ha-CRP and BMI example, a partial (age-adjusted) correlation between ha-CRP and BMI can be computed. For the Pearson correlation, this is done by regressing ha-CRP on age, regressing BMI on age, and computing the Pearson correlation of the 2 sets of residuals, ie, the component of ha-CRP that is unrelated to age and the component of BMI that is unrelated to age. Similarly, an age-adjusted slope for BMI can be estimated by adding age as a predictor to the linear regression model. A regression model

30

# Analysis



## Multiple Comparisons

- ❖ A one-way ANOVA indicates significant differences between groups
- ❖ Perform pairwise tests to locate where the difference lies

## Multiplicity of testing

- ❖ Probability of obtaining at least one significant finding by chance increases with the number of tests performed
- ❖  $1 - (1-\alpha)^k$ , k tests

## Bonferroni correction

- ❖ Adjusts the  $\alpha$  level or P-value
- ❖  $\alpha/k$ ; k tests

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

# Statistical Computing and Graphics

## Where's Waldo? Visualizing Collinearity Diagnostics

Michael FRIENDLY and Ernest KWAN

Collinearity diagnostics are widely used, but the typical tabular output used in almost all software makes it hard to tell what to look for and how to understand the results. We describe a simple improvement to the standard tabular display, a graphic rendition of the salient information as a "tabletop", and graphic displays designed to make the information in these diagnostic methods more readily understandable. In addition, we propose a visualization of the contributions of the predictors to collinearity through a "collinearity plot", which is simultaneously a biplot of the smallest dimensions of the correlation matrix of the predictors,  $R_{XX}$ , and the largest dimension of  $R_{XX}^{-1}$ .

Later, David Hosley wrote "A Guide to Using the Collinearity Diagnostics" (Hosley 1991a), which seemed to promise a solution for visualizing these diagnostics. For context, it is worth quoting the abstract in detail.

The description of the collinearity diagnostics as presented in Hosley, Kuhn, and Welsch's *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, is principally formal, leaving it to the user to implement the diagnostics and learn to digest and interpret the diagnostic results. This paper is designed to overcome this shortcoming by describing the different graphical displays that can be used to present the diagnostic information and more involve

## Common pitfalls in statistical analysis: The perils of multiple testing

Pritya Rangnathan,  
C. S. Pramesh,  
Marco Buisson<sup>1,2</sup>

<sup>1</sup>Department of Anesthesiology,  
Tata Memorial Centre, Department  
of Critical Care Medicine, Division of  
Thoracic Surgery, Tata Memorial  
Centre, Mumbai, Maharashtra, India  
<sup>2</sup>International Drug Development  
Center, San Francisco, California,  
USA; <sup>3</sup>Department of Biostatistics,  
Harvard University, Boston, MA, USA

Address for correspondence:  
Dr Pritya Rangnathan,  
Department of Anesthesiology, Tata  
Memorial Centre, Parel Road,  
Mumbai, Maharashtra - 400 012,  
India.  
E-mail: drprityar@tmc.gov.in

## ABSTRACT

Multiple testing refers to situations where a dataset is subjected to statistical testing multiple times - either at multiple time-points or through multiple subgroups or for multiple end-points. This amplifies the probability of a false-positive finding. In this article, we look at the consequences of multiple testing and explore various methods to deal with this issue.

Key words: Biostatistics, data interpretation, multiplicity, statistical significance

31



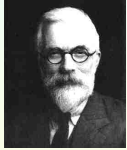
## Conclusions: Statistical Significance

P-value is the probability of obtaining a value of the test statistic at least as extreme as the one computed from the sample data if the null hypothesis is true

### 5% level

### Standard Reporting

- ◆ “If the probability of such an event were sufficiently small - say 1 chance in 20 - then one might regard the results as significant” (Fisher)
- ◆ No sharp boundary between significant and non significant; there is no practical difference between 0.049 and 0.051
- ◆ Increasing evidence as P decreases
- ◆ NS  $p > 0.05$  no evidence against  $H_0$  ‘non significant’
- ◆ \*  $0.01 < p < 0.05$  evidence against  $H_0$  ‘significant’
- ◆ \*\*  $0.001 < p < 0.01$  strong evidence against  $H_0$  ‘highly significant’
- ◆ \*\*\*  $p < 0.001$  very strong evidence against  $H_0$  ‘extremely significant’



Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

32

## Conclusions: Statistical versus Practical Significance

- ◆ **Statistical significance** is a phrase that is commonly used in hypothesis testing
- ◆ Decision: **Reject** the null hypothesis (e.g. null hypothesis no difference between the mean systolic blood pressure for smokers and non-smokers) then the results are said to be **statistically significant**
  - ❖ Data support rejection of the null hypothesis
  - ❖ Data are inconsistent with the null hypothesis
  - ❖ Statistical significance says **nothing about clinical (practical) importance**; it does not say that the difference is important

### Interpreting Significance: The Differences Between Statistical Significance, Effect Size, and Practical Importance

Pavel Kuhnawski, BSc and Fiona Fisher, PhD

It is a common misconception that statistical significance indicates a large and important effect. In fact, the three concepts—statistical significance, effect size, and practical importance—are distinct from one another and, broadly speaking, are not directly related to one another. In this article, we explain the differences between these concepts and distinguish between them. Finally, we propose reporting confidence intervals as a step toward distinguishing these concepts.

**Keywords:** Confidence intervals, Effect size, Clinical significance, Misconceptions

Researchers often hesitate to report a “significant result” and null hypothesis significance testing (NHST) is overvalued. In the most common statistical analysis in most social, life, and behavioral sciences. For example, in 10 leading psychology journals, NHST was reported in more than 97% of articles. An analysis of the last 10 issues of *Statistical Science* and *Journal of the Royal Society* shows that the proportion of articles reporting NHST is more than half (50% and 57% respectively) in each journal. The remainder were not reported articles or reported descriptive statistics such as means only and did not provide any inferential test. Unfortunately, NHST is the subject of great controversy, and many practical and misconceptions have been associated with its use. In this article, we review four of the most common errors associated with this practice and recommend reporting confidence intervals (CIs) as a solution.

**What is Statistical Significance?**  
Before we introduce our four common errors, let us review the definition of statistical significance. A result is deemed “significant” if the P-value produced from a NHST is below a certain value, traditionally 0.05. P is a probability value; specifically, the probability of incorrectly rejecting the null hypothesis when in fact it is true. In other words, it is the probability of a “false positive.” Think of a pregnancy test. P is analogous to the probability of the test either being positive when it really is not or vice versa. There is a small probability that we know, but equally useful is the probability of a false negative. We would never give a pregnant test to a woman who reports a false negative rate, and yet, many researchers make the equivalent error. Statistical power measures the false-negative rate (power = 1 - false negative rate). In the context of detecting a statistically significant P-value, if the effect truly exists, the effect size is a measure of how good a test is; we have over a reasonable chance of finding a statistically significant result. Although we understand that the false-positive error rate is 0.05 or 0.1, as mentioned above, false-negative error rates have historically gone unmentioned. More recent surveys of the statistical power of published research show little improvement. In fact, there were no reported levels of statistical power in our analysis of the last 10 *NATURE* journals.

The crucial message here is that the P-value is a very limited piece of information, relating to false-positive error rates only. Statistical significance merely to certify a statement about the P-value relative to an arbitrary cutoff, is to too often only to false-positive errors. There is much more to know about a set of empirical data.

One particularly common problem is that statistical significance is used ambiguously to our best survey of NHSTs we found above ten-thousands of articles (1970-2010) used the term significance in an ambiguous way. By this, we mean that it was responsible to tell from the reports whether the author meant to refer to the statistical significance or the practical importance of their result. The authors were not even given a legend to explain the difference. In this article, we propose reporting confidence intervals as a step toward highlighting the practical importance of their results. Perhaps, they are assuming that the test is significant. Surely, we will explain why statistical significance, effect size, and practical

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

33

## Conclusions: Statistical versus Practical/Clinical Significance

- ◆ We can estimate the size of any effect (e.g. difference between mean systolic blood pressure for smokers and non-smokers) along with a **95% confidence interval** for our estimate
- ◆ This allows us to assess the **practical or clinical importance**, implications of the effect of interest
- ◆ The use of **confidence intervals** facilitates the distinction between statistical significance and practical/clinical significance

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

34

## Conclusions: Point Estimation vs Confidence Interval

- ◆ Single numerical value
- ◆ Computed from the sample data
- ◆ Estimate of the population parameter

### Examples

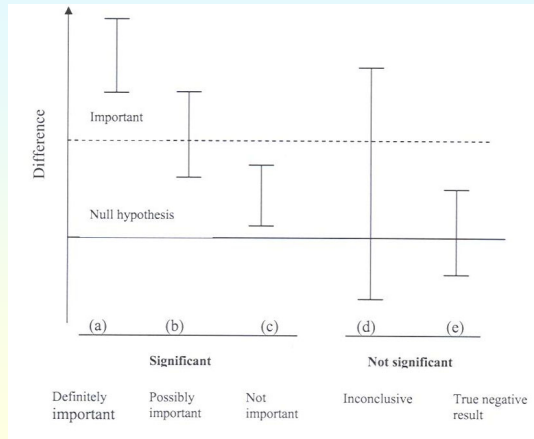
- ❖ Sample mean is a point estimate of the population mean
- ❖ Sample standard deviation is a point estimate of the population standard deviation
- ❖ Sample proportion is a point estimate of the population proportion

- ◆ Range of values that is expected to contain the true value
- ◆ Confidence intervals
  - ❖ Computed from the sample data
  - ❖ Certain certainty (confidence level) that it will contain the true value
- ◆ Confidence Levels
  - ❖ 90%, 95%, 99%,
  - ❖ Symmetric, centred about the sample statistic

Kathleen O'Sullivan, Statistics, School of Mathematical Sciences, UCC; UCC Post Doc Development Hub; Sept 10, 2021

35

## Conclusions: Statistical versus Practical Significance



## Conclusions

- ◆ Drawing **conclusions** consider:
  - ❖ Sampling technique
  - ❖ Statistical power/Sample size
  - ❖ Use of appropriate statistical methods

## Other

- ◆ Provide a description of **statistical methods** used, under Statistical Analysis
- ◆ What **aspects** of statistical analyses conducted are important to present?
- ◆ How to **tabulate** the required results?
- ◆ What is the best **chart** to illustrate the results?
- ◆ Most importantly anyone reading your report should be able to **replicate** your study and statistical analysis
  - ❖ Sufficient detail should be provided so that this can be achieved

## Summary

- ◆ Clarify your research question(s)
- ◆ Choose an appropriate sampling method
- ◆ Consider sample size
- ◆ Examine peer-reviewed literature for statistical analyses conducted
- ◆ Describe your data; begin with descriptive analysis of your data
- ◆ Ensure that the statistical analyses conducted address your research questions and are correct
- ◆ Check assumptions underpinning statistical techniques
- ◆ Provide statistical evidence
- ◆ Discuss practical/clinical significance by assessing 95% CI
- ◆ Consider multiplicity of testing - can find significant results by chance
- ◆ Comment on any statistical or methodological limitations

## Contact Details

- ◆ Name: Kathleen O'Sullivan  
Lecturer, Department of Statistics
- ◆ Location: Room 1-58 Western Gateway Building  
Department of Statistics  
School of Mathematical Sciences  
University College Cork
- ◆ Tel: 021 420 5812/5817
- ◆ Email: kathleen.osullivan@ucc.ie

