PMIM102

Scientific Computing & Health Care

Introduction to Statistics (Statistics I)

Alan Watkins (a.watkins@swansea.ac.uk)

Contents

Underlying concepts & guiding principles
Descriptive statistics - numerical & graphical
Two basic tests illustrated

and

$$\frac{\partial^2 l}{\partial c^2} = -\frac{n}{c^2} - (k+1)\sum_{i=1}^{n}\frac{X_i^c(\log X_i)^2}{(1+X_i^c)^2}. \quad (5)$$

3. EXPECTATIONS IN DERIVATIVES

To consider the regularity of the log-likelihood for complete samples from eqn (1), we take expectations of the first- and second-order partial derivatives of the log-likelihood. This requires the expectations of

$$\log X,\ \log(1+X^c),\ \frac{X^c\log X}{1+X^c} \text{ and } \frac{X^c(\log X)^2}{(1+X^c)^2}.$$

3.1. *Some useful results and two expectations*

There are two basic results: the first is that $Y = 1 + X^c$ follows a Pareto distribution with probability density function $ky^{-(k-1)}$ for $y \geq 1$, so that $\log Y$ has a negative exponential distribution with mean k^{-1}. Thus, we immediately have

$$E[\log(1+X^c)] = k^{-1}. \quad (6)$$

The second basic result is

$$E[X^r] = \frac{\Gamma\left(\frac{r}{c}+1\right)\Gamma\left(k-\frac{r}{c}\right)}{\Gamma(k)} \quad (7)$$

where Γ is the usual gamma function, with first and second derivatives Γ', Γ''. Differentiating eqn (7) with respect to r, we obtain

$$E[X^r\log X] = \frac{\Gamma'\left(\frac{r}{c}+1\right)\Gamma\left(k-\frac{r}{c}\right) - \Gamma\left(\frac{r}{c}+1\right)\Gamma'\left(k-\frac{r}{c}\right)}{c\Gamma(k)} \quad (8)$$

and evaluating this expectation at $r = 0$ gives

$$E[\log X] = \frac{\Gamma'(1)\Gamma(k) - \Gamma(1)\Gamma'(k)}{c\Gamma(k)} = -\left[\frac{\gamma+\psi(k)}{c}\right] \quad (9)$$

where $\psi = \Gamma'/\Gamma$ is the digamma or psi function, and γ is Euler's constant. This gives the second of the required expectations. A third useful result is obtained by differentiating eqn (8) with respect to r; this leads to an expression for $E[X^r(\log X)^2]$, which, evaluated at $r = 0$, yields $E[(\log X)^2]$ as

$$\frac{\Gamma''(1)\Gamma(k) - 2\Gamma'(1)\Gamma'(k) + \Gamma(1)\Gamma''(k)}{c^2\Gamma(k)} = \frac{\left[\frac{\pi^2}{6} + \gamma^2 + 2\gamma\psi(k) + (\psi(k))^2 + \psi'(k)\right]}{c^2} \quad (10)$$

where ψ' is the derivative of the digamma function (and is also known as the trigamma function).

3.2. *Two further results on expectations*

We first note that we can obtain the thi expectation by an appropriate manipulation of eq (8) or (9). Writing expectation as E_k to emphasize tl role of k in the probability density function (1), tl third expectation is

$$E_k\left[\frac{X^r\log X}{1+X^c}\right] = \frac{k}{k+1}E_{k+1}[X^r\log X]$$

on exploiting the form of eqn (1). Using eqn (8) wi $r = c$ and k replaced by $k+1$, we now obtain a expression for the third expectation as

$$E_k\left[\frac{X^c\log X}{1+X^c}\right] = \frac{k}{k+1}\left[\frac{\Gamma'(2)\Gamma(k) - \Gamma(2)\Gamma'(k)}{c\Gamma(k+1)}\right] = \frac{1-\gamma-\psi(k)}{c(k+1)}. \quad (1$$

It is of some interest to note that we can also wri

$$E_k\left[\frac{X^c\log X}{1+X^c}\right] = E_k[\log X] - E_k\left[\frac{\log X}{1+X^c}\right] = E_k[\log X] - \frac{k}{k+1}E_{k+1}[\log X]$$

and then use eqn (9); the required result [eqn (11 is obtained via the recurrence relatic $\psi(k+1) = \psi(k) + k^{-1}$.

Similar manipulations may be used to obtain tl final expectation from suitable expressions f $E[X^r(\log X)^2]$ or $E[(\log X)^2]$; the expectation may b written as $k(k+2)E_{k+2}[X^r(\log X)^2]$, or as

$$E_k\left[\frac{(\log X)^2}{1+X^c}\right] - E_k\left[\frac{(\log X)^2}{(1+X^c)^2}\right] = \frac{k}{k+1}E_{k+1}[(\log X)^2] - \frac{k}{k+2}E_{k+2}[(\log X)^2]. \quad (1$$

Substituting eqn (10) into (12) yields an expressic for the expectation in terms of ψ, ψ' at $k+1$ an $k+2$; further simplification is again possible, sinc $\psi'(k+2) = \psi'(k+1) - (k+1)^{-2}$, and we obta the required expectation as

$$k\left[\frac{\pi^2}{6} + \gamma^2 - 2\gamma + 2(\gamma-1)\psi(k+1) + \{\psi(k+1)\}^2 + \psi'(k+1)\right]/[c^2(k+1)(k+2)]. \quad (1$$

4. REGULARITY AND EXPECTED FISHER INFORMATION

Using eqn (6), we therefore have, from eqn (2)

$$E\left[\frac{\partial l}{\partial k}\right] = \frac{n}{k} - nE[\log(1+X^c)] = \frac{n}{k} - \frac{n}{k} = 0$$

TABLE IV

PARAMETER ESTIMATES FOR THE PRICE MODELS

	EACD(2,2)			WACD(2,2)		
	Estimate	t Statistic	Robust t Statistic	Estimate	t Statistic	
ω	.1150	4.95	3.13	.1250	3.31	
α_1	.0703	3.66	2.30	.0745	2.36	
α_2	.1983	7.59	5.96	.2180	4.94	
β_1	.1520	2.51	1.55	.1246	1.33	
β_2	.4985	8.76	5.28	.4872	5.53	
γ				.7650	12.40	(H_0: $\gamma = 1$)
c_{11}	794.739	—	—	763.080	—	
d_{11}	476.460	.42	.25	459.687	.26	
d_{21}	−195.196	−.08	−.04	−101.911	−.02	
d_{22}	275.513	.21	.10	261.194	.12	
d_{23}	238.016	.15	.08	524.297	.21	
d_{24}	−609.202	−.37	−.24	−668.336	−.24	
d_{25}	16.744	.01	.01	86.048	.03	
d_{26}	−264.385	−.09	−.06	−54.562	−.01	
d_{27}	−3379.156	−.90	−1.15	−3807.942	−.57	
d_{31}	−69.153	−.05	−.02	−127.715	−.06	
d_{32}	−67.975	−.08	−.03	−76.041	−.05	
d_{33}	−319.247	−.29	−.18	−485.941	−.28	
d_{34}	437.515	.43	.28	478.477	.28	
d_{35}	−106.821	−.12	−.09	−180.527	−.12	
d_{36}	1176.874	.29	.22	1089.532	.15	
d_{37}	2788.492	.54	.77	3031.447	.33	
Statistics from Residuals						
Mean	.9999			Mean	.9999	
Std Dev	1.2696			Std Dev	.9943	
Ljung-Box	7.05			Ljung-Box	6.35	
Excess Disp.				Excess Disp.		
Test Statistic	7.93			Test Statistic	−.14	

Notes:

$$\psi_t = \omega + \sum_{j=1}^{p}\alpha_j\tilde{x}_{t-j} + \sum_{j=1}^{q}\beta_j\psi_{t-j} \quad \text{where} \quad \tilde{x}_{t-1} = \frac{x_{t-1}}{\phi(t_{i-1})},$$

$$\phi(t_{i-1}) = \sum_{j=1}^{K} I_j[c_j + d_{1,j}(t_{i-1} - k_{j-1}) + d_{2,j}(t_{i-1} - k_{j-2})^2 + d_{3,j}(t_{i-1} - k_{j-1})^3]$$

where I_j is the indicator variable for the jth segment of the spline ($I_j = 1$ if $k_{j-1} \leq t_{i-1} < k_j$, = 0 otherwise). For $j > 1$ $c_{1,j}$ and $d_{1,j}$ are restricted by the usual differentiability conditions. c_{11} is normalized by restricting the unconditional mean of the diurnal factor to equal the observed sample mean.

] - IBM SPSS Statistics Data Editor

Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1 Visible: 10 of

D	Programme	Group	Gender_FM	Gender	Age_Year	Age_Decade	Endurance_Before	Endurance_After	Endurance_Change
1	Standard	0	Female	1	48	2	239.0	269.3	30.3
2	Standard	0	Female	1	21	0	213.6	313.4	99.8
3	Enhanced	1	Female	1	36	1	434.4	513.7	79.2
4	Standard	0	Female	1	26	0	343.6	416.5	72.9
5	Enhanced	1	Female	1	23	0	676.3	780.1	103.8
6	Enhanced	1	Male	0	25	0	738.6	822.9	84.3
7	Standard	0	Female	1	29	0	282.5	345.9	63.5
8	Enhanced	1	Male	0	48	2	502.5	598.0	95.5
9	Enhanced	1	Female	1	30	1	462.4	554.7	92.3
10	Enhanced	1	Female	1	34	1	766.1	853.2	87.1
11	Standard	0	Female	1	44	2	647.9	754.2	106.3
12	Standard	0	Female	1	33	1	424.8	515.7	90.9
13	Enhanced	1	Female	1	21	0	493.9	601.8	107.9
14	Standard	0	Female	1	35	1	550.0	644.9	94.9
15	Standard	0	Male	0	35	1	671.7	757.0	85.4
16	Standard	0	Male	0	33	1	463.7	558.6	94.9
17	Standard	0	Female	1	46	2	640.2	711.9	71.7
18	Standard	0	Male	0	31	1	351.3	404.8	53.5
19	Standard	0	Male	0	24	0	209.8	271.7	61.9
20	Enhanced	1	Male	0	39	1	213.6	252.4	38.8
21	Standard	0	Female	1	39	1	756.4	843.3	86.9
22	Standard	0	Male	0	48	2	712.8	782.6	69.8
23	Standard	0	Female	1	40	2	794.9	866.2	71.3
24	Enhanced	1	Female	1	42	2	229.5	298.8	69.3
25	Standard	0	Male	0	36	1	593.2	658.1	64.8
26	Enhanced	1	Female	1	44	2	459.5	518.6	59.1

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Peconi_PhD_age_sex

Filter

	age age provided by NHSDW	sex	testm	testf	agem	agef
1	42	2	NA	42	NaN	42
2	49	2	NA	49	NaN	49
3	71	1	71	NA	71	NaN
4	4	1	4	NA	4	NaN
5	24	1	24	NA	24	NaN
6	81	2	NA	81	NaN	81
7	82	2	NA	82	NaN	82
8	42	2	NA	42	NaN	42
9	32	2	NA	32	NaN	32
10	60	1	60	NA	60	NaN
11	95	1	95	NA	95	NaN
12	25	1	25	NA	25	NaN
13	5	2	NA	5	NaN	5
14	11	1	11	NA	11	NaN
15	27	2	NA	27	NaN	27
16	27	1	27	NA	27	NaN
17	2	1	2	NA	2	NaN
18	91	2	NA	91	NaN	91

Showing 1 to 19 of 416,819 entries

Console Terminal

//tawe_dfs/users_staff/sfs5/a.watkins/Documents/

```
R version 3.2.5 (2016-04-14) -- "Very, Very Secure Dishes"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(haven)
> Peconi_PhD_age_sex <- read_sav("//tawe_dfs/users_staff/SFS5/A.Watkins/Documents/spss/Peconi_PhD_age_sex.sav")
> view(Peconi_PhD_age_sex)
>
```

Environment History Connections

Import Dataset | List

Global Environment

Data

Peconi_PhD_age_sex	416819 obs. of 6 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home

Name	Size	Modified
#Data Analysis Plan_Saiful_V0 2_AW.docx	66.8 KB	Aug 3, 2015, 4:21 PM
.Rhistory	609 B	Oct 16, 2017, 1:38 PM
161213 Action plan SAFER 2 Final Report-1-AW.docx	23.7 KB	Dec 17, 2013, 10:54 AM
2015 09 22 LETTERfromBuddugRees AW.docx	727.1 KB	Sep 22, 2015, 10:46 AM
2015UKCRCreg HH 02042015-GH-CE150408ntc-AJW.docx	202.1 KB	Apr 9, 2015, 1:33 PM

A changing health informatics landscape?

Evolving features

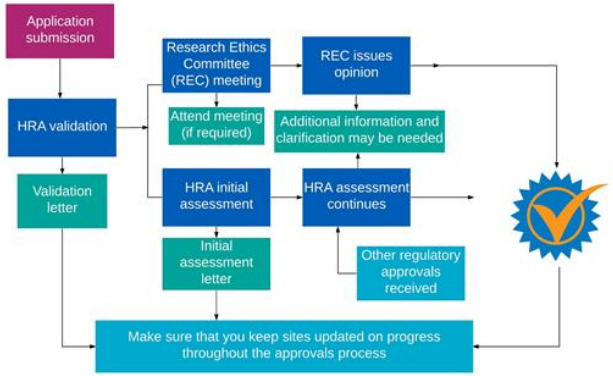
- **GDPR/REC, CAG & HRA/MHRA**
- **NHS Digital/SAIL routine data/e-trials**
- **Long-term follow-up as default**
- **Enhanced PROMs input**

Familiar landmarks

- **IT developments (REDCap cloud)**
- **Funding ('hard' core/'soft' external)**
- **Data validation/oversight requirements**
- **Consolidation/Collaboration**
  - **within Wales/UK/Internationally**

Success requires solid foundations

- **Good ideas & creative/novel methods**
- **Rigorous, well-defined studies**
- **High-quality robust outputs with Impact**



REDCap
Research Electronic Data Capture

Three concepts (building blocks)

CONSORT Statements

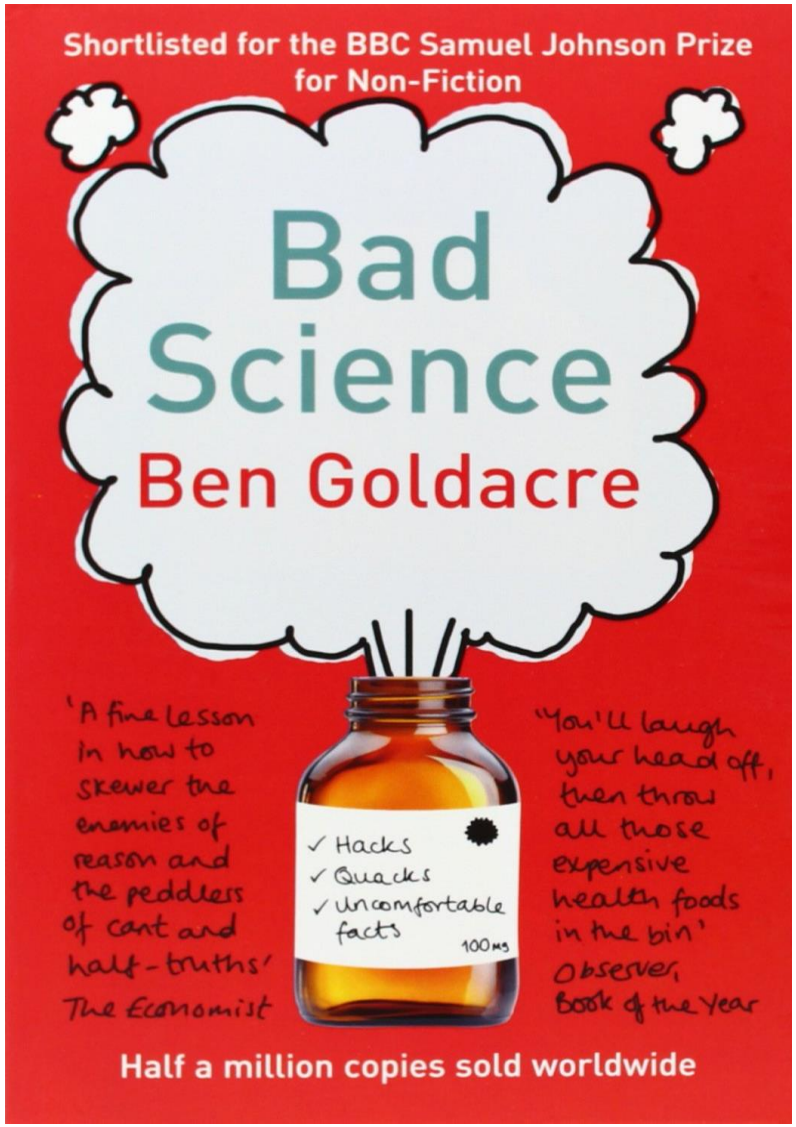
Descriptive Statistics - summarise the data

Inferential Statistics - analyse data & interpret the results

Two guiding principles

Repeatable analyses

Understandable reporting (as per *Lancet* guidelines)



GERIATRICS/ORIGINAL RESEARCH

Paramedic Assessment of Older Adults After Falls, Including Community Care Referral Pathway: Cluster Randomized Trial

Helen A. Snooks, PhD*; Rebecca Anthony; Robin Chatters; Jeremy Dale, PhD; Rachael T. Fothergill, Dr (Clinical); Sarah Gaze; Mary Halter, PhD; Ioan Humphreys; Marina Koniotou; Phillipa Logan, PhD; Ronan A. Lyons, PhD; Suzanne Mason, PhD; Jon Nicholl, PhD; Julie Peconi, PhD; Ceri Phillips, PhD; Alison Porter, PhD; Aloysius Niroshan Siriwardena, PhD; Mushtaq Wani; Alan Watkins, PhD; Lynsey Wilson; Ian T. Russell, PhD

*Corresponding Author. E-mail: h.a.snooks@swansea.ac.uk, Twitter: @HSRSwansea.

Study objective: We aim to determine clinical and cost-effectiveness of a paramedic protocol for the care of older people who fall.

Methods: We undertook a cluster randomized trial in 3 UK ambulance services between March 2011 and June 2012. We included patients aged 65 years or older after an emergency call for a fall, attended by paramedics based at trial stations. Intervention paramedics could refer the patient to a community-based falls service instead of transporting the patient to the emergency department. Control paramedics provided care as usual. The primary outcome was subsequent emergency contacts or death.

Results: One hundred five paramedics based at 14 intervention stations attended 3,073 eligible patients; 110 paramedics based at 11 control stations attended 2,841 eligible patients. We analyzed primary outcomes for 2,391 intervention and 2,264 control patients. One third of patients made further emergency contacts or died within 1 month, and two thirds within 6 months, with no difference between groups. Subsequent 999 call rates within 6 months were lower in the intervention arm (0.0125 versus 0.0172; adjusted difference −0.0045; 95% confidence interval −0.0073 to −0.0017). Intervention paramedics referred 8% of patients (204/2,420) to falls services and left fewer patients at the scene without any ongoing care. Intervention patients reported higher satisfaction with interpersonal aspects of care. There were no other differences between groups. Mean intervention cost was $23 per patient, with no difference in overall resource use between groups at 1 or 6 months.

Conclusion: A clinical protocol for paramedics reduced emergency ambulance calls for patients attended for a fall safely and at modest cost. [Ann Emerg Med. 2017;■:1-11.]

Please see page XX for the Editor's Capsule Summary of this article.

0196-0644/$-see front matter

http://dx.doi.org/10.1016/j.annemergmed.2017.01.006

INTRODUCTION

Background

Falls in older people are an important issue internationally,[1,2] with high human and organizational costs. It is estimated that approximately 30% of home-dwelling people aged 65 years or older fall every year.[3-5] Falls are associated with high mortality, morbidity, and immobility.[5] Recovery from fall injury is often delayed in older people, which increases the risk of subsequent falls.[5] In the United Kingdom, falls account for 3% (approximately $1,312 [£980] million) of total National Health Service (NHS) expenditure,[6] and the prevention of falls in older people has been highlighted as a priority.[7]

Population growth and ageing, the increasing burden of chronic disease, and shortage of health care workers are affecting health care systems in many countries.[8] NHS emergency departments (EDs) are under considerable pressure, and crowding is a major international problem with negative consequences for both patients and providers.

Emergency ambulances (999) are frequently called for older people who have fallen, composing 8% of emergency ambulance attendances in London, UK,[9] with a similar proportion reported in urban emergency medical services in the United States.[10] In the United Kingdom, United States, and Australia,[11] nonconveyance rates are high for this group;

CTU Statistician: Role and responsibilities

1. To lead on development and execution of Statistical Analysis Plans, including sample size considerations and data specification, data validation and processing; summarising the analysis from CONSORT flow chart, demographic and baseline descriptions, and a full assessment (adjusted for case-mix) effect of the intervention under investigation.

2. To liaise on Health Economics to ensure consistency in analytical methods and reported findings

3. To respond to statistically-related comments from colleagues, reviewers and editors.

Statistical Methods: Outline

- Do we even need statistics?
- Can we collect useful (right?) data?
- Where does the analysis fit?
  - Question ----------→Answer
  - Question>**Data** ---------→Answer
  - Question>***Data***>**Statistics**>Answer

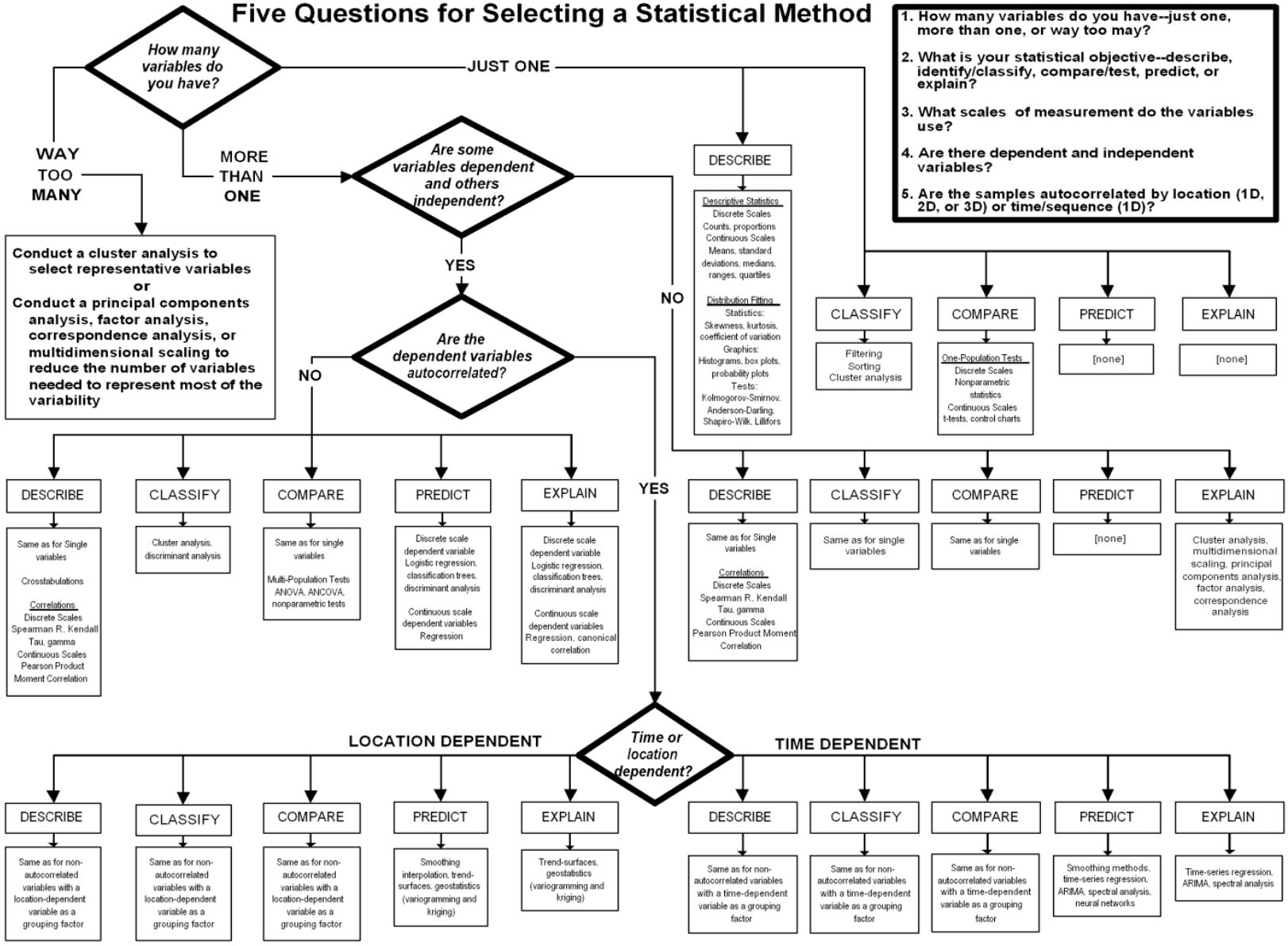
Data Management/Statistical Analysis Plans/Syntax or scripts

Key Statistical Principles …

- Draw some pictures
  - choose appropriate graphical tools for data
- Summarise the amount of data and the data itself
  - choose appropriate numerical measures
  - interpret these summaries correctly
- Understand the basic concepts (& limitations) of statistical analysis
  - interpret (possibly, calculate) confidence intervals and p-values for key hypotheses
  - flow-charts can help, up to a point ….

Five Questions for Selecting a Statistical Method

1. How many variables do you have--just one, more than one, or way too may?
2. What is your statistical objective--describe, identify/classify, compare/test, predict, or explain?
3. What scales of measurement do the variables use?
4. Are there dependent and independent variables?
5. Are the samples autocorrelated by location (1D, 2D, or 3D) or time/sequence (1D)?



For discussion [1]

- See three diagrams taken from published work.
- Rate each of the three diagrams as **Poor, Acceptable** or **Good**
  - If **Poor**, could you suggest some necessary improvements?
  - If **Good**, could you summarise its message?
  - If **Acceptable**, could you summarise its message **and** suggest possible improvements?

Broman et al

American Journal of Human Genetics (1998) 63:861-869

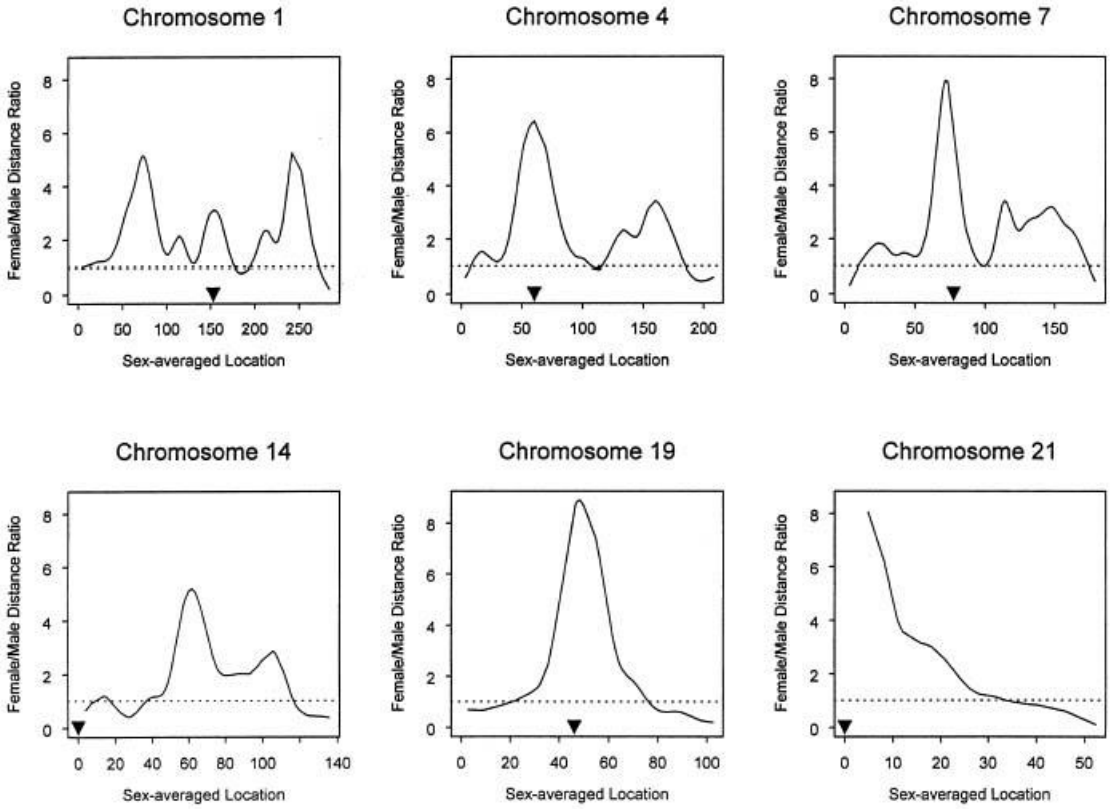


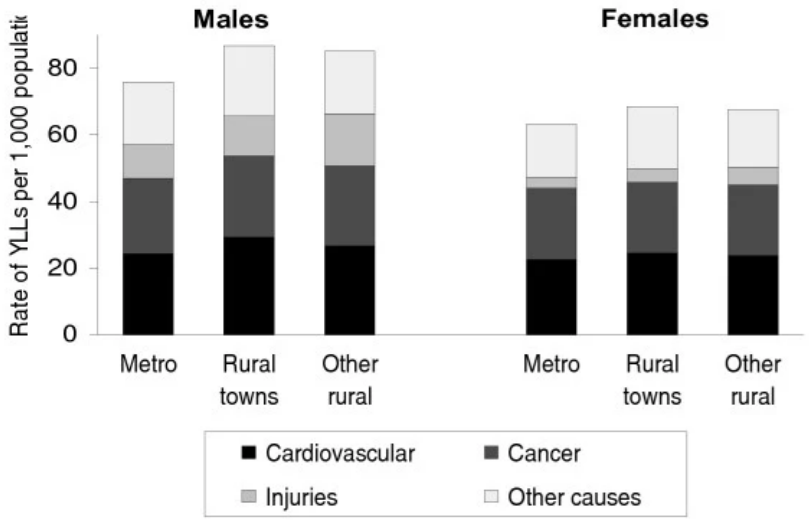
Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

Muscatello et al

BMC Medicine (2006) 4:33

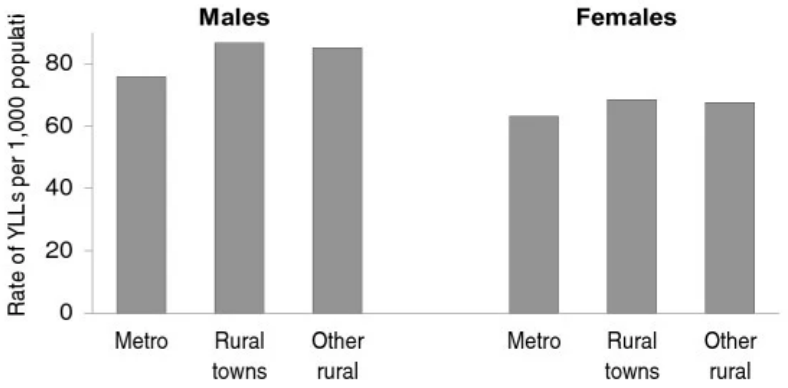
Control graph

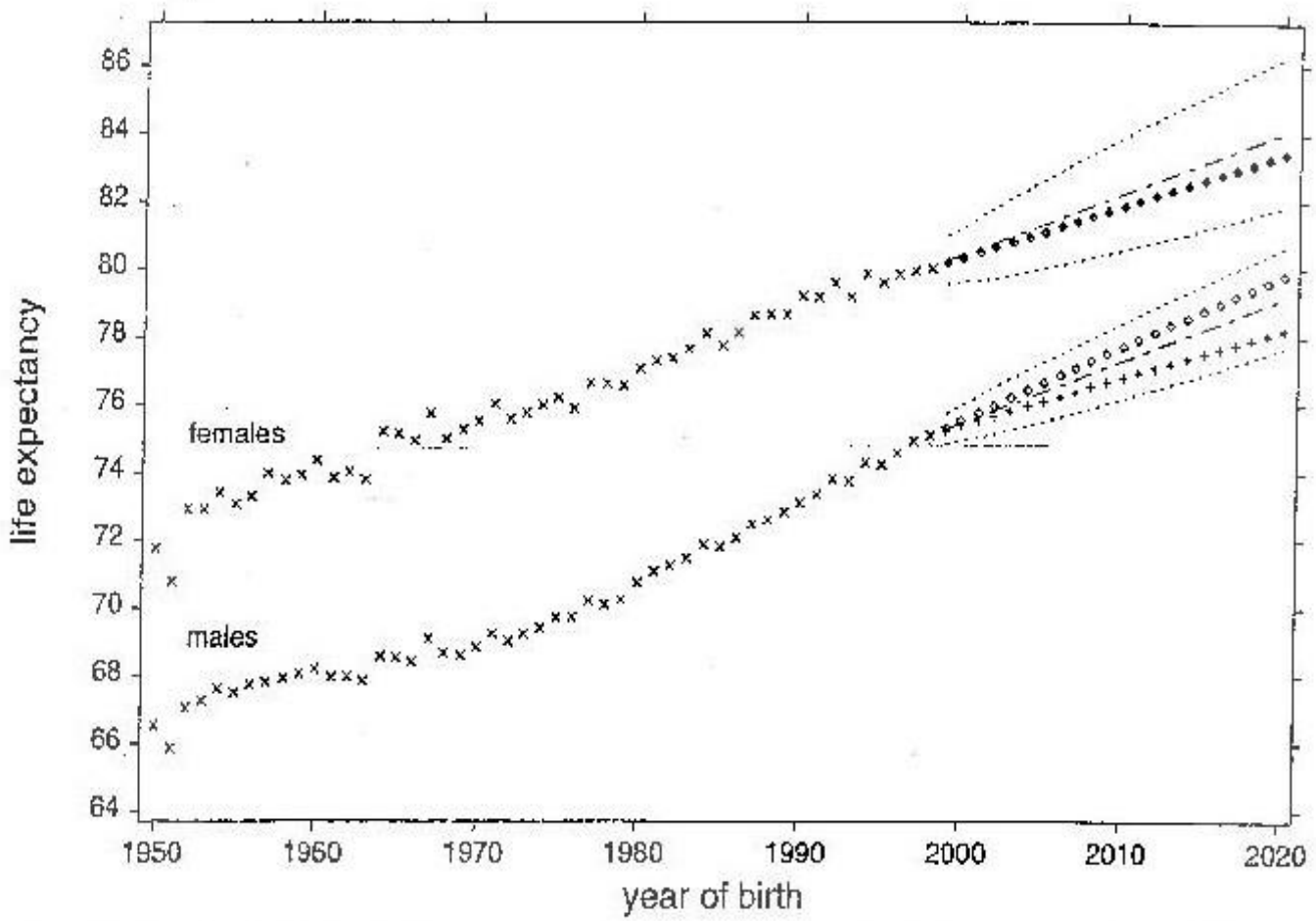
Illustration D: Rates of YLLs by Rurality Status, Sex and Major Causes of Death



Intervention graph

Illustration D: Rates of YLLs by Rurality, Status and Sex





. 5. Life expectancy at birth, by year, with forecasts, for each gender: - — - —, ARIMA(1,1,0); ·····, MA limits; ◇, GLM; +, LC; ×, empirical

For the female experience, in which log-linearity over time is a dominating influence (as dis-sed in Section 4), the forecasts based on both the LC and the GLM models are, for all

Wainer

The American Statistician (1984) 38:2

How to Display Data Badly

HOWARD WAINER*

Methods for displaying data badly have been developing for many years, and a wide variety of interesting and inventive schemes have emerged. Presented here is a synthesis yielding the 12 most powerful techniques that seem to underlie many of the realizations found in practice. These 12 (the dirty dozen) are identified and illustrated.

KEY WORDS: Graphics; Data display; Data density; Data-ink ratio.

1. INTRODUCTION

The display of data is a topic of substantial contemporary interest and one that has occupied the thoughts categorized. This article is the beginning of such a compendium.

The aim of good data graphics is to display data accurately and clearly. Let us use this definition as a starting point for categorizing methods of bad data display. The definition has three parts. These are (a) showing data, (b) showing data accurately, and (c) showing data clearly. Thus, if we wish to display data badly, we have three avenues to follow. Let us examine them in sequence, parse them into some of their component parts, and see if we can identify means for measuring the success of each strategy.

2. SHOWING DATA

Obviously, if the aim of a good display is to convey information, the less information carried in the display,

CONSORT 2010 Key Documents

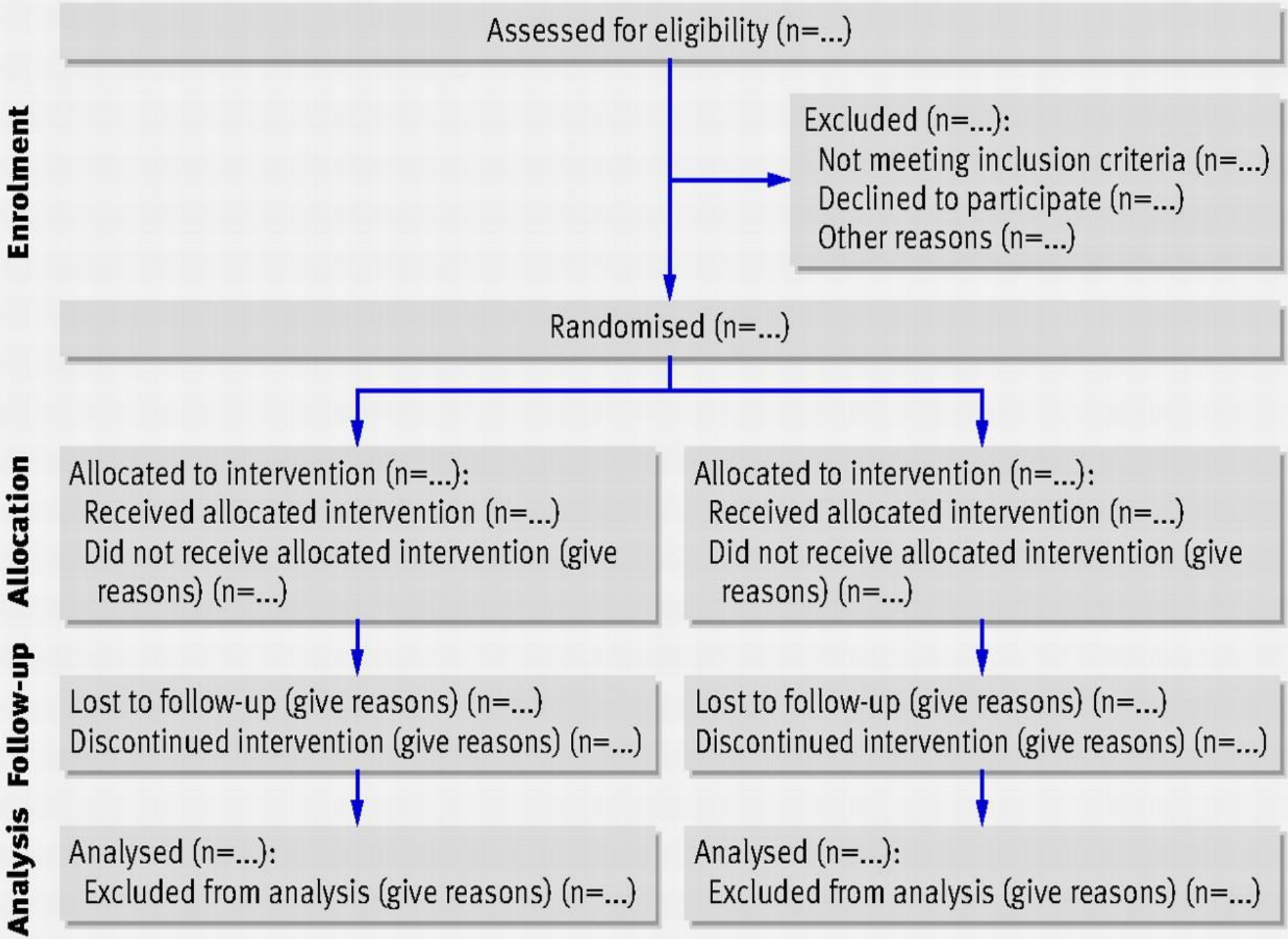
- CONSORT 2010 Checklist
- CONSORT 2010 Flow Diagram
- CONSORT 2010 Statement
- CONSORT 2010 Explanation and Elaboration Document

Welcome to the CONSORT Website

CONSORT stands for Consolidated Standards of Reporting Trials and encompasses various initiatives developed by the CONSORT Group to alleviate the problems arising from inadequate reporting of randomized controlled trials.

The CONSORT Statement

The main product of CONSORT is the CONSORT Statement, which is an evidence-based, minimum set of recommendations for reporting randomized trials. It offers a standard way for authors to prepare reports of trial findings, facilitating their complete and transparent reporting, and aiding their critical appraisal and interpretation.


Enrolment
Assessed for eligibility (n=...)
Excluded (n=...):
Not meeting inclusion criteria (n=...)
Declined to participate (n=...)
Other reasons (n=...)
Randomised (n=...)
Allocation
Allocated to intervention (n=...):
Received allocated intervention (n=...)
Did not receive allocated intervention (give reasons) (n=...)
Allocated to intervention (n=...):
Received allocated intervention (n=...)
Did not receive allocated intervention (give reasons) (n=...)
Follow-up
Lost to follow-up (give reasons) (n=...)
Discontinued intervention (give reasons) (n=...)
Lost to follow-up (give reasons) (n=...)
Discontinued intervention (give reasons) (n=...)
Analysis
Analysed (n=...):
Excluded from analysis (give reasons) (n=...)
Analysed (n=...):
Excluded from analysis (give reasons) (n=...)

The CONSORT Industry: Checklists & extensions….

1. Reporting check-lists (25 items, sometimes more…);
2. For cluster trials;
3. For feasibility studies;
4. For abstracts;
5. For stepped wedge designs…

Required for submission to many journals (*BMJ, Lancet*, …)

See the airAware paper

Other checklists are available…

Kwakkenbos et al. Research Integrity and Peer Review (2018) 3:9
https://doi.org/10.1186/s41073-018-0053-3

Research Integrity and Peer Review

STUDY PROTOCOL Open Access

Protocol for the development of a CONSORT extension for RCTs using cohorts and routinely collected health data

Linda Kwakkenbos[1], Edmund Juszczak[2], Lars G Hemkens[3], Margaret Sampson[4], Ole Fröbert[5], Clare Relton[6], Chris Gale[7], Merrick Zwarenstein[8,9], Sinéad M Langan[10], David Moher[11], Isabelle Boutron[12,13,14], Philippe Ravaud[12,13,14], Marion K Campbell[15], Kimberly A Mc Cord[3], Tjeerd P van Staa[16,17], Lehana Thabane[18], Rudolf Uher[19], Helena M Verkooijen[20,21], Eric I Benchimol[22,23,24], David Erlinge[25], Maureen Sauvé[26,27], David Torgerson[28] and Brett D Thombs[29,30,31,32,33,34*]

Abstract

Background: Randomized controlled trials (RCTs) are often complex and expensive to perform. Less than one third achieve planned recruitment targets, follow-up can be labor-intensive, and many have limited real-world generalizability. Designs for RCTs conducted using cohorts and routinely collected health data, including registries, electronic health records, and administrative databases, have been proposed to address these challenges and are being rapidly adopted. These designs, however, are relatively recent innovations, and published RCT reports often do not describe important aspects of their methodology in a standardized way. Our objective is to extend the Consolidated Standards of Reporting

STROBE Statement

Strengthening the reporting of observational studies in epidemiology

Home
Aims
News
Available checklists
Publications
Translations
Commentaries
Discussion forum
STROBE group
Endorsement
Contact
Links

What is STROBE?

STROBE stands for an international, collaborative initiative of epidemiologists, methodologists, statisticians, researchers and journal editors involved in the conduct and dissemination of observational studies, with the common aim of **STrengthening the Reporting of OBservational studies in Epidemiology**.

The STROBE Statement is being endorsed by a growing number of biomedical journals. Click here for full list.

For STROBE-related entries in PubMed click here.

What's new in the STROBE Initiative?

Observational Studies: Getting clear about transparency

New guidelines for observational studies in PLOS Medicine

Read more

01.09.2014

New article of interest

A Review of Published Analyses of Case-Cohort Studies and Recommendations for Future Reporting

Read more

01.07.2014

Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): an extension of the STROBE

Data management & processing

- Collate and code data

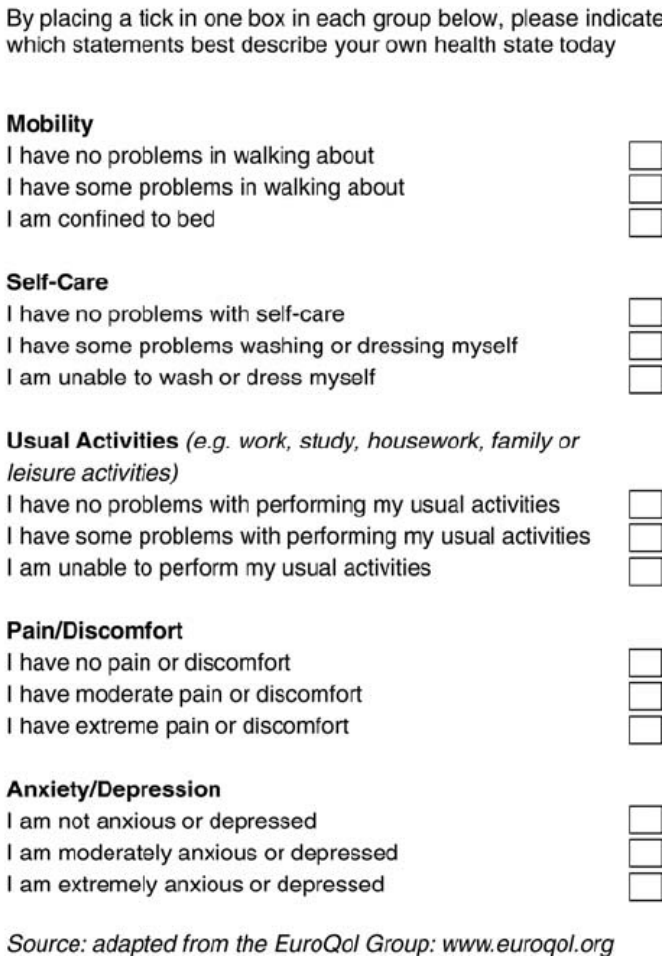
- Compute new variables from old
  - BMI from height & weight
  - Patient Reported Outcome Measures (PROMS)
    - EQ5D, SF12 (general)
    - EORTC, CUCQ (condition specific)
  - Quality Adjusted Life Years (QALYs)
  - Indicators of events; time between events

Quality of Life ~ EQ5D (3 level)

- Score each question 1,2,3
- Convert 5 scores into a single number

- 11111 1.0
- 22222 0.516
- 12321 0.329
- 21223 0.222
- 23322 0.079
- 33332 -0.429

- (Dead = 0)



By placing a tick in one box in each group below, please indicate which statements best describe your own health state today

Mobility
I have no problems in walking about ☐
I have some problems in walking about ☐
I am confined to bed ☐

Self-Care
I have no problems with self-care ☐
I have some problems washing or dressing myself ☐
I am unable to wash or dress myself ☐

Usual Activities *(e.g. work, study, housework, family or leisure activities)*
I have no problems with performing my usual activities ☐
I have some problems with performing my usual activities ☐
I am unable to perform my usual activities ☐

Pain/Discomfort
I have no pain or discomfort ☐
I have moderate pain or discomfort ☐
I have extreme pain or discomfort ☐

Anxiety/Depression
I am not anxious or depressed ☐
I am moderately anxious or depressed ☐
I am extremely anxious or depressed ☐

Source: adapted from the EuroQol Group: www.euroqol.org

Data management & processing

Collect and prepare data with planned analysis in mind

Aim to produce “clean” & full rectangular data set to import into R, SPSS, Stata, SAS, MLWin,…

- (but, ideally, *NOT* Excel! – why?)

Safari File Edit View History Bookmarks Window Help Mon 17:53 Alan Watkins

Microsoft Remote Desktop

PCs Workspaces Search

Search AW

Search

bbc.co.uk

BBC - Home | Sign in to your account | Home - Londo...ck Exchange | Portfolio | int...ctive investor | Apple | iCloud | Google | Yahoo | Wikipedia | Facebook | Twitter | BBC | Sign In | Inbox - Outloo...light version

My Collection - Discogs

Excel: Why using Microsoft's tool caused Covid-19 results to be lost - BBC News

BBC Sign in Home News Sport Weather iPlayer Sounds CBBC CBeebies More Search

NEWS

Home | Coronavirus | US Election | UK | World | Business | Politics | Tech | Science | Health | Family & Education | More

Technology

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

6 days ago

Coronavirus pandemic

Top Stories

Liverpool at 'very high' level in Covid tier system

The city's pubs and betting shops will close on Wednesday as part of a three-tier system in England

3 minutes ago

Johnson details three-tier Covid rules for England

11 hours ago

Nightingale hospitals told 'get ready for Covid'

1 PC

] - IBM SPSS Statistics Data Editor

Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1 Visible: 10 of

D	Programme	Group	Gender_FM	Gender	Age_Year	Age_Decade	Endurance_Before	Endurance_After	Endurance_Change
1	Standard	0	Female	1	48	2	239.0	269.3	30.3
2	Standard	0	Female	1	21	0	213.6	313.4	99.8
3	Enhanced	1	Female	1	36	1	434.4	513.7	79.2
4	Standard	0	Female	1	26	0	343.6	416.5	72.9
5	Enhanced	1	Female	1	23	0	676.3	780.1	103.8
6	Enhanced	1	Male	0	25	0	738.6	822.9	84.3
7	Standard	0	Female	1	29	0	282.5	345.9	63.5
8	Enhanced	1	Male	0	48	2	502.5	598.0	95.5
9	Enhanced	1	Female	1	30	1	462.4	554.7	92.3
10	Enhanced	1	Female	1	34	1	766.1	853.2	87.1
11	Standard	0	Female	1	44	2	647.9	754.2	106.3
12	Standard	0	Female	1	33	1	424.8	515.7	90.9
13	Enhanced	1	Female	1	21	0	493.9	601.8	107.9
14	Standard	0	Female	1	35	1	550.0	644.9	94.9
15	Standard	0	Male	0	35	1	671.7	757.0	85.4
16	Standard	0	Male	0	33	1	463.7	558.6	94.9
17	Standard	0	Female	1	46	2	640.2	711.9	71.7
18	Standard	0	Male	0	31	1	351.3	404.8	53.5
19	Standard	0	Male	0	24	0	209.8	271.7	61.9
20	Enhanced	1	Male	0	39	1	213.6	252.4	38.8
21	Standard	0	Female	1	39	1	756.4	843.3	86.9
22	Standard	0	Male	0	48	2	712.8	782.6	69.8
23	Standard	0	Female	1	40	2	794.9	866.2	71.3
24	Enhanced	1	Female	1	42	2	229.5	298.8	69.3
25	Standard	0	Male	0	36	1	593.2	658.1	64.8
26	Enhanced	1	Female	1	44	2	459.5	518.6	59.1

karela_polyps.sav [DataSet5] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : Group | 3 | Visible: 20 of 20 Variables

	Group	EOTAXIN	LNeo	RANTES	LNran	IL_5	LNil5	TOTIGE	LNtot	PHAD	MX2	SPT	LundMAC
1	3	55.92	4.02	261.84	5.57	52.75	3.97	31.30	3.44	1	1	1	15
2	1	.	.	.	.	.	.	61.95	4.13	1	1	.	6
3	2	132.20	4.88	530.52	6.27	37.98	3.64	24.35	3.19	0	0	0	.
4	2	8.19	2.10	43.74	3.78	3.86	1.35	4.75	1.56	0	0	0	5
5	2	22.10	3.10	1154.69	7.05	54.96	4.01	14.00	2.64	0	0	.	6
6	3	165.79	5.11	993.46	6.90	26.86	3.29	118.50	4.77	1	1	1	.
7	3	66.64	4.20	1548.31	7.34	52.48	3.96	18.70	2.93	0	0	0	12
8	2	50.11	3.91	505.11	6.22	26.35	3.27	123.30	4.81	1	1	.	14
9	2	88.33	4.48	1388.41	7.24	21.83	3.08	34.90	3.55	1	1	.	12
10	1	22.71	3.12	1375.48	7.23	28.17	3.34	13.90	2.63	.	0	.	8
11	3	24.83	3.21	1169.80	7.06	13.66	2.61	446.00	6.10	1	1	.	18
12	1	12.31	2.51	673.24	6.51	9.62	2.26	40.75	3.71	0	1	0	6
13	1	.	.	.	.	.	.	6.10	1.81	0	0	.	.
14	1	15.72	2.75	1735.41	7.46	4.18	1.43	24.25	3.19	1	1	1	6
15	2	17.54	2.86	2073.91	7.64	11.09	2.41	28.50	3.35	.	1	0	7
16	2	10.69	2.37	1664.62	7.42	19.95	2.99	71.15	4.26	1	1	1	14
17	3	11.39	2.43	545.93	6.30	12.76	2.55	7.20	1.97	0	0	0	12
18	1	51.87	3.95	1306.23	7.17	38.89	3.66	38.25	3.64	1	1	1	9
19	3	83.49	4.42	869.11	6.77	17.20	2.84	56.05	4.03	1	1	1	12
20	3	61.41	4.12	2019.14	7.61	23.02	3.14	135.00	4.91	0	1	0	.
21	1	26.15	3.26	1392.78	7.24	29.19	3.37	494.00	6.20	1	1	.	12
22	2	32.43	3.48	1087.67	6.99	12.73	2.54	141.00	4.95	0	1	0	9
23	0	.	.	.	.	.	.	21.60	3.07	0	0	0	.
24	0	.	.	.	.	.	.	46.80	3.85	0	0	1	.
25	0	.	.	.	.	.	.	15.10	2.71	0	0	1	.
26	4	.	.	.	.	.	.	44.90	3.80	1	1	1	6
27	0	.	.	.	.	.	.	50.40	3.92	0	0	0	.
28	1	.	.	.	.	.	.	10.30	2.33	0	0	0	3
29	4	.	.	.	.	.	.	97.10	4.58	1	1	1	.
30	1	.	.	.	.	.	.	46.00	3.83	1	1	1	6
31													
32													
33													
34													
35													
36													

Data View | Variable View

SPSS Processor is ready

Data Types

- Indicator/Nominal: different categories
  - Often binary
    - Absent/Present (coded as 0,1)
    - Male/Female
    - Control/Intervention
  - Red/Green/Blue (coded ?)
- Ordinal: ordered categories:
  - Small/Medium/Large (coded ?)
  - Likert scales (5 or 7 or 9 points)
    - Very sure/Somewhat sure/Neither sure or unsure/Somewhat unsure/Very unsure
- Measurement: discrete/count
  - Number of operations
- Measurement: continuous/intervals
  - not necessarily a whole number
  - time between 999 call and ambulance arrival

Descriptive statistics & beyond …

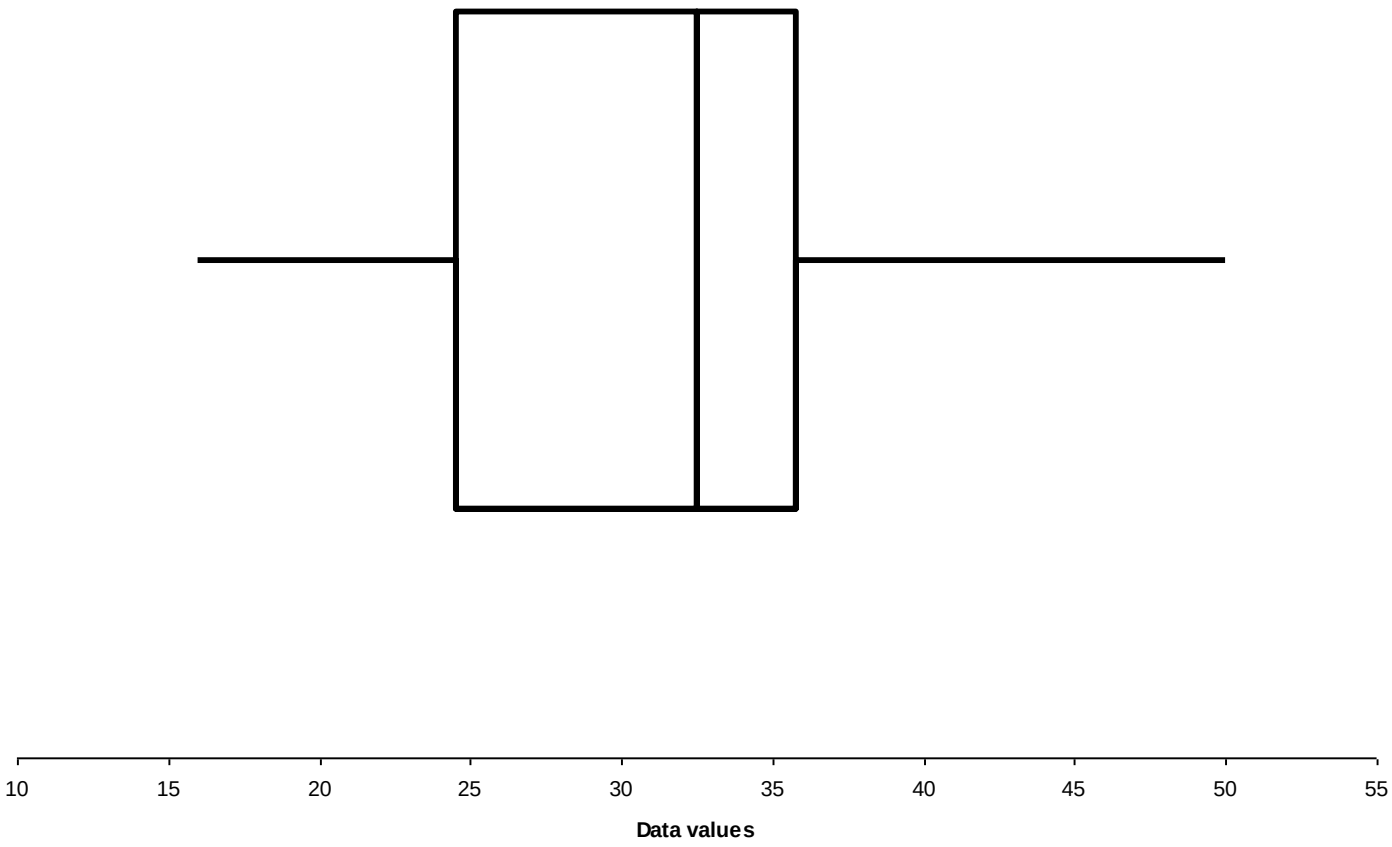
- Twin aims of statistics
  - Produce graphical and numerical summaries
    - Reflecting nature & quantity of data
  - Study and explain location and spread
    - Systematic variation versus randomness
    - Patterns -> Equations -> Models ->Effect size -> Predictions -> Cost Savings
    - Randomness: perhaps nothing more to be said, from the present data
    - Equations (*"models"*) reflect assumptions on data

Graphical Representations

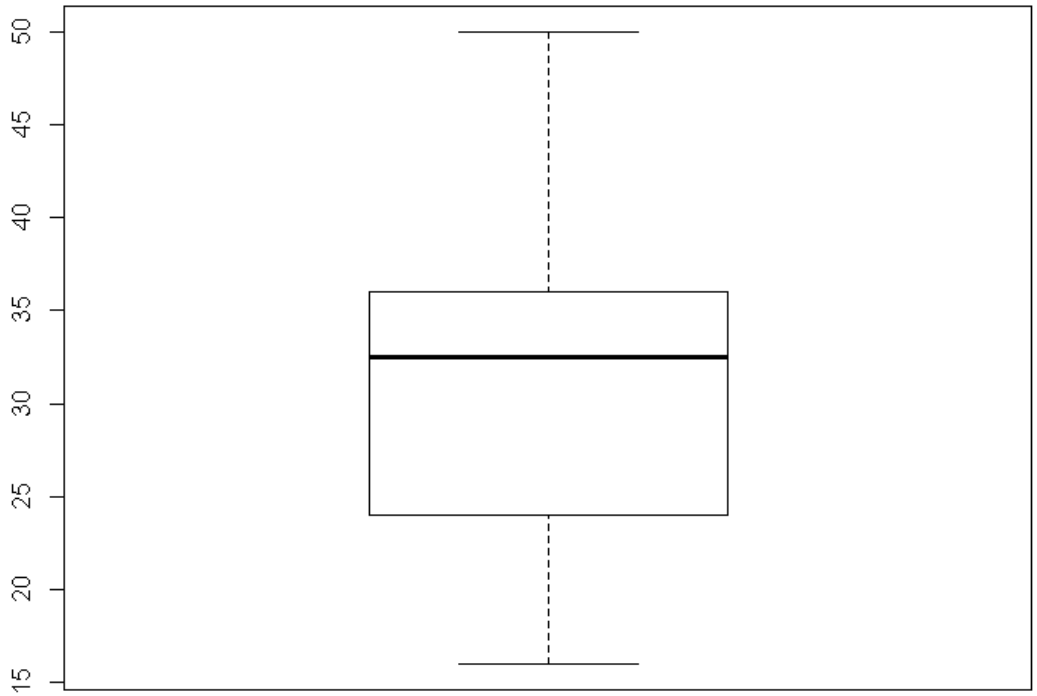
- Draw some pictures that go beyond basic bar and pie charts
  - Box Plots
  - Cumulative Frequency Plots
  - Histograms
- Box (or *Box & Whisker*) Plots
  - shows spread & location
  - box from first to third quartile
  - line in the box to show median
  - whiskers from box to minimum and maximum values
  - extreme values marked as potential outliers

Box-plot:

(data values: 16,31,34,24,42,19,36,26,50,35)

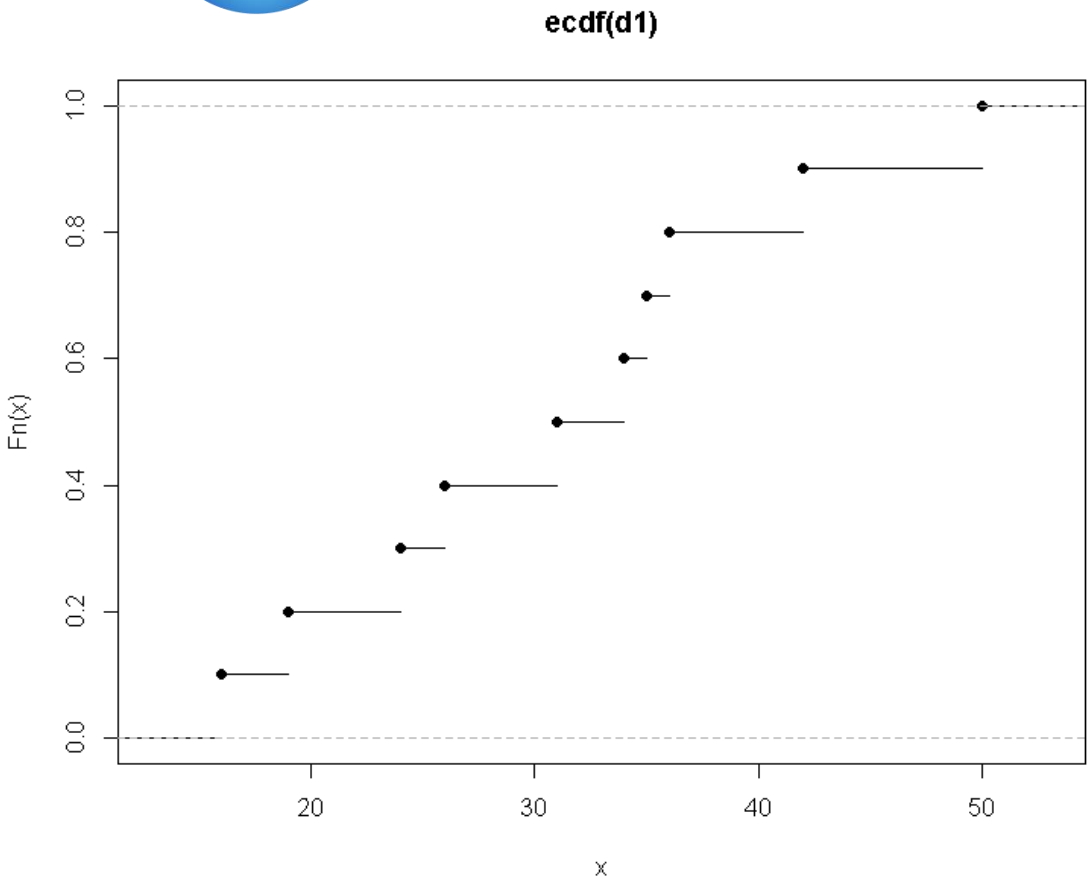


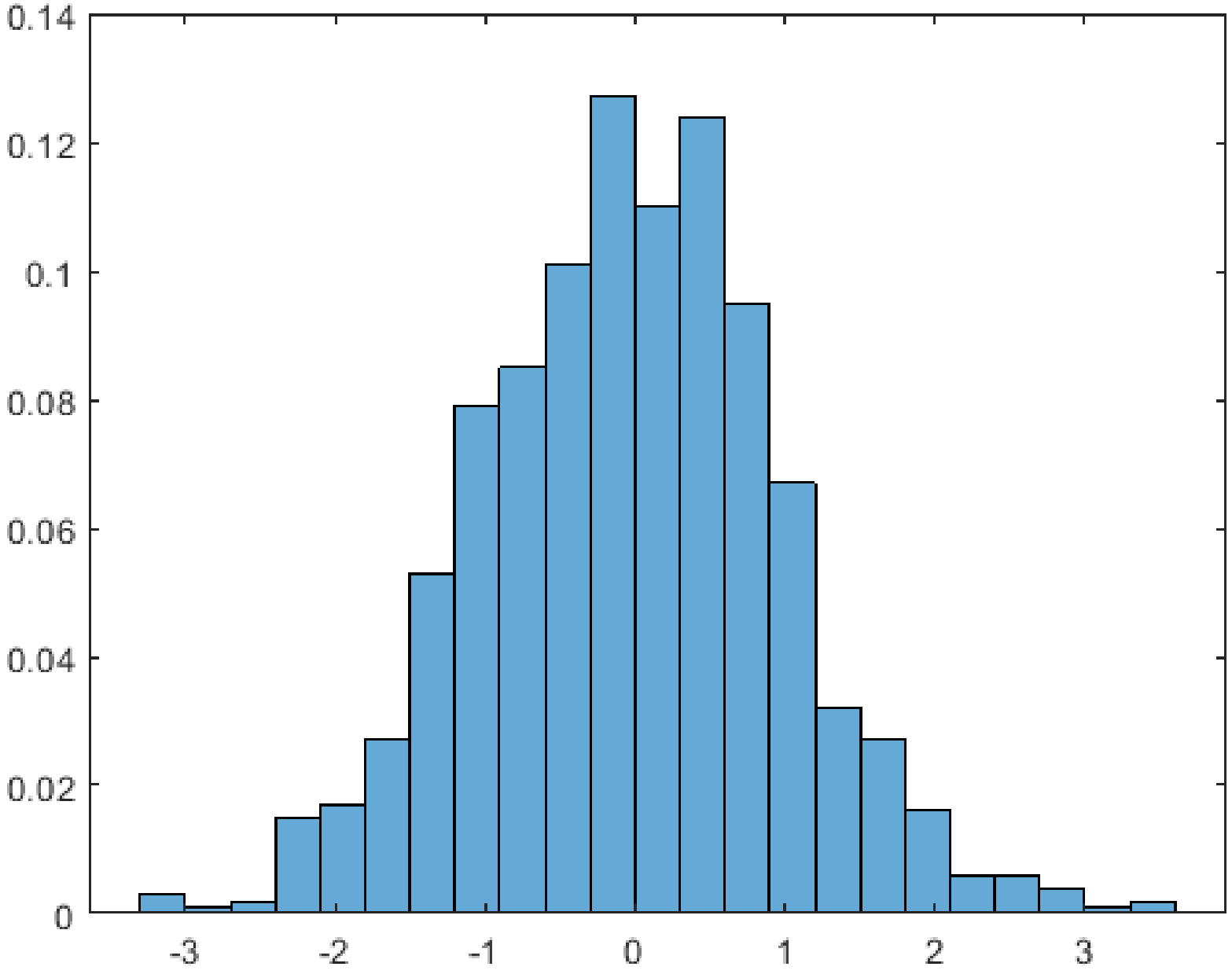
```
d1<-c(16,31,34,24,42,19,36,26,50,35)
boxplot(d1)
```



Cumulative distribution function

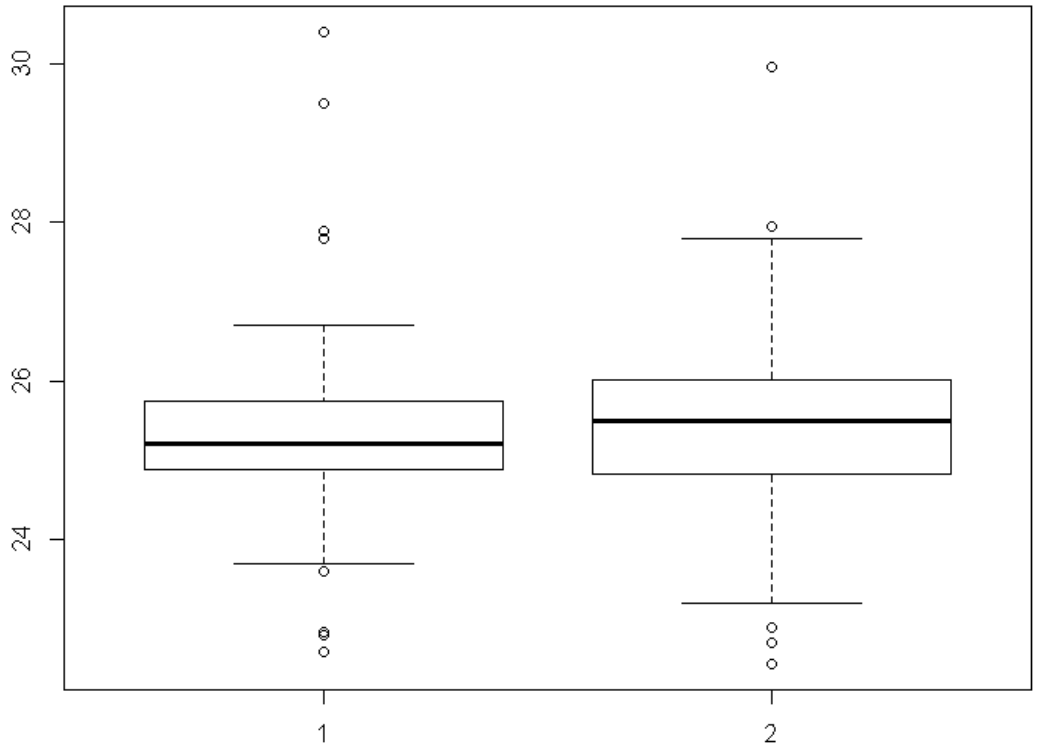
plot(ecdf(d1))




0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-3
-2
-1
0
1
2
3

Compare two groups

boxplot(d2~group)



Numerical summaries

d1<-c(16,31,34,24,42,19,36,26,50,35)

- Then: mean(d1)
  - sd(d1) — *standard deviation (spread)*
  - range(d1) — *minimum and maximum*
  - min(d1) & max(d1) — *or separately*
  - median(d1) — *middle value*
    - quantile(d1,0.5) — *!note spelling!*
  - quantile(d1,0.25) — *first quartile (middle of lower half)*
  - quantile (d1,0.75) — *third quartile*
    - quantile(d1,c(0.25, 0.5, 0.75))
  - summary(d1)
    - Gives Min/1st Quartile/Median/Mean/3rd Quartile/Max
    - 16.00 24.50 32.50 31.30 35.75 50.00 (for d1,as above)

Descriptive & Inferential Methods

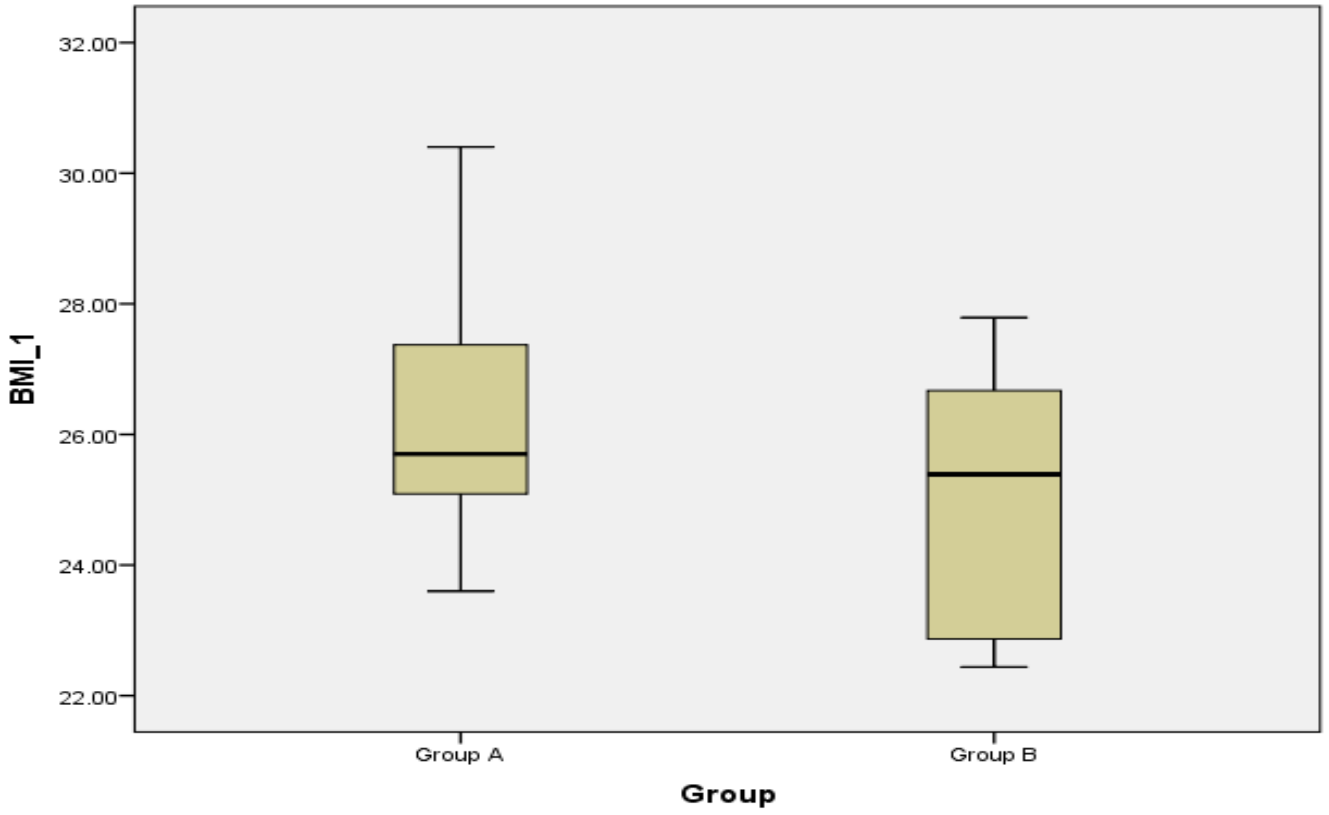
- ***Description***: summarise & illustrate data
- ***Inference***: to explain what we see or what can be shown in data
  - Analysis takes account of sampling variation
  - Analyse the sample that you have, while taking into account the samples you *might* have collected

Some Statistical Jargon

- Populations, Normality, effect size
- Null & alternative hypotheses
- α (significance level) & β (power)
- Type I & type II errors

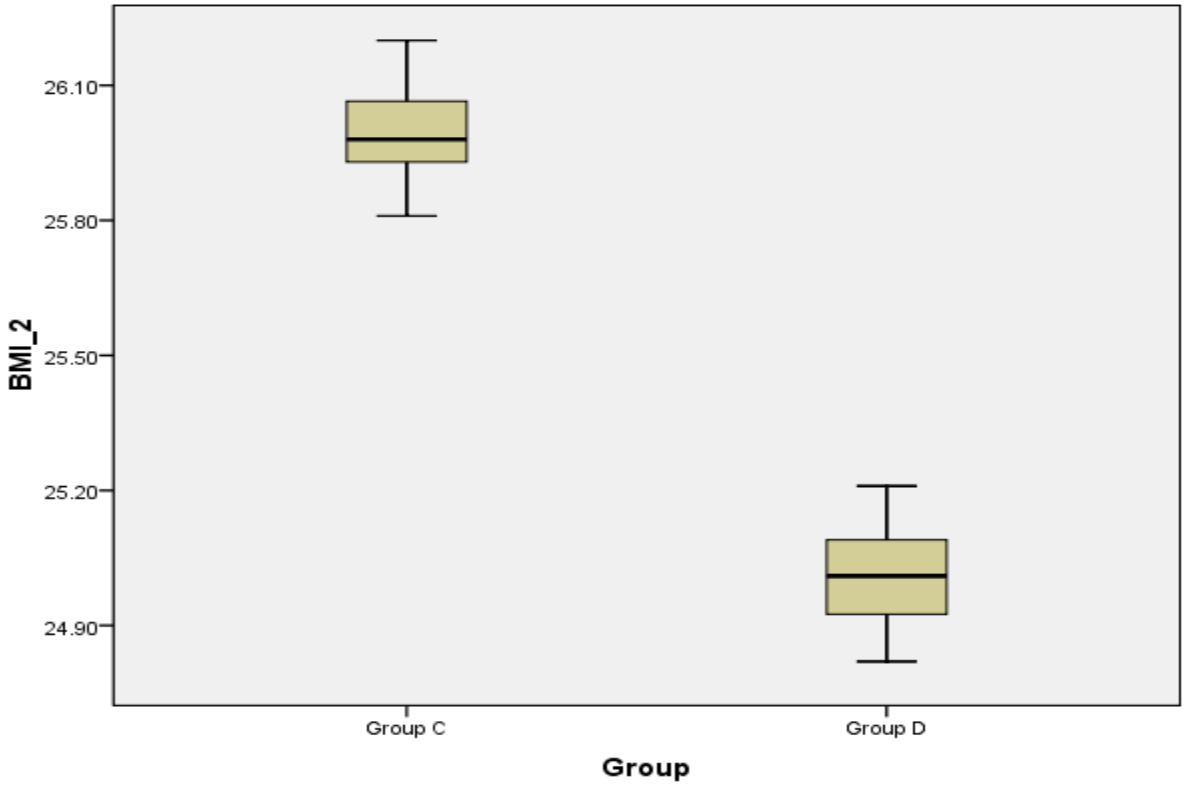
Focus on location and variation

- Side by side box-plots of BMI
  - immediate comparison of two groups (A & B)



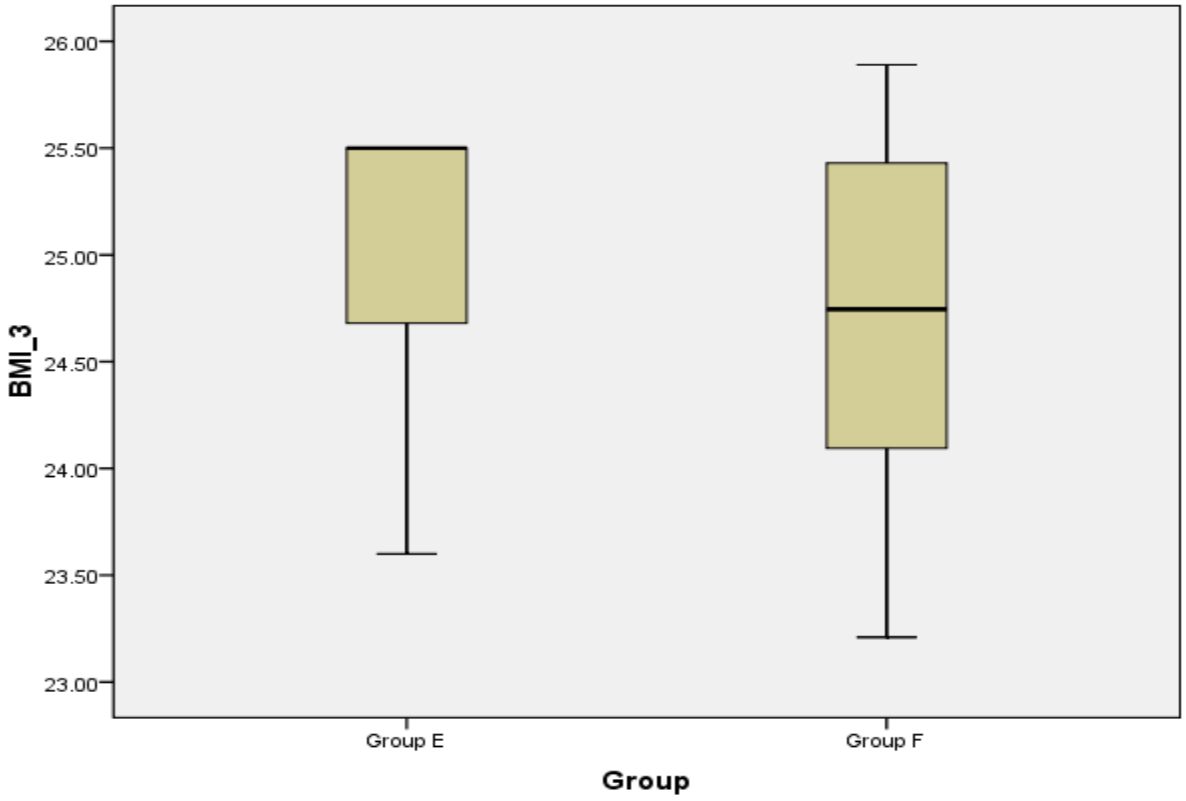
Different patterns of variation

- Box-plots of BMI for two more groups C & D



Variation & Its Many Forms

- more BMI box-plots … (groups E & F)



For discussion [2]

- Consider box-plots A & B
  - what (if any) tentative conclusions may be reached on similarities (or differences) between the two groups?
- What about C & D? E & F?
- Under what circumstances (or, in what trials) might we obtain a sample similar to C? E?

Univariate tests

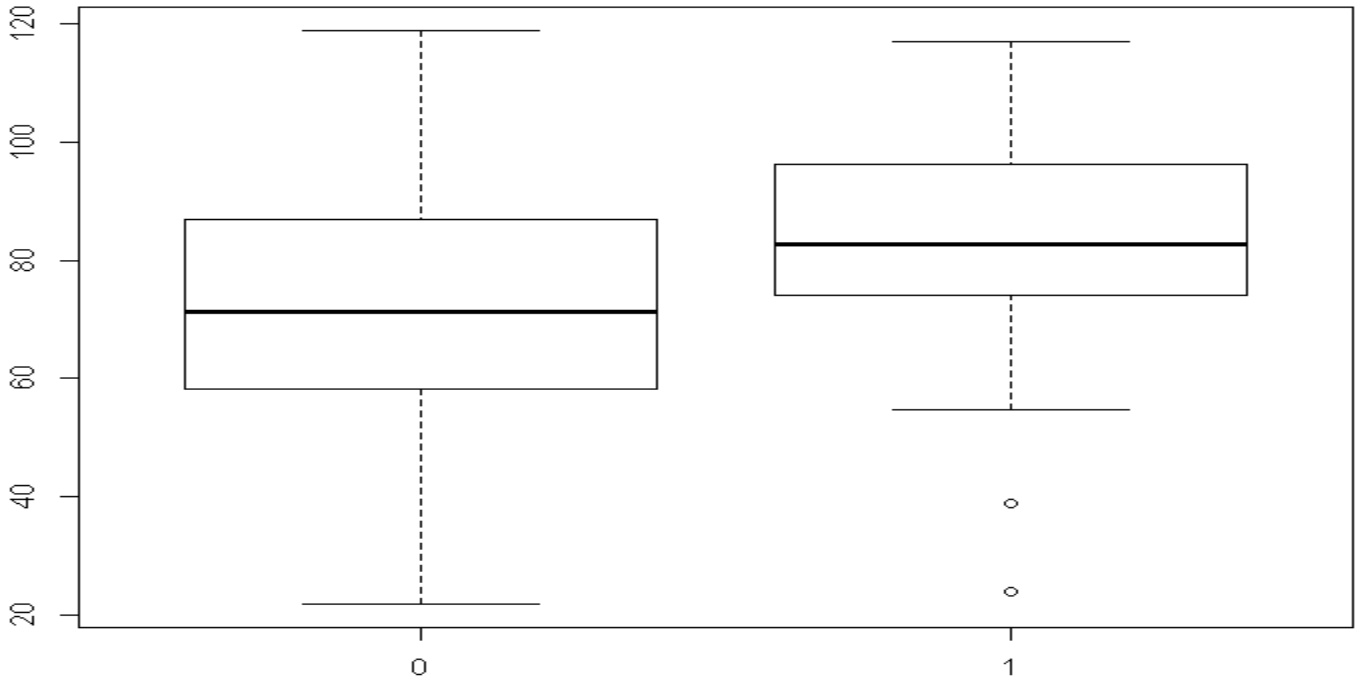
- Assess whether observed differences can be regarded as *statistically significant* taking within group variation into account.
  - *Use t-tests to compare means of measurement variables;*
  - *Use chi-squared to compare proportions of binary outcomes.*
  - *Assumptions (such as Normality) now appear*
  - *Alternative methods make different assumptions*
- Only context (Group membership) to explain why such differences may have arisen
  - For instance, A=control, B=intervention.

Confidence intervals & p-values

- Confidence interval (usually 95%)
  - A guide to plausible values of the true (unknown) population difference centered on the observed (sample) difference
  - Focus on whether this interval includes 0 or not
- p-value
  - Assuming the true difference is 0, the probability of observing a difference greater than that actually observed
  - Focus on whether this is less than 0.05 or not
- Relation
  - The 95% CI includes 0 is equivalent to $p > 0.05$
  - The 95% CI excludes 0 is equivalent to $p < 0.05$

Two sample t-test

- boxplot(change~group) to produce side by side box-plots
- gives immediate comparison of two groups



t.test(change~group)

- Welch Two Sample t-test data:
  - change by group
    - t = -3.7411, df = 123.15, p-value = 0.0002796
  - alternative hypothesis:
    - true difference in means is not equal to 0
  - 95 percent confidence interval:
    - -19.579558 -6.029615
  - sample estimates:
    - mean in group 0 mean in group 1
    - 71.30981 84.11440

Corresponding SPSS Output

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Endurance_Change	Standard	63	71.3098	19.99166	2.51871
	Enhanced	63	84.1144	18.39502	2.31756

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Endurance_Change	Equal variances assumed	.877	.351	-3.741	124	.000	-12.80459	3.42271	-19.57910	-6.03008
	Equal variances not assumed			-3.741	123.151	.000	-12.80459	3.42271	-19.57956	-6.02961

Results and Interpretation

- Summary statistics
- t-test statistic, df (degrees of freedom) and corresponding p-value
- Confidence interval for difference in means

- The observed difference is **statistically significant**

Equivalent forms of analysis

Statistical methods mean we can obtain essentially the same analysis in various ways:

- Two sample t-test
- Linear Model
- Regression

Some approaches lead more naturally to useful generalisations.

Categorical Data

- Data comprises two observed categorical responses
- Example: gender, and survival (an event during follow-up or not)
  - Variable 1: gender: male or female
  - Variable 2: survival: event during follow-up period or not
- Data

	female	male
Event in follow-up	28	48
Event-free	72	102

table(event, gender)

Chi-squared analysis of Cross-tabulations

chisq.test(table(event,gender),correct=FALSE)

- Pearson's Chi-squared test
- data: table(event, gender)
- X-squared = 0.45372, df = 1, p-value = 0.5006

- Compare this with output from chisq.test(table(event,gender),**correct=TRUE**)
  - Slightly different assumptions -> different answers

Remember that …

- Statistics generally appear both at the start (design) and end (analyis) of a trial.
  - Good statistical analysis can't completely overcome limitations of bad design
    - For instance, sample size too small
- Poor statistical analysis can almost ruin an otherwise well-designed trial.

For discussion [3]

- See four extracts taken from published work.
- Can you determine trial aim, sample size, study design?
  - Primary outcome? Secondary outcomes?
  - Detail of analysis undertaken? Findings?
  - Is the data available?
  - **Given the data, could one repeat the reported analyses?**

Miller et al: *BMJ* 2012

Abstract

Objectives To test the impact of provider performance pay for anaemia reduction in rural China.

Design A cluster randomised trial of information, subsidies, and incentives for school principals to reduce anaemia among their students. Enumerators and study participants were not informed of study arm assignment.

Setting 72 randomly selected rural primary schools across northwest China.

Participants 3553 fourth and fifth grade students aged 9-11 years. All fourth and fifth grade students in sample schools participated in the study.

Interventions Sample schools were randomly assigned to a control group, with no intervention, or one of three treatment arms: (*a*) an information arm, in which principals received information about anaemia; (*b*) a subsidy arm, in which principals received information and unconditional subsidies; and (*c*) an incentive arm, in which principals received information, subsidies, and financial incentives for reducing anaemia among students. Twenty seven schools were assigned to the control arm (1816 students at baseline, 1623 at end point), 15 were assigned to the information arm (659 students at baseline, 596 at end point), 15 to the subsidy arm (726 students at baseline, 667 at end point), and 15 to the incentive arm (743 students at baseline, 667 at end point).

Main outcome measures Student haemoglobin concentrations.

Results Mean student haemoglobin concentration rose by 1.5 g/L (95% CI −1.1 to 4.1) in information schools, 0.8 g/L (−1.8 to 3.3) in subsidy schools, and 2.4 g/L (0 to 4.9) in incentive schools compared with the control group. This increase in haemoglobin corresponded to a reduction in prevalence of anaemia (Hb <115 g/L) of 24% in incentive schools. Interactions with pre-existing incentives for principals to achieve good academic performance led to substantially larger gains in the information and incentive arms: when combined with incentives for good academic performance, associated effects on student haemoglobin concentration were 9.8 g/L (4.1 to 15.5) larger in information schools and 8.6 g/L (2.1 to 15.1) larger in incentive schools.

Conclusions Financial incentives for health improvement were modestly effective. Understanding interactions with other motives and pre-existing incentives is critical.

Trial registration number ISRCTN76158086.

Introduction

Inexpensive, efficacious technologies and services exist for improving human health in developing countries, but

Gussekloo et al

Abstract

The aim was to investigate the incidence rate of dementia for community residents aged 85 years and over.

It was a two wave community study of 224 subjects (community residents including those residing in a nursing home) older than 85 years, restudied 4·1 years after a community prevalence study. A two stage method was used, comprising the mini mental state examination followed in a stratified sample by the geriatric mental state schedule (A3)/AGECAT. Incidence rates were based on person-years at risk.

The overall incidence of dementia was 6·9 (95% confidence interval (95% CI) 4·8–9·1) per 100 person-years at risk. The incidence was significantly higher for women than for men; respectively 8·9 (95% CI 5·9–11·9) *v* 2·7 (95% CI 0·5–4·9) per 100 person-years at risk.

In the fastest growing age group seven out of 100 persons develop dementia each year. Women, who constitute two thirds of the oldest old, seem to have a higher risk. Further research is needed into the risk factors for dementia in this age group.

(*J Neurol Neurosurg Psychiatry* 1995;59:507–510)

Sheth et al (2012)

ABSTRACT

Objectives: To assess the immunization status of children of Gandhinagar (Rural) district and to compare it with the NFHS3/DLHS3 coverage results.

Materials & Methods: A Multi-Indicator Cluster Survey (MICS) was planned and community-based cross-sectional survey was conducted in April 2008. The Study was conducted using 30 cluster technique. Proforma designed by UNICEF, modified by experts and approved for uniform use by department of health & family welfare, Government of Gujarat was used as a study tool.

Statistical analysis used: Simple proportions and Chi-square test.

Results: Coverage for BCG, OPV3, DPT3 & Measles were 92.04%, 85.23%, 83.71% & 82.20% respectively. BCG scar was seen in 83.95% of children out of those who received BCG. The proportions of fully immunized children were 79.55%. Unimmunized children were 4.16%. Dropout rate was 9.05% for BCG-DPT_3, 10.69% for BCG-Measles & 7.53% for DPT_1-DPT_3. Compared to NFHS3 (2005-06) as well as DLHS3 (Gandhinagar district, 2007-08) the current survey shows higher coverage for all vaccines except measles which was higher in DLHS3 (87.3%). Gender wise difference in the coverage of different vaccines or various dropout rates was not statistically significant.

Conclusions: Although the vaccination coverage shows higher coverage than previous studies, it is still below the minimum targets set as national goal.

Key-words: MICS, children 12-23 months, vaccination status, coverage evaluation survey

Chipman, M. L. and Morgan, P. (1975). *British Journal of Preventive and Social Medicine,* **29, 190-195. The role of driver demerit points and age in the prediction of motor vehicle collisions.** The records of drivers, selected from the file of licensed drivers in Ontario, were reviewed to study the relationship between demerit points, other driver characteristics, and the frequency or risk of future collisions and traffic convictions. A stratified sample of 500–600 drivers from each of five levels of demerit points was selected. Low-point drivers differed significantly from high-point drivers in age, sex, and class of licence; estimates of risk of collision or conviction in each demerit point group had to take account of these differences. Discriminant analysis was used to identify drivers likely to be involved in collisions or to be given traffic convictions, and to identify accidents involving injury or fatality. Of the traits considered (demerit points, age, sex, class of licence, history of previous accidents), demerit points represented the only variable of importance in predicting future collision involvement. Since it is the only one of these variables which can be altered by driver behaviour it offers an opportunity to prevent accidents.

Further Reading ...

- www.consort-statement.org
- An Introduction To Medical Statistics
  - Martin Bland, OUP.
- Practical Statistics For Medical Research
  - D.G. Altman, Chapman & Hall.

Alan Watkins a.watkins@swansea.ac.uk

Alan Watkins a.watkins@swansea.ac.uk