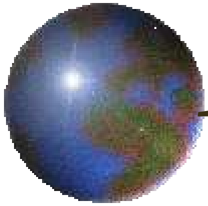


Quality Issues with Data (it Happens More than You Think) and How to Deal with Them

by Roberto Rivera (UPRM, CAAEPR)

Introduction

- No doubt that making decisions based on data can be extremely beneficial
  - Vaccines
  - Census
  - Big Data: Walmart and hurricane Frances
- Some believe we are in the midst of a golden age of data Big Data / Data analytics / Data Science buzzwords
- Data can be found anywhere and there’s a drive to turn data into knowledge
- Aim: Good data/information principles overall, tips on dealing with data
- No knowledge of computers, data analytics, statistics needed for this talk

Evidence based management (EBM)

- *In the past many managerial decisions were made based on anecdote (e.g. CEO or director’s opinion)*
- *EBM is a principle that argues that organizations are better off making decisions based on evidence*
- *Evidence is often obtained from data*

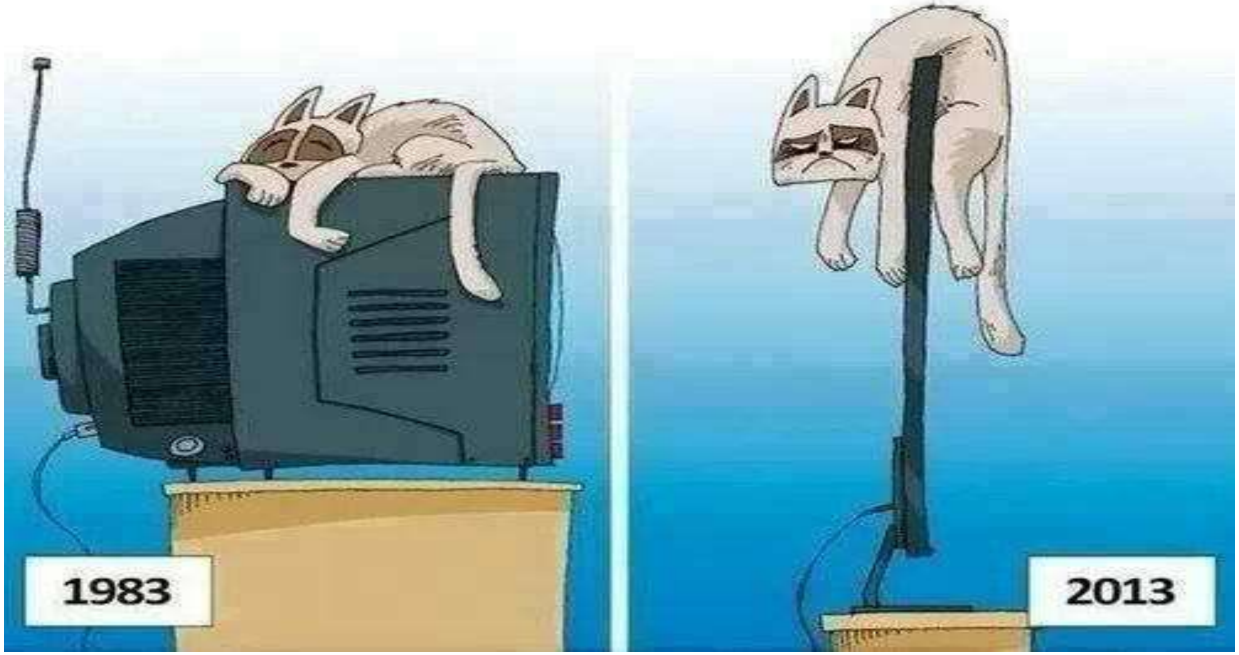
Evidence based management (EBM)

TimoElliott.com

"You may well have data, Smithers, but I have strong opinions, and I pay your wages"

Evidence Based Management

- Organizations implement statistical methods to keep up with the rest of world



Evidence based management (EBM)

- *What’s the EBM Outlook in Puerto Rico?*
- *18% growth for statistics profession in Puerto Rico from 2012-2022 (we will take any positive growth these days)*
- *But 22% growth in the U.S. We are lagging now, we will lag more in 2022*☹

Data case 1: Billing data

- *In recent years, there’s been an interest in using insurance billing data to make decisions*
- *Billing data uses international codes (ICD) to classify condition for which a patient is receiving a treatment for*
- *A collaboration was established to determine demographics of atrial fibrillation (AF) patients in a region, treatment effectiveness and potential adverse effects of treatments*

Data case 1: Billing data

- *Billing transactions for patients with AF were provided for several regions.*
- *For example region 1, had 14 million transactions over a two year span*
- *Data included: gender, age, ICDs (sometimes treated for many conditions), medicine name, medicine dosage, place of service, etc.)*
- *With data this large it is tempting to immediately jump to summaries*
- *Experience suggest data should be screened first.*

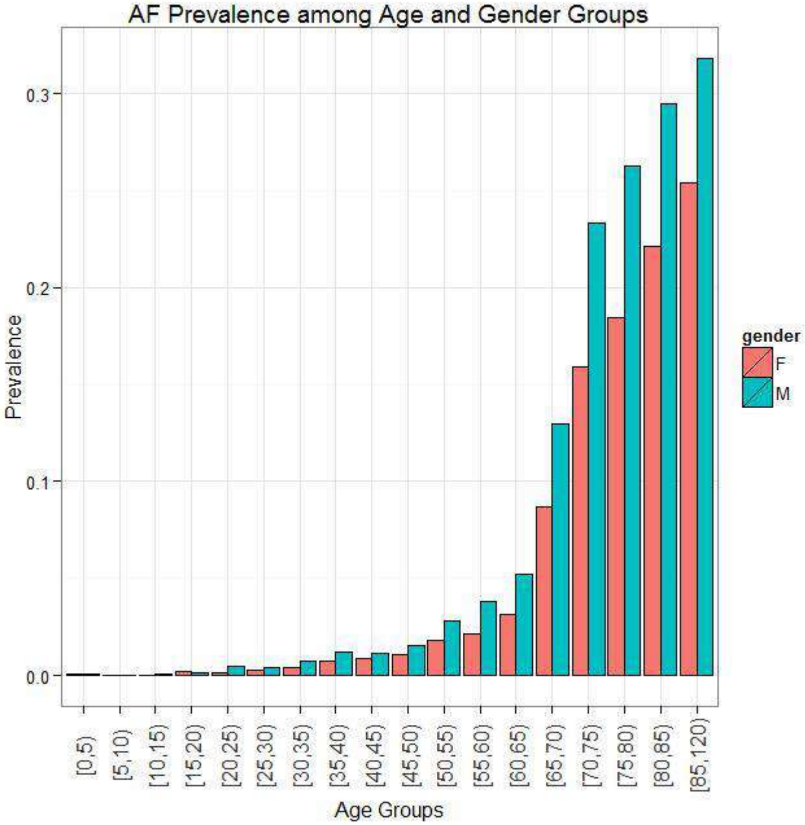
Data case 1: Billing data (screening)

- *Multiple ICD codes for same condition.*
  - *427.31 oficial ICD for AF, but data may show 42731*
  - *427.3 was also available in data*
- *An ICD code does not mean a person certainly had the condition (sometimes patient being screened for condition)*
  - *Though this is not a data issue, it could go undetected*
- *Of the over 14 million transactions for region 1, 1 million were DUPLICATES (7%)*

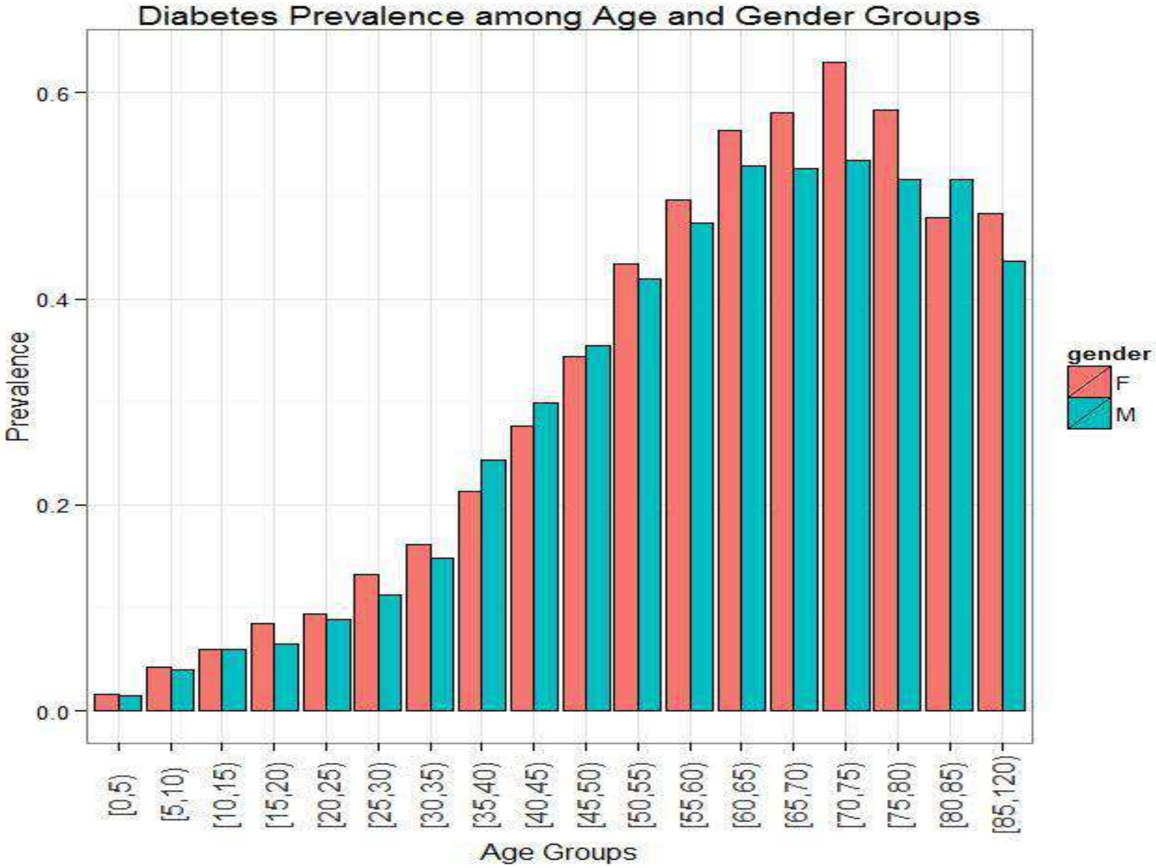
Data case 1: Billing data (screening)

- *Although we were told data comes from AF patients summaries show all sorts of ages in the data, but AF occurs mostly at older age*
- *Literature suggests 0.1% - 2% prevalence of AF in people <50 years old*
- *For región 1 77% of unique patients in data are <50 years old*
- *All regions have over 50% unique patients who are <50 years old*

Data case 1: Billing data (screening)



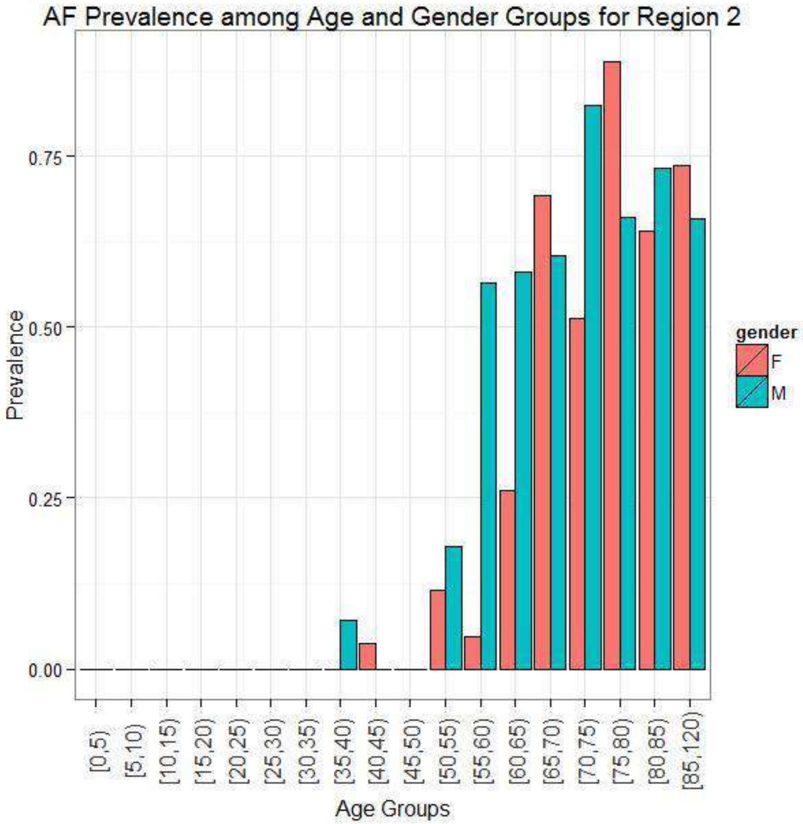
Data case 1: Billing data (screenina)



Data case 1: Billing data (screening)

- *Region 1 summaries suggest data is representative of a more general population, not AF patients*
- *However, analogous Figures for some regions simply do not make any sense!*

Data case 1: Billing data (screening)



Data case 1: Billing data (screening, some other issues)

- *Age and age category that didn't match (which one is wrong when this happens?)*
- *Sometimes admission date will be after dismissal date*
- *Pregnant men!! 500 of them in region 1*
- *However, we shouldn't be a purist (500 transactions out of 12.5 million is very small)*
- *Billing data is administrative data and not intended for research purposes.*
- *Data may still useful for preliminary studies*

Data case 1: Billing data

- *Some collaborators want regions to be merged, analysis conducted anyway*
- *Two scenarios: 1. Data is useless; 2. Represents a different population, but regions issues*
- *Before analysis we must go back to the drawing table*
  - *What population is being represented?*
  - *Why are some regions giving strange summaries?*
- *Without addressing these questions, the point of the study is lost: being objective*

Study: Ebola cured in chickens

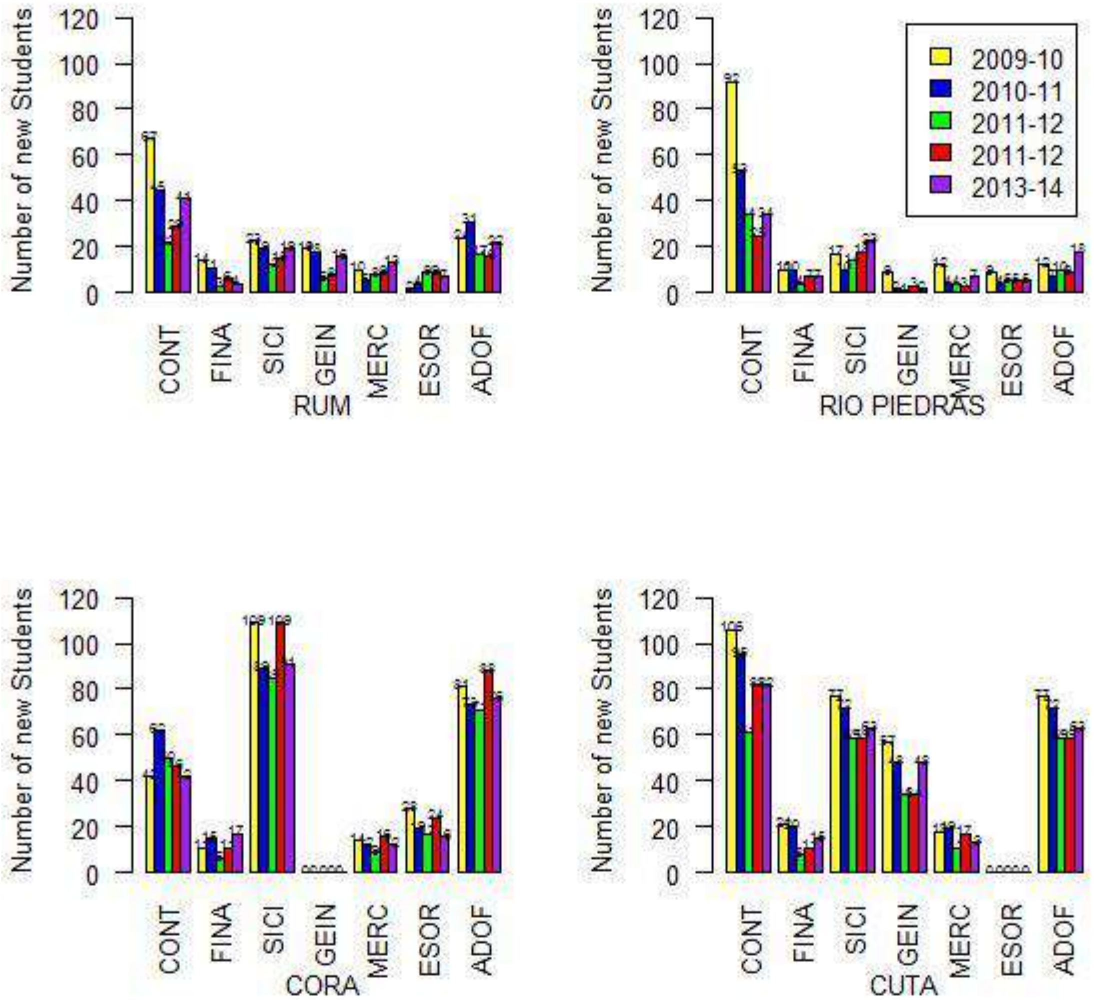
- *A scientist happily gives the news that he was able to cure Ebola in chickens*
- *33% of the chickens showed improvement*
- 1/3 showed no change
- *Unfortunately the 3rd chicken got away*
- *Obviously big enough data is important! (one of the reasons Big data is so popular now)*

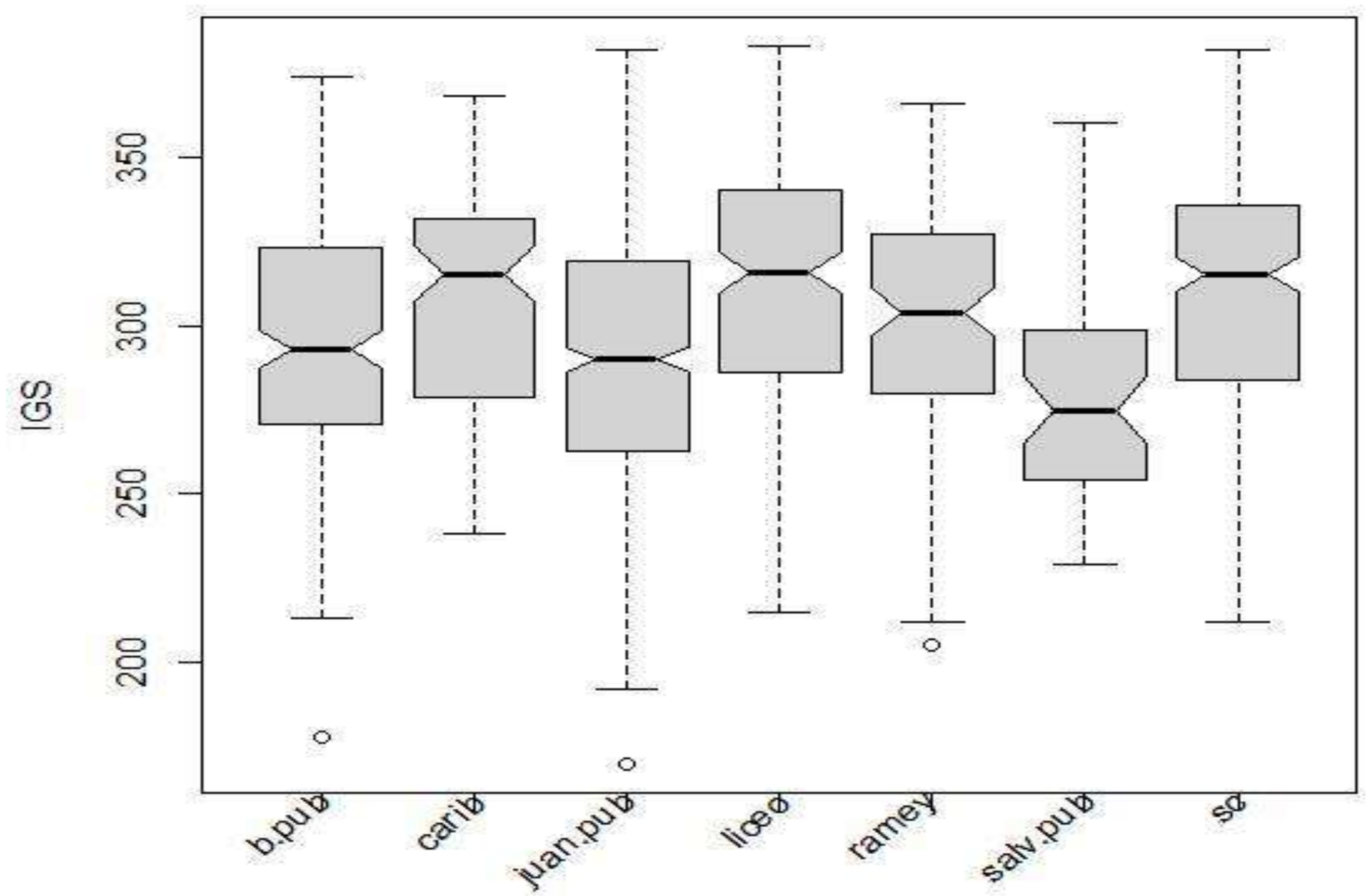
Big data

- Walmart and hurricane Frances
- Many examples of analyzing text
  - Translations (Google Translate)
  - Twitter feeds to forecast elections, crime or Academy Award winners
- As part of the Golden age of data, many cities and regions now share all sorts of data openly online on all sorts of topics: traffic accidents (NYC), teaching staff (EU), UPR admissions (PR)

Data case 2: UPR admissions

- Available in data.pr.gov it provides admission data for UPR campuses from 2009 to 2013.
  - Includes: GPA, IGS, High school, gender, municipality
- Over 30,000 observations
- University administrators can: determine profile of students, minimum IGS decisions
- Parents can use it to choose right High School for child (e.g. is private better than public?)
- Schools can use it to check if its performance has improved


Number of new Students
120
100
80
60
40
20
0
67
41
31
CONT
FINA
SICI
GEIN
MERC
ESOR
ADOF
RUM
92
17
2009-10
2010-11
2011-12
2011-12
2013-14
RIO PIEDRAS
109
109
17
24
81
CORA
106
77
57
17
77
CUTA


IGS
200
250
300
350
b.pub
carib
juan.pub
liceo
ramey
salv.pub
sc

Data case 2: UPR admissions

- Data errors (e.g. typos) are not a big problem in this case
- But, where does the data come from?
- Recently updated from over 30,000 to almost 70,000 records (for the same school periods)
- date of admission data not kept fixed for all campuses
- Do admissions match University records?

Data case 2: UPR admissions

- Data.pr.gov admissions vs OIIP admissions

School Year	RUM Portal Admissions (data.pr.gov)							RUM OIIP Admissions						
	CONT	FINA	SICI	GEIN	MERC	ESOR	ADOF	CONT	FINA	SICI	GEIN	MERC	ESOR	ADOF
2009-10	125	35	38	32	36	6	41	105	23	31	29	26	5	30
2010-11	98	26	37	38	39	9	55	78	20	35	29	29	9	44
2011-12	58	15	28	23	25	17	31	45	12	9	12	18	4	14
2012-13	60	19	29	37	39	12	34	45	8	22	24	27	11	23
2013-14	82	15	34	38	50	13	40	66	11	27	31	39	12	29

Data case 2: UPR admissions

- Compared to the Billing data, issues with this data appear to be minimal
- Still valuable to compare HS performance and within campus department performance or profiles
- Only some type of analysis may be affected (e.g. comparing total admissions of campuses since admission dates are not kept fixed)
- Other among campus comparisons (mínimum IGS) perform fine

Data case 3: Epidemiology

- A study aims to assess the association between two diseases. Subjects were invited to participate in the study in 2005. Data was gathered. 3 years later (2008) subjects were invited to a follow up for more measurements.

Data case 3: Epidemiology

	2005 data				**2008 data**			
Subject		Sex	age	smoke		Sex	age	smoke
1		F	48	Current		F	51	Past
2		F	50	Past		M	52	Never
3		M	33	Past		F	37	Current
4		F	50	Never		F	47	Never
5		M	57	Current		F	57	Never

Other issues with data in general

- Errors with coding
  - Example determining the status of a disease
- Corrupted file (variables get shifted)
- Sometimes a variable is measured as an integer, others measured as a real number
- Data variables are poorly defined.
- Secondary data
  - Easy to obtain secondary data nowadays: Twitter, Facebook, etc. But how reliable is this data?
- Missing data, especially when nonrandom

Are Data issues common?

- The data issues raised here are very common
- In a 1960 census study showed 62 women, aged 15 to 19 with 12 or more children.
- In a recent survey of Top 500 companies, 75% of respondents reported big problems resulting from Big Data

Information quality

- (Wikipedia) Quality of the content of information systems
- Information quality is more general than data quality
- Low information quality is bad, low data quality is not necessarily bad
- Quality is subjective and dependent on goals
- Old saying: Garbage in, garbage out!

What is considered Quality information anyway

- Simply put, the better the data represents the process of interest the better the quality of the information
- We cannot determine how well the data represents the process of interest directly since the main traits we are interested in are unknown
- By screening the data to determine how much sense it makes, we can infer how reliable the data is

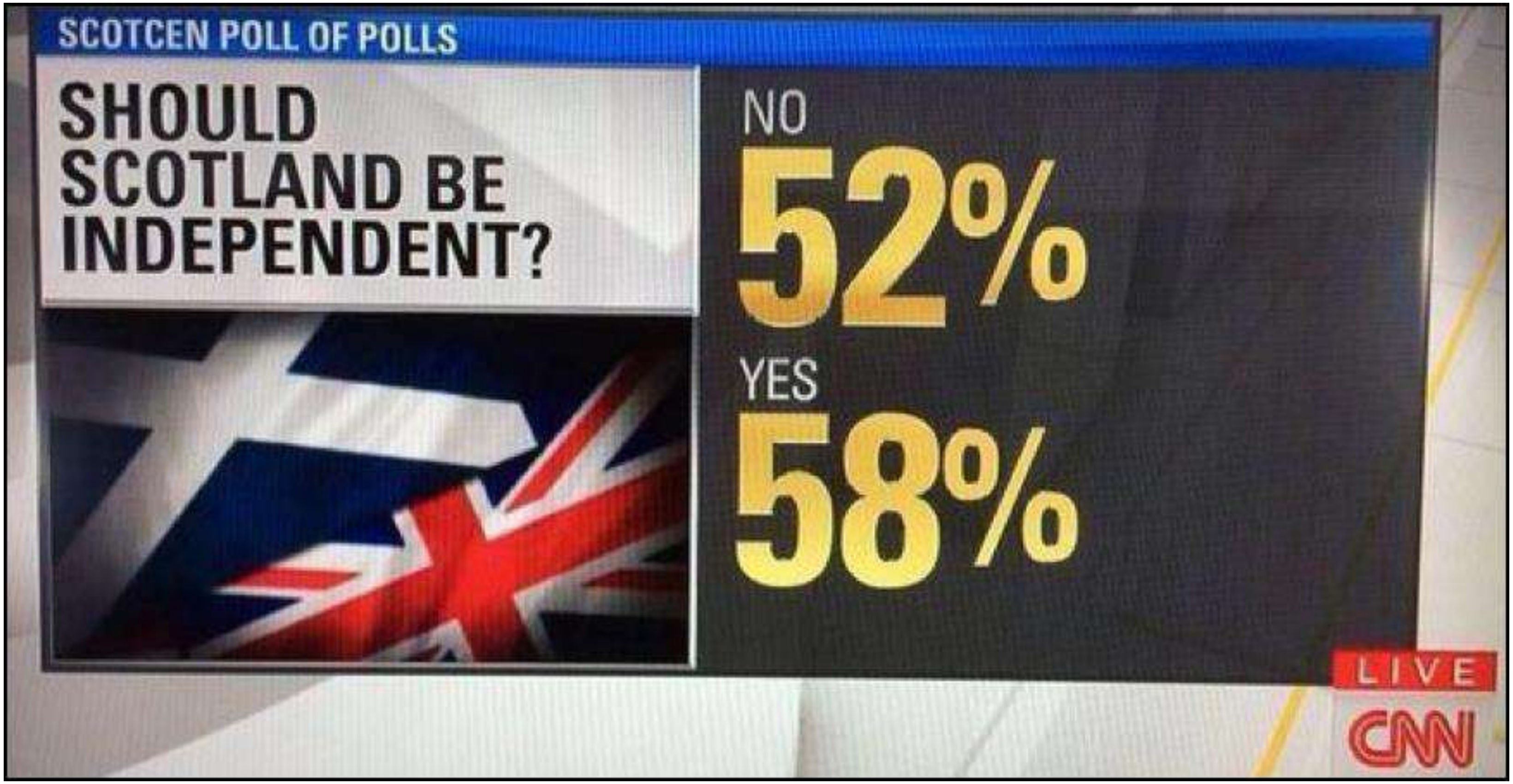
How Information Quality affects statistical procedures

- Well it depends
- In the billing data, duplicates will not affect statistics of unique patients, but may affect statistics of overall transactions.
- Specifically, bad information quality may
  - bias estimates
  - Affect variance of variables
  - When issue causes both bias and variance inference is affected the most

How Information Quality affects statistical procedures

- Sometimes all it takes is one data issue to potentially make your data useless!!
- Information Quality is dependent of goal: biased regression coefficients are a problem for explaining a process, but don’t matter for prediction

CNN poll results over Scotland independence



SCOTCEN POLL OF POLLS

SHOULD SCOTLAND BE INDEPENDENT?

NO

52%

YES

58%

Some stats on Information Quality

- Not much research evidence on it
- Some consultants provide some estimates
  - Bad data costs the electro industry $1.2 Billion annually
  - Data clean up routinely accounts for 20-25% of an organization’s budget
  - Data clean up routinely takes at least 65% of the project effort
- De Veaux and Hand (2005) How to Lie with Bad Data. Statistical Science, 20, 231-238

Information Quality in statistical education and research

- Remarkably, Information Quality plays virtually no role in statistical education
- Statisticians acknowledge the importance of Information Quality, but the topic is hard to express in a general way (i.e. mathematically) since data issues are highly dependent of the situation at hand and quality is subjective
- Little research on the topic performed by statisticians

Plan A to handle poor IQ: Data Management Program

- Collection
- Entry
- Data Screening,
- Storage, Analysis are also part of this program but they’ll be emphasized in Plan B

Data collection

- State clear objectives, and consider resources and constraints to determine best way to meet objectives.
  - Designed experiment or observational data?
  - Type of measurements
  - Develop steps to meet objectives
- Use data management protocol!!!
- If survey used, design survey carefully
  - Length of questions, spacing, language
  - Make pilot/test runs of survey to ensure participants answer questions effectively

Data entry tips

- Use code when possible.
  - For example, if age will be recorded multiple times, enter date of birth and date of measurement to determine age.
    - since less individual entries are needed when entering date of measurement instead of actual age, human error is reduced.

Data entry tips

- Use appropriate software for data entry to minimize errors (e.g. Double data entry).
- It’s imperative to create a variable dictionary or codebook.
  - Saves time!
  - prevents relying on memory
  - makes it easier to share the data
  - helps build convenient, consistent codification, based on similar variables for future measurements.

Data entry tips

- Create variable dictionary or codebook.
  - Dictionary should include,
    - variable abbreviation used,
    - definition of variable (continuous/discrete?) what does the abbreviation represent,
    - If quantitative, unit of measurement.
    - For dummy categorical variables what each coded value represents (e.g. M –male, F – female, or 1- female, 0- male)
    - For some categorical variables explanation of how categories were defined (many disease classifications do not have gold standards).

Data entry tips

- Use intuitive names for variables.
- Limit use of dummy variables of categories of variables. (save for analysis)
    - If one chooses to use dummy variables, keep it consistent.
- Define missing values appropriately
    - careful in treating missing values as a category, it may lead to bias

Data sharing

- Define variables properly (i.e. codebook)
- Specify who collected data and how data was collected (e.g. UPR admissions)
- Data sharing- privacy and confidentiality must be protected. Only share necessary data (exclude information that identify patients).
- Secondary data (not collected by you):
  - how was the data collected? Does it match your needs? What are the limitations of the data? Any alarm signs about the quality of the data?

Data sharing

- Proprietary data (Twitter, Facebook, etc)
  - In science, transparency and reproducibility are traits expected from data based procedures
  - Companies sometimes will share their data, but without providing details on how the data was acquired.
  - Reasons? Competitiveness, people ‘gaming the model’
  - Without this information it is difficult to quantify the quality of the data (e.g. data may be biased)

Screening data tips (often forgotten step)

- Start screening data early in the project.
    - change patterns of ineffective data entry, data collection or
    - correcting error by contacting respondents.
- Conduct data screening ‘regularly’.
    - Audits
    - Check missing value or skipping patterns by study subjects (may indicate privacy concern), etc.
    - Any flaws or concerns should be properly documented.

Screening data tips

- Check if observations are feasible (trivial vs nontrivial)
- do summaries make sense? (Including maximum and minimum)
- Do classifications make sense? This is difficult to do with large data
- Calculate error rates (variable wise and dataset wise)
- Count categories of a non-changing categorical variables at different time points.
  - E.g. one can check if the total of males and females at t=0 match with the total of males and females at follow up(s).
  - WARNING; assuming no missing values in gender, if the counts don’t match there is an issue, but if they do match this doesn’t necessarily mean there is no issue (multiple imputation errors may cancel each other out).

Screening data tips

- For numerical data, check with a priori predictable changes to detect issues. (ages issue in epidemiology)
- Refer to original data collection forms when in doubt
- Do categorical entries at different time points make sense?
- Plot the data (dotplots, Histograms, scatter plots)
  - For example, suppose data includes an age variable and years of education variable (both quantitative). You shouldn’t expect, say a 15 year old to have over 20 years of education. A scatterplot would be able to detect this type of issue

Screening data tips

- Use search. Search for rows of observations for unlikely combination of values in variables.
  - E.g. a male subject cannot be pregnant.
- If multiple people are in charge of data entry, ensure use of dictionary and communication between personnel.

Plan B: Deal with it

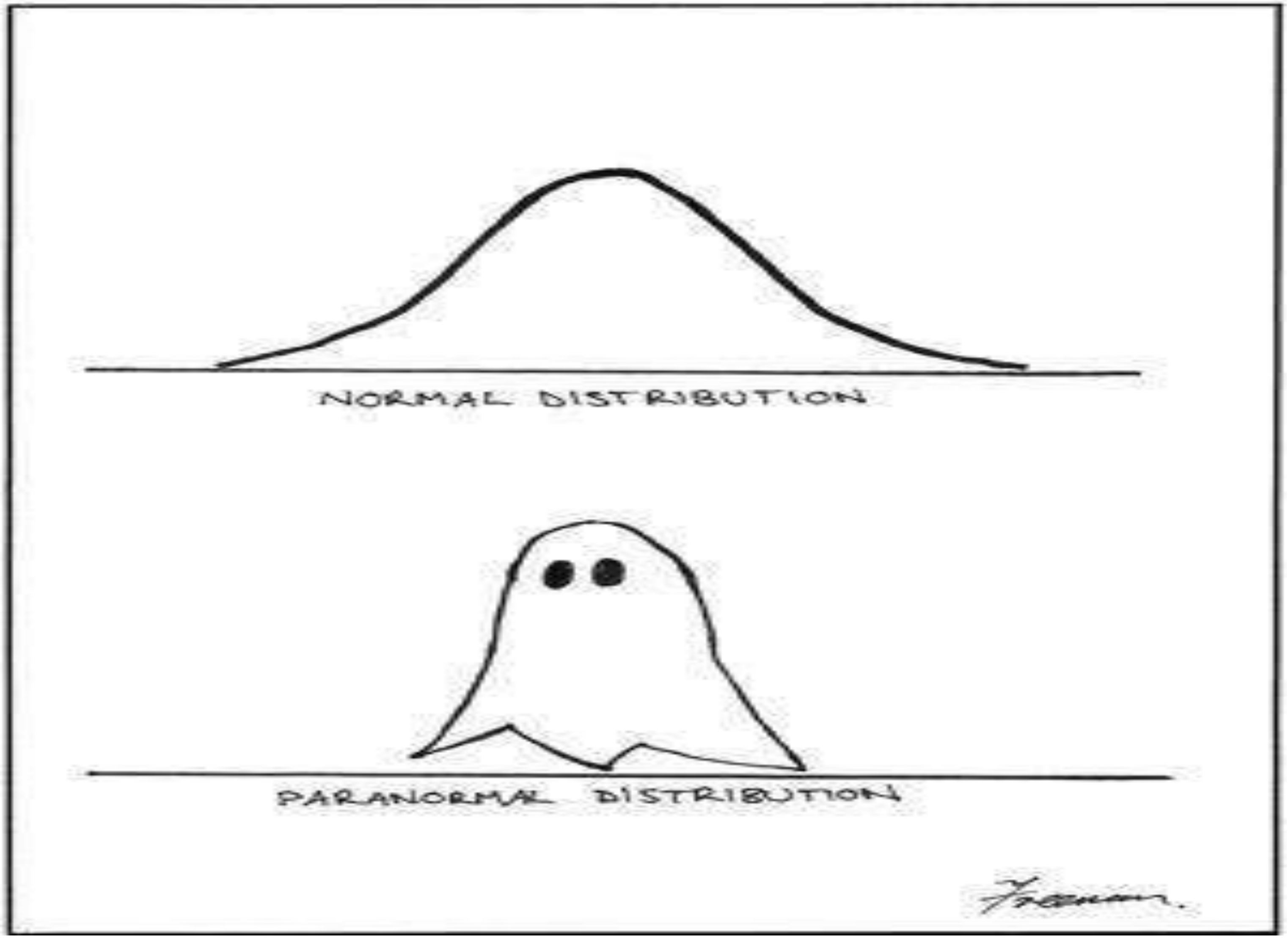
- A Data Management Program is not always possible or may not eliminate major issues
- Statistics come to the rescue…..
- And some critical thinking
- With Statistics (don't forget data screening);
  - Use robust statistics (M, R, or S estimators for regression) and compare results with traditional statistics. These may not perform well for multivariate parameters
- Filter data based on constraints

More research needed

- Not much work lately on use of statistics for information quality
- Decision theory could be key
  - If D_T is the true data, D_a is the acquired data, D_c is the cleaned up data we need to find a way to measure the difference in statistical inference using D_a or D_c relative to D_T.
  - But is D_T unknown (Bayesian methods to the rescue?)

Data analysis tips

- When coding categories based on some criteria (for example, when determining the diagnosis of a disease), place next to each other the criteria for categories with the obtained categorization to ensure proper categorization
- Informed Skeptic approach works best
- Avoid black-box algorithms
- Rely on experts!!!


NORMAL DISTRIBUTION
PARANORMAL DISTRIBUTION

Summary

- Bad Data occurs frequently (Bad Information occurs less frequently)
- We should strive to ensure the best data is gathered, but bad data will always be lurking
- The goal is not to eliminate bad data but to minimize its impact
- A multidimensional approach (managerial, systems, statistics) works best
- Data Quality is task that needs attention at all stages of a Project
- Objective of Project is key (don’t be a purist)

Questions?