Data Collection and analysis

Chris Mainey

Senior Statistical Intelligence Analyst
Healthcare Evaluation Data (HED)

University Hospitals Birmingham NHS FT

chris.mainey@uhb.nhs.uk
www.hed.nhs.uk

Healthcare Evaluation Data (HED)

- Online hospital benchmarking system
- Statistical models and analysis tools
- Activity, Mortality, Readmissions, Length-of-Stay, Marketshare etc.
- Used by ~60 NHS and other organisations
- Training and support

- **Using national NHS hospital data**

www.hed.nhs.uk

Overview

- What are data?
- How do we generate them?
- What types of data are there:
  - Qualitative
  - Quantitative
- Methods of analysis with Quantitative data
  - Summary statistics
  - Plotting
  - Uncertainty

What are data?

Data, Information, Knowledge, Wisdom?

Data

- Symbols representing properties, product of observation.
- "Know nothing"

Information

- Contained in descriptions and answers to questions.
- "Know what"

Knowledge

- Hot to transfer information into instructions.
- "Know how"

Wisdom

- Ability to increase effectiveness. Adds value, which requires judgement
- "Know why"

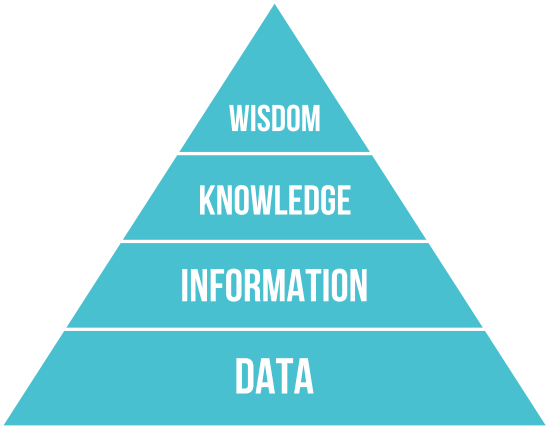


image credit: Longlivetheux / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0) Source: https://en.wikipedia.org/wiki/DIKW_pyramid

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, **33**(2), 163–180. https://doi.org/10.1177/0165551506070706
R.L. Ackoff, (1989). From data to wisdom, *Journal of Applied Systems Analysis* **16** 3–9.
M. Zeleny, (1987). Management support systems: towards integrated knowledge management, *Human Systems Management* **7**(1) 59–70.

How do we generate data?

Primary

- We deliberately collect data about a thing:
  - Occurrences of an event
  - Duration of something

Secondary

- We generate data as a by-product, or reuse it for another purpose:
  - Using patient care records to build risk models
  - Using patient data

Question:

Is the quality of care, in hospital X, poor?

- How would you answer this question?
- Where / how could you access data for it?
- What are the issues with this question, and potential answers?

Types of information

Various theoretical frameworks for this, and we are not doing an in-depth study here. We will summarise into:

Qualitative

- Understanding the meaning, concepts, or how something is appears in the 'real world'

Quantitative

- Methods for discovering measurable 'facts'
- Methods for dealing with 'hard' evidence, quantifiable usually in numbers

Both involve measurement, interpretation, proxy effects etc.

Qualitative Analysis

Qualitative Data

...qualitative researchers study things in their natural settings, attempting to make sense of, or interpret, phenomena in terms of the meanings people bring to them. Qualitative research involves the studied use and collection of a variety of empirical materials – case study, personal experience, introspective, life story, interview, observational, historical, interactional, and visual texts – that describe routine and problematic moments and meanings in individuals' lives ***(Denzin & Lincoln, 2005)***

- Interview, survey, feedback etc.
- Observational: e.g. Ethnography

Examples:

- Patient experience
- Ethnographic study of behaviour in hospitals
- Understanding what patients want from care

Denzin, Norman K., and Yvonna S. Lincoln. 2005. Introduction. The discipline and practice of qualitative research. In The Sage handbook of qualitative research, ed. Norman K. Denzin and Yvonna S. Lincoln, 1–32. Thousand Oaks: SAGE Publications.

Qualitative analysis

- Understanding your philosophical position
  - Realism, Constructivism etc.
- Identifying your data collection method
  - Sampling strategy
- Record data, e.g. transcripts
- Analysis methods
  - Formal, e.g. constant comparative methods
  - Will you 'code' and how will you do it?

Example Tools:

- Nvivo
- MAXQDA
- ATLAS.ti

Quantitative analysis

Data types

Data can take various forms: E.g. measurements, grouping factors, estimates, observations etc.

- Different in terms of:
  - Storage
  - Methods for summary/processing
  - Interpretation
- A few major groups of data types:
  - Numeric
  - Binary
  - Categorical

Data types (2)

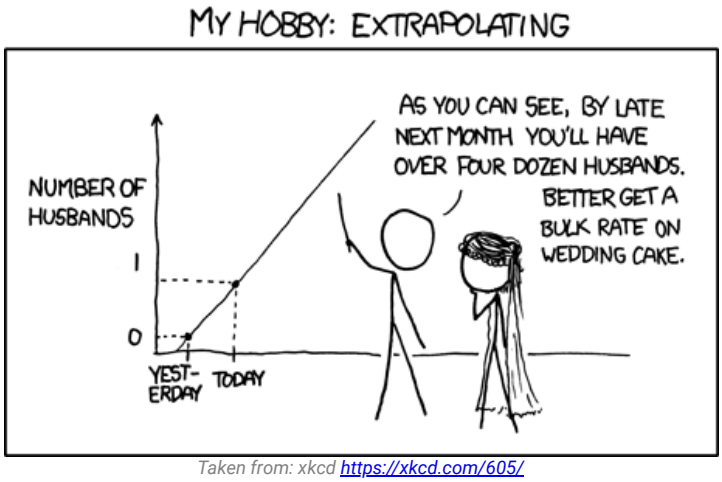
Numeric :

- **Continuous:** values that can be constantly divided with a possible number in between
  - E.g. height of a person could be 172, 173 or 172.5 cm
  - Examples in HI: physiological measurements like blood pressure
- **Discrete:** values that can only take whole numbers, usually obtained by counting.
  - E.g. Number of patients seen in a clinic could be 35 or 36 but not 35.5
  - Examples in HI: counts of patients, waiting time measured in whole minutes, length of stay measured in days (like HED/HES), number of patient episodes

Data types (3)

Binary :

- Exclusive two state variable
  - E.g. 0/1, yes/no, TRUE/FALSE
  - Examples in HI: Patient dead or alive?, true or false answer to survey, patient status for a genetic marker



Taken from: xkcd https://xkcd.com/605/

Data types (4)

Categorical :

- **Nominal:** Categories without any notion of order
  - E.g. Hair Colour, Brand of car, Country of residence
  - Examples in HI: Ethnicity, Admission method, Treatment speciality

- **Ordinal:** Categories with order, but not linear like numeric
  - E.g. Survey answers 'Good, OK & Bad'. There is order, but 'OK' ≠ 'Bad' x 2 and 'Good' ≠ 'Bad' + 'OK'
  - Examples in HI: Cancer stage, self-assessed patient answers like 'is your health poor, OK or good,' Age-groups <1, 1-16, 17-40 etc.

Quantitative analysis

- Scientific method
- Identifying your data collection method
  - Sampling strategy is statistical, to represent population
- Measureable data, e.g.:
  - Attendances at clinic
  - Readmission to hospital
- Analysis methods
  - Visualise / tabulate
  - Transform/summarise
  - Statistical tests/modelling

Example Tools:

- Relational Databases ("SQL" systems)
- Excel
- R, Python

Mixed Methods?

Certain questions might best be answered by a mixture of methods.

E.g. Evaluating a patient safety programme:

- Interviews with managers and healthcare staff
- Ethnographic observations on wards
- Casenote review for adverse events
- Patient outcomes: adverse events, mortality
- Patient satisfaction

Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation (2011); ***BMJ*** *342 doi: https://doi.org/10.1136/bmj.d195*

Methods and considerations for quantitative data

How can we go about answering a question?

Example, summarising data

"Dear informatics, I would like age of all patients admitted as an emergency to general medicine in December?"

OK, here:

```
##   [1] 75 81 59 70 64 67 66 54 68 72 80 66 70 76 75 52 59 52 86 56 51 59 72 61 53
##  [26] 72 75 69 64 55 74 54 61 74 86 53 68 69 76 58 59 79 59 69 91 55 59 68 58 70
##  [51] 68 60 89 54 85 76 56 56 84 91 90 87 90 85 54 76 91 79 53 62 72 69 75 76 76
##  [76] 76 63 85 76 85 67 63 91 63 64 69 63 60 57 83 69 60 58 70 59 85 68 85 56 79
## [101] 85 76 76 73 60 87 57 67 72 92 58 55 54 71 90 55 58 59 63 77 85 77 53 66 73
## [126] 53 79 70 70 77 56 65 85 64 74 66 74 59 68 79 66 56 68 63 66 66 68 70 64 72
## [151] 56 83 53 69 67 77 68 63 73 57 86 52 75 78 76 61 71 77 64 62 77 69 69 66 85
## [176] 65 61 72 69 73 53 77 54 56 72 70 69 67 62 78 58 54 69 76 86 59 80 84 56 78
## [201] 75 57 68 91 91
```

How would you answer that question

- What is the question?
- Sending a list of numbers doesn't answer the question
- How might we show this in a better way?
- Summary figures?

Visualise:

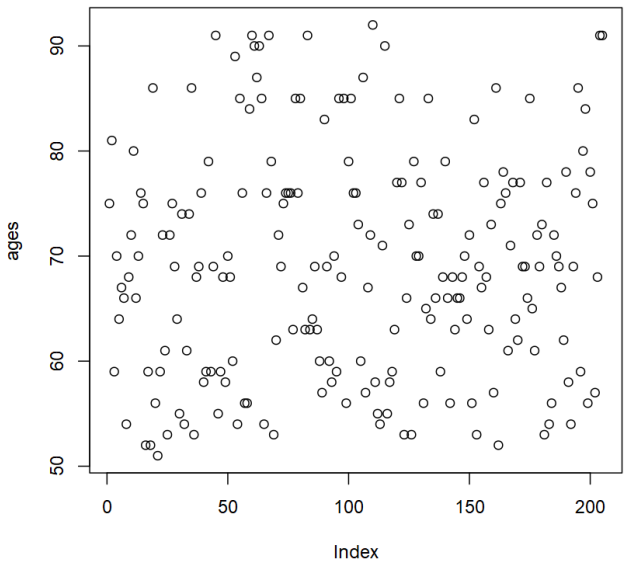
- Scatter plots
- Histogram or Kernel Density (sounds more impressive than it is)
- Box plot

Scatter plots

Plots and 'x' variable by a 'y' variable by point

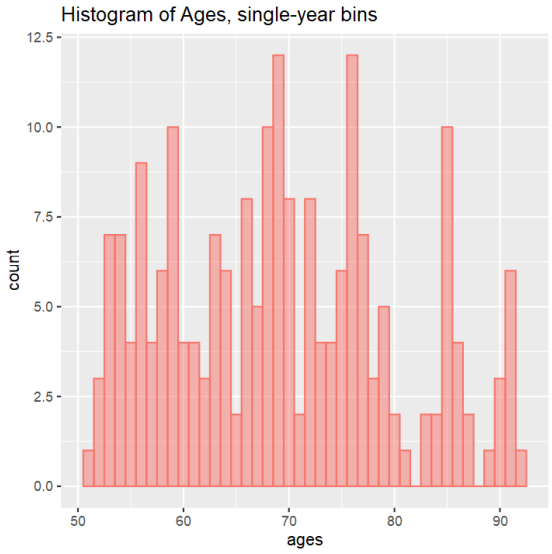
Why doesn't this help?

- We've only got one variable, not two
- No summary information
- We want to see some kind of distribution

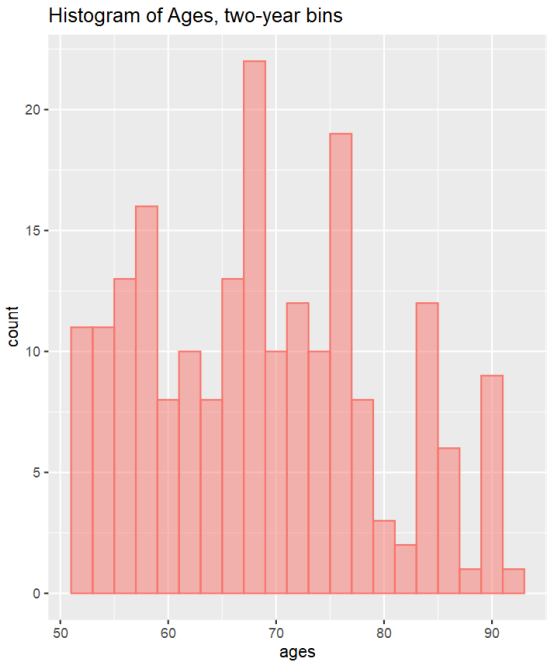


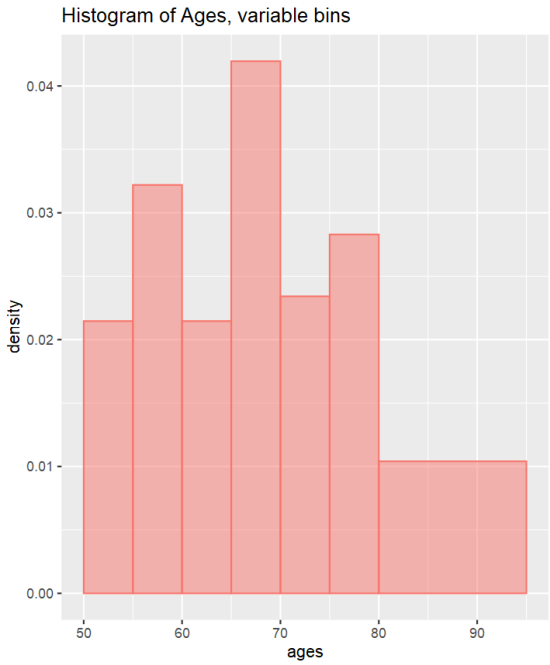
Histograms and Bar Charts

- Plot of binned counts
- Good way to visualise distribution
- Bin sizes can vary & do not have to be equal
- Bar charts are related, but do not share the 'binning' idea. Can be used with categorical

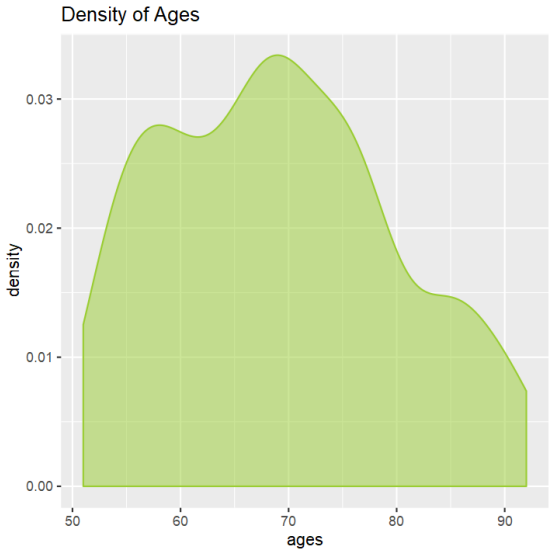


Histograms and Bar Charts (2)





Kernel Density

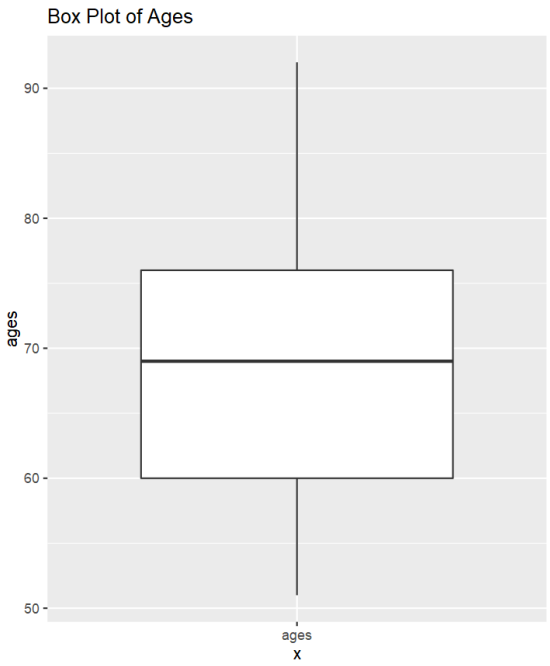


- Similar to a smoothed histogram
- Plots the probability density of data rather than counts values
- Conceptually harder, but nicer visualisation

Box Plots

Box range is (“hinges”):

- 25th percentile
- 75th percentile
- Line is the median (50th percentile)
- Whiskers extend hinge ± 1.5 * IQR
- Outliers (further points) are represented
- Terms will be explained in the following slides



Summarising data and distributions

Question:

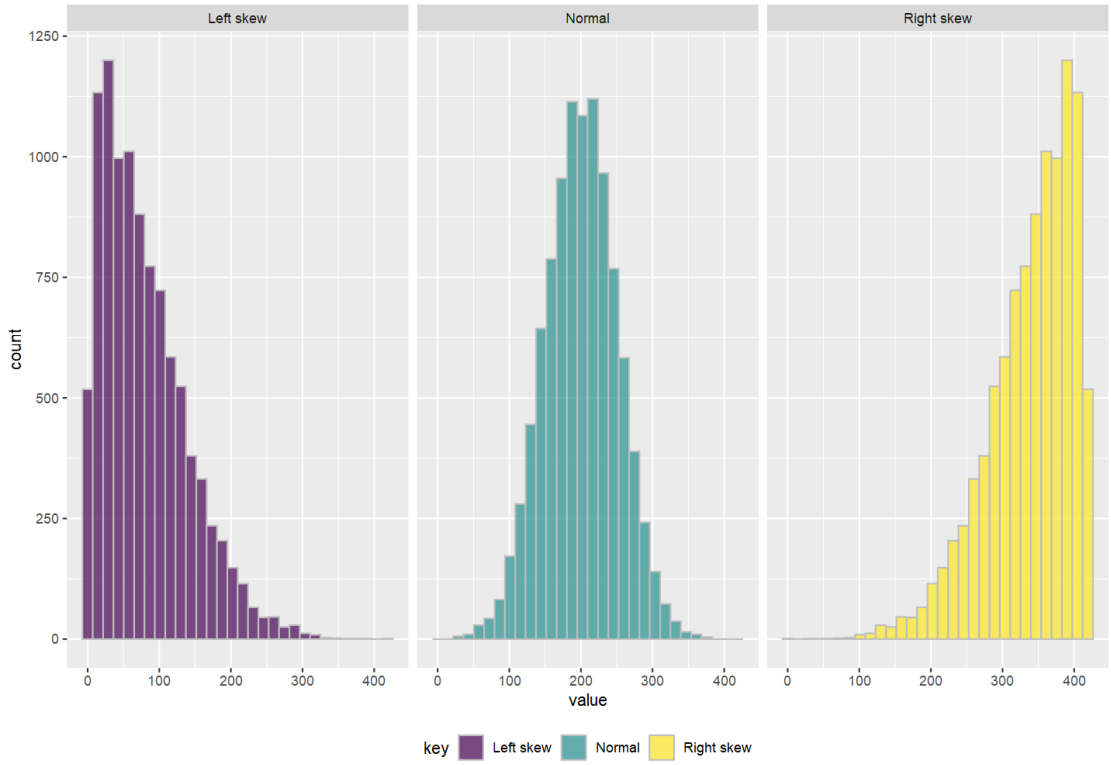
How can we describe the same data in words or numbers?

Centre point

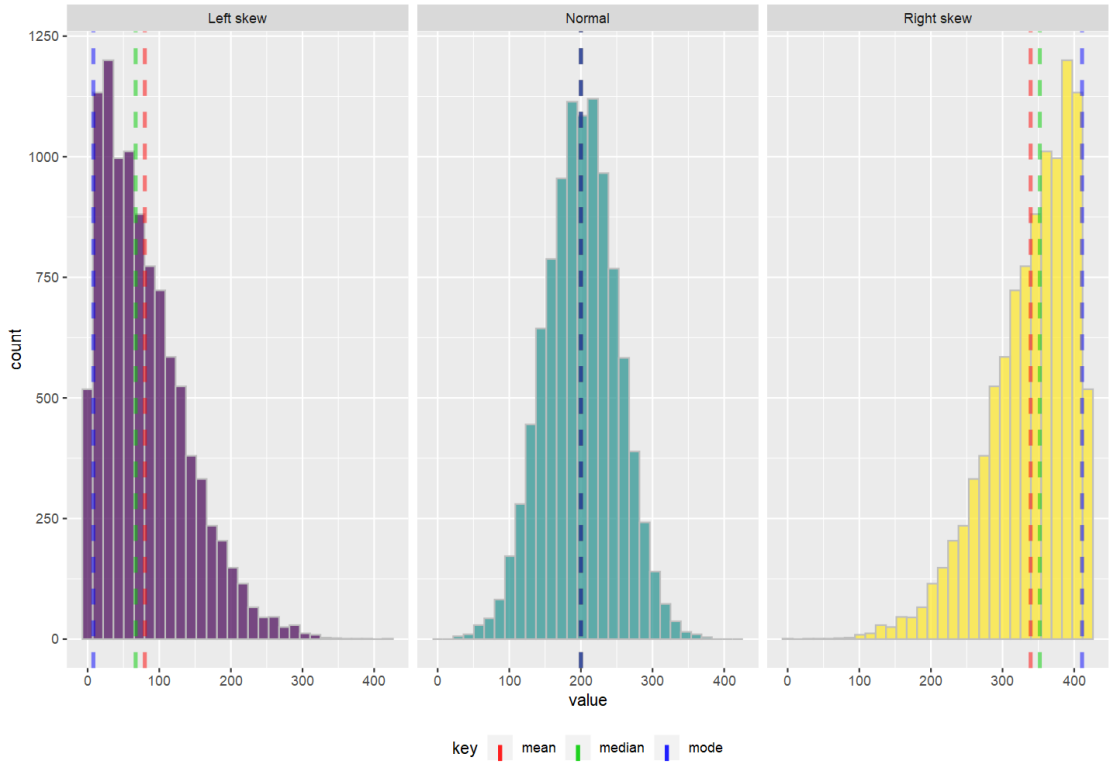
Few ways to describe this:

- Mean - sum / count
- Median - Middle value or ordered data
- Mode - The most common value

Distributions



Distributions and centre points



Uncertainty in data

Uncertainty

Knowing a number is not necessarily enough. How certain are our estimates?

Question:

If you toss a balanced coin, how often do you get "heads"?

- If we try twice, and both are heads, what does our evidence say?
  - 100% coin toss = heads
  - What if we do it 10 times? Are you more certain?

Sample size and uncertainty

- Larger sample size reduces uncertainty
- Commonly illustrated by:
  - Standard deviation
  - Standard error / confidence intervals

Implications:

- Case-series / case report
- Clinical Trials:
  - "Power" / sample size calculation
  - Statistical analysis methods
- Meta-analysis

Summary

- Data are not the goal, used to derive knowledge, information etc.
- Primary collected for reason, secondary is reuse of data
- Qualitative data used to understand meaning and context
  - Commonly interviews, observations of feedback
  - Often 'coded' by researchers
- Quantitative data used to measure effects
  - Often summarising observations
  - Mean & Median commonly used to describe mid-point
- The number of data points affects our certainty about estimates.
  - Larger sample sizes preferred