Data analysis

Ben Graham

MA930, University of Warwick

October 5, 2015

Data Analysis

- George Box: All models are wrong, some models are useful.
- Corollary: This model is wrong, therefore it is useful.
- Peter Norvig (not): All models are wrong, and increasingly you can succeed without them.

Statistics

- What is a statistic?
- Why is this course not called “statistics”?
- Data
  - Designed experiments
  - Observed data
  - Big data
  - Small data
- Summary statistics (mean, median, mode, min, max, ...)
- Graphs
- Probabilistic models
  - Seek the model parameters
    - Frequentist statistics
    - Bayesian statistics

Key principles of statistics

- Taking averages is good.
- Correlation does not imply causation (missing covariates, Simpson's paradox).
- Interpolation good; extrapolation bad.
- Can you play catch?

Machine learning *

- Supervised learning
  - Learning to approximate high dimensional functions
  - Boring? Includes a huge range of problems.
  - Neural networks, decision trees, random forests, support vector machines, bagging and boosting.
- Unsupervised learning
  - Clustering
  - PCA, LLE, RBMs
  - Dimensionality reduction
  - Simplify correlation structures of data?

Problems with English

From the FT:

> *Linda is single, outspoken, and deeply engaged in social issues. Which of the following is more likely?*

1. That Linda is a bank manager.
2. That Linda is a bank manager who is an active feminist.

Set theory

Definition 1.1.1. The set, S, of all possible outcomes of a particular experiment is called the *sample space* for the experiment.

- Coin toss
- Sequence of coin tosses
- Two children, at least one of them a boy.
- Waiting time at a red traffic light.
- Waiting time passing a traffic light.

Events

Definition 1.1.2 An Event is any collection of possible outcomes of an experiments, that is any subset of S.
Includes $\emptyset$, $\{x\}$ for every $x \in S$, and S.

How many events when you

- toss a coin
- roll a die

De Morgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c$$

Proof?

Disjoint events

Definition 1.1.5 Two events A and B are disjoint if $A \cap B = \emptyset$.
Definition 1.1.6 If $A_1, A_2, \ldots$ are a collection of pairwise disjoint events, and if

$$\cup_i A_i = S$$

then $A_1, A_2, \ldots$ form a partition of S.

Axioms of Probability

Def 1.2.1 A collection of events is called a σ-algebra, denoted $\mathcal{B}$ if

1. $\emptyset \in \mathcal{B}$
2. If $A \in \mathcal{B}$ then $A^c \in B$ (closed under complements)
3. If $A_1, A_2, \cdots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ (closed under countable unions).

Examples

- Toss a coin
- Roll a die
- Roll a die to see if you get a 6.

Probability space

Def 1.2.4 Given S and $\mathcal{B}$, a probability function is a function $P : \mathcal{B} \to [0, 1]$ s.t.

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$
2. $P(S) = 1$
3. If $A_1, A_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_i P(A_i)$.

Examples

- Toss a coin
- Roll a die to see if you get a 6.
- Circle $\{(x, y) : (x - 0.5)^2 + (y - 0.5)^2 \leq 1\}$ in the unit square $[0, 1]^2$.

National Lottery counting

- There are $49! = 1 \times 2 \times \cdots \times 48 \times 49$ ways to pick 49 balls in order (without replacement).
- If we only pick 6 balls, there are

$$\frac{49 \times 48 \times \cdots \times 44}{6 \times 5 \times \cdots \times 1}$$

possibilities.

Def 1.2.17 Binomial coefficients $\binom{n}{r}$ = n choose r =

$$\frac{n!}{r!(n-r)!}$$

ways of picking r objects from n objects.

Conditional probability

Def 1.3.2

- Events $A, B \in \mathcal{B}$.
- $P(B) > 0$.
- The conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$P(\cdot \mid B)$ satisfies the axioms for being a probability measure!

Bayes Rule

By the definition of conditional probability:

$$P(A \mid B) = P(B \mid A)\frac{P(A)}{P(B)}$$

Theorem 1.3.5: Let

- $A_1, A_2, \ldots$ partition the sample space S,
- Let B be any event ($P(B) > 0$),

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B \mid A_j)P(A_j)}$$

Or if $A_1 = A$, $A_2 = A^c$,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}$$

Independence

Def 1.3.7: Two events A and B are independent if

$$P(A \cap B) = P(A)P(B).$$

Can an event be independent of itself?
Is A independent of B^c?

Def 1.3.12: A collection of events $A_1, A_2, \ldots$ is independent if for every n element subset $A_{i_1}, \ldots, A_{i_n}$

$$P\left(\cap_{j=1}^{n} A_{i_j}\right) = \prod_{j=1}^{n} P\left(A_{i_j}\right).$$

Do not confuse this with "pairwise independence"!
Do not think about "pairwise independence"!!!

"Random variables"

Def 1.4.1: A random variable is a function $X : S \to \mathbb{R}$.

- Represent something random like rolling a die
- Not actually random themselves,
- also not actually variables, on account of being functions.

Examples

- Toss a coin (1 for H, 0 for T)
- Toss n coins and count the number of H
- Toss a coin repeatedly: count how many H before the first T.

CDF - cumulative distribution function

Def 1.5.1

$$F_X(x) = P(X \leq x)$$

- cadlag: continue à droite, limite à gauche
- Left limit 0
- Right limit 1
- non-decreasing

Examples

- Roll a die
- Traffic lights waiting time.
- Radioactive decay

Density and mass functions

Def 1.6.1: Discrete r.v. - probability mass function

$$f_X(x) = P(X = x) \text{ for all x}$$

Def 1.6.3: Continuous r.v. probability density function $f_X(x)$ satisfies

$$P(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(x)dt$$