Aston Business School

Week 5 Statistics and Probability

BN2255 – Business Analytics in Practice

Topics of this presentation

- Descriptive statistics
  - Measures of location
  - Measures of dispersion
  - Coefficient of Variation
- Basics of Probability

‘Descriptives’

- First thing to do in any type of statistical analysis is to look at the ‘descriptives’
  - short-form for a range of statistical measures that ‘describe’ the data
    - namely, measures of location and dispersion
- Most important measures of location:
  - Average (mean), Mode and Median
- Most important measures of dispersion
  - Standard deviation (STDEV), Minimum (MIN), Maximum (MAX) and Range
- Descriptive statistics can be calculated ‘manually’ or automatically through the **‘Data analysis’ plugin**
  - As always, manual offers greater control and leaves an audit trail

Measures of location

- Average (arithmetic mean): add all values and divide by however many they are
  - usually taken as the 'expected' value of a data series
    - Eg. average annual salary for a business school graduate is £22,000
    - represents our 'best guess'
    - sometimes this is misleading!
- Median: the mid-point value of a data series that separates the data into two equal halves
  - also a measure of what is 'expected'
    - sometimes more telling than the average
      - what does it mean when the median is larger (smaller) than the mean?
- Mode: the common value in the data series
  - can also represent an 'expected' value, but not often used for that purpose

Measures of Dispersion

- Minimum and Maximum values
  - first indication of how ‘spread’ the data series is
  - sometimes useful in detecting ‘unusual’ values
    - counter-intuitive values that can be data errors
  - the difference between the maximum and the minimum values is the range
- Standard Deviation = a measure of distance between each observation and the mean
  - Distance is expressed as the difference of each observation to the mean, **squared**
  - Most common way to measure the variability of a data series
    - High standard deviation, high variability – data are more disperse

Coefficient of variation

- But how can we tell if the calculated standard deviation is high?
  - Compare it with the mean!
- Coefficient of Variation = STDEV()/AVERAGE()
  - a measure of how large is a standard deviation relative to the mean
- CoV is measure of *relative* consistency
  - High values (eg. greater than 1) suggest that the data is widely dispersed
    - That implies that the average of the series might not be representative of the majority of observations
      - Might not be accurate if used as a forecast!
- High (or low) CoV values are neither ‘good’ or ‘bad’!
  - They are just a feature of the data

Probability

- Probability = how likely it is that an event will occur
- Notion of probability is necessary for rational decision making!
  - most business decisions are fraught with uncertainty
  - need to develop a framework that allows us to account for uncertainty when planning
  - no uniquely ‘correct’ way of doing that!
    - very difficult to assess ‘correctness’ since there is no counterfactual
  - but the most widely used approach uses probabilities to derive expected value
    - More on the notion of expected value on the next lecture

Probability Axioms

Kolmogorov Axioms
[Ω = sample space, E = any event in Ω, P = probability measure]

First Axiom:
The probability of an event is a non-negative real number:
$\mathbf{P(E) \in \mathbb{R},\ P(E) \geq 0, \quad \forall E \in \Omega}$

Second Axiom:
There are no elementary events outside the sample space.
$\mathbf{P(\Omega) = 1.}$

Third Axiom:
Any countable sequence of disjoint events E1, E2, … satisfies
$\mathbf{P(E_1 \cup E_2 \cup \ldots) = \Sigma\, P(E_i)}$

- In simple terms:
- Probability of an event is between 0 and 1
  - 1 means that the event will always happen
  - 0 means that the event will never happen
- When you add the probabilities of all possible events, the sum needs to be equal to 1
  - something will happen!

Other rules of probability

Ratio Rule:

$P(A) = m/n$, m = number of all outcomes when A is true
n = number of all possible outcomes

Addition Rule:

$P(A \text{ or } B) = P(A) + P(B)$, if A and B are mutually exclusive

$= P(A) + P(B) - P(A \text{ and } B)$, if they are not

Complement Rule:

$P(A) + P(\sim A) = 1$, where "~A" = "A is not true"

Multiplication Rule:

$P(A \text{ and } B) = P(AB) = P(A) \times P(B)$, if A and B are independent events

$= P(A) \times P(B|A)$, if A and B are dependent events

Conditional Probability Rule:

$P(B|A) = P(A \text{ and } B) / P(A)$