Lesson 10 – Statistics for Exploratory Data Analysis

**Questons for Mentor:**

**Learning From Data:**

* Chapter 3: Why are we looking at Data anyway? Populations and Measurement
  + Inductive Inference: taking particular instances and trying to work out general conclusions
  + Induction is generally uncertain as opposed to deduction which is logically certain
  + Random sampling soup example – you don’t need to eat all the soup to find out if it needs more seasoning. You can just taste a spoonful, *provided you’ve given it a good stir.*
  + External validity
  + When we want to use the data to draw broader conclusions about what is going on around us, then the quality of the data becomes paramount
  + Mean and standard dev are statistics when describing a set of data, and parameters when describing a population
  + Populations
    - Literal population: identifiable group such as when we pick a person at random when polling
    - Virtual population: we frequently take measurements using a device such as taking someone’s blood pressure
    - Metaphorical population: There is no larger population at all, we didn’t do any samplint, we have all the data and there is no more we could collect, like number of murders each year, exam results for a class etc
* Chapter 4: What Causes What?
  + Correlation does not imply causation
    - Just because 2 things are correlated, doesn’t mean one thing caused the other
  + What is causation?
    - Doesn’t mean X causes Y 100% of the time
    - It means X happening increases the chances of Y happening
      * Smoking causing lung cancer example
  + Must have a diverse population for an effective study
  + Separate the population using real drug and placebo and don’t let them know which group theyre in
  + Confounder – a factor that affects multiple things similarly
    - Increase in ice cream sales and increase in drowning deaths are both due to the weather
  + Simpson’s paradox
    - Example with admission data from Cambridge where men had higher acceptance rate overall but women had higher acceptance rate in each subject
* Chapter 5: Modelling Relationships using Regression
  + Linear regression - Least squares model
    - Best fit line going through middle of cloud of points in scatter plot indicating the mean values for the heights of fathers and sons but doesn’t follow diagonal line of ‘equality’
  + Regression = any process of fitting lines or curves to data came to be called ‘regression’
    - In regression, dependent variable is what we want to predict or explain – y-axis
  + Statistical models have two main components
    - A mathematical formula that expresses a deterministic, predictable component
  + Multiple linear regression
    - Multiple explanatory variables (independent variables)
  + Logistic regression
    - Curve cannot go above 100% or below 0%
  + 4 broad modeling strategies
    - Rather simple mathematical representations for associations, such as the linear regression analyses in this chapter, which tend to be favoured by statisticians.
    - Complex deterministic models based on scientific understanding of a physical process, such as those used in weather forecasting, which are intended to realistically represent underlying mechanisms, and which are generally developed by applied mathematicians.
    - Complex algorithms used to make a decision or prediction that have been derived from an analysis of huge numbers of past examples, for example to recommend books you might like to buy from an online retailer, and which come from the world of computer science and machine learning. These will often be ‘black boxes’ in the sense that they may make good predictions, but their internal structure is somewhat inscrutable — see the next chapter.
    - Regression models that claim to reach causal conclusions, as favoured by economists
* Chapter 6: Algorithms, Analytics and Prediction
  + Algorithms contain:
    - Classification – the situation we’re facing
    - Prediction – tells us what’s gonna happen
  + Find patterns with clusters – unsupervised learning
  + Feature engineering – reduce raw data down to manageable dimensions due to large *p*
  + Classification tree – takes a few variables and uses yes/no tree to classify
  + Mean squared error – average of the squares of the errors, and is analogous to the least-squares criterion from regression analysis
  + Don’t over-fit a model
    - If you fit too closely, your model becomes based on such a small population of records that it’s unbiased but it lacks variability and is not a reliable estimate
  + Other models
    - Random forests: comprise large number of trees, each produce a classification with final classification decided by majority vote
    - Support vector machines: to find linear combinations of feature that best split diff outcomes
    - Neural networks: layers of nodes, each node depending on previous layer by weights. Aka deep learning models.
    - K-nearest-neighbour: classifies according to majority outcome among close cases in training set
* Chapter 7: How sure can we be about what’s going on? Estimates and intervals
  + Uncertainty is part of statistics and must be accepted
  + Sample size is key to accuracy
  + Bootstrapping is taking multiple samples from the population
  + Central limit theorem – distribution of sample means tends to trend towards a normal distribution
  + Variability in stats is based on samples
  + Shape of distribution of statistics does not depend on shape of distribution of the original distribution where the data points are drawn
* Chapter 8: Probability Theory: The language of uncertainty and variability
  + Classical probability: how we learn in school with coins, dice and packs of cards
  + Enumerative probability: colored socks in a drawer example. What’s the probability of drawing a certain color sock?
  + ‘long run frequency’ probability: proportion of times an event occurs in an infinite sequence of identical tests
  + Propensity or ‘chance’: idea that there is some objective tendency of the situation to produce an event
  + Subjective of ‘personal’ probability: somebody’s feeling about a particular situation
* Chapter 9: Putting probability and statistics together
  + Formulae from probability theory result in insight and convenience and don’t rely on a simulation
  + Aleatory uncertainty – before it happens, the uncertainty (i.e. before coin is flipped)
  + Epistemic uncertainty – after it happens, the uncertainty (i.e. scratch ticket outcome is decided, we just don’t know it yet)
  + Statistics are used when we have epistemic uncertainty about some quantity of the world
  + Confidence interval is the range of population parameters for which our observed statistic is a plausible consequence
  + 95% confidence interval is generally 2 standard deviations
* Chapter 10: Answering Questions and Claiming Discoveries
  + Hypothesis – proposed explanation for a phenomenon
    - Not absolute truth, but a provisional, working assumption
  + The null hypothesis is what we’re willing to assume is the case until proven otherwise
  + **P-value** – the probability of getting a result at least as extreme as we did, if the null hypothesis (and all other modelling assumptions) were really true
    - A measure of the compatibility of the data with some preformed hypothesis
    - Can be one-tailed or two-tailed depending on if we observe such an extreme value in one of two variables (i.e. men and women folding their arms with 7% difference in either direction)
    - If a P-value is small enough, then we say the results are statistically
  + Chi-squared test of association – measures the total discrepancy between the observed and expected counts given the null hypothesis of no-association true
    - Overall measure of the dissimilarity between the observed and expected counts
  + T-statistic – tells us whether the association between an explanatory variable and the response is statistically significant
    - T-values >2 or <-2 correspond to P < 0.05
  + **Type I Error**: is made when we reject a null hypothesis when it is true
  + **Type II Error**: is made when we do not reject a null hypothesis when in fact the alternative holds
* Chapter 11: Learning from Experience the Bayesian Way
  + Bayes theorem – provides a formal mechanism for learning from experience
    - The initial odds for a hypothesis x the likelihood ratio = the final odds of the hypothesis
  + Likelihood ration – probability of true / probability of false
  + Scientifically correct way to change our mind based on the introduction of new evidence
  + Bayes method is more subjective and less scientific than other methods
* Chapter 12: How things go wrong
  + The fight to be statistically significant may be hurting statistical community
  + Faulty statistical methods can drive major issues
* Chapter 13: How we can do statistics better
  + Change must be driven by the scientists, media and consumers