STATG006: Introduction to Statistical Data Science

Thomas Honnor
t.honnor@ucl.ac.uk

Department of Statistical Science
University College London

Chapter 1

- Statistical Data Science
- Overview of key concepts

Statistical Data Science

- Data Science
- Statistical Data Science
- Relationship to Machine Learning
- This module
- Quantity of material
- A note on examples

Data Science

- A non-specific title which will mean different things to different people
  - Wikipedia: “[A]n interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insight from data in various forms”
  - Also Wikipedia: “used as a sexed-up term for statistics”
- In short: The study of processes for extraction of information from data
- How does this differ from Statistics?
- How does this differ from Machine Learning, Data Mining, Predictive Analytics etc. etc.?

Statistical Data Science

- Is the science and engineering behind a Google query also “Data Science”?
  - Yes, why not?

- In Statistics we are concerned with **modelling uncertainty**

- As part of this *Probability* will play an important role in either the **design of models**, or in understanding the **properties of such models**

Relationship to Machine Learning

- Machine Learning has *Artificial Intelligence* (a.k.a. *autonomous systems*) as a key motivation

- In principle, it tries to remove humans from the decision making loop

- Applications may be driven by humans
  - Spam filtering
  - Image recognition (see example)
  - Advertising

- Process may also be semi-autonomous
  - Detection of tumours in medical images

This module

- We will provide an overview of the major statistical ideas

- The intersection with Machine Learning will be clear but the emphasis of this course is different
  - We will put less emphasis on prediction than in Machine Learning
  - More emphasis on modelling and analysing the properties of the models we develop
  - Prediction remains important

- Data Science is an iterative process and the idea is to mimic this in class

This module

- As a compulsory (Introduction) module this course aims to more self-contained than most you will encounter
  - Though other modules will develop further some of the topics we cover
- The emphasis is less on Mathematics, more on learning by example
  - This will be reflected in the ICA and examination
- However, there is no Statistics without Probability and the two will be intertwined
  - A degree of quantitative maturity is assumed, as well as some working knowledge of Probability

Quantity of material

- From current descriptions this might seem to be a lot of material
  - It is!

- Our syllabus is broad and somewhat shallow
  - The degree of Mathematical detail is kept to a minimum
  - Future careers in Data Science will at times require the ability to “talk Gaussian”

- To keep up you will need to make use of all available resources
  - Moodle (including forums)
  - Office hours

A note on examples

- Many of the examples introduced in lectures will use R code

- Some students will see R in more detail in other courses
  - If not, don't fret
  - There will be no need to write R code in this module, but some examples will be more interesting if you can
  - The same examples could also be produced in MATLAB/Python/Julia etc. if you have alternative experience

- Regardless, I highly encourage you to load up example code in R, run through it and try to understand what is happening and why
  - Change something, break it, fix it/reload and try something else

Overview of Key Concepts

- Statistical Inference
- Samples and Populations
- Random Variables
- Probability Densities
- Fitting
- Summary

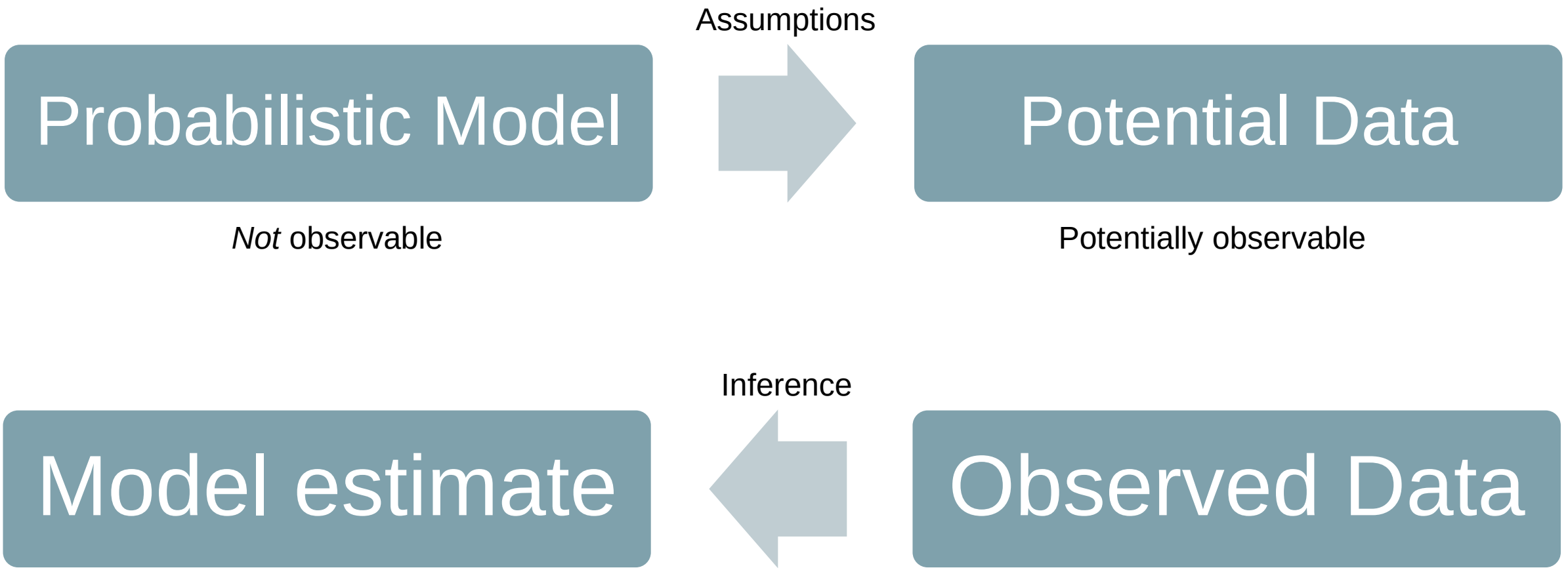
Outline

- A first taste of Statistical Modelling:

From assumptions to data, from data to conclusions

1. Case study: human height and other measures
2. Tools: plots, simple probability models
3. A taster for statistical evaluation of models
4. A taster for computational aspects of model fitting
5. Conditional probability and estimation

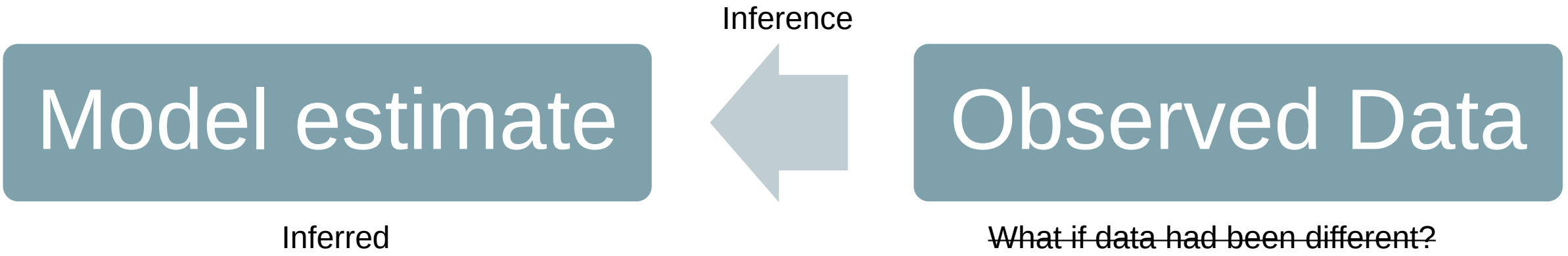
A sketch of Statistical Inference



A sketch of Statistical Inference

- Without going into detail, in this module we focus on **Frequentist Inference**
  - We consider what would happen had the data been different, and the long run consequences

- STATG004 will teach some of you about **Bayesian Inference**



An illustrative example

- The Third National Health and Nutrition Examination Survey
  - "designed to provide national estimates of the health and nutritional status of the United States … population"
  - https://www.cdc.gov/nchs/nhanes/index.htm

- We will use this illustrate some of the following concepts
  - Focus in particular on measurements of sex, age, weight and height
  - Data (following some pre-processing) and code available on Moodle

Activity

- Requirements (available from Moodle):
  - chapter1.R
  - nhanes.dat

- We will:
  - Examine heights
  - Investigate how heights differ according to subgroups
  - Consider how to model height data
  - Consider how to assess the variability of the proposed model

Samples and populations

- There is a group of people which we collected data from (the **sample**)
  - There is also a further group of people which we did not collect data from
  - All of these potential people form a **population**

- Our aim is to carry out **Statistical Inference**
  - We want to **characterise the population**
  - Using the information we have from the sample
  - What does this mean in an era where we can collect data much more cheaply than ever?

Samples and populations

- Sometimes we can collect “all” of the data
  - Is this then the population?
  - Is inference just descriptive in such a case?

- Populations can be infinite, even if you believe you have access to “all” of the data
  - A population can include future units
    - People to be born
    - New products to be recommended
    - New spam emails to be intercepted
    - Political campaigns to be run

Samples and populations

- Within the field of Statistics a *statistic* is just a function of the data, that is, any summary of the sample

- While this may seem like a technicality, it emphasises an important distinction:
  - Statistics are based upon quantities you directly observe
  - If an object depends upon unknown quantities then it is not a statistic

Samples and populations

- While we observe and make inference upon a sample, **judicious assumptions are necessary to relate a sample to a population**

- One such assumption is the basis by which the sample was selected from the population
  - Why some units ended up being in the sample and why some did not
  - More detail on this step is provided in the course STATG002: Statistical Design of Investigations

Samples and populations

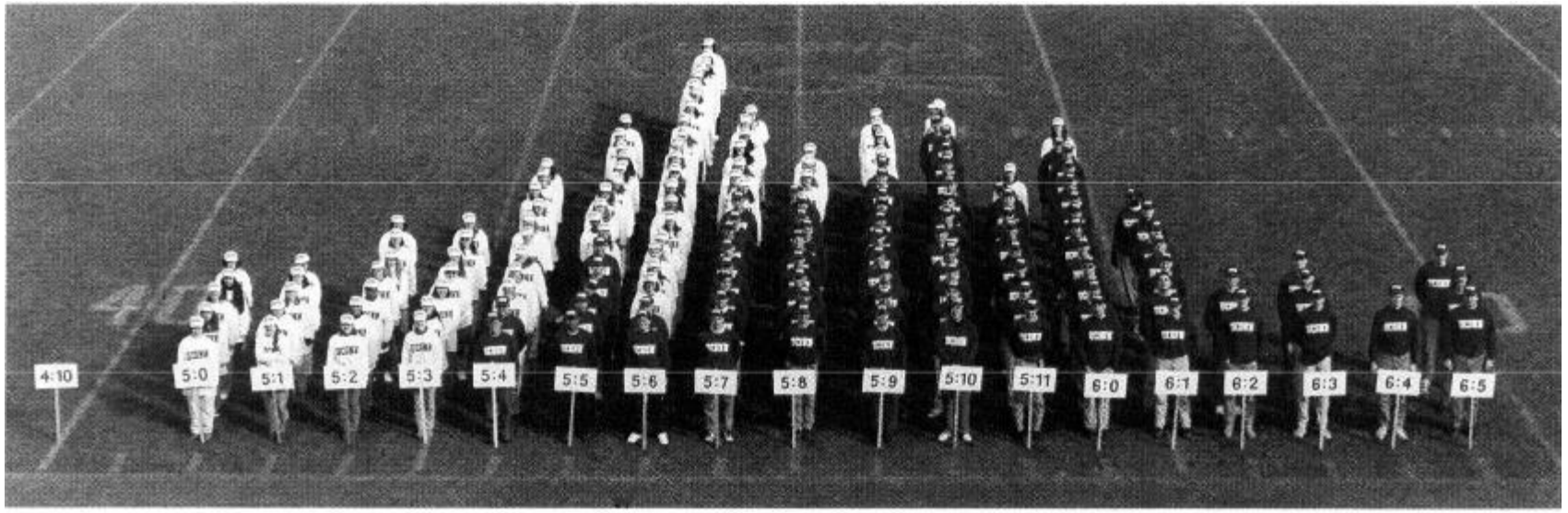


Figure 7. Living histogram of 143 student heights at University of Connecticut.

Schilling et al., (2002) "Is human height bimodal?"
The American Statistician Volume 56, Issue 3, 2002

- We can produce a similar histogram plot in R using the NHANES data

“I take no prisoners, I make no assumptions”

- Some people may be of the opinion that “data speaks for itself”
  - That opinion is the antithesis of the approach required to successfully investigate data within the Frequentist framework
- Data does not speak for itself
  - There is only so much information about the population which a sample can provide
  - In some cases the population is infinite, while samples are most certainly finite
- We do make assumptions
  - Which are not always correct
  - Which are **not** the same as blind belief

Our first link to probability

- We assume that the variability in our data in generated probabilistically

- Heights may be considered to be real, continuous numbers
  - What can we possibly mean by “the probability of that person’s height being 1.90 metres is 0.002”?
  - Consider the alternative “the probability of that person’s height being between 1.89 and 1.91 metres”
  - By what framework can we describe statements like this more generally?

Randomness

- Our data may be interpreted as a collection of **random variables**

- There is a formal mathematical definition of a random variable
  - But, we aren't covering all of the mathematical details
  - It suffices to say that it is a quantification of random **events**
  - A series of random events gave you your current height
  - These random events came from some **sample space**
  - Your measurement corresponding to the outcome of this process is a random variable

Random variables

- Random variables can be manipulated algebraically in the same manner as ordinary variables

- If we define *X* to be an individuals weight (in kg), *Y* to be their height (in m), we may define the individuals body mass index, *Z*:

$$Z = \frac{X}{Y^2}$$

- In this scenario *Z* itself is a random variable as a function of random variables

Notation

- Throughout this course we will use a notational convention adopted by many statistical texts

- Upper case letters will be used to denote random variables
  - X, Y, Z, etc.

- Lower case letters will be used to denote values taken by random variables
  - x, y, z etc.

Scope

- We will typically divide random variables into two main classes
  - **Discrete**
  - **Continuous**

- A discrete random variable takes values from a discrete set
  - Discrete sets may be infinite
  - Discrete random variables may also be used to represent categories eg. “male”, “female”, “heads”, “tails”
  - In such cases they are formally encoded as numbers eg. “heads” = 1, “tails” = 0

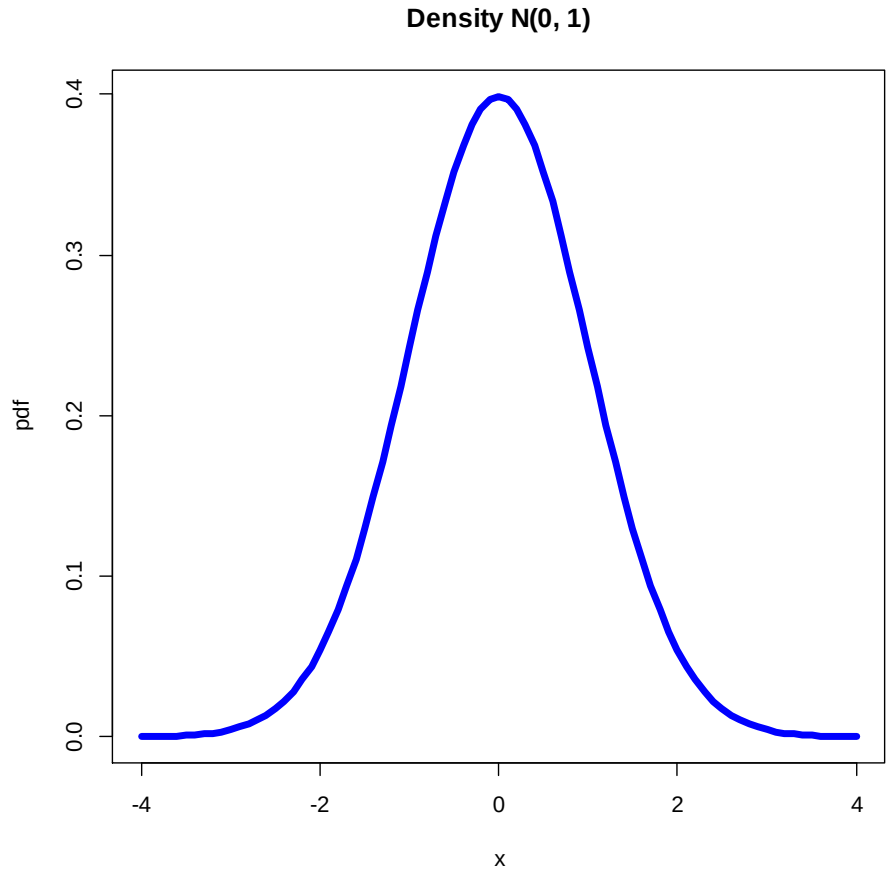
Example

- Let us define *Y* to be the height of an individual in metres

- We will simulate plausible randomness by **subsampling**
  - Begin with the NHANES collection of 20 050 individuals heights
  - From this data sample with replacement *n* times
  - The resulting collection of *n y*'s may be considered a random sample of the heights of *n* individuals

- We may investigate how our information about *Y* changes with each sample and with the sample size, *n*, in R

Probabilities, densities and the Gaussian (or Normal) Distribution



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

- An explanation of p(x) is given in the following slides

Density Functions

- Probability distributions are often described by a corresponding **Probability Density Function** (PDF)
  - However, we can't apply probabilities to point outcomes in continuous spaces
  - We instead change our thinking towards the events of finding outcomes in a particular interval

- This leads to consideration of the **Cumulative Distribution Function** (CDF)

$$F(x) \equiv P(X \leq x)$$

From CDFs to PDFs

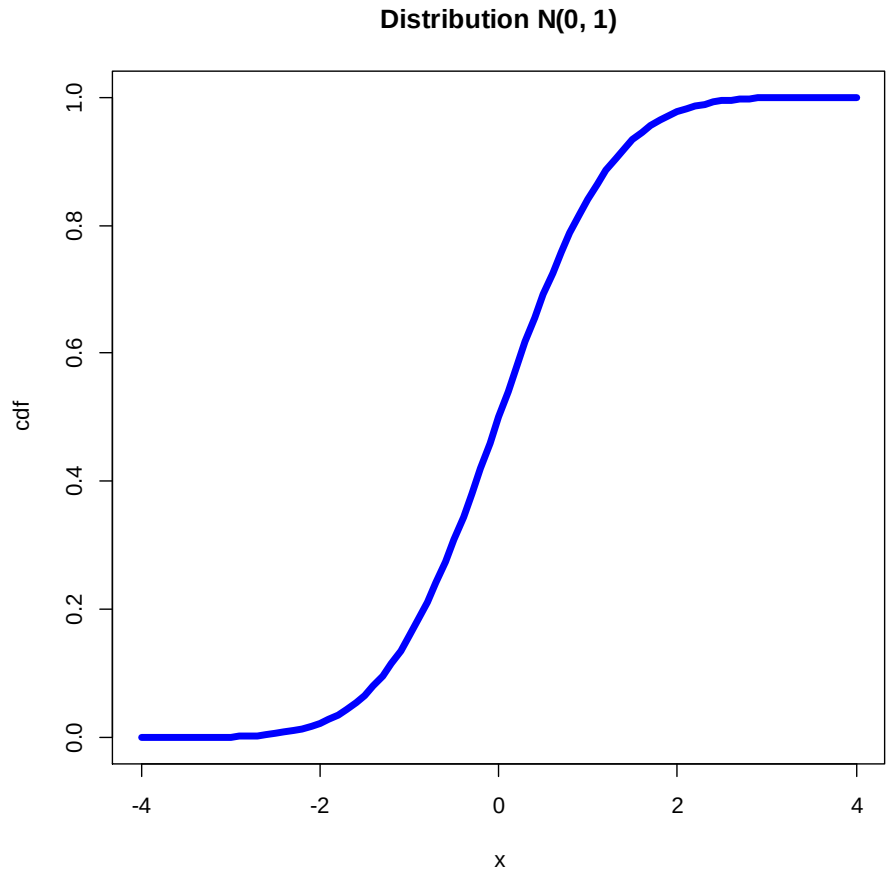
- For continuous data the density function (PDF) is just the derivative of the distribution function (CDF)

$$p(x) \equiv \frac{dF(x)}{dx}$$

  - In the majority of cases it is easier to think in terms of PDFs than CDFs

- We will consider discrete random variables later
  - We describe the Probability Mass Function (PMF), the probability of a discrete X taking a particular value

For illustration purposes



- This is a plot of the CDF for a Normal(0,1) random variable
- The exact shape is specific to this distribution
- More general features are common to all CDF
  - Monotonicity
  - Bounds at 0 and 1

The Normal (or Gaussian or Bell Curve)

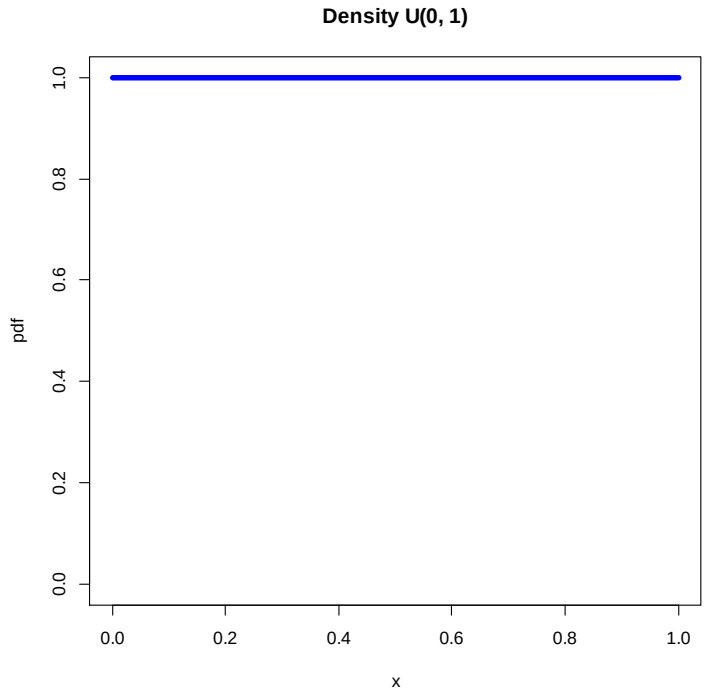
- The motivation for considering the Normal distribution is its ubiquity

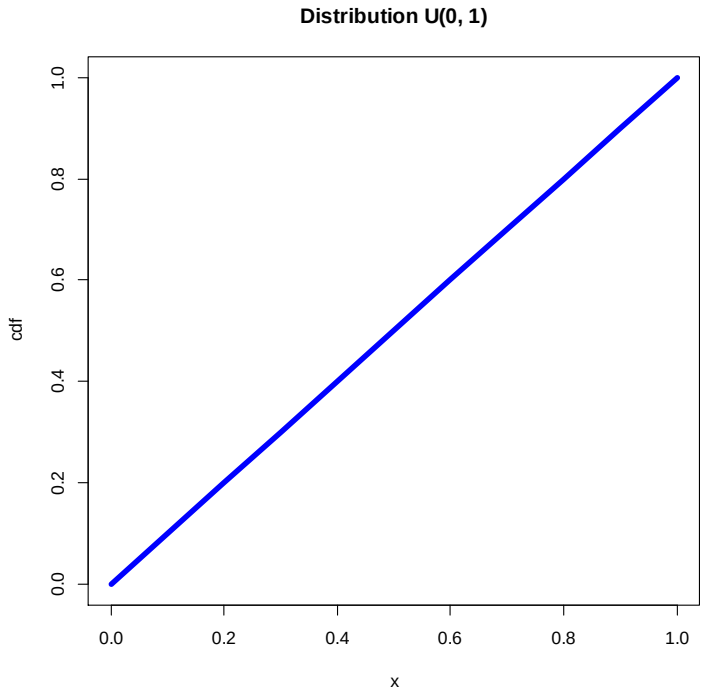
- If we average many random variables their distribution will converge to a Gaussian
  - The more random variables we average over, the closer the resulting distribution will be to the Normal distribution

- This property is known as the **Central Limit Theorem** and there are several variations of it
  - More details can be found in the books by Rice and Wasserman and others

Demonstration

- Consider, *X*, following the Uniform distribution on [0,1] (the support)





$$X \sim U[0,1]$$
$$0 \leq x \leq 1$$

$$p(x) = 1$$
$$F(x) = P(X \leq x)$$
$$= x$$

Demonstration

- Now, consider a vector of *n* independent and identically distributed random variables $X^{(i)}$ with $i = 1, 2, 3, \ldots, n$

$$X^{(1)}, X^{(2)}, X^{(3)}, \ldots, X^{(n)}$$

$$X^{(i)} \sim U[0, 1]$$

Demonstration

- On some occasions in the future we will use bold face to denote random vectors

$$\mathbf{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \\ \cdots \\ X^{(n)} \end{bmatrix}$$

- To proceed, presume that you don't see ***X***
  - Rather, we only observe the average which we will denote by *Y*

$$Y = \frac{1}{n} \sum_{i=1}^{n} X^{(i)}$$

Demonstration

- Finally, let's say that you have measured Y_1, Y_2, Y_3, …, Y_p means of p sets, each containing n Uniform random variables
  - What then is the distribution of these Y?

$$Y_j \sim ?, \qquad j = 1, 2, 3, \ldots, p$$

- We can investigate the PDF of Y using R
  - Sample np Uniform random variables
  - Calculate the p sets of means (p = 1 000)
  - Plot a histogram of these means (n = 1, 2, 3, …, 10)
  - What happens as n increases towards infinity?

Returning to the previous example

- Last lecture we introduction a dataset of individuals heights and made references to the normal distribution

- We return to both of these and assume that human height follows a normal distribution
  - There are physical processes which determines height, but we abstract them away
  - Probabilities are used to summarise our ignorance

- This is a model
  - All models are approximations
  - Heights cannot be negative, normal random variables can be

Which normal distribution?

- The normal distribution has two parameters
  - Mean: μ
  - Variance: σ^2 (Standard deviation: σ)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

Model parameters

- The mean is the **location** parameter
  - For the normal distribution it tells you where the “peak” of the density will be located, p(x) is maximised when x = μ
- The variance is the **scale** parameter
  - The variance controls the degree to which the probability mass/density “spreads” around the mean
- R demonstration of the normal distribution density for a range of parameters

Model fitting

- In Machine Learning model fitting is also referred to as ‘learning’
- We may adjusted the free parameters such that the **implied model** in some sense “matches” the data
- There are multiple **scores** by which model fit may be quantified
  - There are corresponding **algorithms** to optimise scores for best fit
  - We will introduce some of these later
- R demonstration on model fitting and changing score

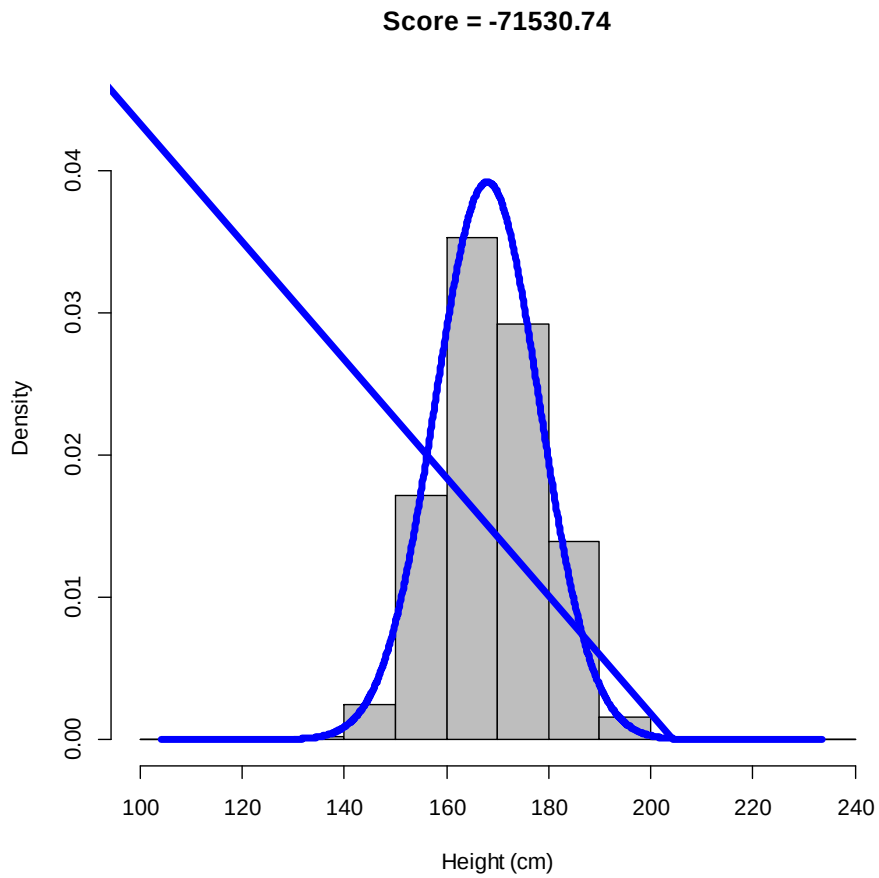
Best fitting model

- From the height data available we may conclude that the best fitting model **estimate** for one choice of score has parameters

$$\hat{\mu} = 168$$

$$\hat{\sigma} = 10.2$$

- An explanation for the derivation of these values will be given later



Verifying model fit

- We have proposed a normal model and determine the parameters which obtain the best fit

- But, is the normal distribution a “good” model?

- We may compare the normal CDF against the **empirical CDF** for observed data, $x^{(1)}$, $x^{(2)}$, …, $x^{(n)}$, given by

$$F_n(x) \equiv \frac{1}{n} \sum_{i=1}^{n} I(x^{(i)} \leq x)$$

Verifying model fit

- We can use R to plot the empirical CDF and normal CDF to check for agreement

- We might wish to see whether the similarity between distributions is by chance or not
  - Our data is assumed to have been generated by a probabilistic process
  - The data itself could therefore be different if we re-ran the experiment
  - Part of analysis is to qualify how likely it is for data to “look normal” even if the population is not normally distributed
  - We will return to this

Quantile-Quantile (Q-Q) plot

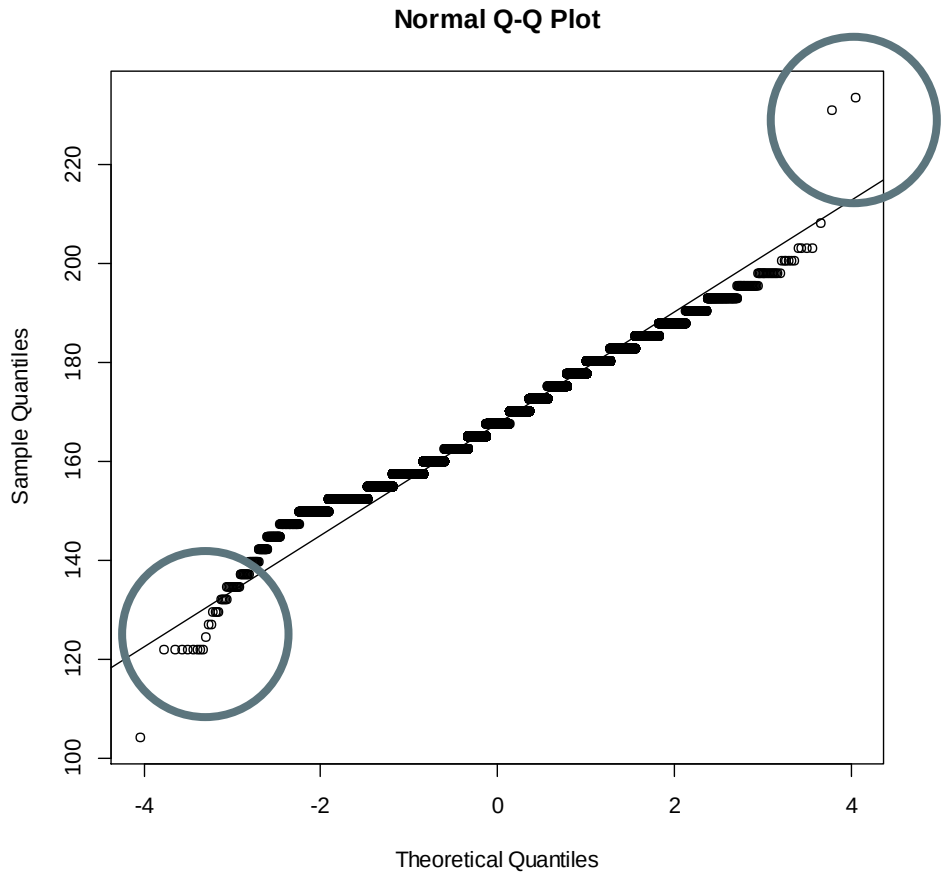
- A further method for visually assessing model fit is the quantile-quantile (Q-Q) plot

- A quantile is simply the inverse of the CDF
  - The 0.9 quantile is the value of x such that $F(x) = 0.9$, i.e. $x = F^{-1}(0.9)$
  - The median is the 0.5 quantile
  - The upper and lower quartiles are the 0.25 and 0.75 quantiles

- The Q-Q plot is similar to comparing CDFs

Q-Q Plot

- We can plot the Q-Q plot in R
- Good fit is indicated by a strong linear relationship
- Our normal assumption is reasonable
  - The model fits worse at the **tails** of the distribution



What next?

- We have learned an estimate of the approximate distribution of heights from a population

- We could then make use of this information
  - The range of clothing sizes to be kept in stock in a shop

- Snapshots vs. trends
  - We could collect data across time periods assuming changes in the population will result in interesting comparisons

Comparison of distributions

- We might compare distributions through summary statistics
  - A common choice is the **expectation** of the distribution

$$E[X] \equiv \int xp(x)\ dx$$

- The expectation or **mean** can be thought of as a weighted average or “centre of mass” of the distribution

- We could estimate the expectation of the distribution
  - The expectation is then our **estimand**, a feature of the distribution which is a quantity of interest

Expected value of functions of random variables

- As we have already stated, a function of random variables is itself a random variable

- We may therefore take the expectation of such functions of random variables by the following

$$E[f(X)] = \int f(x)p(x)\ dx$$

The use of expectations as summaries

- The value of expectations as summaries of distributions depends upon the chosen distribution

- For example, the expected value of a normally distributed random variable is the location parameter, μ

$$X \sim N(\mu, \sigma^2)$$

$$\Rightarrow E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} dx$$

$$= \mu$$

When may means not be useful

- Cases when data is highly skewed (asymmetric) or multi-modal (multiple peaks)

- Classic example is historical average life expectancy
  - In Classical Rome average life expectancy was between 20 and 30
  - However, given survival to age 10 average life expectancy is as high as 47.5
  - (https://en.wikipedia.org/wiki/Life_expectancy)
  - In STATG015 and STATG016 survival analysis, a more advanced method of modelling life expectancy, is discussed

NHANES data

- We have determined that for the particular normal model the expected height is 1.68m

$$\hat{\mu} = 168$$

- What if the data had been different?
  - We sampled 20 050 individuals and recorded 19 129 heights
  - What if a different set of 20 050 individuals had been sampled?
  - An upcoming discussion of confidence intervals will shed more light on this issue

Dual use of probability

- Probability therefore plays two roles in our analysis
- We will use it to **express assumptions about the population**
- It will also be used to analyse the **sampling properties** of our **estimators**

Summaries other than the mean

- The normal distribution is parameterised by its mean and variance parameters
  - However, mean and variance also refer to general summaries of a probability distribution

- For example, the variance of a random variable, *X*, is a measure quantifying dispersion (how "spread out" *p(x)* is)
  - There are other measures of dispersion

$$\mathrm{Var}(X) \equiv E\left[(X - E[X])^2\right] = \int (x - E[X])^2 p(x)\, dx$$

Further summary statistics

- The nth moment of a random variable, *X*, is simply the summary statistic given by

$$E[X^n]$$

- For example the mean is the first moment

- The variance may be written as a function of first and second moments

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2$$

Multivariate data

- We began discussion of the NHANES data by considering two measurements, sex and height

- These two measurements are not independent
  - We can see from earlier histograms that the data are distributed with different locations and shapes

- Multivariate models may be used to model such dependencies
  - Such models become vital for prediction tasks

Prediction in action

PNAS

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski[a,1], David Stillwell[a], and Thore Graepel[b]

[a]Free School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and [b]Microsoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

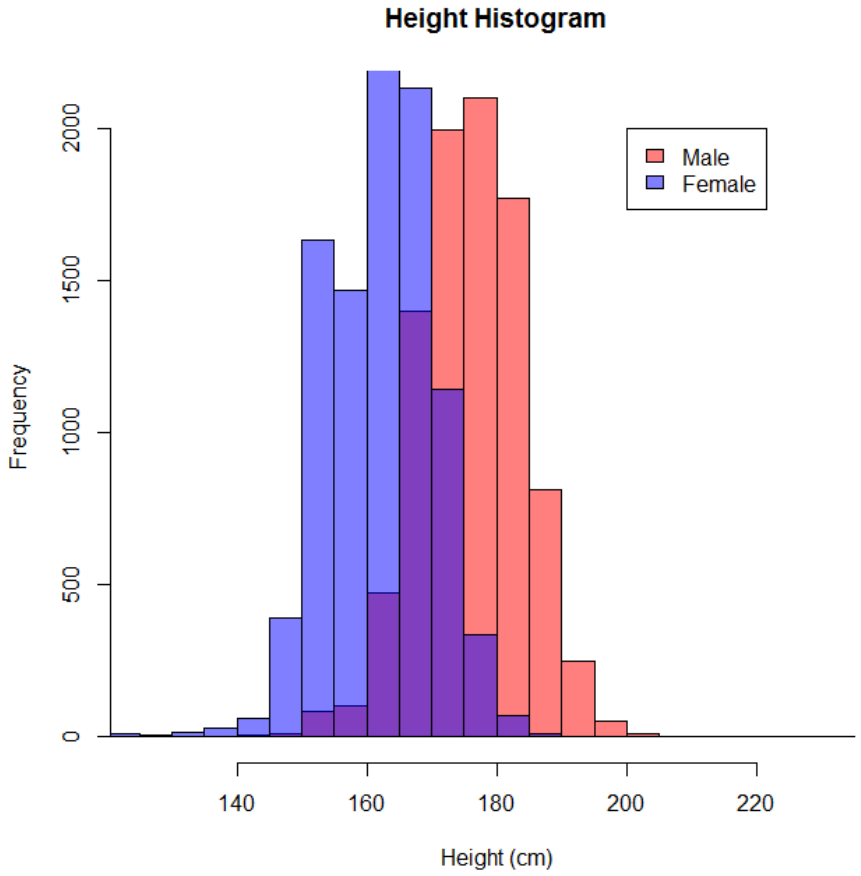
We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for

browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically

(Proceedings of the National Academy of Sciences, 2013)

Could we predict sex from height?



More than two dimensions

- We can visualise across two dimensions how an individuals sex affects both their height and weight
  - Height and weight are themselves associated and so they should not be treated as independent pieces of information
  - The resulting plot can be seen in the R demonstration

- In Chapter 2 we will see how to formalise association and prediction problems more precisely

Summary of Chapter 1

- Probability vs. statistical inference
  - Models to data vs. data to models
- We have seen two simple models so far
  - Normal distribution
  - Uniform distribution
- We have explored notions of parameter estimation
  - Explaining the data by matching the model to the data
- We have briefly touched on multivariate data, associations and prediction
- Chapter 2 goes on to discuss testing and confidence intervals