MPO1: Quantitative Research Methods
Session 1: Data and Probability

Thilo Klein

University of Cambridge
Judge Business School

CAMBRIDGE Judge Business School

Timetable

Timetable

- Lectures
  12, 19, 26 Oct, 2, 9, 16, 22, 29 Nov, 2-4pm @ Mill Lane 2

- Lab Sessions R and RExcel software
  Michael Freeman, http://frmn.me
  20 Oct, 2-4pm; 3, 17 Nov, both 11-1pm; 30 Nov, 9-11am, all @ Computer Lab
  Jerry He
  20 Oct, 2-4pm @ W2.02; 3 Nov, 12.30-2.30pm @ W4.03; 17 Nov, 11-1pm @ W2.02; 30 Nov, 9-11am @ KH107

CAMBRIDGE Judge Business School

Deadlines

Contest (Multiple Choice Exercises)

	Sheet 1	Sheet 2	Sheet 3	Sheet 4
Submit on	25 Oct	1 Nov	8 Nov	15 Nov
Weight	8 %	10 %	12 %	14 %

	Sheet 5	Sheet 6	Sheet 7	–
Submit on	22 Nov	29 Nov	6 Dec	–
Weight	16 %	18 %	22 %	–

Assessment (Workbooks)

	Book 1	Book 2
Handed out	30 Nov	12 Dec
Submit on	12 Dec	17 Jan
Weight	40 %	60 %

CAMBRIDGE Judge Business School

Objectives

Objectives of the module

- "Introduction" to *applied* statistical methods
- Mathematical sophistication $\sim$ simpler research journal papers in finance/strategy/marketing ...
- Learning by doing - do many exercises
- Should enable you to estimate useful, insightful and *exciting* regression models and make careful inferences.

Motivation

Patterns and relationships in Finance / Management / Economics with important strategic and policy implications:

- Do financial intermediaries reduce information asymmetries on online lending platforms?
- Does management advice improve productivity and performance of firms?
- Does microfinance reduce poverty?
- How much are people willing to pay for different hospital care packages?
- Does smoking lead to lung cancer?
- Do smaller class sizes lead to better test score performance?

CAMBRIDGE Judge Business School

Causal Effects

Using data to measure causal effects (1)

- Ideally we should do experiments:
  - e.g., experiment to estimate effect of access to microcredit on small enterprise revenue / household consumption / savings, etc.
- But almost always have to make do with *observational* (non-experimental) data
  - At best, data from "natural experiments"
  - Increasingly, behavioural finance, economics, management data come from class room experiments

Causal Effects

Using data to measure causal effects (2)

- What is the difficulty in using observational data when we wish to estimate causal effects?
  - Notion: Data generating process: empirical observations are outcome of (natural) "experiments"
  - The same experiment performed by "nature", leads to different outcomes (some randomness)
  - And, we have no control over the experiment of interest
    We need to:
    - identify the causes and factors relevant to the outcome of interest
    - To disentangle effects of the different causes on the outcome
    - To come to conclusions about these effects with some assurance about their level of accuracy, i.e., quantifying our uncertainty about conclusions

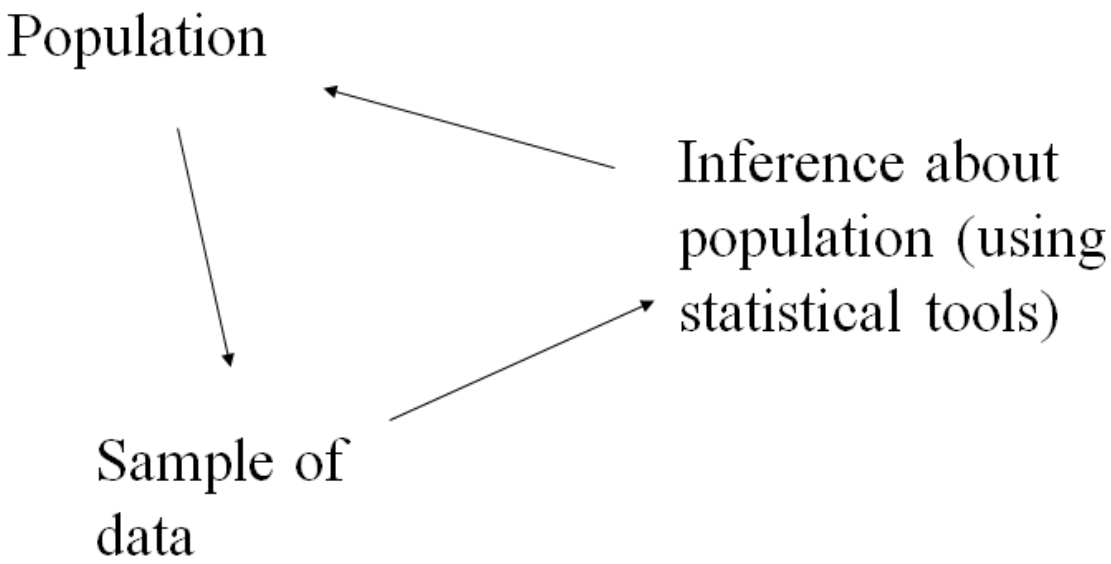
Learning points

You will (learn) ...

- Statistics studies *sets* of objects/entities/things (firms, individuals, households ...)
- Statistics studies "causes of variation": If there is no variation, one individual describes the population
- You will learn
  - How to exploit variation (between observations in data) to estimate causal effects
  - Hands-on experience of regression with focus on applications - theory only as needed
  - How to evaluate the other people's analysis - understand empirical papers critically

CAMBRIDGE Judge Business School

Quantitative research



CAMBRIDGE Judge Business School

Paradigm

So, you (should) have a *useful* theory about the phenomenon of interest. You need to solve:

1. the *Specification* problem - specify a model from (your) theory. The mathematical form you think governs the population. You do not know (and will never *know*) the *parameters* of this
2. the *Estimation* problem - choose methods to *estimate* the unknown parameters governing the population, using sample data
3. the *Inference* problem - quantify the degree of uncertainty attached to these estimates, given that they are based on just one (random) sample

Method

Step by step

- Formulate a model (based on hypotheses about the population)
- Gather data - sample
- Estimate the model - estimate population parameters
- Make inferences - test hypotheses about the population
- Interpret results, in terms of the theory

Review

Topics today:

- Data description: *Statistics* that summarise data - these are always "estimates" of the unknown population parameters
- Probability principles: how can the world be described in terms of random variables and probability distributions (i.e., probability models)
- Next: Introduction to statistical inference: drawing conclusions about the *population* from only one sample, using probability principles

Review (cont'd)

Further on...

- Estimation procedures for regression models: why and how they work
- Inference after regression: how to test hypothesis

CAMBRIDGE Judge Business School

First task

Reducing Data

- Data (set of observations from some measurement exercise) are the raw material
  Data sourced from:
  - Questionnaires/surveys/tests/measurements
  - Other peoples data: web, government records, databases. (was the data collected carefully?)
  - Planned/designed experiments
- Need to reduce data: summarize patterns of variation in data
- Capture the shape of data

Nature of Data (1)

Variable: characteristic/trait/attribute/measurement that can vary (have two or more different values.)

- Nominal / Categorical: must be classified into categories
  - Binary (dichotomous): Only two possible values, e.g., "Yes/No"
  - Multinomial: More than two categories, e.g., "red, green, blue"
- Ordinal scale: values based on rank order, e.g., Likert scales

Nature of Data (2)

Variable: characteristic/trait/attribute/measurement that can vary (have two or more different values.)

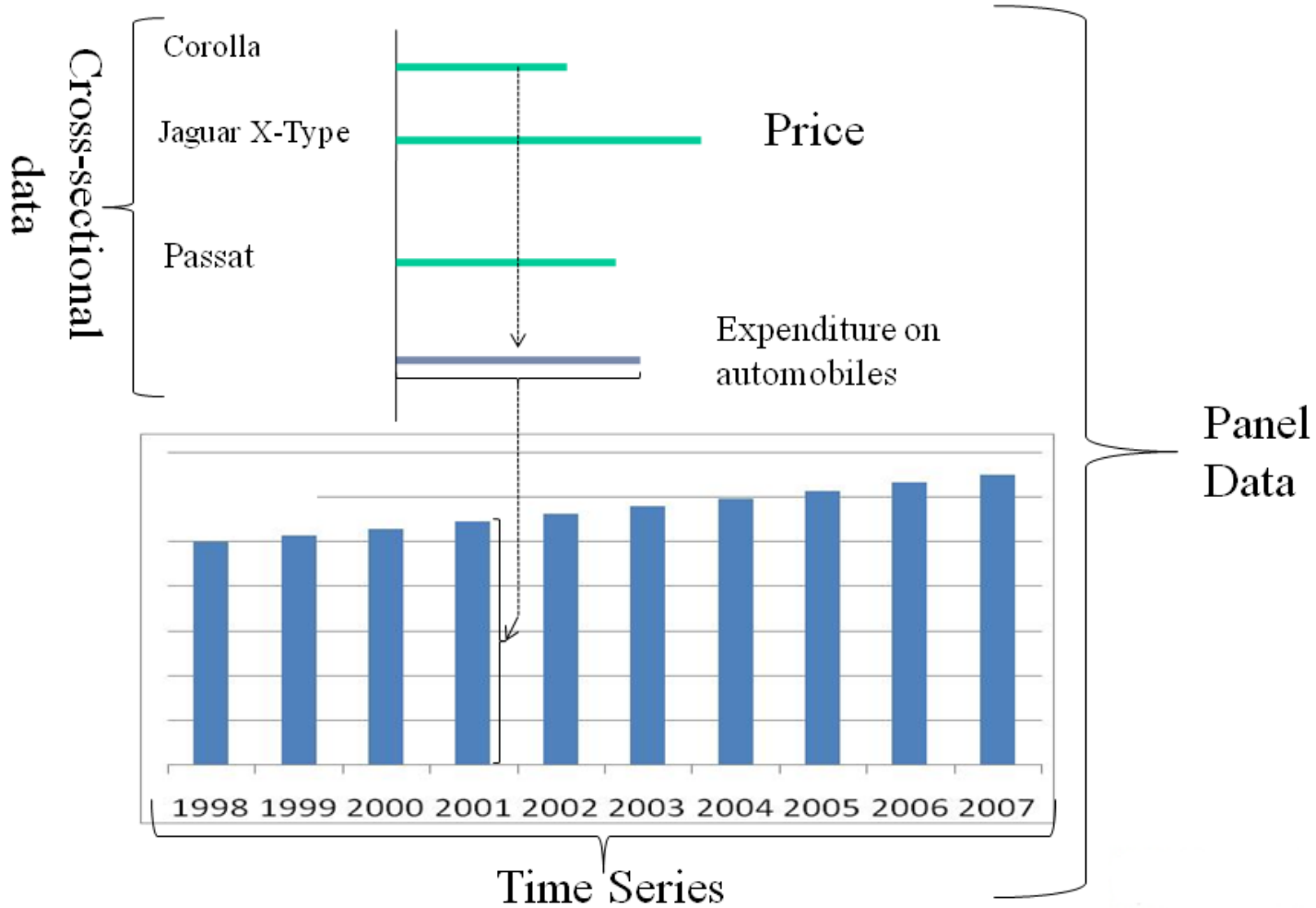
- Interval scale: distance along the scale means the same anywhere on the scale, but 0 does not represent absence; only difference between values has meaning. e.g., intelligence scores
- Cardinal, Ratio scale: same as interval scale, but 0 represents absence of the "thing", e.g., revenue
  - discrete (countable number of values), e.g., number of brands in a category
  - continuous (uncountably infinite number of values), e.g., profitability

Nature of Data (3)

Observations

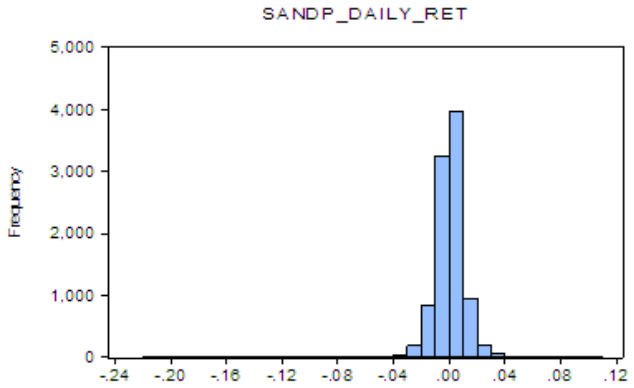
- Cross section data: Registrations by car "models" in the UK in September 2010
- Time series data: Cash rebate offered on Picasso each month 2000-2009
- Panel data: Cash rebates on different car models, monthly, 2000 to 2009

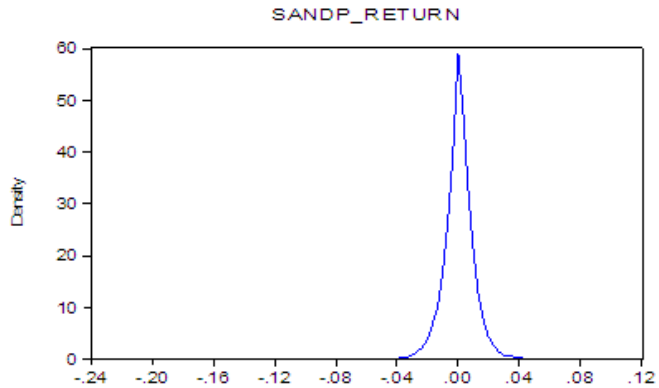
Types of Data

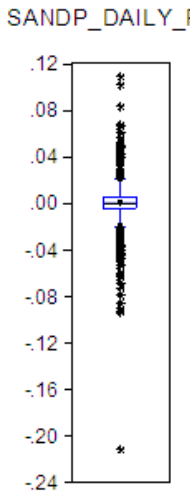


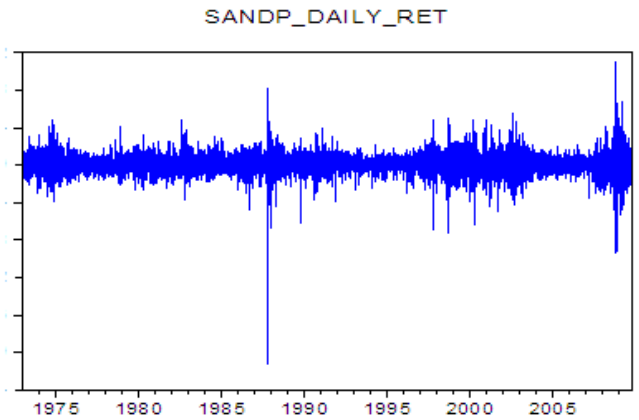
Variety of graphical representation

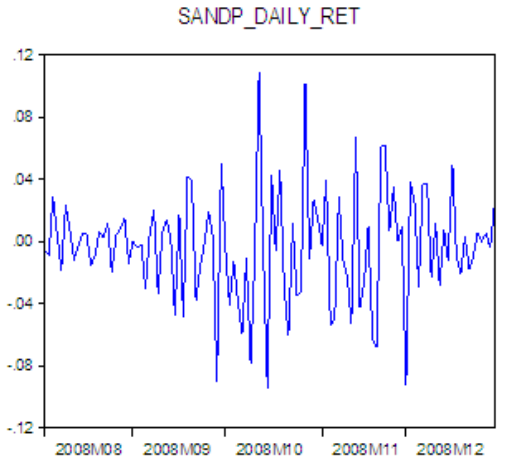
S and P daily returns, 1973 to 2009







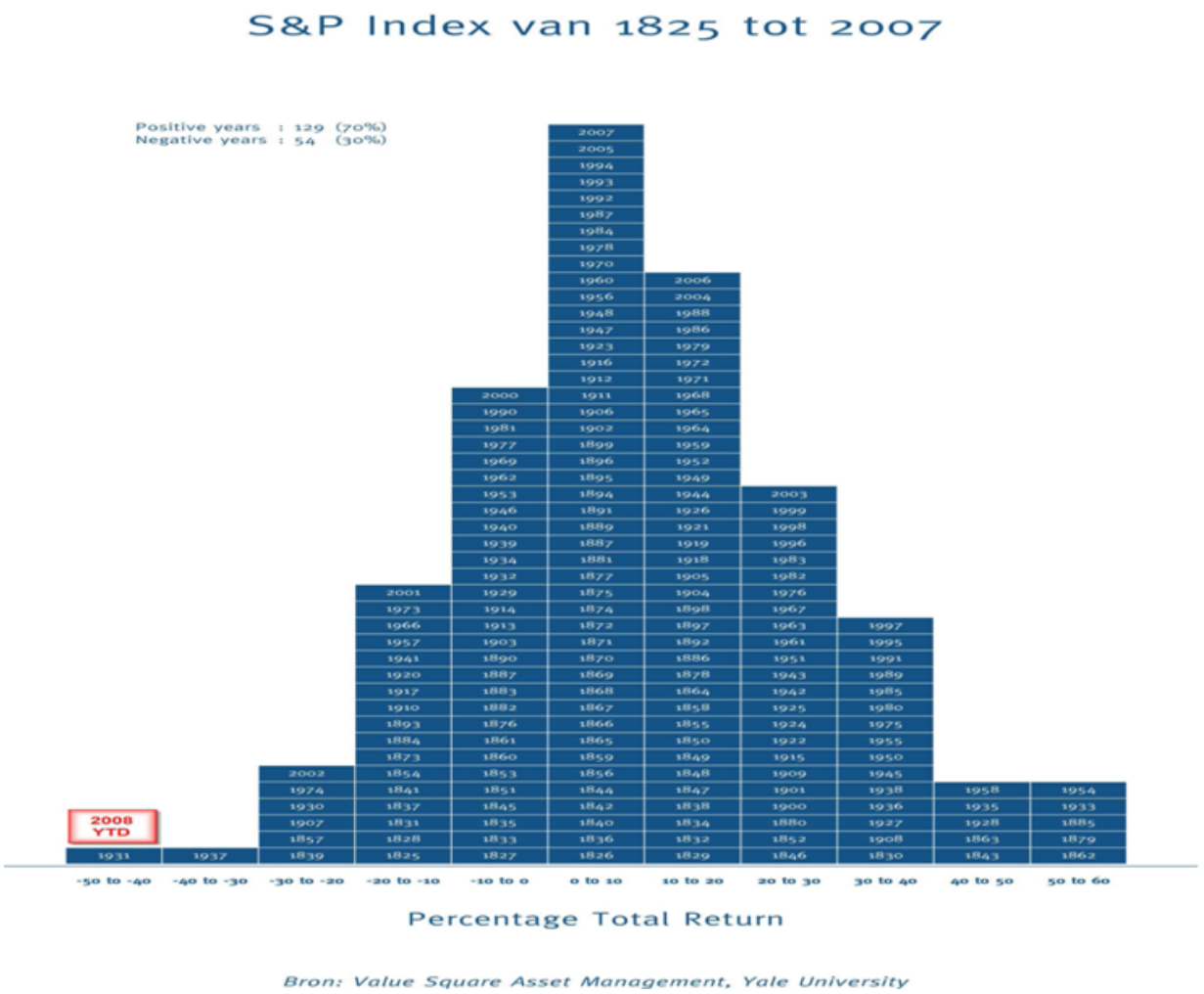




CAMBRIDGE Judge Business School

Variation in data

Histogram of S and P daily returns



Bron: Value Square Asset Management, Yale University

CAMBRIDGE Judge Business School

Summary Statistics

Summary (or Sample) Statistics; and corresponding Population parameters

- Characterize variation in data, reduce data bulk
- Use *moments*, i.e., quantitative measures of shape of a set of data points
  - Central tendency
    Dispersion
    Skewness
    Kurtosis
- Using related summary statistics
  - Covariance / Correlation (between two or more variables)
  - Regression coefficients (in cross section, time series, panel, limited dependent variable, and other regression contexts)
  - Post regression diagnostic statistics

Notation

Notation – Bling!

- X : Random variable
- x : A specific value of the r.v. X
- N : Number of observations
- $i = 1, 2, ..., N$: Index for observations in the data
- $\{x_i\}, \quad i = 1, 2, ..., N$: Data set
- $\sum_{i=1}^{N} x_i$: Summation
- $x_{(1)}, x_{(2)}, ..., x_{(N)}$: Ordered observations, smallest to largest

CAMBRIDGE Judge Business School

Central tendency

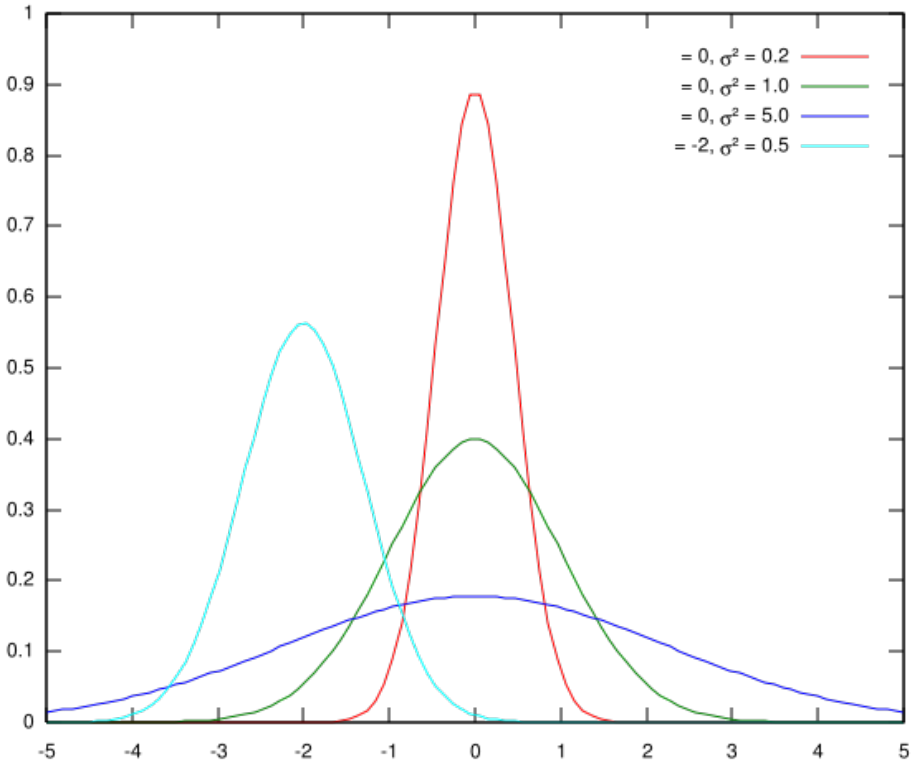
Central tendency and Quantiles

- $\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$: Sample mean μ_X: Population mean
- Quantiles and Percentiles

  p^{th} percentile : value of X such that p percent of observations fall at or below it.
  - Median, M = the 50^{th} percentile.
    - $M = x_{\left(\frac{N+1}{2}\right)}$
    - $M = \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N+1}{2}\right)}\right)$
    - Relative advantages: Median and Mean?
  - First quartile $Q_1 = 25^{th}$ percentile
  - Third quartile $Q_3 = 75^{th}$ percentile

Unimodal data (distribution)

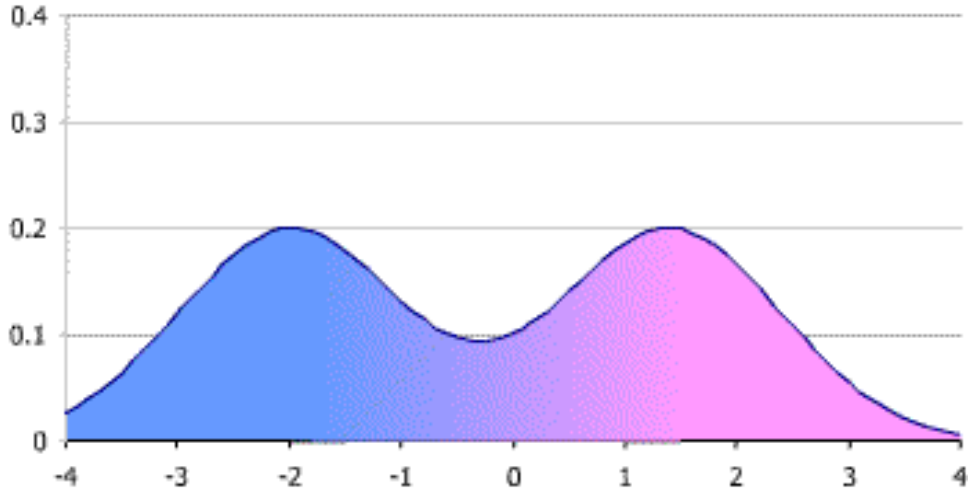
Mode: example of unimodal data (distribution)



CAMBRIDGE Judge Business School

Bimodal data (distribution)

Mode: example of bimodal data (distribution)



CAMBRIDGE Judge Business School

Dispersion

Measures of Dispersion

- Variation in a set of data: spread about the typical value. Are observations clustered or dispersed in the around a central tendency?
- $s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$: Sample Variance
  σ^2: Population Variance
- $s^2 = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - \overline{x}^2$: Sample Variance again
- s : Sample Standard deviation
  σ: Population Standard deviation
- $x_{(N)} - x_{(1)}$: Range
- $Q_3 - Q_1$: Inter quartile range
- $\frac{s}{\overline{x}}$: Sample Coefficient of variation
  $\frac{\sigma}{\mu}$: Population Coefficient of variation.

CAMBRIDGE Judge Business School

Standardising data (1)

z scores

- Standardizing: scaling different datasets to a common scale
  z-score : subtract mean and divide by standard deviation.

$$z_i = \frac{x_i - \overline{x}}{s}$$

- The number of standard deviations that a particular data point is from the sample mean
- Q:
  - What does a z-score of 2 mean?
  - What is the sample mean of z-values?
  - What is the sample standard deviation of z-values?

Standardising data (2)

Symmetric, mound shaped data, and z-scores

- For nearly symmetric, mound shaped data:
  - 68 % of the data lie within one st. dev. of the mean, approximately
  - 95 % of the data is within two st. devs. of the mean, approximately
- Ponder:
  - On October 10, 2008, % change in FTSE100 was 7 st.devs downward
  - On November 24, 2008, % change in the index was 7 st.devs upward.
  - On Sep 19, 1987, the drop was by 22 st. devs.

Aside

Chebyshev's inequality

- Chebyshev's (or Tchebysheff's) inequality states that in **any** data sample (or probability distribution):
  - *no more than* $1/k^2$ *of the values are more than* k *standard deviations away from the mean*
- Let X be a random variable with mean μ and finite variance σ^2
  Then for any real number $k > 1$:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

CAMBRIDGE Judge Business School

Skewness

Skewness (Sk): Converse of symmetry of the distribution

Sample:

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^3}{s^3}$$

Population:

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3}$$

- Sk Positive: long tail to the right
- Sk Negative: long tail to the left

Kurtosis

Kurtosis (Kt): Peakedness of the distribution

Sample:

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^4}{s^4}$$

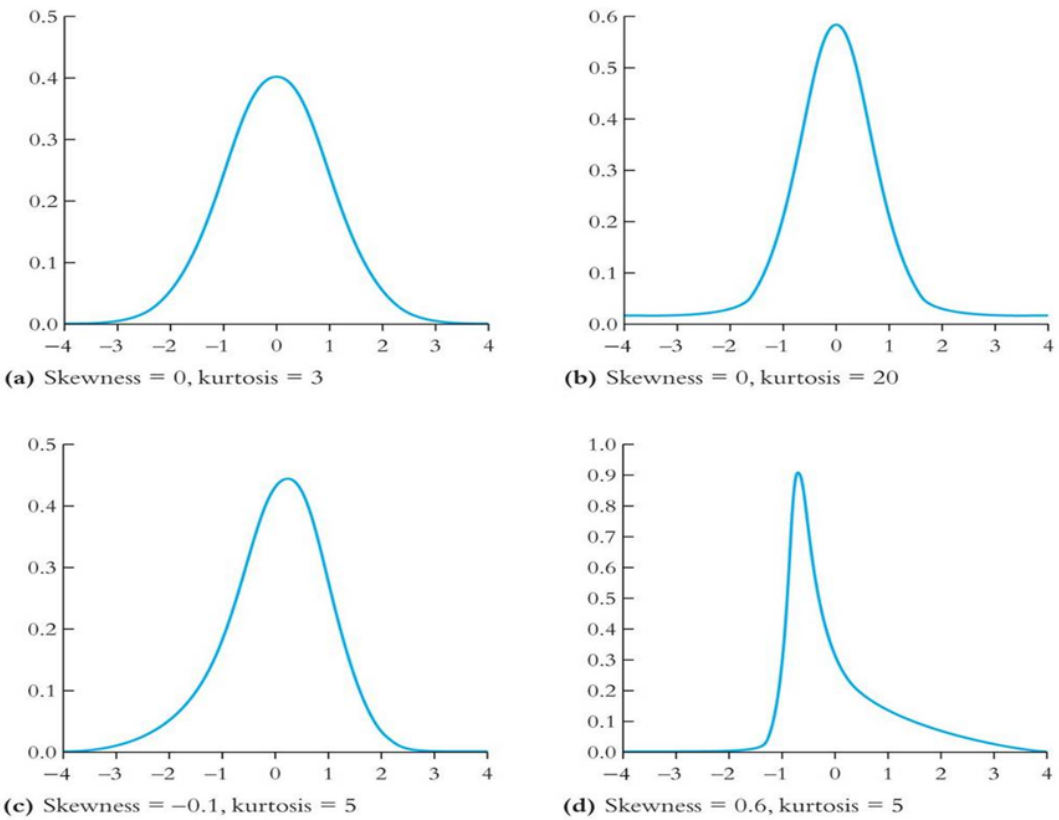
Population:

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4}$$

- measure of mass in tails (of probability of large values)
- Leptokurtic $Kt > 3$ (highly peaked)
  Mesokurtic $Kt = 3$ (medium peaked)
  Platikurtic $Kt < 3$ (flat)

Shapes of distributions

Differing skewness and kurtosis



(a) Skewness = 0, kurtosis = 3

(b) Skewness = 0, kurtosis = 20

(c) Skewness = −0.1, kurtosis = 5

(d) Skewness = 0.6, kurtosis = 5

CAMBRIDGE Judge Business School

Asymmetry

Q: Mean, Median, Mode and Asymmetry

- In a unimodal distribution, if the distribution is exactly symmetric, mean, median and mode are the same
- If the distribution is skewed, the three measures differ
  - What is the ordering if positively skewed?
  - What is the ordering if negatively skewed?

Population

Population: The infinitely large collection of all **possible** entities of interest

- Attributes measured using some (numerical) scale
- Random Variable: Numerical measure of a feature of interest (eg. test score, growth rate, stock price) that can vary upon repeated "experiments"
- ... many (unobserved / unobservable) factors govern the observed outcome..
- (Random) differences in outcomes of the same experiment are governed by *probability*

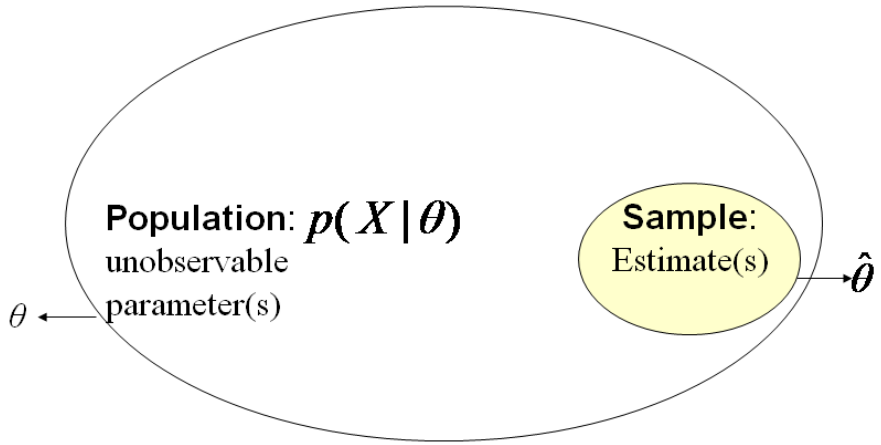
Samples

Samples: Subsets of a population

- Another source of randomness: we observe only a sample, one subset of the population
  - Q: How is the sample selected? How should it be?
- Issue: What can we say about a population *parameter* from a sample *statistic*?
  - Probability theory

Inference problem

Sampling and the inference problem



- Model: Probability distribution of a random variable: $P(X|\theta)$
- Parameter: θ
- Inference: come to conclusion about unknown θ given data: $x_1, \cdots, x_N$
- X being a random variable, inference must be based on the framework of probability theory

Random Variables (1)

Introduction to Random Variables (1)

- Everything is a Random variable!
  - Any observation is the numerical outcome of a (random) experiment.
- Repetition of the same experiment can lead to different outcomes:
  - too many (unobserved / unobservable) factors govern the observed outcome..
- (Random) differences in outcomes of the same experiment governed by probability
- Probability: numerical measure of the *likelihood of occurrence* of an event of interest.
  - ranging from impossible (0) to certain (1)

Random Variables (2)

Introduction to Random Variables (2)

- Data Generating Process: $P(X|\theta)$
- Outcome ($X =?$) unknown prior to the experiment
- You may have a conjecture on the probability distribution over the different potential x values that X may take
- Note the notion of *potential* repetition of experiment:
  - If the experiment were repeated many times the outcomes would be governed by probabilities underlying the experimental process
- What specific probability distribution characterizes the outcome (random) variable?
- If we can "estimate" this distribution, we have a fix on the process generating the phenomenon. How can we do this?
- First, the notion of probability: numerical measure of the *likelihood of occurrence* of an event of interest, ranging from impossible (0) to certain (1).

Gambling consult

First, a gambling consult

- What is more likely:
  - Rolling at least one 6 in four throws of a single die,
  - or rolling at least one double 6 in 24 throws of a pair of dice?
- Question - Chevalier de Mere to Blaise Pascal (1623-1666) that led to modern probability theory
- Concepts:
  - A random experiment is the process of observing the outcome of a chance driven process
  - Elementary outcomes are all possible results of the random experiment
  - The Sample space is the set of all elementary outcomes

Probability notation

Notation

- S : Sample space - Set of all possible elementary outcomes
- A : Event, an elementary outcome, or a set of elementary outcomes
- $P(A)$: Probability of event A occurring

$$P(A) = \frac{|A|}{|S|}; \qquad P(A) = \lim_{n \to \infty} \frac{N_A}{N}$$

- Objective, frequentist, interpretation of the concept of probability
  - Relative frequency of an event in infinite trials
  - Distinct from "subjective" probability

Probability Rules

Rules

- $1 \geq P(A) \geq 0$; $P(S) = 1$
- Can combine events to make other events using logical operations: A and B, A or B, not A
- Probability of event A or B:
  Addition Rule: $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$
- If the events are A and B mutually exclusive:
  $P(A \bigcup B) = P(A) + P(B)$
- For any event, $P(A) = 1 - P(\text{not } A)$

Conditional Probability

Rules

- Conditional Probability of Event A given Event B has occurred:

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

- If A and B are mutually exclusive,

$$P(A|B) = 0$$

- Rearranging the first: $P(A|A) = 1$
- Also note:

$$P(A \bigcap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Independence

Rules

- Event A and Event B are independent if:

$$P(A|B) = P(A)$$

- Probability of Event A and B (Intersection of A and B) if A and B are independent - Multiplication Rule:

$$P(A \bigcap B) = P(A) \cdot P(B)$$

Optional homework

Stock Market Experiment

- When? next week Monday, 8:00 till 8:50pm
- Where? wherever you can connect to the internet
- What? www-experiment
- How?
  - 8:00pm – you'll receive invitation link by email
  - 8:00-8:30pm – create your account & read instructions
  - 8:31pm – the first of 20 trading periods finishes
  - in every subsequent trading period you have 1 minute to make your decision
  - 8:50pm – end of the experiment.