Mathematical Basics

Ashwin Srinivasan

Assumed Knowledge

A basic familiarity will be assumed in the following areas:

1. Sets and Counting
2. Logic
3. Relations and Orderings
4. Functions
5. Linear Algebra
6. Calculus
7. Statistics
8. Probability
9. Mathematical thinking
   - 9.1 Proofs
   - 9.2 Approximations
   - 9.3 Data Visualisation

The slides that follow provide an introduction on what will assumed to be known in these areas

Sets and Counting

Sets I

- Fundamental concept in mathematics
  - A set S contains *elements* ($S = \{a, b, \ldots\}$); elements are *members* of a set ($a \in S$).
  - Two sets S, T are equal ($S = T$) iff they contain precisely the same elements, otherwise $S \neq T$.
  - A set T is a subset of S ($T \subseteq S$) if every member of T is a member of S. If $T \subseteq S$ and $S \subseteq T$ then $S = T$. $T \subseteq S$ means $S \supseteq T$.
  - If $T \subseteq S$ and S contains an element not in T then T is a proper subset of S ($T \subset S$) $T \subset S$ means $S \supset T$.
- Intersection of two sets S, T
  - The set with elements in common to sets S and T, denoted by $S \cap T$ or ST or $S \cdot T$. $ST \subseteq S$ and $ST \subseteq T$ for all S, T.
  - If S and T are disjoint, ST is denoted by the unique set having no members ($\emptyset$). $\emptyset \subseteq S$ for all S and $\emptyset \cdot S = \emptyset$ for all S.
- Union of two sets S, T

Sets II

  - The set with elements which belong to at least S or T, denoted by $S \cup T$ or $S + T$. $S \subseteq S + T$ and $T \subseteq S + T$ for all S, T. $S + \emptyset = S$ for all S. The set consisting of the union of all possible sets is called the universal set U.
  - If S and T are disjoint, then the number of elements in $S \cup T$ is equal to the sum of the number of elements in S and T
- Generalised intersection $\bigcap_{i=1}^{n} S_i$ and and union $\bigcup_{i=1}^{n} S_i$
- Equivalence of two sets S, T
  - If there is a $1 - 1$ correspondence between members of S and members of T (every member of S corresponds to just one member of T and every member of T corresponds to just one member of T) then $S \sim T$.
  - If there is a $T \subset S$ and $S \sim T$ then S is said to be infinite, otherwise S is said to be finite. The set of natural numbers $\mathcal{N}$ is of particular importance. $\mathcal{N}$ is infinite, and any set $S \sim \mathcal{N}$ is said to be countable.
- Representing sets:

Sets III

- Enumeration of elements. $S = \{a, b, c, d\}$
- Set-builder notation. $S = \{n : n \text{ is a natural number } > 5\}$
- Graphically, using Venn diagrams

Algebra of Sets

Law	Identity	Dual Identity
Identity	$A \cup \emptyset = A$ $A \cup U = U$	$A \cap U = A$ $A \cap \emptyset = \emptyset$
Idempotent	$A \cup A = A$	$A \cap A = A$
Commutative	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Associative	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributive	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	
Complement	$A \cup \overline{A} = U$	$\emptyset^c = U$
DeMorgan	$\overline{(A \cup B)} = \overline{A} \cap \overline{B}$	

Sets of Sets

- Power set: set of all subsets of a set. The power set of a set $S = \{1, 2, 3\}$ is the set $\{\ \emptyset, \{1\}, \{2\}, \{3\},\ \{1, 2\}, \{1, 3\}, \{2, 3\},$ $\{1, 2, 3\}\ \}$
  - The power set of a set with n elements will have 2^n elements
- Partition set: set of mutually disjoint subsets $S_1, S_2, \dots, S_k$ of a set S such that $S = S_1 \cup S_2 \cup \cdots \cup S_k$

Cartesian Product

- An ordered tuple $(a_1, a_2, \ldots, a_n)$ is different to a set. In this, the position of the element in the ordering is important.
- Given sets A and B, the Cartesian product $A \times B$ is the set of all elements (a, b) where $a \in A$ and $b \in B$
  - If $A = \{1, 2\}$ and $B = \{a, b, c\}$, then $A \times B = \{(1, a), (1, b), (1, c), (2, a), (2, b), (2, c)\}$
- A subset of $A \times B$ is called a (binary) relation from set A to set B

Extension: Multisets

- QUESTION: How can sets be used to represent elements that are not distinct (e.g. names of your friends)?
  - ANSWER: They cannot
- Extend element notation to a form *n.e* to represent *n* occurrences of an element *e*
  - Thus the set $S = \{3.1, 2.2\}$ represents the "multiset" $\{1, 1, 1, 2, 2\}$

Induction on the set $\mathcal{N}$)

- The set of natural numbers $\mathcal{N} = \{1, 2, 3, \ldots\}$
- The principle of mathematical induction:
  - Let P be some function that returns either *TRUE* or
  - Let $P(1)$ be *TRUE*
  - Let $P(n+1)$ be *TRUE* whenever $P(n)$ is *TRUE*
- P is *TRUE* for every element of $\mathcal{N}$

A Chess Story I

Once upon a time there was a King who loved to play games, and was so good at them that he had gotten bored of the games that he had. He said, "To the man who can invent a new game that is better than all these, I will give whatever he wants."

One day, a man came to his court and said, "I have a game that Your Majesty will never completely master." The game had two armies each lead by a King who commanded the army to defeat the other by capturing the enemy King. It was played on an 8x8 square board. The King loved this game so much that he offered to give the man anything she wished for. "I would like one grain of rice for the first square of the board, two grains for the second, four grains for the third and so on doubled for each of the 64 squares of the game board.", he said. "Is that all?" asked the King, and turned to his Vizier and said, "Arrange for the reward.". Whereupon, the Vizier left for the Royal Granary. Several hours later, he returned with a worried look. "Your Majesty, I do not think we would be able to fulfil the lady's request."

- How much rice was actually needed?
  - The number of grains needed is $2^1 + 2^2 + \cdots + 2^{64}$.
  - This is $18,446,744,073,709,551,615 \approx 18 \times 10^{18}$. Suppose about 100 grains of rice weigh 1g. Then, 1kg of rice ($= 10^3$ grams) will contain about 10^5 grains. So, about 18×10^{13} $= 1.8 \times 10^{14}$ kg of rice was needed.
  - Total rice production in the world is about 800 million tonnes $= 800 \times 10^6 \times 10^3 = 8 \times 10^{11}$ kg

Basics of Counting I

- The number of elements in a set S will be denoted by $|S|$
- Two basic counting principles:

  Sum Rule. If X is the union of disjoint non-empty sets $S_1, S_2, \ldots, S_n$ then $|X| = |S_1| + |S_2| + \cdots + |S_n|$

  - If $E_1, E_2, \ldots, E_n$ are mutually exclusive events and E_1 can happen e_1 ways and E_2 can happen e_2 ways *etc*., then E_1 *or* E_2 *or* $\cdots$ E_n can happen in $e_1 + e_2 + \cdots + e_n$ ways
  - If an object O_1 can be selected in o_1 ways and an object O_2 can be selected independently in o_2 ways *etc.* Then O_1 *or* O_2 *or* $\cdots$ O_n can be selected in $o_1 + o_2 + \cdots + o_n$ ways

  Product Rule. If X is the Cartesian product $S_1 \times S_2 \times \cdots, S_n$ of non-empty sets, then the $|X| = |S_1| \cdot |S_2| \cdot |S_n|$

Basics of Counting II

- If $E_1, E_2, \ldots, E_n$ are mutually events and E_1 can happen e_1 ways and E_2 can happen e_2 ways *etc*., then the sequence of events E_1 *and* E_2 *and* $\cdots$ E_n happens in $e_1 \times e_2 \times \cdots \times e_n$ ways
- If an object O_1 can be selected in o_1 ways and an object O_2 can be selected in o_2 ways *etc.* Then O_1 *and* O_2 *and* $\cdots$ O_n can be selected in $o_1 \times o_2 \times \cdots \times o_n$ ways

- Exanples:
  - In how many ways can we get a card with a heart or a spade when drawing from a pack of cards?
  - If two distinguishable dice are thrown, how many different ways can they fall?
  - If 100 distinguishable dice are thrown, hown many different ways can they fall?

Basics of Counting III

- How many 3 digit numbers can be formed with the numerals $1, 2, 3, 4, 5, 6$? How many such numbers can be formed if repetitions are not allowed?

Counting by Correspondence

Question. How many subsets S_n are there of a set with n elements?

Answer (by induction). $S_n = 2^n$. We could prove this using mathematical induction: (a) It holds trivially for $n = 0$; (b) Suppose the number of subsets of a set with k elements is 2^k; (c) We want to show that the number of subsets of a set S with $k+1$ elements is 2^{k+1}. Let $S = S_0 \cup \{a\}$ where a denotes some arbitrary element and S_0 is a k-element set. Then all subsets of S_0 are subsets of S. In addition, for each subset X of S_0, $\{a\} \cup X$ will be a subset of S. That is, there are $S_{k+1} = 2^k + 2^k = 2^{k+1}$.

Answer (by correspondence). We establish a 1-1 correspondence between the subsets of a set $S = \{x_1, x_2, \ldots, x_n\}$ and n-digit binary sequences $(y_1, y_2, \ldots, y_n)$ as follows. Let a subset X correspond to a binary sequence B_X such that if $x_i \in X$ then $y_i = 1$ in B_X. Clearly this is

Combinations and Permutations I

- A *combination* of n objects taken r at a time is an unordered selection of r objects
- A *permutation* of n objects taken r at a time is an ordered selection of r objects
- Example:
  - Suppose the n objects are a, a, a, b, c. Combinations of 3 objects chosen from these are: aaa, aab, aac, and abc
  - The 3 permutations are aaa, aab, aba, baa, aac, aca, caa, abc, acb, bac, bca, cab, and cba
- There are no restrictions on selection: repetitions are allowed
- We could force selection to be such that an object x_k can be selected up to n_k times
- We will use a list notation to denote restrictions on selection
  - $[3 \cdot a, 3 \cdot b, 1 \cdot c]$ denotes that a and b can appear upto 3 times, but c can appear at most once in a combination

Combinations and Permutations II

  - Given $[n_1 \cdot x_1, n_2 \cdot x_2, \cdots, n_k \cdot x_k]$, $n_i = \infty$ denotes that there is no restriction on the number of times x_i can occur. If all the n_i's are ∞ then we are considering combinations or permutations with unlimited repetitions

- Even more restrictions could be placed using constraints written as statements in first-order logic

$$Selection([n_1 \cdot a, n_2 \cdot b, 1 \cdot c]) \leftarrow Prime(n_1) \wedge Even(n_2)$$

- We will largely be concerned with combinations and permutations with either <u>no</u> repetitions, or <u>unlimited</u> repetitions

Enumeration Formulæ(no repetitions) I

- The number of permutations of r objects from n without repetition is given by

$$P(n,r) = \frac{n!}{(n-r)!}$$

  Note that if $r = n$, then $P(n,n) = n!$. When we refer to "permutations of n objects", we mean $P(n,n)$

- The number of combinations r objects from n without repetition is given by

$$C(n,r) = \frac{P(n,r)}{r!} = \frac{n!}{r!(n-r)!}$$

- Examples:

Enumeration Formulæ(no repetitions) II

- There are $P(10,3)$ 3-digit numbers with leading 0's that contain no repeated digits. If no leading 0's are allowed, there are $9 \times P(10,2)$ such numbers
- Suppose there are 7 apples and 3 oranges, that have be placed in a row such that the oranges are always together
    - The 3 oranges can be arranged together in $P(3,3) = 3!$ ways. For each of these, there $P(8,8) = 8!$ ways in which the apples and orange-group could be arranged. By the product rule the total number of ways in which the apples and oranges can be arranged together is $3! \times 8!$
- The number of ways in which a jury of 12 people can be selected from a group of 20 is $C(20,12)$

Enumeration Formulæ(with repetitions) I

- The number of permutations of r objects from n with repetition is given by

$$P_{\infty}(n, r) = n^r$$

- Examples:
    - The number of different ways to answer an exam with 10 true or false questions is 3^{10}
    - A cricket match can end in one of 3 ways for the home side: *win*, *lose*, *draw* (we will treat a tie a a draw). How many ways can you predict the outcomes of a 5-match series so that exactly 3 predictions are correct? 3 of the 5 matches can be correctly predicted in C(5,3) ways. The remaining 2 predictions are wrong. Here is what can happen:

Predicted	Actual		
	HomeWin	HomeLose	Draw
HomeWin	√	×	×
HomeLose	×	√	×
Draw	×	×	√

Enumeration Formulæ(with repetitions) II

Each remaining prediction can be wrong in 2 ways. The total number of ways to predict outcomes is thus $C(5,3)2^2$

- The number of ways of selecting r objects from n distinct objects with unlimited repetitions
  - Let the objects $\{a_1, a_2, \ldots, a_n\}$. Let there be x_1 repetitions of a_1, x_2 repetitions of a_2, and so on. Clearly, for any values of $x_1, \ldots, x_n$, $x_1 + x_2 + \cdots + x_n = r$
  - The total number of ways of selecting r objects is equal to the number of solutions to the equation $x_1 + x_2 + \cdots + x_n = r$
  - This is the same as the number of ways of distributing r indistinguishable balls into n distinct boxes
  - How is this to be done? Place x_1 balls into box 1. Represent box 1 by a binary number with x_1 0's followed by 1. Thus, 3 repetitions of a_1 will be represented by 0001. When all boxes are represented in this manner, there will be r 0's and $n-1$ 1's
  - Thus, each distribution of r balls into n boxes is represented by a binary number with $n-1$ 1's and r 0's

Enumeration Formulæ(with repetitions) III

- The total number of ways of distributing r balls into n boxes is the number of binary numbers with $n-1$ 1's and r 0's
- This is equal to $C(n-1+r, r)$

Logic

Propositions

- Formal logic is concerned with statements of fact, as opposed to opinions, commands, questions, exclamations *etc.* Statements of fact are assertions that are either true or false, the simplest form of which are called *propositions*
- Propositions can be assigned a truth value of either *true* (t, or 1) to propositions.or *false* (f, or 0) but not both:

  The earth is flat.
  Humans are monkeys.
  $1 + 1 = 2$

  These are not propositions:

  Who goes there?
  Eat your broccoli.

Connectives

- Propositions are usually represented by letters like $P, Q, \ldots$, or $p, q, \ldots$. *Compound statements* can be combining two or more propositions with *logical connectives* (or simply, connectives). The connectives we will look at here will allow us to form sentences like the following:

  It is not the case that P
  P and Q
  P or Q
  P if Q

- The connectives are usually represented by symbols like $\wedge$ (and), $\vee$ (or), and $\neg$ (not).

Syntax I

- Every language needs a *vocabulary*. For the language of propositional logic, we will restrict the vocabulary to the following:

Propositional symbols:	$P, Q, \ldots$
Logical connectives:	$\neg, \wedge, \vee, \leftarrow$
Brackets:	$(,)$

- The next step is to specify the rules that decide how legal sentences are to be formed within the language. For propositional logic, legal sentences or *well-formed formulæ* (wffs for short) are formed using the following rules:
    1. Any propositional symbol is a wff;
    2. If α is a wff then $\neg\alpha$ is a wff; and
    3. If α and β are wffs then $(\alpha \wedge \beta), (\alpha \vee \beta)$, and $(\alpha \leftarrow \beta)$ are wffs.

Syntax II

Wffs consisting simply of propositional symbols (Rule 1) are sometimes called *atomic* wffs and others *compound* wffs. Informally, it is acceptable to drop outermost brackets. Here are some examples of wffs and 'non-wffs':

Formula	Comment
$(\neg P)$	Not a wff. Parentheses are only allowed with the connectives in Rule 3
$\neg\neg P$	P is wff (Rule 1), $\neg P$ is wff (Rule 2), $\therefore$ $\neg\neg P$ is wff (Rule 2)
$(P \leftarrow (Q \wedge R))$	P, Q, R are wffs (Rule 1), $\therefore$ $(Q \wedge R)$ is a wff (Rule 3), $\therefore$ $(P \leftarrow (Q \wedge R))$ is a wff (Rule 3)
$P \leftarrow (Q \wedge R)$	Not a wff, but acceptable informally
$((P) \wedge (Q))$	Not a wff. Parentheses are only allowed with the connectives in Rule 3
$(P \wedge Q \wedge R)$	Not a wff. Rule 3 only allows two symbols within a pair of brackets

Normal Forms

- Every formulae in propositional logic is equivalent to a formula that can be written as a conjunction of disjunctions. That is, something like $(A \vee B) \wedge (C \vee D) \wedge \cdots$. When written in this way the formula is said to be in *conjunctive normal form* or CNF
- There is another form, which consists of a disjunction of conjunctions, like $(A \wedge B) \vee (C \wedge D) \vee \cdots$, called the *disjunctive normal form* or DNF
- In general, a formula in CNF can be written as:

$$F = \bigwedge_{i=1}^{n} \left(\bigvee_{j=1}^{m} L_{i,j} \right)$$

and a formula in DNF as:

$$G = \bigvee^{n} \left(\bigwedge^{m} L_{i,j} \right)$$

Semantics I

- An *interpretation* is simply an assignment of either *true* or *false* to all propositional symbols
- For example, given the wff $(P \leftarrow (Q \wedge R))$ here are two different interpretations:

	P	Q	R
I_1 :	true	false	true
I_2 :	false	true	true

- For a formula with N propositional symbols, there can never be more than 2^N possible interpretations. For example, the 3 propositional symbols P, Q and R here are are all $8 (= 2^3)$ interpretations:

Semantics II

	P	Q	R
I_1 :	false	false	false
I_2 :	false	false	true
I_3 :	false	true	false
I_4 :	false	true	true
I_5 :	true	false	false
I_6 :	true	false	true
I_7 :	true	true	false
I_8 :	true	true	true

- Truth or falsity of a wff only makes sense given an interpretation (by the principle of bivalence, any interpretation can only result in a wff being either *true* or *false*)

Semantics III

- Thus, the wff P is *false* in interpretation I_1 and *true* in interpretation I_5. To obtain the truth-value of compound wffs like $(P \leftarrow (Q \wedge R))$ is obtained using the semantics of the connectives.
- The semantics of the connectives are summarised in truth-tables:

P	Q	$P \wedge Q$	$P \vee Q$	$\neg P$	$\neg Q$
f	f	f	f	t	t
f	t	f	t	t	f
t	f	f	t	f	t
t	t	t	t	f	f

- One more truth-table is of interest. This concerns the connective $\leftarrow$ (if). The statement $P \leftarrow Q$ is to be read as "if Q then P".

Semantics IV

P	Q	$P \leftarrow Q$
f	f	t
f	t	f
t	f	t
t	t	t

If you have not seen this before, it may be surprising, e.g.

flatworld	*humanmonkeys*	*flatworld* $\leftarrow$ *humanmonkeys*
f	f	t

- Note: $P \leftarrow Q \equiv P \vee \neg Q \equiv \neg Q \vee P$
- Every interpretation (that is, an assignment of truth-values to propositional symbols) that makes a well-formed formula *true* is said to be a *model* for that formula

	P	Q	R
I_1 :	true	false	true
I_2 :	false	true	true

Semantics V

- Interpretation I_1 is a model for $(P \leftarrow (Q \wedge R))$; and that I_2 is not a model for the same formula. In fact, I_1 is also a model for several other wffs like: P, $(P \wedge R)$, $(Q \vee R)$, $(P \leftarrow Q)$, *etc*. Similarly, I_2 is a model for Q, $(Q \wedge R)$, $(P \vee Q)$, $(Q \leftarrow P)$, *etc*
- A wff may be such that *every* interpretation is a model. An example is $(P \vee \neg P)$. Since there is only one propostional symbol involved (P), there are at most $2^1 = 2$ interpretations possible. The truth table summarising the truth-values for this formula is:

	P	$\neg P$	$(P \vee \neg P)$
I_1 :	false	true	true
I_2 :	true	false	true

  $(P \vee \neg P)$ is thus *true* in every possible 'context'. Formulæ like these, for which every interpretation is a model are called *valid* or *tautologies*

Semantics VI

- A wff may be such that *none* of the interpretations is a model. An example is $(P \wedge \neg P)$. Again there is only one propostional symbol involved (P), and thus only two interpretations possible. The truth table summarising the truth-values for this formula is:

	P	$\neg P$	$(P \wedge \neg P)$
I_1 :	false	true	false
I_2 :	true	false	false

  $(P \wedge \neg P)$ is thus *false* in every possible 'context'. Formulæ like these, for which none of the interpretations is a model are called *unsatisfiable* or *inconsistent*.

- Consider the following:

Statement	Formally
The rabbit either went down Path A or Path B.	$P \vee Q$
It did not go down Path A.	$\neg P$
Therefore it went down Path B.	$\therefore Q$

Semantics VII

- Here, we want to establish that if the first two statements are true, then the third follows. In general, what we are trying to establish is that some well-formed formula α is the *logical consequence* of a conjunction of other well-formed formulæ Σ (or, that Σ *logically implies* α). This relationship is usually written thus:

  $$\Sigma \models \alpha$$

  Σ being the conjunction of several wffs, and logical consequence can therefore also be written as the following relationship between a pair of wffs:

  $$((\beta_1 \wedge \beta_2) \ldots \beta_n) \models \alpha$$

Semantics VIII

- What we want is the following: whenever the statements in Σ are true, α must also be true. In formal terms, this means: $\Sigma \models \alpha$ *if every model of* Σ *is also model of* α
  - Recall that a model for a formula is an interpretation (assignment of truth-values to propositions) that makes that formula *true*;
  - Therefore, a model for Σ is an interpretation that makes $((\beta_1 \wedge \beta_2) \ldots \beta_n)$ *true*. Clearly, such an interpretation will make each of $\beta_1, \beta_2, \ldots, \beta_n$ *true*;
  - Let $I_1, I_2, \ldots, I_k$ be all the interpretations that satisfy the requirement above: that is, each is a model for Σ and there are no other models for Σ (recall that if there are N propositional symbols in Σ and α together, then there can be no more than 2^N such interpretations);
  - Then to establish $\Sigma \models \alpha$, we have to check that each of $I_1, I_2, \ldots, I_k$ is also a model for α (that is, each of them make α *true*).

Semantics IX

- Look again at:

Statement	Formally
The rabbit either went down Path A or Path B.	$P \vee Q$
It did not go down Path A.	$\neg P$
Therefore it went down Path B.	$\therefore Q$

	P	Q	$(P \vee Q)$	$\neg P$	$((P \vee Q) \wedge \neg P)$
I_1 :	false	false	false	true	false
I_2 :	false	true	true	true	true
I_3 :	true	false	true	false	false
I_4 :	true	true	true	false	false

- Closely related to logical consequence is the notion of *logical equivalence*. A pair of wffs α and β are logically equivalent means:

$$\alpha \models \beta \quad \textit{and} \quad \beta \models \alpha$$

Semantics X

- This means the truth values for α and β are the same in all cases, and is usually written more concisely as:

$$\alpha \equiv \beta$$

More on the Conditional I

- Recall the truth table for the conditional:

α	β	$(\alpha \leftarrow \beta)$
false	false	true
false	true	false
true	false	true
true	true	true

- There is, therefore, only one interpretation that makes $(\alpha \leftarrow \beta)$ *false*. This may come as a surprise. Consider for example the statement:

$$\text{The earth is flat} \leftarrow \text{Humans are monkeys}$$

More on the Conditional II

An interpretation that assigns *false* to both 'The earth is flat' and 'Humans are monkeys' makes this statement *true* (line 1 of the truth table). In fact, the only world in which the statement is false is one in which the earth is not flat, and humans are monkeys

- The unusual nature of the conditional is due to the fact that it allows premises and conclusions to be completely unrelated. This is not what we would expect from conditional statements in normal day-to-day discourse
- Consider now the truth table for $(\alpha \vee \neg\beta)$:

α	β	$\neg\beta$	$(\alpha \vee \neg\beta)$
false	false	true	true
false	true	false	false
true	false	true	true
true	true	false	true

More on the Conditional III

It is evident from these truth tables that every model for $(\alpha \leftarrow \beta)$ is a model for $(\alpha \vee \neg\beta)$ and vice-versa. Thus:

$$(\alpha \leftarrow \beta) \equiv (\alpha \vee \neg\beta)$$

- Note the following related statements:

Conditional	$(\alpha \leftarrow \beta)$
Contrapositive	$(\neg\beta \leftarrow \neg\alpha)$

It should be easy to verify the following equivalence:

$$\text{Conditional} \equiv \text{Contrapositive}$$
$$(\alpha \leftarrow \beta) \equiv (\neg\beta \leftarrow \neg\alpha)$$

- Errors of reasoning arise by assuming other equivalences. Consider for example the pair of statements:

S_1 : Fred is an ape ← Fred is human
S_2 : Fred is human ← Fred is an ape

More on the Conditional IV

Are these two statements equivalent?

Clauses I

- The conditional:

  $$(\text{Fred is human} \leftarrow (\text{Fred walks upright} \wedge \text{Fred has a large brain}))$$

  is equivalent to:

  $$(\text{Fred is human} \vee \neg\, (\text{Fred walks upright} \wedge \text{Fred has a large brain}))$$

- Using De Morgan's Law and dropping some brackets for clarity:

  $$\text{Fred is human} \vee \neg\, \text{Fred walks upright} \vee \neg\, \text{Fred has a large brain}$$

Clauses II

- Each of the premises on the right-hand side of the the original conditional (Fred walks upright, Fred has a large brain) appear negated in the final disjunction; and the conclusion (Fred is human) is unchanged
- We will use the term *clauses* to denote formulæ that contain propositions or negated propositions joined together by disjunction ($\vee$).
- We will also use the term *literals* to denote propositions or negated propositions. Clauses are thus disjunctions of literals.

Proof Theory

- Proof theory considers the “derivability” of a sentence given a set of inference rules $\mathcal{R}$
  - The sentences given initially are called the *axioms*, and those derived are *theorems* (syntactic consequences)
- A sentence is derivable from a set of axioms S using $\mathcal{R}$: $S \vdash_{\mathcal{R}} s$
  - Axioms can be *logical* (valid sentences of logic) or *non-logical* (problem specific sentences in logic)
- Axioms $+ \mathcal{R} =$ Inference system
- Axioms $+$ all theorems $=$ Theory
  - A theory is consistent iff there is no sentence s s.t. the theory contains both s and $\neg s$

Soundness and completeness I

- We would like theorems derived to be logical consequences of the axioms provided
  - We can then be sure of the correctness of the theorem in the intended model for the axioms
  - Remember, logical consequences of the axioms are true in all models for the axioms
- This property depends entirely on the inference rules chosen, and those that have this property are called *sound*
  - if $S \vdash_{\mathcal{R}} s$ then $S \models s$
  - Examples of sound inference rules:

$$\begin{array}{ll} \textit{modus ponens} & \{q, p \leftarrow q\} \vdash p \\ \textit{modus tollens} & \{\neg p, p \leftarrow q\} \vdash \neg q \end{array}$$

- We would also like to derive *all* logical consequences, and rules with this property are said to be *complete*
  - if $S \models s$ then $S \vdash_{\mathcal{R}} s$
- That is $S \models s \equiv S \vdash_{\mathcal{R}} s$

Proof procedures

- Axioms and inference rules are not enough. We need a strategy to apply the rules.
  - Inference system + strategy = Proof procedure
- For logic programs:
  - 1 inference rule: *resolution*
  - Strategy: **S**elected **L**inear **D**efinite (SLD)
  - Proof procedure: SLD-resolution

Resolution

- Consider the clauses:

  C_1: $is\_dangerous \ \leftarrow \ is\_cheetah$

  C_2: $is\_cheetah \ \leftarrow \ is\_carnivore, has\_tawny\_colour, has\_dark\_spots$

  - The *resolvent* of C_1, C_2 is the clause:

  C: $is\_dangerous \ \leftarrow \ is\_carnivore, has\_tawny\_colour, has\_dark\_spots$

  - Remember

    C_1: $is\_dangerous \lor \neg is\_cheetah$

    C_2: $is\_cheetah \lor \neg is\_carnivore \lor \neg has\_tawny\_colour \lor \neg has\_dark\_spots$

    C: $is\_dangerous \lor \neg is\_carnivore \lor \neg has\_tawny\_colour \lor \neg has\_dark\_spots$

  - C_1, C_2 are called the *parent* clauses, and *is_cheetah* is the the literal that is resolved upon

Soundness of resolution

- A single resolution step does the following:
  - From $p \leftarrow q$ and $q \leftarrow r$
  - Infer $p \leftarrow r$
- Since resolution is sound, we can always add the clauses inferred to the original program

Completeness of resolution

- Resolution has these properties
  - Consider a set of clauses s.t. each clause has *at most* 1 positive literal. Such clauses are called *Horn* clauses
  - If a set of Horn clauses is unsatisfiable then resolution will derive the empty clause. Resolution is thus "refutation complete"
  - However, it is not "affirmation complete". That is, if $P \models s$, then it need not follow that $P \vdash s$ using resolution

    $$\{p \leftarrow, q \leftarrow\} \models p \leftarrow q$$
  - But, if $P \cup \{\neg s\} \ \vdash \Box$ using resolution then $P \cup \{\neg s\} \models \Box$ or $P \models s$

The Deduction theorem

- Let $P = \{s_1, \ldots s_n\}$ be a set of clauses and s be a sentence

  Theorem. $P \models s$ iff $P - \{s_i\} \models (s \leftarrow s_i)$

  - Implication is preserved if we remove any sentence from the left and make it a condition on the right

    $P - s_1, \ldots, s_i \models (s \leftarrow s1 \wedge \ldots \wedge s_i)$

    $\emptyset \models (s \leftarrow s1 \wedge \ldots \wedge s_n)$
  - That is, every model of $\emptyset$ is a model of $s \leftarrow s1 \wedge \ldots \wedge s_n$
  - $s \leftarrow s1 \wedge \ldots \wedge s_n$ is valid
- Now consider $P \models q$

  $p \leftarrow q \;\equiv\; \sim q \leftarrow \sim p$ and

  $q \leftarrow \;\equiv\; q \leftarrow TRUE \;\equiv\; FALSE \leftarrow \sim q$

  $P \models q \;\equiv\; P \models (q \leftarrow)$ iff:

  $P \models (FALSE \leftarrow \sim q)$ iff:

  $P \cup \{\sim q\} \models FALSE$
- That is $P \models q$ iff $P \cup \{\sim q\}$ is unsatisfiable

  Logical consequence can be checked by Refutation

Introduction to First-Order Logic I

- Suppose you wanted to express logically the statement: ‘All humans are apes.’ One of two ways can be used to formalise this in propositional logic. We can use a single proposition that stands for the entire statement, or with a well-formed formula consisting of a lot of conjunctions:
  Human1 is an ape $\wedge$ Human2 is an ape $\ldots$
- Using a single proposition does not give any indication of the structure inherent in the statement (that, for example, it is a statement about two sets of objects—humans and apes—one of which is entirely contained in the other). The conjunctive expression is clearly tedious in a world with a lot of humans. Things can get worse. Consider the following argument:

  Some animals are humans.
  All humans are apes.
  Therefore some animals are apes.

Introduction to First-Order Logic II

It is evident that this is valid: yet it is beyond the power of propositional logic to establish it. If, for example, we elected to represent each of the statements with single propositions then all we would end up with is:

Statement	Formally
Some animals are humans.	P
All humans are apes.	Q
Therefore some animals are apes.	$\therefore R$

This is clearly not what we want. hat is needed is in fact something along the following lines:

Statement	Formally
Some animals are humans.	Some P are Q
All humans are apes.	All Q are R
Therefore some animals are apes.	$\therefore$ some P are R

Here, P, Q, and R do not stand for propositions, but for terms like *animals*, *humans* and *apes*. The use of terms like these related to each other by the expressions 'some' and 'all' will allow us to form sentences like the following:

Introduction to First-Order Logic III

All P are Q
No P are Q
Some P are Q
Some P are not Q

The expressions 'some' and 'all' are c:.alled *quantifiers*, which when combined with the logical connectives introduced in connection with proposition logic $(\neg, \wedge, \vee, \leftarrow)$, results in the powerful framework of first-order or *predicate logic*

First-order logic: Syntax I

- The language of predicate logic introduces many new constructs that are not found in the simpler, propositional case:

  Constants. It is conventional in predicate logic to use lowercase letters to denote proper names of objects. For example, in the sentence 'Fred is human', Fred could be represented as *fred*.

  Variables. Consider the statements:

  All humans are apes
  Some apes are not human

  Using the letter x as a variable that can stand for individual objects, these can be expressed as:

  For all x, if x is human then x is an ape
  For some x, x is an ape and x is not human

First-order logic: Syntax II

Quantifiers. The language of predicate logic introduces the symbol $\forall$, called the *universal quantifier*, to denote 'for all.' The symbol $\exists$, called the *existential quantifier*, is used to denote 'for some' or, more precisely, 'for at least one.' The sentences above can therefore be written as:

$$\forall x \text{ (if } x \text{ is human then } x \text{ is an ape)}$$
$$\exists x \text{ (} x \text{ is an ape and } x \text{ is not human)}$$

Predicates. In their simplest case, these are are symbols used to attribute properties to particular objects. It is conventional in logic (but ungrammatical in English) to write the subject after the predicate. Thus the sentence 'Fred is human' would be formalised as $Human(fred)$ More generally, predicate symbols can be used to represent relations between two or more objects. Thus, 'Fred likes bananas' can be represented as: $Likes(fred, bananas)$. The general form is therefore a predicate symbol, followed by one or more *arguments* separated by commas and enclosed by brackets. The number of arguments

First-order logic: Syntax III

is sometimes called the *arity* of the predicate symbol, and the predicate symbol is often written along with its arity (for example, *Likes*/2). Formalising sentences like those above would result in quantified variables being arguments:

$$\forall x \text{ (if } Human(x) \text{ then } Ape(x))$$
$$\exists x\ (Ape(x) \text{ and not } Human(x))$$

Or, using the logical connectives that we have already come across:

$$\forall x(Ape(x) \leftarrow Human(x))$$
$$\exists x(Ape(x) \wedge \neg Human(x))$$

First-order logic: Syntax IV

Functions. Consider the statement: 'The father of Fred is human.' Although we have not named Fred's father, it is evident that a a unique individual is being referred to, and it possible to denote him by using a *function* symbol. One way to formalise the statement is: $Human(father(fred))$. Here, it is understood that $father(fred)$ denotes Fred's father. A function symbol is one which, when attached to one or more terms denoting objects produces an expression that denotes a single object. It is important that that the result is unique: a function symbol could not be used to represent, for example, 'parent of Fred.' As with predicates, the number of arguments of the function is sometimes called its arity. Function symbols are normally considered an extension to the basic vocabulary of predicate logic.

First-Order Logic: Semantics

Too complicated to get into here

Relations and Orderings

Relations and Operations I

- Finite sequence: a set of n elements placed in a $1-1$ correspondence to the set $\{1, \ldots, n\}$ arranged in order of succession. An *ordered pair* is a sequence of 2 elements.
- Dyadic or binary relation R over a set S: a set of ordered pairs (x, y) where $x, y \in S$. If $(a, b) \in R$ then aRb means "a is in the relation R to b" or "relation R holds between a and b."
  - An important relation on S is the equality relation that consists of pairs (x, x)
  - Another important relation on the set of numbers consist of the the pairs (x, y) where x is less than y
- In general, a binary relation from set S to T is a subset of $S \times T$
  - The domain of a relation R is the set of elements $x \in S$ such that there is an element $y \in T$ with $(x, y) \in R$. The image of R is the set of all elements $y \in T$ such that there is an $x \in S$ with $(x, y) \in R$

Relations and Operations II

  - The inverse relation R^{-1} is a relation from T to S and consists of ordered pairs (y, x) where $x \in S$ and $y \in T$ and $(x, y) \in R$
- Finitary operation in a set S: let $s_n = (x_1, \ldots, x_n)$ be sequences of n elements. To each such sequence, associate just one element $y \in S$. The set P of ordered pairs (s_n, y) is a finitary operation in S. $s_n P y$ denotes a n-ary operation and is denoted by $P(x_1, \ldots, x_n) = y$. If $n = 1$, then P is a dyadic relation over S.
  - Let $S = \mathcal{N}$. Then addition $(+)$, subtraction $(-)$ etc. are examples of binary operations in S.
- If a n-ary operation P is defined for every sequence s_n of n elements of A, then S is *closed* wrt P. A set S closed wrt one or more finitary operations is called an *algebra*. A *subalgebra* is a subset of an algebra S which is self-contained wrt to the operations.
  - $\mathcal{N}$ is closed wrt the binary operations of $+$ and $\times$, and $\mathcal{N}$ along with $+, \times$ form an algebra.

Relations and Operations III

  - The set $\mathcal{E}$ of even numbers is a subalgebra of algebra of $\mathcal{N}$ with $+, \times$. The set $\mathcal{O}$ of odd numbers is not a subalgebra.
- Composition of relations:
  - Let R_1 be a relation from S to T and R_2 be a relation from T to V. Then the composite relation $R = R_2 \circ R_1$ consists of pairs (x, z) where $(x, y) \in R_1$ and $(y, z) \in R_2$ for some $y \in T$.

Representing relations

- Binary relations on sets of numbers can be represented as a set of points in the Cartesian plane (more generally, they can be viewed as functions in some multi-dimensional space)
- Binary relations on finite sets of elements can be represented as entries in a matrix
- Binary relations can be represented using arrows between elements in diagrams of the two sets
- A relation on a set S can be represented by a "directed graph"

Types of Relations

Reflexive. Let $A = \{1, 2, 3\}$ and $R = \{(1, 1), (1, 2), (2, 2), (2, 3), (3, 3)\}$. R is reflexive, because $(a, a) \in R$ for all $a \in A$

Symmetric. The relation $\perp$ on the set of lines in a plane such that $(l_1, l_2) \in\perp$ is symmetric because if $(a, b) \in\perp$ then $(b, a) \in\perp$

Antisymmetric. The relation $\leq$ on the set of numbers N is antisymmetric because if $a \leq b$ and $b \leq a$ then $a = b$

Asymmetric. The relation $<$ on the set of numbers N is aysmmetric because if $a < b$ then $b \not< a$

Transitive. The relation $\subseteq$ on the powerset $P(A)$ of set $A = \{1, 2, 3\}$ is transitive since if $a \subseteq b$ and $b \subseteq c$ then $a \subseteq c$ for $a, b, c \in P(A)$

Equivalence Relations I

- An equivalence relation E over a set S is a dyadic relation over S that satisfies the following properties:

  Reflexive. For every $a \in S$, aEa

  Symmetric. If aEb then bEa

  Transitive. If aEb and bEc then aEc

  - Let $S = \mathcal{N}$ and aEb iff $a + b$ is even. That is, E consists of all ordered pairs (a, b) whose sum is even. This makes all even numbers equivalent, and the odd numbers equivalent
  - Let $S = \{a, b, c, d\}$ and xEy if $x, y \in \{a, b\}$ or $x, y \in \{c, d\}$. This makes a, b equivalent to each other and c, d equivalent to each other

  Theorem. Any equivalence relation E over a non-empty set S results in a partition of S into disjoint non-empty subsets which contain all the members of S.

Equivalence Relations II

- The subsets are called "equivalence classes" or "blocks of the partition". Some special partitions: every block contains exactly 1 element (zero partition); at most 1 block contains more than 1 element (singular partition); and 1 block contains all the elements (unity partition).

Theorem. Any partition of a set S into disjoint subsets such that every member of S is in some subset results in an equivalence relation E over S.

Partial Order I

- Gven an equality relation $=$ over elements of a set S, a partial order $\preceq$ over S is a dyadic relation over S that satisfies the following properties:

  Reflexive. For every $a \in S$, $a \preceq a$

  Anti-Symmetric. If $a \preceq b$ and $b \preceq a$ then $a = b$

  Transitive. If $a \preceq b$ and $b \preceq c$ then $a \preceq c$

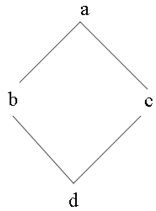
  - If $a \preceq b$ and $a \neq b$ then $a \prec b$
  - $b \succeq a$ means $a \preceq b$, $b \succ a$ means $a \prec b$
  - If $a \preceq b$ or $b \preceq a$ then a, b are comparable, otherwise they are not comparable
  - A set S over which a relation of partial order is defined is called a *partially ordered set*
  - It is sometimes convenient to refer to a set S and a relation R defined over S together by the pair $< S, R >$
  - Examples of partially ordered sets $< S, \preceq >$:
    - S is a set of sets, $S_1 \preceq S_2$ means $S_1 \subseteq S_2$

Partial Order II

    - $S = \mathcal{N}$, $n_1 \preceq n_2$ means $n_1 = n_2$ or there is a $n_3 \in \mathcal{N}$ such that $n_1 + n_3 = n_2$
    - S is the set of equivalence relations $E_1, \ldots$ over some set T, $E_L \preceq E_M$ means for $u, v \in T$, uE_Lv means uE_Mv (that is, $(u, v) \in E_L$ means $(u, v) \in E_M$).
  - Given a set $S = \{a, b, \ldots\}$ if $a \prec b$ and there is no $x \in S$ such that $a \prec x \prec b$ then b *covers* a or a is a *downward cover* of b
  - Given a set S let S_{down} be a set of downward covers of $b \in S$. If for all $x \in S$, $x \prec b$ implies there is an $a \in S_{down}$ s.t. $x \preceq a \prec b$, then S_{down} is said to be a *complete* set of downward covers of b.

- Diagrammatic representation of a partially ordered set

Partial Order III



Lattice I

- A lattice is a partially ordered set $< S, \preceq >$ in which every pair $a, b \in S$ has a greatest lower bound ($a \sqcap b$ or ab or *meet*) in S and a least upper bound ($a \sqcup b$ or $a + b$ or *join*) in S

  Theorem. A lattice is an algebra with the binary operations of $\sqcap$ and $\sqcup$

- Properties of $\sqcap$ and $\sqcup$
  - $a \sqcap b = b \sqcap a$, and $a \sqcup b = b \sqcup a$
  - $a \sqcap (b \sqcap c) = (a \sqcap b) \sqcap c$, and $a \sqcup (b \sqcup c) = (a \sqcup b) \sqcup c$
  - If $a \preceq b$ then $a \sqcap b = a$, and $a \sqcup b = b$
  - $a \sqcap (a \sqcup b) = a$, and $a \sqcup (a \sqcap b) = a$
- Example
  - Let S be all the subsets of $\{a, b, c\}$, and for $X, Y \in S$, $X \preceq Y$ mean $X \subseteq Y$, $X \sqcap Y = X \cap Y$ and $X \sqcup Y = X \cup Y$. Then $< S, \subseteq >$ is a lattice.

Bounded Lattice I

- A lattice L has a lower bound $\bot$ if for every element $x \in L$ $\bot \preceq x$. Similarly, L has an upper bound $\top$ if for every $x \in L$, $x \preceq \top$. If L has both an upper and lower bound then the lattice is bounded
  - The lattice of subsets of the set $S = \{a, b, c, d\}$ is a bounded lattice with $\bot = \emptyset$ and $\top = S$
- Bounded lattices have the following properties:

$$a \sqcup \bot = a \text{ and } a \sqcap \bot = \bot$$
$$a \sqcup \top = \top \text{ and } a \sqcap \top = a$$

Theorem. Every finite lattice $\{a_1, a_2, \ldots, a_n\}$ is bounded

Complementary Elements

- An element a in a bounded lattice L has a complementary element $x \in L$ if $a \sqcup x = \top$ and $a \sqcap x = \bot$
  - Complements need not always exist. Also, if they exist, they need to be unique, except in a special kind of lattice called a distributive lattice

- If every element of a lattice L has a complement then it is called a complemented lattice

Quasi-order I

- A quasi-order Q in a set S is a binary relation over S that satisfies the following properties:

  Reflexive. For every $a \in S$, aQa

  Transitive. If aQb and bQc then aQc

  - Differs from equivalence relation in that symmetry is not required
  - Differs from partial order in that no equality is defined, therefore anti-symmetry property cannot be defined

Theorem. If a quasi-order Q is defined on a set $S = \{a, b, \ldots\}$, and we define a dyadic relation E as follows: aEb iff aQb and bQa, then E is an equivalence relation.

Quasi-order II

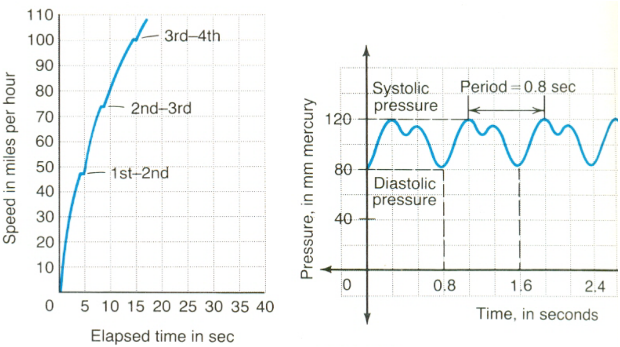
Theorem. Let E partition S into subsets $X, Y, \ldots$ of equivalent elements. Let $T = \{X, Y, \ldots\}$ and $\preceq$ be a dyadic relation in T meaning $X \preceq Y$ in T iff xQy in S for some $x \in X, y \in Y$. Then T is partially ordered by $\preceq$.

- A quasi-order order Q over a set S results in a partial ordering over a set of equivalence classes of elements in S

Functions

Graphs I

$x^b \qquad e^x \qquad \log(x)$



From: Miller, Heeren, Hornsby Jr. *Mathematical Ideas*

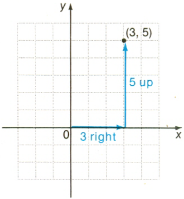
- These are *graphs* showing the dependence of one variable (y) on another (x)

Graphs II

- *Algebra* was usually concerned with relationships between a variables, expressed as thematical equations like $y = x^2$. The rules of algebra tell us how to find the values of y (or x), given the values of x (or y)
- *Geometry*, tells us, given definitions of lines, points, and angles, how to analyse figures in a plane (curves, polygons, circles and so on), and in higher dimensions (polyhedrons, cones, spheres and so on)
- It is not obvious that these two areas a related.
- A major achievement in the history of mathematics was the unification of algebra and geometry, largely due to Descartes and Pascal, in the form of *coordinate* or *analytic* geometry

Graphs III

- Coordinate geometry is what allows us to draw graphs of equations containing variables, by establishing a correspondence between ordered pairs of numbers (x, y) and points in a coordinate system



From: Miller, Heeren, Hornsby Jr. *Mathematical Ideas*

Drawing Graphs

- We can draw a graph by first computing a table of ordered pairs; finding the corresponding points in the coordinate system; and then drawing a curve that passes through those points

$$y = 4 - 2x$$

x	-1	0	1	2	3	4
y	6	4	2	0	-2	-4
(x, y)	$(-1, 6)$	$(0, 4)$	$(1, 2)$	$(2, 0)$	$(3, -2)$	$(4, -4)$

- It is sometimes easier to do this approximately, by looking at behaviour at extremes (what happens to y when $x = 0$ and $x = \pm\infty$). This is called the graphs *asymptotic behaviour*

$$y = \frac{a^3}{x^2 + a^2}$$

Functions I

- A mathematical function F is a set of ordered pairs (x, y) such that there is no pair with the same first component. The set of all first components (the x's) is the *domain* of the function, and the set of all second components (the y's) is *range* of the function
- When it is possible to do so, it is more compact to write F not as a set of ordered pairs $\{(x_1, y_1), (x_2, y_2), \ldots, \}$, but as an equation between x and y. This is especially so if F is an infinitely large set
  - For example, the function defined by the equation $y = 4 - 2x$ represents the infinite set $F = \{\ldots, (-1, 6), (0, 4), (1, 2), \ldots\}$
- It is common to shorten "the function defined by the equation $y = \ldots$" to "the function $y = \ldots$"

Functions II

- Some important functions for us are: power functions ($y = x^b$), exponentials ($y = a^x$), logarithmic ($y = \log(x)$), and factorial functions ($y = n!$)
- We will sometimes show functions as $f : x \to \dots$. For example, the function $y = x^b$ will be shown as the function $f : x \to x^b$
- Given sets A and B, a function $f : A \to B$ assigns an element of A to exactly one element of B
  - A subset of $A \times B$ such that there is exactly one ordered pair (a, b) for every $a \in A$.
  - A function is thus a special kind of relation.
  - A is called the domain of f and B is called the co-domain of f. If $f(a) = b$ then b is image of a and a is the pre-image of b. The set of images of elements of A is the range of f
- Here are some useful functions:

  $\lfloor x \rfloor$: the floor function

Functions III

$\lceil x \rceil$: the ceiling function

$|x|$: absolute value function

$ln(x)$: the natural logarithm function

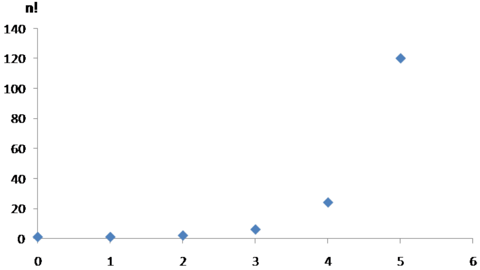
$e^{(x)}$: the exponential function

The Factorial Function I

- The function $y = f(n) = n!$ is a function defined over the natural numbers ($n = 0, 1, 2, \ldots$) and is a function that increases very quickly

n	$n!$	2^n
0	1	1
1	$1 \times 0! = 1$	2
2	$2 \times 1! = 2$	4
3	$3 \times 2! = 6$	8
4	$4 \times 3! = 24$	16
5	$5 \times 4! = 120$	32
6	$6 \times 5! = 720$	64

The Factorial Function II

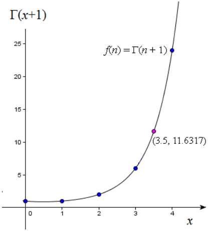


- What about the points in between? That is, can we draw a curve, and calculate 3.5!?
  - The answer is "No", because $f(n) = n!$ is only defined over the positive integers and 0
  - But there is a curve that can be drawn joining these points, and representing a function that is defined over the real numbers

The Factorial Function III

- This function is called the *Gamma* function, $\Gamma(x)$ for $x \in \Re$. The Gamma function and the factorial function are related in this way:

$$f(n) = n! = \Gamma(n+1)$$



(We can use this in reverse: if we want to calculate $f(100)$, then we can do this by finding out $\Gamma(101)$)

The Gamma Function I

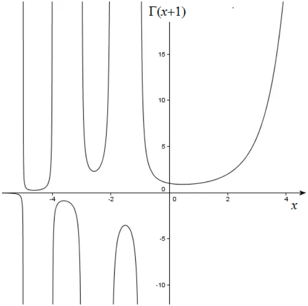
- The Gamma function is defined as:

  $$\Gamma(x+1) = \int_0^\infty t^x e^{-t} dt$$

  So, $\Gamma(x)$ is the area under the curve formed by the product of a negative exponential and a power curve
- You should be able to work out that $\Gamma(1) = \int_0^\infty e^{-t} dt = 1$

The Gamma Function II

- In fact, this function is not defined for the negative integers $(x = -1, -2, \ldots)$



- Like the factorial function, $\Gamma(1) = 1$. But, slightly different to the factorial function, $\Gamma(n+1) = n\Gamma(n)$. So:

$$\Gamma(n) \; = \; (n-1) \times (n-2) \cdots \times 1 \; = \; (n-1)!$$

Linear Algebra

Basics

- Revise these:
  - Vectors
  - Matrices
- You should be able to add and subtract vectors; add and subtract and multiply matrices, etc.
- The best book for all this is G.Strang, *Introduction to Linear Algebra*, The 5^{th} edition was released in 2016.

Scalars, Vectors, and Matrices I

- A scalar is a quantity represented by a number
- A vector is a sequence of numbers usually represented thus:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$$

- The point specified by the special vector in which all entries as 0 is called the origin (and is usually denoted by $\mathbf{0}$)

Scalars, Vectors, and Matrices II

- A $m \times n$ matrix is a 2-D arrangement of numbers into m rows each of n columns usually represented thus:

$$A_{m,n} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ldots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

- If $m = n$ then the matrix is called a *square* matrix. A special square matrix with just 1's along the diagonal, and 0's elsewhere is called the *identity* matrix

$$I_{n,n} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Scalars, Vectors, and Matrices III

We will often drop the index-subscripts. So $A_{m,n}$ will simply be written as A, and $I_{n,n}$ as I. Also if $m = n$, sometimes $A_{m,n}$ will be written as A_n and $I_{n,n}$ as I_n

- A *diagonal* matrix is a matrix in which only the elements along the diagonal $a_{i,i}$ are non-zero. The sum of elements along the diagonal is called the *trace* of the matrix.
- An upper triangular matrix is a one in which $a_{i,j} = 0$ for $i < j$. A lower triangular matrix is one for which $a_{i,j} = 0$ for $i > j$.
- It is sometimes useful to think of a scalar v as a matrix $V_{1,1}$ with a single element

$$V_{1,1} = \begin{bmatrix} v \end{bmatrix}$$

Scalars, Vectors, and Matrices IV

and a vector $\mathbf{v}$ with m entries is a matrix $V_{m,1}$:

$$V_{m,1} = \begin{bmatrix} v_{1,1} \\ v_{2,1} \\ \vdots \\ v_{m,1} \end{bmatrix}$$

and a matrix as an arrangement of vectors:

$$A = \begin{bmatrix} | & | & \cdots & | \\ \mathbf{v_1} & \mathbf{v_2} & \cdots & \mathbf{v_n} \\ | & | & \cdots & | \end{bmatrix}$$

m is called the dimension of the vector. We will continue to refer scalars and vectors as v and $\mathbf{v}$ respectively, but sometimes use matrix operations on them.

Scalars, Vectors, and Matrices V

- Matrices can also be broken into smaller matrices called *blocks*. For example this matrix consists of two I_2 blocks:

$$A_{2,4} \ = \ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

  Blocks do not have to be square.
- The transpose of a matrix A^T is obtained by swapping rows and columns. Thus, if $B = A^T$, then $b_{i,j} = a_{j,i}$.
  - Clearly the matrix of a scalar v is equal to its transpose
  - A matrix is said to be *symmetric* if $A^T = A$; and *skew-symmetric* if $A^T = -A$
- Some combinations of scalars and vectors are useful to remember:
  - The combination $a\mathbf{u}$ denotes points along a line. It is a vector in the direction of $\mathbf{u}$ scaled by a
  - The combination $a\mathbf{u} + b\mathbf{v}$ denotes points in 2-D space
  - The combination $a\mathbf{u} + b\mathbf{v} + c\mathbf{w}$ denotes points in 3-D space

Products I

- If matrix $C_{m,p}$ is the product of matrices $A_{m,n}$ and $B_{n,p}$ then:

$$c_{i,j} = \sum_k a_{i,k} b_{k,j}$$

- Matrix multiplication is distributive ($A(B + C) = AB + AC$) and associative ($A(BC) = (AB)C$, but not commutative ($AB \neq BA$)
- For matrices partitioned into blocks, multiplication can be done over blocks (pretend the blocks are elements)
- It is easy to check that a special property of the identity matrix I is

$$IA = A$$

  - If A is a square matrix then $AI = IA$

Products II

  - Closely related to the identity matrix is the *permutation* matrix P. $P_{i,j}$ is the identity matrix with rows i and j swapped. For example

$$P_{1,3} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

  - Check that the product $P_{i,j}A$ swaps the rows i and j of a matrix A

- The *inverse* of a matrix A is defined as the matrix B such that:

$$BA = A^{-1}A = I$$

  A matrix A may not always have an inverse, but if it exists, it is unique (can you prove the inverse is unique?)

- If A and B have an inverse, then AB has an inverse (can you prove this is $B^{-1}A^{-1}$?)

Products III

- The transpose of products is the product of the transposes in reverse order ($(AB)^T = B^T A^T$)
- The dot product of a pair of vectors $\mathbf{u}$ and $\mathbf{v}$ with the same dimension m is the matrix product $U_{m,1}V_{m,1}$. It is a number (i.e. it is a scalar)
    - This is usually denoted by $\mathbf{u} \cdot \mathbf{v}$ and we will simply say $\mathbf{u} \cdot \mathbf{v} \ = \ \mathbf{uv}$, when matrix multiplication is understood on the r.h.s.
    - So, entry $C_{i,j}$ in the matrix multiplication of matrices A and B is the a dot product between row i of A and column j of B
    - Unlike general matrix multiplication, the dot product is commutative ($\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$)
    - The length of a vector $\|v\|$ is equal to $\sqrt{\mathbf{v} \cdot \mathbf{v}}$. A unit vector is a vector of length 1. The standard unit vectors along the x and y axes are denoted $\mathbf{i}$ and $\mathbf{j}$
        - In general, a unit vector that makes an angle θ is $cos\theta\mathbf{i} + sin\theta\mathbf{j}$

Products IV

- It is easy to get a unit vector in the direction of any vector $\mathbf{v}$ by re-scaling the length. The vector $\mathbf{v}/\|v\|$ has length 1 and is in the direction of $\mathbf{v}$
- For a pair of vectors $\mathbf{v}, \mathbf{w}$ with an angle θ between them, it can be shown:

$$\frac{\mathbf{v} \cdot \mathbf{w}}{\|v\|\|w\|} = cos\theta$$

It follows immediately that if $\theta = 90^\circ$ then $\mathbf{v} \cdot \mathbf{w} = 0$

Subspaces I

- A subspace V of $\Re^m$ is a subset of $\Re^m$ s.t. (a) if $\mathbf{v}, \mathbf{w} \in V$, then $\mathbf{v} + \mathbf{w} \in V$; and (b) if $\mathbf{v} \in V$ then $\alpha \mathbf{v} \in V$
- Let V be the set of vectors

$$\mathbf{v} = \begin{bmatrix} x \\ 0 \\ 0 \end{bmatrix}$$

  Show that V is a subspace
- Let V be the set of vectors

$$\mathbf{v} = \begin{bmatrix} x \\ 0 \\ 1 \end{bmatrix}$$

  Show that V is not a subspace

Subspaces II

- Let $V = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}\}$ be a subset of vectors in $\Re^m$. For $x_i \in \Re$, show that the set S consisting of vectors:

$$x_1\mathbf{v_1} + x_2\mathbf{v_2} + \cdots + x_n\mathbf{v_n}$$

  is a subspace. S is called the span of V.

- What is the span of the vectors $\mathbf{u} = [100]$ and $\mathbf{v} = [010]$ and $\mathbf{w} = [001]$?

Linear Transformations I

- Let $f : \Re^n \rightarrow \Re^m$ be a function with the following properties:
  1. $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$
  2. $f(a\mathbf{x}) = af(\mathbf{x})$ for any scalar a

  Then f is said to be a linear transformation or simply "f is linear"
- Show the following transformation is not linear:

$$f\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = \begin{bmatrix} x^2 \\ y^2 \\ z^2 \end{bmatrix}$$

- Let $A_{m,n}$ be a $m \times n$ matrix. Show that for $\mathbf{x} \in \Re^n$: (a) The function $f_A(\mathbf{x}) = A\mathbf{x}$ is a function $f : \Re^n \rightarrow \Re^m$; and (b) f_A is a linear

Linear Transformations II

- Write the following linear transformation as a matrix function $A\mathbf{x}$:

$$f\left(= \begin{bmatrix} x \\ y \\ z \end{bmatrix} \right) = \begin{bmatrix} (x+y+z)/3 \\ (x+y+z)/3 \\ (x+y+z)/3 \end{bmatrix}$$

- In general, for every linear transformation $f : \Re^n \to \Re^m$ there exists a matrix A s.t. if $f(\mathbf{x}) = \mathbf{y}$ then $A\mathbf{x} = \mathbf{y}$
  - Linear transformations are matrix equations

Equations I

- A set of linear equations:

$$\begin{array}{rcl} // a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n & = & b_1 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots & = & \vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n & = & b_m \end{array}$$

is more concisely written using matrices as:

$$A_{m,n}X_{n,1} \ = \ B_{m,1}$$

or, simply:

$$A\mathbf{x} = \mathbf{b}$$

So, the set of linear equations is just a linear transformation from $\Re^n$ to $\Re^m$

Equations II

- Show that if A^{-1} exists, then the value of $\mathbf{x}$ that satisfies:

$$A\mathbf{x} = \mathbf{b}$$

is:

$$\mathbf{x} = A^{-1}\mathbf{b}$$

(A is often called the *coefficient* matrix)

- Show that if $\mathbf{u}$ and $\mathbf{v}$ are solutions to $A\mathbf{x} = \mathbf{b}$ then any combination $\alpha\mathbf{u} + (1-\alpha)\mathbf{v}$ is also a solution for $\alpha \in \Re$
  - So, if a set of linear equations has more than one solution, it has infinitely many solutions
  - The main interest therefore is to find out if: the equations have 0 solutions; 1 solution; or more than 1 solution.

Equations III

- Returning to the original equations:

$$\begin{array}{rcl} //a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n & = & b_1 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots & = & \vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n & = & b_m \end{array}$$

  Looking at this column-wise, the equations are of the form:

$$x_1\mathbf{a_1} + x_2\mathbf{a_2} + \cdots + x_n\mathbf{a_n} = \mathbf{b}$$

  This is a linear combination of the column vectors of the matrix A (these are vectors in $\Re^m$.

- These are 2 different ways of looking at solving a set of equations

Equations IV

  - The *row* view looks at each equation as a hyperplane in $\Re^n$. The solution is the $n-$dimensional point of intersection of these hyperplanes
  - The *column* view looks at the equations as a linear combination of vectors. The solution is the combination that produces the vector on the r.h.s.
  - Both views are valid: the row view is probably more familiar from 2-D equations in school. For higher dimensions, the column view becomes easier.

- In general, all the vectors obtainable from the linear combination of a set of vectors is called the *span* of the set
  - The span (*i.e.* the linear combination) of a set of vectors in $\Re^m$ is a *subspace* of $\Re^m$
  - The equations $A\mathbf{x} = \mathbf{b}$ has a solution if $\mathbf{b}$ is in the span of the column vectors of A (this is called the *column space* of A

Equations V

  - Since $\mathbf{b}$ is an m-dimensional vector, it is a point in $\Re^m$. So, if any point in $\Re^m$ is not in the column space of A, then that point has no solution
  - This means that A has to have at least m columns. WHY? Let $m = 3$ and A have 2 columns. So, $\mathbf{x}$ will be a 2-dimensional vector, and the column-space will be the vectors reachable using a linear combination of the

- So, a set of linear equations

$$A\mathbf{x} = \mathbf{b}$$

  has a solution means that there is some linear combination of the columns of A that results in $\mathbf{b}$

- This will mean all of the following: (a) A has and inverse; (b) The columns of A are independent; and (c) the deteriminant of A is not 0
  - Review these ideas from Strang

Equations VI

- A necessary (but not sufficient) condition for a solution to exist is to return to the linear transformation view of $Avxx = \mathbf{b}$. In this, the matrix A transforms a vector from $\Re^n$ to $\Re^m$. For this to be possible uniquely, $n \geq m$.
    - The number of columns n of A should be $\geq$ the number of rows m
    - The column space of A spans $\Re^m$
- Why is this condition not sufficient?
    - Not all the m columns may be independent of each other.
    - So, we need the matrix to have at least 1 set of m linearly independent columns
    - If $n \geq m$, then there can be more than one set of m independent columns. If so, the solution will not be unique.
    - So, for a unique solution, we need exactly 1 set of m linearly independent columns. This is enforced if $n = m$ (that is, the matrix is square)

Equations VII

- A square matrix that does not satisfy this is called a *singular* matrix.

Solving $A\mathbf{x} = \lambda\mathbf{x}$ I

- We now look at a solving a special equation: $A\mathbf{x} = \lambda\mathbf{x}$ for square matrices A. For any square matrix A, the values of $\mathbf{x}$ for which this is true are called the *eigenvectors* of A and the corresponding values of λ are called *eigenvalues* of A
    - If $A\mathbf{x} = \lambda\mathbf{x}$ then $(A - \lambda I)\mathbf{x} = \mathbf{0}$. That is, $(A - \lambda I)$ is not invertible (it is singular), or equivalently, the determinant of $(A - \lambda I)$ is 0
    - If A is an $n \times n$ matrix, then it will have n eigenvalues
- So, to compute the eigenvalues and eigenvectors:
    - Get the determinant of $A - \lambda I$ ($detA - \lambda I$) If A is a $n \times n$ matrix then this will be a polynomial in λ^n.
    - Find solutions to the equation $\det(A - \lambda I) = 0$. If A is a $n \times n$ matrix will have n roots. Each root is an eigenvalue. (Quick checks: the sum of eigenvalues is equal to the trace of A; and the product of eigenvalues is $\det(A)$)
    - With each eigenvalue λ_i find the corresponding eigenvector $\mathbf{x_i}$ by solving $(A - \lambda_i I)\mathbf{x_i} = \mathbf{0}$

Solving $A\mathbf{x} = \lambda\mathbf{x}$ II

- NOTE: If λ is an eigenvalue of A then λ^n is an eigenvalue of A^n. There is no change in eigenvector. (That is, $A^n\mathbf{x} = \lambda^n\mathbf{x}$)
- ALSO: if λ is an eigenvalue of A with corresponding eigenvector $\mathbf{x}$, then any scaled vector $\mathbf{sx}$ is also an eigenvector of A with the same eigenvalue λ. (Can you prove this?)
  - So it is sufficient to look at unit-length eigenvectors
- AND: If the solution of $(A - \lambda_i I)\mathbf{x_i} = \mathbf{0}$ is $\mathbf{x_i} = \mathbf{0}$ then λ_i isn't an eigenvalue (each eigenvalue must result in a non-zero eigenvector)

Diagonalisation I

- One important use of eigenvectors is to make a matrix diagonal (diagonalisation). Why is this useful? We will see that soon
- A $n \times n$ matrix A is "diagonalisable" if it has n eigenvalues, resulting in n linearly independent eigenvectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$
- The diagonalisation of a matrix A is as follows:
  - Let S be the matrix formed with the eigenvectors of A as columns

$$S = \begin{bmatrix} \mathbf{x_1} & \mathbf{x_2} & \cdots & \mathbf{x_n} \end{bmatrix}$$

  - The AS is:

$$AS = \begin{bmatrix} \lambda_1 \mathbf{x_1} & \lambda_2 \mathbf{x_2} & \cdots & \lambda_n \mathbf{x_n} \end{bmatrix}$$

Diagonalisation II

- So, $AS = S\Lambda$ where

  $$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

  is a diagonal matrix comprised of the eigenvalues of A. Λ is also written sometimes as $\text{diag}(\lambda)$ where λ is the column vector formed by the eigenvalues $\lambda_1, \ldots, \lambda_n$

- S and Λ allow us to re-express A:

  $$AS = S\Lambda \quad \text{therefore} \quad A = S\Lambda S^{-1} \quad \text{and} \quad S^{-1}AS = \Lambda$$

  (S^{-1} exists because the columns of S are the eigenvectors which are linearly independent)

Diagonalisation III

- The expression $A = S\Lambda S^{-1}$ is a *decomposition* of A into 3 factors (S, Λ and S^{-1}). This is called an *eigendecomposition* of A, and is only defined for square matrices. Specifically, a $n \times n$ matrix has an eigendecomposition if it has n distinct eigenvalues.
- A more general decomposition of a non-square matrix is called a *singular value decomposition* or SVD. The SVD of a matrix $m \times n$ matrix A is defined as:

$$A \;=\; UDV^T$$

Calculus

Basics

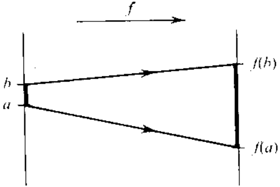
- Revise these:
  - Differential calculus
  - Integral calculus
- You should be able to compute: maximum and minimum points of a one-dimensional function, areas under curves *etc*

Scale Factors and Derivatives I

- Suppose with every point x in the domain, we associate another point $3x$ (that is, we are considering the function $f : x \to 3x$). In effect, we are uniformly scaling the function every point $x = a$ by 3
- Now suppose with every point x in the domain, we associate another point x^2 (that is, $f : x \to x^2$). Now, the scaling (or "stretching") of points is not uniform. Points between 1 and 2 get mapped to points between 1 and 4; those between 2 and 3 get mapped to points between 4 and 9; and so on. So, all we can say is that the *average* scale factor between 1 and 2 is $(4-1)/(2-1 = 3$; betweeen 2 and 3 is $(9-4)/(3-2) = 5$; and so on

Scale Factors and Derivatives II

- In general:



The average scale factor for the interval $[a, b]$ is $\frac{f(b)-f(a)}{b-a}$

- Now, for each element $x = a$, we can, in fact calculate the actual scale factor at $x = a$, by looking at intervals above and below $x = a$. So, for $x = 3$ and the squaring function say, the average scale factor over $[3, b]$ is $(b^2 - 9)/(b - 3) = (b + 3)$ (for $b \neq 3$). Over the interval $[a, 3]$, this is $(9 - a^2)/(3 - a) = (3 + a)$ (for $a \neq 3$). So, as a and b get closer to 3, the average scale factor gets closer to 6. This is called the *local scale factor* or *derivative* at 3

Scale Factors and Derivatives III

- For any function $f(x)$, is possible to associate with each value $x = a$ its derivative value at that point. This mapping is itself a function of x, and is usually denoted by $f'(x)$. Thus, if $f : x \rightarrow x^2$, you can verify $f' : x \rightarrow 2x$ (we found this above for $x = 3$)
- The process of finding f' given f is known as *differentiation*. In general, using the average scale factor (A.S.F.) in the interval $[a, b]$ for the function f, we can calculate $f'(x = a)$ as the limiting value as b approaches a of the A.S.F. This is usually written as $f'(a) = \lim_{b \rightarrow a} \frac{f(b) - f(a)}{b - a}$
- Sometimes b is taken to be $a + h$, and

$$f'(x) = \frac{f(x+h) - f(x)}{h}$$

Numerical Derivatives I

- We can attempt to approximate the derivative of f as a linear combination of values near x. We will use the points $f(x-h)$, $f(x)$ and $f(x+h)$
- The Taylor expansion of the points of these values about x are:

$$\begin{aligned} //f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + \cdots \\ f(x) &= f(x) \\ f(x-h) &= f(x) - hf'(x) + \frac{h^2}{2!}f''(x) - \frac{h^3}{3!}f'''(x) + \cdots \end{aligned}$$

Numerical Derivatives II

- These equations can be used to give approximations to $f'(x)$. An approximation with $O(h^2)$ error is:

  $$f'(x) \ = \ \frac{f(x+h) - f(x-h)}{2h}$$

  This is the *central difference* approximation for $f'(x)$

- Similarly, an approximation with $O(h^2)$ error for $f''(x)$ can be obtained by taking the first two terms of the Taylor expansion:

  $$f''(x) \ = \ \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

- The same technique can be used to approximate $f'(x)$ using more points
  - Derive an approximation for $f'(x)$ in terms of $f(x)$, $f(x+h)$ and $f(x+2h)$

Numerical Derivatives III

- BUT: what if we do not have a closed form for the function f, and consequently are not able to calculate $f(x+h)$ *etc*.
  - We will have to approximate the value of f by fitting a function between points around $f(x)$. For example, a linear interpolation would use a straight line approximation between known points near x
  - How to we get the values for f near x? We will assume we have "black-box" that can give us the value of $f(x)$, given any x as input

Multi-dimensional Functions I

- So far, we have only looked at functions of one-dimensional functions $y = f(x)$
- Often we have functions of more than 1 variable:

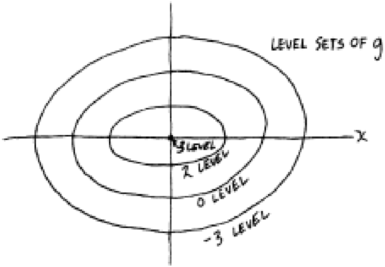
$$f(x_1, x_2) = 2x_1 + 3x_2 \qquad f(x_1, x_2, x_3) = e^{-x_1} + 2x_2x_3$$

- In general, you know how to draw graphs of functions of one variable (like $y = x^2$), and even some attempt can be made of functions of 2 variables. For example, the function $y = 3 - {x_1}^2 - 2{x_2}^2$, can be reconstructed by looking at various fixed values of y.
  - At $y = 3$, the function is ${x_1}^2 + 2{x_2}^2 = 0$
  - At $y = 0$ the function is ${x_1}^2 + 2{x_2}^2 = 3$
  - and so on

  This gives us cross-sections that are … (draw them)

Multi-dimensional Functions II

- But what is to be done with functions of 3 or more variables? We will need a generalisation of the above, called *level sets*
- Let $g(x_1, x_2) = 3 - x_1^2 - 2x_2^2$. The points at which the value of g is constant at a particular level (say -6) is the ellipse $x_1^2 + 2x_2^2 = 9$. This is called the *level set* at the -6 level.



(From: C. Ash, R.B. Ash (1993), *The Calculus Tutoring Book*)

Multi-dimensional Functions III

- The level set of a function with 2-variables are pictures in 2-dimensions. Similarly, level sets of a function with 3-variables are usually surfaces in 3-dimensions $(F(x_1, x_2, x_3) = c)$
  - Draw the level sets for $f(x_1, x_2, x_3) = 2x_1 + 3x_2 + 6x_3 - 10$
- Level sets of a function are like cross-sections (discretised version) of the graph (which is continuous). The level-set of a function in 2-variables is a picture in $2-$space, but the graph is a 3-dimensional surface (recall the ellipsoid and the 2-dimensional ellipses)
- Level-sets need not be only for functions. The equation $x_1^2 + x_2^2 + x_3^2 = 4$ is a sphere. It is not a function of the form $x_3 = f(x_1, x_2)$, but the sphere 4-level set of $h(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2$ and the 0-level set of $g(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - 4$

Partial Derivatives I

- A function $f(x_1, x_2)$ has two first-order *partial derivatives* $\frac{\partial f}{\partial x_1}$ and $\frac{\partial f}{\partial x_2}$
    - $\frac{\partial f}{\partial x_1}$ is the derivative of f w.r.t. x_1 with x_2 treated as a constant. Similarly for $\frac{\partial f}{\partial x_2}$
    - The partial of $x_1^2 x_2^3$ w.r.t to x_1, the function behaves like cx_1^2 where c is the constant x_2^3. So, $\frac{\partial (x_1^2 x_2^3)}{\partial x_1}$ is $2x_1x_2^3$. Similarly $\frac{\partial (x_1^2 x_2^3)}{\partial x_2} = 3x_1^2x_2^2$
    - What is $\frac{\partial f}{\partial x_1}$ for $f(x_1, x_2, x_3) = x_3 e^{2x_1 + 3x_2 + 4x_3})$
- Higher-order partial derivatives are defined using the lower order partials. Thus $\frac{\partial^2 f}{\partial x_1 \partial x_2}$ is really $\frac{\partial}{\partial x_1} \frac{\partial f}{\partial x_2}$
    - For $f(x_1, x_2) = x_1^3x_2^5 + x_1^3 + x_2^4 + 7$, find $\frac{\partial f}{\partial x_1}$, $\frac{\partial f}{\partial x_2}$, $\frac{\partial^2 f}{\partial x_1^2}$, $\frac{\partial^2 f}{\partial x_1 \partial x_2}$, $\frac{\partial^2 f}{\partial x_2 \partial x_1}$

Partial Derivatives II

- The last two are called *mixed* partials. We will assume (without proof) that mixed partials will always be equal. This is also the case for $\frac{\partial f}{\partial x_1 \partial x_1}$.
- So, what matters is the number of times the differentiation is done, not the order
- Sometimes the notation f_{xx}, f_{xy} *etc.* f_{xy} normally means $(f_x)_y$, which is different to the $\frac{\partial}{\partial}$ ordering (but this does not matter) are used for partials

Partial Derivatives, Graphs and Level Sets I

- We often want to know what the value of the partial derivative is at a specific point:

$$\left.\frac{\partial f}{\partial x_1}\right|_{x_1 = a x_2 = b}$$

- Consider what this means graphically:
  1. The graph of $f(x_1, x_2)$ is a surface in 3-space.
  2. $x_2 = b$ is a plane in 3-space, and $f(x_1, b)$ is a curve obtained by the intersection of the surface $f(x_1, x_2)$ and the plane $x_2 = b$
  3. $\frac{\partial f}{\partial x_1}|_{x_2 = b}$ is the slope of the curve obtained as x_1 varies
  4. $\left.\frac{\partial^2 f}{\partial x_1 \partial x_2}\right|_{x_1 = a, x_2 = b}$ is the slope of the curve at the point $(x_1 = a, x_2 = b, f(a, b))$

Partial Derivatives, Graphs and Level Sets II

- With level sets, contour lines are intersected by orthogonal lines denoting increasing values of x_1, x_2 etc. As x_1 increases, if it progressively intersects contours with increasing level set values then $\frac{\partial}{\partial x_1}$ is positive. If instead the level sets decrease, then the partial is negative.
- The partial derivatives of a function of variables x_1, x_2 $\frac{\partial f}{\partial x_1}$ and $\frac{\partial f}{\partial x_2}$ represents the instantaneous rates of change in directions of the vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ (usually x_1, x_2 are called x, y and the vectors $\mathbf{i}, \mathbf{j}$
  - The partial derivative is thus a *directional* derivative
  - For a function of 2-dimensions, the directional derivative of f in the direction of any vector $\mathbf{u} = u_1\mathbf{i} + u_2\mathbf{j}$ is:

$$D_{\mathbf{u}} f \;=\; \frac{\frac{\partial f}{\partial x} u_1 + \frac{\partial f}{\partial y} u_2}{\|\mathbf{u}\|}$$

Maxima and Minima I

- We will say $f(x_1, \ldots, x_n)$ has a *relative* or *local* maximum at $a_1, \ldots, a_n$ iff $f(a_1, \ldots, a_n) \geq f(x_1, \ldots, x_n)$ for all points $x_1, \ldots, x_n$ near $a_1, \ldots, a_n$. Similarly for minima.
- Let us now assume that partial derivatives are defined everywhere and are finite.
- A necessary condition for $f(x_1, \ldots, x_n)$ to have a relative maximum (minimum) at $a_1, \ldots, a_n$ then $\frac{\partial f}{\partial x_i} = 0$ for all the x_i at $a_1, \ldots, a_n$. But this is not sufficient. So, the values for which the partials are 0 include all the relative extrema but may include other points as well (the points where the partials are 0 are called *critical* points)
  - Can you show an example of a critical point that is not a relative maximum or minimum?

Maxima and Minima II

- In general, to find absolute extrema, we will have to find: (A) the critical points; (B) the values of the function at boundary conditions; and (C) if there are any infinite values as $x_i \to \infty$ or $x_i \to -\infty$. values
    - It may be possible that the partials are never equal to 0 in the region of allowed values. Then, we would have to either examine the function's behaviour or compute values in the region to get progressively increasing or decreasing values of f

The Gradient I

- We saw earlier that for a function of two variables the partial derivatives are directional derivatives in the directions $\mathbf{i}, \mathbf{j}$
- That is, $\frac{\partial f}{\partial x}$ is the instantaneous rate of change of f in the direction of $\mathbf{i}$ and $\frac{\partial f}{\partial y}$ is the instantaneous rate of change of f in the direction $\mathbf{j}$
- The vector ∇f, called the *gradient* of f is:

$$\nabla f \;=\; \frac{\partial f}{\partial x}\mathbf{i} \;+\; \frac{\partial f}{\partial y}\mathbf{j}$$

- RECALL: the directional derivative of f in the direction of any vector $\mathbf{u} = u_1\mathbf{i} + u_2\mathbf{j}$ is:

$$D_{\mathbf{u}} f \;=\; \frac{\frac{\partial f}{\partial x}u_1 + \frac{\partial f}{\partial y}u_2}{\|\mathbf{u}\|}$$

The Gradient II

- It is easy to see that this is just:

$$D_{\mathbf{u}}f \;=\; \frac{\nabla f \cdot \mathbf{u}}{\|\mathbf{u}\|}$$

  and that $D_{\mathbf{i}}f = \frac{\partial f}{\partial x}$ and $D_{\mathbf{j}}f = \frac{\partial f}{\partial y}$

- $D_{\mathbf{u}}f$ is the (signed) projection of ∇f onto $\mathbf{u}$. This is a maximum if $\mathbf{u}$ is in the direction of ∇f; a minimum in the direction of $-\nabla f$; and 0 is a direction perpendicular to ∇f
  - Maximum rate of *increase* in f is in the direction of ∇f and the maximum rate of *decrease* in f is in the direction of $-\nabla f$
  - All this is about *instantaneous* change. As soon as you move to a different point there is a new gradient, and new directions of maximum increase and decrease

The Gradient III

- IN PRINCIPLE: you could take small steps in the direction of ∇f at any point and possibly get to a relative maximum. Similarly for $-\nabla f$ for a relative minimum). This is the basis of the numerical procedure of gradient ascent (gradient descent):

  $$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \eta \nabla f_{\mathbf{x}^{(t)}}$$

  where η is some small value and $\nabla f_{\mathbf{x}^{(t)}}$ is the gradient of f at $\mathbf{x}^{(t)}$ (the $+$ sign becomes $-$ for gradient descent)

- For example, let us find the minimum value of the function $z = f(x,y) = \frac{x^2}{a^2} + \frac{y^2}{a^2}$. This is an elliptic paraboloid. Let us take $a = 1$ and $\eta = 0.1$ for this example.
  - The partials are $\frac{\partial f}{\partial x} = \frac{2x}{a^2}$ and $\frac{\partial f}{\partial y} = \frac{2y}{a^2}$
  - The gradient vector at a point (x,y) is $\frac{2x}{a^2}\mathbf{i} + \frac{2y}{a^2}\mathbf{j}$. If $a = 1$, then this is $2x\mathbf{i} + 2x\mathbf{j}$.

The Gradient IV

  - Let us start at the point $(1, 1)$. That is, $\mathbf{x}^{(0)} = \mathbf{i} + \mathbf{j}$. The gradient at this point for $a = 1$ is $\nabla f = 2\mathbf{i} + 2\mathbf{j}$. To find the minimum, we want to move in the direction of $-\nabla f = -2\mathbf{i} - 2\mathbf{j}$
  - A small step with $\eta = 0.1$ in the direction of $-\nabla f$ is $-0.2\mathbf{i} - 0.2\mathbf{j}$. So $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - 0.2\mathbf{i} - 0.2\mathbf{j} = 0.8\mathbf{i} + 0.8\mathbf{j}$. The value of f at this point is $0.64 + 0.64 = 1.28$.
  - Keep iterating until we come to a relative minimum $\mathbf{0}$
- Two immediate issues that comes up: (1) What if we overshoot the mininimum; (2) What if there is a lower minimum value than the local minimum we arrive at?
- A third issue: what if do not have a closed-form equation for f (and therefore ∇f)? ANS: use numerical differentiation

The Gradient V

- To approximate $\frac{\partial f}{\partial x_i}$, give the black box an input which keeps all other variables constant, but changes x_i to $x_i \pm h$. The corresponding output from the black box will give the values of $f(x_i \pm h)$. The central difference approximation will give us an estimate of $\frac{\partial f}{\partial x_i}$

Statistics

Statistical Analyses

- Statistical analyses usually involve one of 3 things: (1) The study of populations; (2) The study of variation; and (3) Techniques for data abstraction and data reduction
- Statistical analysis is more than statistical computation:
  1. What is the question to be answered?
  2. Can it be quantitative (i.e. can we make measurements about it)?
  3. How do we collect data?
  4. What can the data tell us?

Some things you may want to do with data

1. Visualise
2. Summarise
3. Determine the distribution of values
4. Compare groups of instances
5. Identify or describe relationships
6. Identify groups of similar instances
7. Identify groups of variables

Averages

$$\text{Mean} = \frac{\text{Sum of readings}}{n}$$

6, 3, 7, 5, 6, 4, 4, 5, 6, 7, 3, 5,9, 6, 4, 2, 7, 5, 8, 6

$\underline{\text{Mean} = 5.4}$ N = 20)

2, 3,3, 4,4,4, 5,5,5,5, 6,6,6,6,6, 7,7,7, 8, 9

↑

$\underline{\text{Mean} = 5.4}$ N = 20)

Values are roughly symmetrically distributed about the mean with a "hump" in the middle (roughly bell-shaped)

Frequency Distributions

- By counting the number of times a value appears in a set of readings, we get a *frequency distribution*
- Frequency distributions that are approximately bell-shaped are well summarised by the mean (about half the values are above the mean, and about half are below it)
- All other distributions are not

Skewed and U-Shaped Distributions

2, 3,3,3, 4,4,4,4, 5,5,5,5,5,5, 6,6,6, 8, 10, 15

Mean = 5.4 N = 20)

0,0,0,0,0,0,0,0, 1,1,1,1,1, 2,2,2, 3,3, 15, 76

Mean = 5.4 N = 20)

0,0,0,0,0,0,0, 1,1,1, 2, 10, 11,11,11, 12,12,12,12,12

Mean = 5.4 N = 20)

Other Averages: the Median and the Mode

2, 3,3, 4,4,4, 5,5,5,5, 6,6,6,6,6, 7,7,7, 8, 9

Mean = 5.4 N = 20)

Median = 5.5 Mode = 6

0,0,0,0,0,0,0,0, 1,1,1,1,1, 2,2,2, 3,3, 15, 76

Mean = 5.4 N = 20)

Median = 1.0 Mode = 0

0,0,0,0,0,0,0, 1,1,1, 2, 10, 11,11,11, 12,12,12,12,12

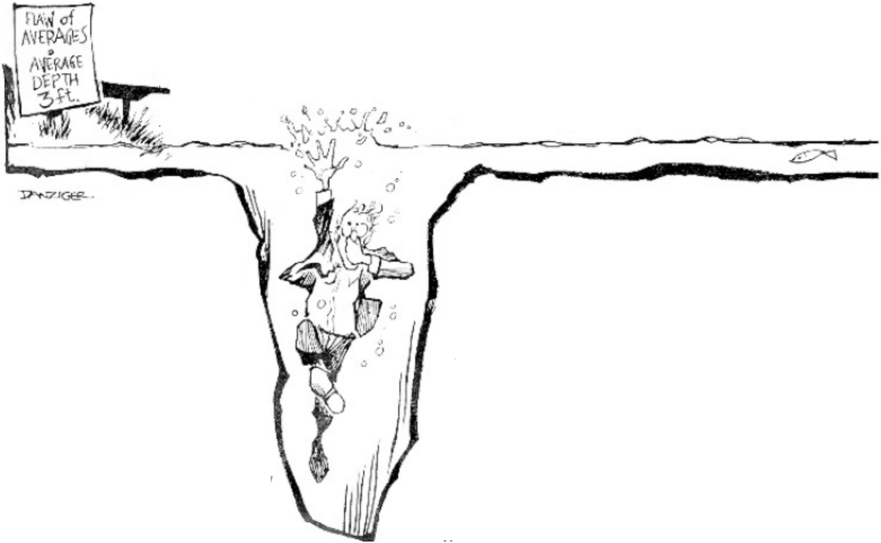
Mean = 5.4 N = 20)

Median = 1.5 Mode = 0 (and 12)

Why Use is the Mean?

- The mean is a good summary of approximately bell-shaped distributions
- It is easy to calculate
- It can be used to compare two datasets, as long as both have roughly the same kind of distribution. This is even if the distributions are skewed or U-shaped.
- It can be used to show-up outliers

The Mean is not Enough



Scatter or Spread

- Scatter refers to the spread of values in a set of data
- It is not enough to say: "The average depth is 3ft." You would also want to include a statement like "Depths vary from 1ft to 12ft."
- There are several quantitative measures of scatter or spread. Of these, the most relevant to us is the *variance* (and the related quantity, the *standard deviation*)

$$\text{Variance} = \frac{\text{Sum of squared deviations from the mean}}{n}$$

(actually, the denominator is $n - 1$)

$$\text{S.d.} = \sqrt{\text{Variance}}$$

What the Variance Tells Us

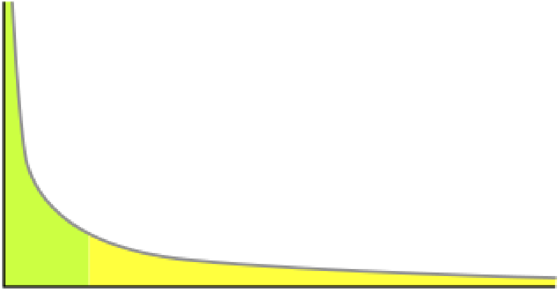
- For approximately bell-shaped distributions, nearly 70% of the observations lie within 1 s.d. of the mean
- Even for skewed distributions, this proportion is over 50%
- Irrespective of the distribution, "nearly all" values are close to the mean: no more than $1/k^2$ of the values will be more than k standard deviations away from the mean (this is known as "Chebyshev's inequality")

Deviation (from the mean)	Proportion of Values
2 s.d	At most 25%
3 s.d.	At most 12%
5 s.d.	At most 4%

When the Variance is Not Enough

- Values far away from the mean are rare, but rare events may be very important
- Unexpected events with high impact can, and often do greatly influence events
- Just as the mean by itself is not enough, the variance by itself may not be enough. We need to know the shape of the distribution as well. With some kind of skewed distributions, knowing the mean and the variance will not really help.

The Long Tail I



- Large number of occurrences in the "tail" portion of the distribution of data
- This means an unusual number of observations that are much greater than the mean
- The success story of some businesses (like Amazon) has been explained by using such distributions

The Long Tail II

- The "tail" consists of sales of unusual or rare books
- Never sell many of any one of these, but there are many of these that are sold (2 copies of "The History of Tractors in the Ukraine", 1 copy of "Science Year Book 1954", 5 copies of "The Dietary Habits of the Natives of the North Congo Basin", and so on)
- The business model is selling small numbers of a very large variety of hard-to-find items customers, rather than selling large volumes of fewer easy-to-find items
- Given enough rare items, a large population of customers, and low stocking/distribution costs, the buying pattern of the population has been shown to result in a long tail distribution
- So, a business can try to exploit this and see how to make money out the tail, while others make money elsewhere

Quick Estimates of the Mean and the Spread I

1. Find the total T of N observations. Estimate the (arithmetic) mean from $m = T/N$.
   - This works very well when the data follow a symmetric bell-shaped frequency distribution (of the kind modelled by "normal" distribution)
   - A simple mathematical expression of this is $m = \frac{1}{N}\sum_i x_i$, where the observations are x_1, $x_2 \ldots x_n$
   - If we can group the data so that the observation x_1 occurs f_1 times, x_2 occurs f_2 times and so on, then the mean is calculated even easier as $m = \frac{1}{N}\sum_i x_i f_i$
   - If, instead of frequencies, you had relative frequencies (i.e. instead of f_i you had $p_i = f_i/N$), then the mean is simply the observations weighted by relative frequency. That is, $m = \sum_i x_i p_i$
2. Calculate the total T and the sum of squares Σ of N observations. The estimate of the standard deviation is $s = \sqrt{\frac{1}{N-1}\sum_i (x_i - m)^2}$

Quick Estimates of the Mean and the Spread II

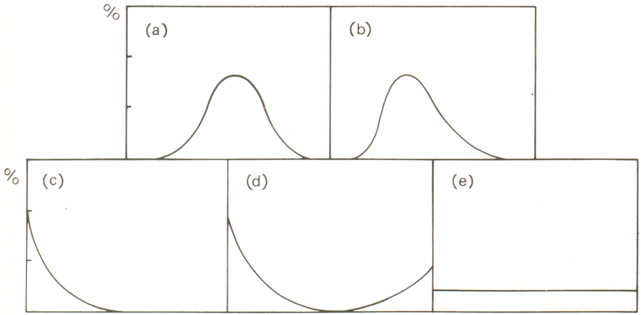
  - Again, this is a very good estimate when the data are modelled by a normal distribution
  - For grouped data, this is modified to $s = \sqrt{\frac{1}{N-1}\sum_i (x_i - m)^2 f_i}$

3. Other quick ways to compute a measure of the mean level: (a) The median: the value exceeded by 50% of the observations ($y(0.5)$); (b) The mid-quartile: the average of $y(0.25)$ and $y(0.75)$
   - Both of these are sensitive to skewed data
4. Other quick ways to compute a measure of the spread: (a) Calculate the difference between the smallest and the largest observation (this is called the *range*); (b) Calculate the mean deviation of the data from $y(0.5)$; (c) Calculate the difference between the means of the largest 5% and the smallest 5% of the data
   - Again, these are sensitive to deviations from data modelled by the normal distribution

Summarising Data

- Summarising data thus involves:
  - Presenting tables or graphs in an understandable form
  - Showing the number of data points *n*
  - Calculating the average (usually the mean)
  - Calculating the scatter (usualy the standard deviation)
  - Describing the shape of the distribution

Frequency Distributions

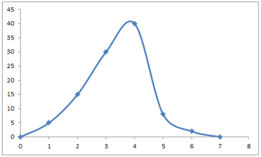


From: Ehrenberg (1986)

Frequency Distributions from Observations

- Observed readings do not come in any order
- To find the shape the observations have requires us to order them by size and count the frequency of occurence of each value.

	Observed Data							
	0	1	2	3	4	5	6	7
Freq.	0	5	15	30	40	8	2	0
	Mean = 3.3 N = 100)							

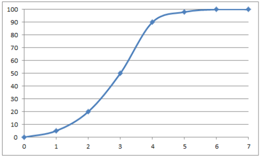


(Free-hand curve)

Cumulative Frequencies

- *Relative frequencies* are the observed frequencies as proportions of the total number of instances
- If we add the relative frequencies, from the lowest value then we get the *cumulative frequencies*. This tells us the proportion of instances less than or equal to a value

	Observed Data							
	0	1	2	3	4	5	6	7
Freq.	0	5	15	30	40	8	2	0
C.F.	0	5	20	50	90	98	100	100



(Free-hand curve)

Percentiles

- The cumulative frequency tells us the proportion of instances less than or equal to a value
- If 75% of the instances are less than or equal to 35, then 35 is called the 75th percentile
- The median is therefore the 50th percentile; and the top decile is the 90th percentile and so on
- A quartile is the value separating 25 percent of the instances. The *inter-quartile range* is the difference between the 75th and the 25th percentile

Standardised Variables

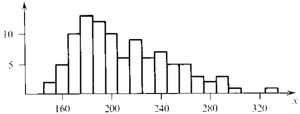
- It is often easier to "standardise" an observed reading X by:
  1. Subtracting the mean $\overline{X}$ from the reading X
  2. Dividing the result by the standard deviation s
- The standardised value Z is therefore:

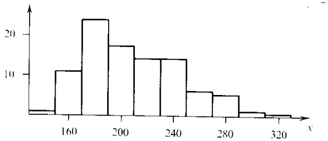
$$Z = \frac{(X - \overline{X})}{s}$$

- From a table of X values, we can calculate a table of Z values. This table of Z values will have a mean value $\overline{Z}$ of 0 and a standard deviation of 1
- Standardisation is particularly helpful when we compare different sets of observations
- They are also used to deal with theoretical models of frequency distributions

Frequency Diagrams and Histograms I

- There is a general difficulty with visualisation of frequencies

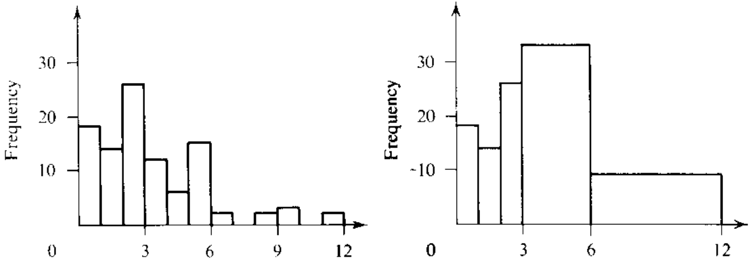




- The same data can give rise to two different looking diagrams based on the choice of interval length

Frequency Diagrams and Histograms II

- Here are frequency diagrams compiled from two different tabulations of the same data (the second uses a different interval length for some part of the data):



- Now, unless you are careful, the data look completely different

Frequency Diagrams and Histograms III

- Histograms represent frequencies corresponding to each interval not by the height of the rectangle, but by the area. That is, $Freq. = Area = IntervalWidth \times RectangleHeight$. So, the height of a rectangle represents $Freq/IntervalWidth$, or a *frequency density*
- Histograms allow us to estimate the probability distribution of a continuous variable

Where do the Data come from? (Sampling)

- For groups (populations) that are fairly homogeneous, we do not need to collect a lot of data. (We do not need to sip a cup of tea several times to decide that it is too hot.)
- For populations which have irregularities, we will need to either take measurements of the entire group, or find some way of get a good idea of the population without having to do so
- *Sampling* is a way to draw conclusions about the population without having to measure all of the population. The conclusions need not be completely accurate
- All this is possible if the sample closely resembles the population about which we are trying to draw some conclusions

What We Want From a Sampling Method

- No systematic bias, or at least no bias that we cannot account for in our calculations
- The chance of obtaining an unrepresentative sample can be calculated. (So, if this chance is high, we can choose not to draw any conclusions.)
- The chance of obtaining an unrepresentative sample decreases with the size of the sample

Simple Random Sampling

- Each element of the population is associated with a number
- Shuffle all the numbers and put them into into a hat
- Draw a sample of n numbers from the hat and get the corresponding elements of the population

Usually, there are no hats, and we will be using a computer to generate n numbers that are approximately random.

In addition, the computer will use a mathematical relationship between elements of the population and the set of numbers. Inverting this relationship using the n random numbers will then give the elements of the population.

Multi-Stage and Stratified Sampling

- It may be impractical to try and number each element of the population
- Instead, consider the population as being comprised of groups (for example, districts rather than individual streets of a town). Number the groups, and randomly select some groups in the first stage. Follow this with randomly select an individual from each of the groups selected.
- It may be helpful to partition the population into "strata". These are groups that are non-overlapping subsets of the population that together comprise the entire population. Some fixed number of individuals are then selected from each stratum.
- Stratification only makes sense if each stratum is somewhat more homogeneous thant the overall population

Probability Sampling

- In effect, numbers drawn using simple random sampling (in a single stage or more) use a uniform probability distribution over the numbers. That is, the probability of getting any number from $1 \ldots n$ from the hat is $1/n$.
- A more general form of this is to use any kind of probability distribution over $1 \ldots n$. For example, a distribution could make larger numbers are more likely than smaller numbers. This is a skewed distribution
- For example, take a 2-stage sampling procedure in which households are grouped according to size, and the probability of selecting larger households is higher. A household is selected and then an individual is selected from that household. This gives a greater chance of selecting individuals from larger households
- Once again, it is relatively straightforward to do this form of probability-based sampling using a computer

Other Kinds of Sampling

- Cluster sampling
- Weighted sampling
- Variable sampling
- Area-based sampling
- Random-walk sampling
- Opportunity sampling
- Panel sampling

Sampling Issues: Replacement

- Question: Once a number is drawn from the hat, should the number be put back into the hat or not?
- When there are many numbers in the hat, it does not really matter. But if the numbers are small, then it can make a difference (if the number is not put back in, there are only $n-1$ numbers after the first draw, $n-2$ after the second, and so on)
- In practice, sampling is often done *without* replacement, but to make calculations easier, it is assumed to have been done *with* replacement

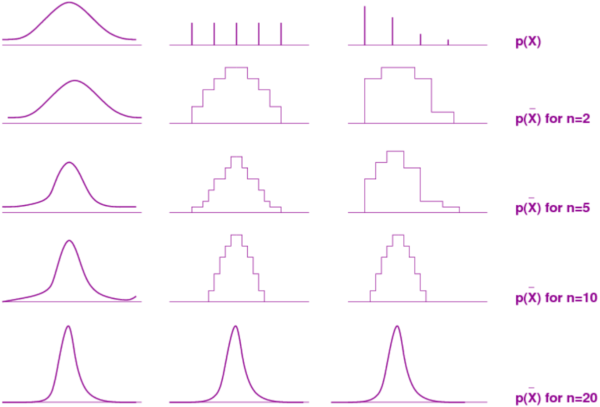
Sampling Issues: The Population

- It is not a trivial task to obtain a complete list of the population from which a sample is drawn (this list is usually called the *sampling frame*)
- Sometimes, a sample is drawn from a "population" that is unrepresentative of the population about which we want to draw some conclusions
- TV phone-in polls, "random sample of consumers", convenience samples of people walking in the street, "informed sources" are all examples of biased samples, or of cases where no explicit sampling actually took place
- They do not provide any statistical basis for generalisation

Sampling Issues: Sample Variation

- Samples provide an approximate summary of the population. There is bound to be some difference. For example, the mean of observations in a sample may be close, but nor the same as the population mean μ
- In fact, if we take more than one sample, then the means of these two samples are also probably different (and different again to the population mean)
- It therefore becomes meaningful to talk about the distribution of sample means, usually called the *sampling distribution of the mean*. One of the consequences of the Central Limit Theorem is that this distribution is approximately Normal.
- The remarkable thing is that this is so irrespective of the distribution of values of the population from which the samples were drawn, as long sample sizes are large enough

The Central Limit Theorem



The Central Limit Theorem (contd.)

- The result is a *theorem*. It assumes: (1) the observations are drawn randomly; and (2) the population distribution has a finite variance
- The result is a *limit* theorem. The probability that some fraction of the sample means fall within an interval converges to the probability that the same fraction of Normally distributed values fall within that interval
- The "central" part refers to to the fact that mean value converges to its average or central value (the population mean μ)

Sample Means Vary

Sample	Observations	Mean
1	3, 0, 0, 4, 2, 0, 2, 0, 12	2.3
2	4, 1, 1, 0, 17, 1, 0, 3, 1, 2	3.1
3	...	2.0
...	...	0.6
...	...	...
...	...	...

- The distribution of means in the last column is approximately Normal
- The mean of this distribution is the same as the mean of the population from which the samples are drawn
- That is, the means are scattered approximately symmetrically about the mean of the population of the population. This scatter or standard deviation (called the *standard error of the mean*) is a scaled-down version of the population s.d.

Approximations with a Single Sample

- The results about the distribution of the means is not very helpful, since we do not know either the population mean or its variance
- We also do not have many samples: we usually have a few, or often only one sample. What can we do in such cases?
- We can only *estimate* the parameters. For example:

$$\text{estimated std. error of the mean} = \frac{s}{\sqrt{n}}$$

- The CLT tells us that distribution of means is approximately Normal provided:
  - The sample size is about 10 or more, if frequencies in the original population are distributed Normally
  - The sample size is about 100 or more, if frequencies in the original population are distributed in a skewed manner

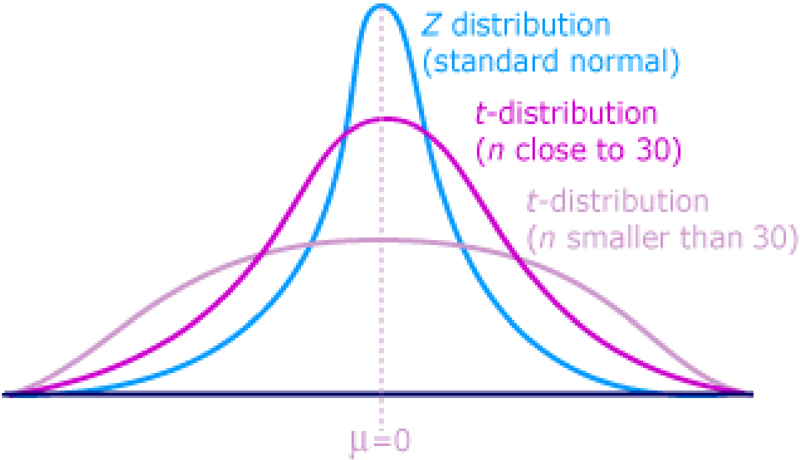
Approximations with a (Small) Single Sample

- The results about approximations to the Normal distribution of the means is not very close when sample sizes are small
- When sample sizes is small, the "*t*-distribution" gives the proportion of times different values of a specific ratio occurs in samples of that size. The ratio is like a standardised variable:

$$t = \frac{(\text{sample mean} - \text{pop. mean})}{\text{estimated std. error of mean}}$$

- Now, for each sample, we can calculate a t value. If the means of the samples is a Normal distribution then the resulting distribution of the *t*-values is a *t*-distribution
- For small samples, the *t*-values follow a *t*-distribution when the population is approximately Normal

The *t*-distribution



Estimation from a Sample

- Estimating some aspect of the population using a sample is a common task. Along with the estimate, we also want to have some idea of the accuracy of the estimat (usually expressed in terms of *confidence limits*)
- Some measures calculated from the sample are very good estimates of corresponding population values. For example, the sample mean *m* is a very good estimate of the population mean μ. But this is not always the case. For example, the range of a sample usually under-estimates the range of the population
- We will have to clarify what is meant by a "good estimate". One meaning is that an estimator is correct on average. For example, on average, the mean of a sample is a good estimator of the mean of the population
- For example, when a number of samples are drawn and the mean of each is found, then average of these means is equal to the population mean

Efficient Estimators

- As well as being correct on average, we would also like the distribution of sample values to have a low scatter
- Estimators can therefore be compared on the basis of the variance of their sample distributions

$$\text{Efficiency of V vs. W} = \frac{\textit{variance of W}}{\textit{variance of V}}$$

- If this value is greater than 1 then V is *more efficient* than W. For example:
  - When samples are drawn from a population that is approximately Normal, the distribution of sample medians has a variance of about 1.6 times the variance of the distribution of the sample means.
  - When samples are drawn from a population that has a specific power law distribution called the Laplace distribution, the distribution of sample medians has a variance of 0.5 times the variance of the distribution of sample means

The sample mean is a less efficient estimator than the median

The Bias-Variance Tradeoff I

- When comparing unbiased estimators, we would like to select the one with minimum variance (that is, the most efficient estimator)
- In general, we would be comparing estimators that have some bias and some variance
- We can combine the bias and variance of an estimator by obtaining the *mean square error* of the estimator, or MSE. This is the average value of squared deviations of an estimated value V from the true value of the parameter θ. That is:

$$\mathrm{MSE} \ = \ \mathrm{Avg.\ value\ of} (V - \theta)^2$$

The Bias-Variance Tradeoff II

- Now, it can be shown that:

$$\text{MSE} \ = \ (\text{variance}) \ + \ (\text{bias})^2$$

- So, we can re-define the efficiency of estimators:

$$\text{Efficiency of V vs. W} \ = \ \frac{\mathsf{\textit{MSE of W}}}{\mathsf{\textit{MSE of V}}}$$

- Since

$$\text{MSE} \ = \ (\text{variance}) \ + \ (\text{bias})^2$$

  The lowest possible value of MSE is 0

The Bias-Variance Tradeoff III

- In general, we may not be able to get to the ideal MSE of 0. Sampling theory tells us the minimum value of the variance of an estimator. This value is known as the *Cramer-Rao* bound. So, given an estimator with bias b, we can calculate the minimum value of the variance of the estimator using the CR bound (say, v_{min}). Then:

  $$\mathrm{MSE} \ \geq \ v_{min} + b^2$$

  The value of v_{min} depends on whether the estimator is biased or unbiased (that is $b = 0$ or $b \neq 0$)
- It is not the case that v_{min} for an unbiased ($b = 0$) estimator is less than v_{min} for a biased estimator. So, the MSE of a biased estimator can end up being lower than the MSE of an unbiased estimator

Monte-Carlo Estimation

- Sampling distributions for statistics are not often well known
- Rather than try to obtain an analytic form of the sampling distribution of a particular statistic simply obtain the distribution empirically by repeatedly drawing samples and calculating the statistic
- For example, suppose we wanted to calculate the sampling distribution of the mode. One way to do this is to find the distribution experimentally
    - Repeatedly sample 5 points and calculate their mode. The frequencies of means obtained in this way can be plotted. The mean and variance can then be obtained empirically.
- This kind of estimation is called "Monte-Carlo estimation". Often it is the only practical way to determin a sampling distribution when the mathematics is intractable

Accuracy of an Estimate (Confidence Intervals) I

- If sample sizes are large enough, then the sampling distribution of a statistic like the mean will be approximately Normal with mean μ and s.d. equal to the standard error
- That is, about 95% of the observations will lie between $2 \times \mathrm{std.err.}$ of μ. Using a single sample, this is turned around to say that we are *95% confident* that:

$$\mu \; = \; \text{sample mean} \pm 2 \times \text{std.err.}$$

- This is actually not a probability statement about μ. With a relative frequency based interpretation, probability statements are only possible about *random variables*: those that can take one of several values. The population mean has a single value

Accuracy of an Estimate (Confidence Intervals) II

- The frequency-based interpretation of a confidence limit is a bit complicated. But in practice the means of most random samples will be somewhat similar, and it is usually good enough to act as though there is a very high chance that the population mean is between $2 \times \mathrm{std.err}$ of the sample mean

Small Sample Confidence Intervals

- The 95% interval for the population mean:

$$\mu \; = \; \text{sample mean} \pm 2 \times \text{std.err.}$$

  has two difficulties: (1) *std.err.* is equal to $\sigma/\sqrt{n}$. So, to calculate this, we need to know the population's standard deviation σ; and (2) The Normal approximation is not very good for small samples

- Both these problems are "solved" using the t distribution. So, the the 95% confidence interval for small samples with unknown σ is:

$$\mu \; = \; \text{sample mean} \pm t_{95\%,n-1} \times \text{sample std.err.}$$

  where $t_{95\%,n-1}$ is a value from the t distribution. For small n this will be a bit larger than 2. The *sample std.err.* is equal to $s/\sqrt{n}$.

What Does a “Confidence Interval” Mean?

- The general approach for constructing a confidence interval for a parameter μ using a sample estimate m requires us to do something line $m \pm k \times s.e.$
- A 95% c.i. $m \pm a$ for some μ does not mean: we are 95% sure that μ lies between $m + a$ and $m - a$. It means in 5 times out of 100, the interval centred on m will not include μ
- To understand this, we must understand that estimates of the mean vary from one sample to the other. If we knew the distribution of how mean-estimates varied, then we could use that to construct the confidence interval

Significance I

- Sometimes, we have a prior hypothesis about the population. For example, female literacy is higher in smaller households. Is this really true?
- One way to find out if this is true, is to check all households. If we cannot do this, we have to take a sample. Now, suppose in the sample we took, there was no difference in female literacy levels in small and large households. Is our hypothesis refuted?
- We could repeat the exercise with a larger sample. Or, we could avoid the extra work using a *test of significance*
- The test gives us the probability that the difference between the sample value and the hypothesised value is a statistical fluke of sampling

Significance II

- If the difference is probably a fluke, then we say that our prior hypothesis cannot be ruled out. The prior hypothesis is usually called the *null hypothesis*
- If the probability of a fluke is very low, then we have reason to think that the null hypothesis is, in fact, not true. Put another way, *if* the null hypothesis was in fact true, then the probability of getting this sample of values is very small
- To get this probability, we will need to know the distribution of the sample statistic

Correlation I

- The *correlation coefficient* is a number between -1 and +1 that indicates whether a pair of variables x and y are associated or not, and whether the scatter in the association is high or low
  - High values of x are associated with high values of y *and* low values of x are associated with low values of y, and scatter is low
  - A value near 0 indicates that there is no particular association and that there is a large scatter associated with the values
  - A value close to -1 suggests an inverse association between x and y
- Only appropriate when x and y are roughly linearly associated (doesn't work well when the association is curved)

Correlation II

- The formula for computing correlation between x and y is:

$$r = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}}$$

  This is sometimes also called *Pearson's correlation coefficient*

- The terms in the denominator are simply the standard deviations of x and y. But the numerator is different. This is calculated as the average of the product of deviations from the mean:

$$\mathrm{cov}(x,y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

- What does "covariance" mean?¿
  1. Case 1: $x_i > \overline{x}$, $y_i > \overline{y}$
  2. Case 2: $x_i < \overline{x}$, $y_i < \overline{y}$
  3. Case 3: $x_i < \overline{x}$, $y_i > \overline{y}$

Correlation III

4. Case 4: $x_i > \overline{x}$, $y_i < \overline{y}$

In the first two cases, x_i and y_i vary together, both being high or low relative to their means. In the other two cases, they vary in different directions

- If the positive products dominate in the calculation of $\text{cov}(x, y)$, then the value of r will be positive. If the negative products dominate, then r will be negative. If 0 products dominate, then r will be close to 0.
- You should be able to show that:

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}}$$

- Computers generally use a short-cut formula:

$$r = \frac{\sum_i x_i y_i \;\; - \;\; n\overline{xy}}{n-1}$$

Correlation IV

- The same kinds of calculations can be done if the data were not actual values but ranks instead (i.e. ranks for the x's and the y's). This is called *Spearman's rank correlation*, but we won't do these calculations here.

What Happens If You Sample? I

- Suppose you have a sample of x, y pairs and you calculate $r = 0.3$. Is this really the case?
- Sampling theory tells us something. If: (a) the relative frequencies observed are well modelled by a special kind of mathematical function (a a “Normal” or Gaussian distribution); (b) the true correlation is 0; and (c) the number of samples is large
- Then:
  - The sampling distribution of the correlation coefficient (that is, how r varies from sample to sample) is also approximately distributed according to the Normal distribution with mean 0 and s.e. of approximately $1/\sqrt{n}$
- We can use this to calculate the (approximate) probability of obtaining the sample if the assumptions were true

What Happens If You Sample? II

- Suppose we calculate $r = 0.3$ from the sample, and that the s.e. is 0.1 say. Then if the sample came from a population with true correlation 0, this would be quite unusual (less than 1% chance)
- We would say instead that the sample was probably from a population with correlation 0.3, with a 95% confidence interval of $\pm 2 \times 0.1$

What Does Correlation Mean? I

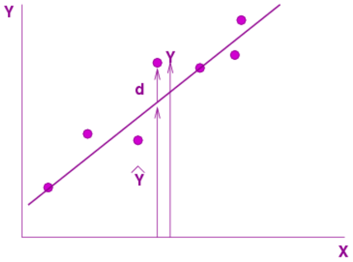
- r is a quick way of checking whether there is some linear association between x and y
- The sign of the value tells you the direction of the association
- All that the numerical value tells you is about the scatter in the data
- The correlation coefficient does not model any relationship. That is, given a particular x you cannot use the r value to calculate a y value
    - It is possible for two datasets to have the same correlation, but different relationships
    - It is possible for two datasets to have the different correlations but the same relationship
- Cannot use correlations to compare datasets. All you can derive is whether there is a positive or negative relationship between x and y

MAXIM #4: Correlation isn't causation

Regression

- Given a set of data points x_i, y_i, what is the relationship between them?
- One kind of question is to ask: are these linearly related in some manner? That is, can we draw a straight line that describes reasonably well the relationship between X and Y
- Remember, the correlation coefficient can tell us if there is a case for such a relationship
- In real life, even if such a relationship held, it will be unreasonable to expect all pairs x_i, y_i to lie precisely on a straight line. Instead, we can probably draw some reasonably well-fitting line. But which one?

Linear Relationship Between 2 Variables I



- GOAL: fit a line whose equation is of the form $\hat{Y} = a + bX$
- HOW: minimise $\sum_i d_i^2 = \sum_i (Y_i - \hat{Y})^2$ (the "least squares estimator")

Linear Relationship Between 2 Variables II

- The calculation for b is given by:

$$b \;=\; \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$$

  where $\mathrm{cov}(x,y)$ is the covariance of x and y, given by $\sum_i (x_i - \overline{x})(y_i - \overline{y})$ as before

- This can be simplified to:

$$b = \sum(xy)/\sum x^2$$

  where $x = (X_i - \overline{X})$ and $y = (Y_i - \overline{Y})$

- $a = \overline{Y} - b\overline{X}$

From Relationships to Generalisations I

- An empirical relationship—like a regression line— obtained under some conditions, cannot constitute the basis of a generalisation that holds under substantially different conditions
- For example, a a linear relationship between $\log(\mathit{weight})$ and *height* of children, obtained from data can only become generally acceptable if it is shown to hold (with approximately the same coefficients) across a range of conditions
  - The result holds irrespective of race, gender, geography, socio-economic status, time *etc.*
- When different sets of data are modelled by the same relationship, then there is a case for the relationship being *lawlike*

Using Relationships I

- Summarisation. The obvious use of a relationship is that it summarises the data from which it was obtained
- Prediction. For data drawn under the same conditions, we would expect to be able to use the relationship for prediction
  - Distinguish here between *prediction* and *extrapolation*
  - We will use the first to mean using the relationship within the operating range; and the second to mean using the relationship outside the operating range
  - When we predict, we expect to do so with high accuracy
  - When we extrapolate, a successful outcome is unexpected and suggests we might have found a lawlike relationship
  - *MAXIM #6: Empirical relationships do not hold universally*
- Understanding. Although empirical relationships do not tell us why something happens, they can form the low-level building blocks for developing a better understanding.

Probability

Probability Functions I

- In this course, we will be especially interested in functions that can be used to model what will be called the *probability* of outcomes or events
- This may be the relative frequency with which an outcome occurs, but there will be other meanings associated with the term "probability".
- A relative frequency distribution, like a histogram, can be used to give us the proportion of x's that take some specific value, or, the proportion of that x's that lie in some range.
- When used in this way, a frequency distribution can also be thought of as a *probability distribution*. Probability distributions refer to the probability than an individual observation will take a particular value (or lie in a range of values)

Probability Functions II

- Rather than a diagram like a histogram, we would like to represent these distributions using mathematical functions

Probability Distributions (contd.)

- Functions representing probability distributions have to satisfy some properties:
  1. They have "parameters", that allow us to change the shape of the distribution
  2. Irrespective of the value of its parameters, the total area under a probability distribution sums to 1, for the same reason that relative frequencies add to 100%
  3. The probability of obtaining values between say x_1 and x_2 is the area under the probability distribution between x_1 and x_2. This probability (and clearly, the area) must lie between 0 and 1

Theoretical Models (contd.)

- The functional form of a distribution can sometimes be obtained using a *probability* or *stochastic* model for what is being observed
- Here is a table of the number of boys in 100 families of of 3 children (from Ehrenberg, 1986):

	Number of Boys				Avg.
	0	1	2	3	% of Boys
% of Families	11.6	36.2	38.3	13.9	51.5

Probability Models and Probability Distributions I

	Number of Boys				Avg.
	0	1	2	3	% of Boys
% of Families	11.6	36.2	38.3	13.9	51.5

- Here are two possible explanations
  1. About 10% of the families can only have girls; and about 15% can only have boys. Of the remaining 75% about half have a slight pre-disposition to boys and the other have a slight pre-disposition to girls.
  2. All families are alike. There is a 51% chance of any baby being a boy, and the number of boys x in a family of n is as if by chance (*i.e.* the same as getting x "heads" when tossing a coin n times with a fixed probability of 51% of landing heads), and unrelated to the other children in the family
- The second explanation is a sequence of "Bernoulli trials" (a Bernoulli process), which necessarily results in a Binomial distribution of frequencies

"As if by Chance"

- Does this mean that the incidence of boys is purely by chance?
- No — physiological and biochemical reasons may underlie this, but these may be so complex and the outcomes irregular enough to be modelled by a chance process
- In statistics, "randomness" usually means a lack of predictability. That is, although the underlying process may be regular, its knowledge may not be sufficient to predict the outcome
- The principal notion to deal with this is the idea of a *random variable*. Informally, a random variable is one that takes one of several pre-defined values. Exactly which value will be taken is not known because of chance effects. A probability distribution describes how likely each value is. That is, one view of a probability distribution is as the relative frequency—in the long run—with which a random variable will take its different values.

Probability Models: Another Example I

- Suppose you conducted an experiment of tossing 6 coins $N = 100$ times and tabulated the following:

Number of Heads (i)	Freq. $f(i)$	Relative Freq. $(f(i)/N)$
0	1	1/100
1	10	10/100
2	23	23/100
3	31	31/100
4	25	25/100
5	8	8/100
6	2	2/100

  To avoid confusion, we will say that the experiment consisted of 100 trials, each consisting of a "6-coin" toss

Probability Models: Another Example II

- The last column tabulates the relative frequency with which $0, 1, 2, \ldots, 6$ heads is observed. We will call this the "probability of obtaining i heads" (where $i = 0, 1, \ldots, 6$) on any single 6-coin toss. Can we model this probability as a function of i? (Why would you want to do this anyway?)
- The answer is "almost". We do this using the Binomial model as before. In this, each 6-coin toss is a Bernoulli process (a sequence of Bernoulli trials), which results in a Binomial distribution of relative frequencies for the number of heads So, a mathematical function that should closely approximate the last column is:

  $$p(i) = \binom{6}{i} a^i b^{(6-i)}$$

  where i is the number of heads (first column), a is the (theoretical) probability of a head for a coin, b is the

Probability Models: Another Example III

(theoretical) probability of a tail on a coin—all 6 coins are taken to be identical—and $p(i)$ is the probability of getting i heads on any one 6-coin toss. Remember

$$\binom{A}{B} = \frac{A!}{B!(A-B)!}$$

- We want $p(i)$ to be close to the relative frequency (last column). Here is a comparison, assuming $a = b = 0.5$:

Number of Heads (i)	Relative Freq. (observed)	$p(i)$ (theoretical)
0	1/100	1/64
1	10/100	6/64
2	23/100	15/64
3	31/100	20/64
4	25/100	15/64
5	8/100	6/64
6	2/100	1/64

Probability Models: Another Example IV

- Notice that the $p(i)$ values add up to 1. That is, in this case $\sum_{i=1}^{6} p(i) = 1$. This will be an important requirement of any function that is used to model probabilities over an exhaustive set of outcomes. Another important requirement is that for all values of i, $0 \leq p(i) \leq 1$.
- For reasons that will become clear later, we will call such a function a *probability mass functions*
- The relative frequency column shows that outcomes are uneven. That is, there are fewer trials 0 heads appears to be less likely than say a sequence with 3 heads
- One of the uses of a probability model (the probability function) is that it will allow us to calculate beforehand what value we can expect for n (the number of heads), without actually having to do the experiment

Probability Models: Another Example V

- For example, the probability model says that on N trials of 6 (fair) coins you should expect 0 heads to occur $N/64$ times, 1 head to occur $6N/64$ times, and so on. So, the (theoretical) mean number of heads per trial will be:

  $$\frac{1}{N} \times (0 \times N/64 + 1 \times 6N/64 + \cdots + 6 \times N/64)$$

  This mean works out to be: ...
- The mean from the observations is: ...
- Recall from the earlier lectures on introductory statistics, the mean of outcomes $x_1, x_2, \ldots$ with frequencies $f(x_1), f(x_2), \ldots$ can be calculated using:

  $$\text{Mean } m \ = \ \frac{1}{N} \sum_k x_k f(x_k)$$

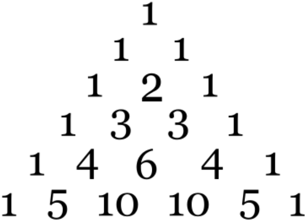
  where $N = \sum_k f(x_k)$

Aside: The Binomial Expansion I

- The term *binomial* comes from algebra, where it is used to refer to the sum of two terms like $a + b$ or $q + p$
- Take, for example, the sum $q + p$ multiplied by itself 6 times. That is $(q + p)^6$. After the first multiplication, we get $q^2 + 2pq + p^2$. Multiplying again by $(q + p)$, we get $q^3 + 3pq^2 + 3p^2q + p^3$, and so on
- If you look at the coefficients of the expansion, you will see that they look like this: $(1, 2, 1)$; $(1, 3, 3, 1)$; $(1, 4, 6, 4, 1)$ and so on. These are

Aside: The Binomial Expansion II

rows of a combinatorial structure usually called *Pascal*'s triangle



1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

- The line after the last one above is $(1, 6, 15, 20, 15, 6, 1)$, which are exactly the the theoretical frequencies of the number of heads, and represent the coefficients of q^6, pq^5, p^2q^4 and so on

Aside: The Binomial Expansion III

- If we take q to denote the relative frequency of getting a tail, then q^6 denotes the relative frequency of getting 6 tails (i.e. 0 heads). This outcome happens just once. Similarly, pq^5 denotes the relative frequency of getting 1 head and 5 tails. There are 6 times this outcome happens; and so on
- What happens if $q = 0.8$ and $p = 0.2$?

Number of Heads	Number of Cases	Rel. Freq.	Cum. Rel. Freq.
0	1	$(0.8)^6 = 0.262$	0.262
1	6	$6(0.8)^5(0.2) = 0.393$	0.655
2	15	$15(0.8)^4(0.2)^2 = 0.246$	0.901
3	20	$20(0.8)^3(0.2)^3 = 0.082$	0.983
4	15	$15(0.8)^2(0.2)^4 = 0.015$	0.998
5	6	$6(0.8)(0.2)^5 = 0.002$	≈ 1
6	1	$(0.2)^6 \approx 0$	≈ 1

Aside: The Binomial Expansion IV

- So, with $p = q = 1/2$, the chance of getting 5 heads was $6/64$, or about 1 in 10. Now it is 2 in a 1000.
- From now on, we will call the relative frequency of an outcome as the *probability* of that outcome
- A trial (like a coin toss) in which there are just two outcomes ("success" and "failure"), each with a fixed probability p, is called a *Bernoulli* trial. We model the outcome by a random variable, $\mathit{Bernoulli}(p)$ that takes the value 1 with probability p and 0 with probability $1 - p$
- A sequence of Bernoulli trials results in a Binomial system or Binomial experiment. That is, if n is a positive integer and $X_1, X_2, \ldots X_n$ are each independent $\mathit{Bernoulli}(p)$ random variables, then the value $X_1 + X_2 + \cdots + X_n$ (the total number of successes in n Bernoulli trials) is a random variable $\mathit{Binomial}(n, p)$

The Binomial Probability Model I

- If the probability of success in a Bernoulli trial is p and the probability of failure is q, then the probability of i successes in a sequence of n trials is given by $B(n,i)p^i q^{(n-i)}$, where $B(n,i)$ is some number that we have so far obtained using Pascal's triangle
- It should be easy to see that $B(n,i)$ is simply $\binom{n}{i}$
- Thus, the probability of i successes and $(n-i)$ failures is:

$$p_i = \binom{n}{i} p^i q^{(n-i)}$$

- It can be shown that $\binom{n}{i} = \frac{n!}{i!(n-i)!}$. So:

$$p_i = \frac{n!}{i!(n-i)!} p^i q^{(n-i)}$$

Mean and Spread of the Probability Model I

- According to the probability model we have used, the average number of heads per toss of 6 coins, was:

$$\frac{1}{N} \times (0 \times N/64 + 1 \times 6N/64 + \cdots + 6 \times N/64)$$

- This "theoretical mean" is called the *expected value* (usually denoted μ). It is easy to see that the expected value is simply a weighted average that multiplies each outcome by its (theoretical) probability. That is, the mean for the probability model over a set of outcomes $x_1, x_2, \ldots$ with probabilities $p(x_1), p(x_2), \ldots$ is:

$$\mu \; = \; E(X) \; = \sum_k x_k p(x_k)$$

Mean and Spread of the Probability Model II

- Compare this with the mean of observations of outcomes $x_1, x_2, \ldots$ with frequencies $f(x_1), f(x_2), \ldots$:

  $$m \ = \ \frac{1}{N}\sum_k x_k f(x_k) \ = \ \sum_k x_k f(x_k)/N$$

  $f(x_k)/N$ is the relative frequency observed for outcome x_k
- So, in a probability model, we replace the relative frequency of and outcome by the probability, to get the theoretical mean of the model. This applies in general for any probability model.

Mean and Spread of the Probability Model III

- Using the calculations of observed variance:

$$s^2 \;=\; \frac{1}{N}\sum_k (x_k - m)^2 f(x_k) \;=\; \sum_k (x_k - m)^2 f(x_k)/N$$

  we are now able to write down directly the variance of any probability model:

$$\sigma^2 = \sum_k (x_k - \mu)^2 p(x_k)$$

- It is not hard to show that

$$\sigma^2 \;=\; E(X^2) - (E(X))^2$$

Continuous Probability Distributions I

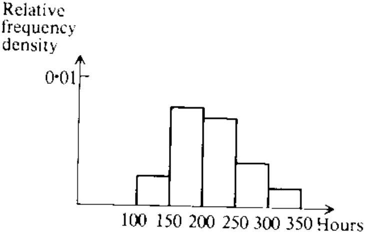
- So far, we have looked mostly at outcomes that have been *discrete*, like the number of heads (this is the same as saying we have been looking at discrete random variables)
- As a result, we have been able to tabulate these values and their relative frequencies. We have been able to think of constructing functions that approximate the relative frequency of an outcome
- For many cases, discrete events may not be identifiable. For example, suppose we wanted to answer questions on the life of a car battery. Battery-life is a continuous variable, and it in fact only makes sense to ask questions like: "what is the probability that the battery life is greater than (or less than) ...?" (That is, it would not really make sense to ask: "what is the probability that the battery-life is exactly 240 hours?")

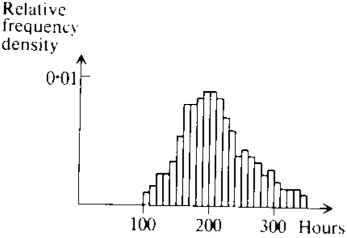
Continuous Probability Distributions II

- To model such probabilities we will have to use a continuous function However, the function will not be a probability mass function (i.e. one that computed the probability of a particular value—like 240 hours), but a probability *density* function (p.d.f.)
- We will have to go back to histograms

Probability Density Functions I

- Recall that histograms were frequency density diagrams. Here are two histograms for battery-life, using different interval widths





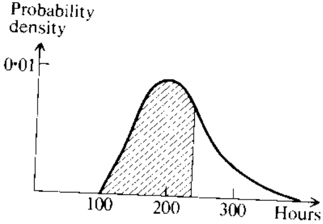
- The small difference to the histograms we looked at before is that the Y-axis is now relative frequency density (that is, the area of a rectangle now represents the proportion of instances that lie in the rectangle)

Probability Density Functions II

  - Now the proportion of instances (= relative frequency) with battery life less than 240 can be calculated by summing up the areas of rectangles to the left of 240 in any one of the two histograms
- The probability density function (p.d.f.) is a curve that approximates this relative frequency density. For any interval, it the area under the curve gives the probability of a data

Probability Density Functions III

instance falling in that interval



- As in the discrete case, we will have two requirements on on any p.d.f. $\phi(x)$:
    1. $\phi(x) \geq 0$ for all values of x;
    2. $\int_{-\infty}^{+\infty} \phi(x) = 1$

Mean and Spread (Continuous Models) I

- It is important that when we fit a continuous probability model to data, we have approximately the same mean and spread
- We saw earlier how we can obtain the mean and spread of a discrete probability function:

$$\mu = \sum_i x_i p(x_i) \qquad \sigma^2 = \sum_i (x_i - \mu)^2 p(x_i)$$

- The natural extension to continuous models is:

$$\mu = \int x\phi(x)dx \qquad \sigma^2 = \int (x - \mu)^2 \phi(x)dx$$

(where the integration is over the domain of the p.d.f.)

Modelling Evenly Scattered Frequency Distributions I

- The (continuous) uniform distribution is a family of curves such that is used to model relative frequencies that are approximately the same in any interval.
- The distribution has two parameters, a and b, which are its minimum and maximum values. The distribution is often abbreviated $U(a, b)$

$$\phi(x) = \begin{cases} \frac{1}{b-a} & \mathrm{if}(x \in [a, b]) \\ 0 & \mathrm{otherwise} \end{cases}$$

Modelling Evenly Scattered Frequency Distributions II

- The mean and spread of this deviation can be easily shown to be:

$$\text{Mean} \;=\; E(X) \;=\; \int_{-\infty}^{+\infty} x\phi(x)dx$$

That is,

$$\begin{aligned}\text{Mean} &= \int_a^b \frac{x}{b-a}dx \\ &= \frac{1}{2}\frac{(b^2-a^2)}{b-a} \\ &= \frac{b+a}{2}\end{aligned}$$

Modelling Evenly Scattered Frequency Distributions III

- Similarly, for the variance:

  $$\text{Variance} \;=\; E(X^2) - (E(X))^2$$

  If you do the calculations correctly, you should find:

  $$\text{Variance} \;=\; \frac{(b-a)^2}{12}$$

- The uniform distribution has another important role: it can be used to *simulate* many other distributions. We will see how to do this later.

Modelling Symmetric, Bell-shaped Frequency Distributions

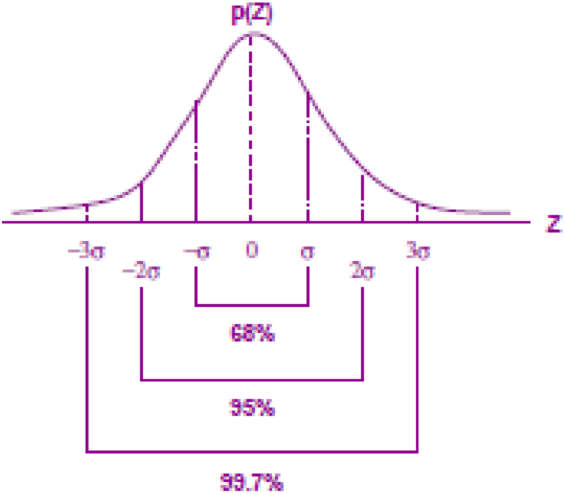
- The Gaussian or Normal distribution has 2 parameters: the mean μ and standard deviation σ. In this distribution the frequency with which a value x occurs is:

  $$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{(x-\mu)^2/2\sigma^2}$$

  This formula is hardly ever needed in practice. What is more useful is to know that:

  - The frequency distribution is symmetric with a bell-shape;
  - About 70% of the x values lie within 1 s.d. of the mean μ;
  - 95% of the observations lie within 2σ of μ; and
  - Nearly all (99.7%) of the observations lie within 3σ of μ
- If this theoretical distribution is used to fit observed data, then these properties must hold (at least approximately: a perfect fit will rarely happen)

The Standard Normal Distribution



Mean and Spread of the Standard Normal Distribution I

- The calculation of the mean uses the expression for expected values:

$$\text{Mean} \;=\; E(X) = \int_{-\infty}^{+\infty} x\phi(x)dx$$

Ignoring the constant for the moment,

$$\begin{aligned}
\text{Mean} &= \int_{-\infty}^{+\infty} xe^{-x^2/2}dx \\
&= \int_{-\infty}^{0} xe^{-x^2/2}dx + \int_{0}^{+\infty} xe^{-x^2/2}dx \qquad (\text{let } t = x^2/2) \\
&= \int_{\infty}^{0} e^{-t}dt + \int_{0}^{\infty} e^{-t}dt \\
&= -\int_{0}^{\infty} e^{-t}dt + \int_{0}^{\infty} e^{-t}dt \\
&= 0
\end{aligned}$$

Mean and Spread of the Standard Normal Distribution II

- Similarly, for the variance:

$$\text{Variance} \;=\; E(X^2) - (E(X))^2$$

- Now

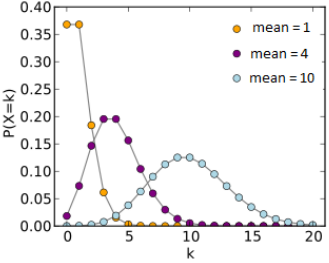
$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx \\ &= \int_{-\infty}^{0} x^2 e^{-x^2/2} dx + \int_{0}^{\infty} x^2 e^{-x^2/2} dx \end{aligned}$$

  With $u = x$ and $v = e^{-x^2/2}$, we get $dv = -xe^{-x^2/2} dx$. So, each of the integrals above is of the form $\int udv$, which can be evaluated by by parts as $\int udv = uv - \int vdu$.

- If you get all the steps right, you will find Variance $= 1$

Modelling Skewed Frequency Distributions

- The Normal distribution is a *continuous* distribution. That is, X values can be any number. Data which come as counts are usually quite skewed and not well described by a Normal distribution
- The Poisson distribution is a skewed distribution that is often used to describe data like these:



(From: the Wikipedia entry on Poisson distributions)

The Poisson Distribution

- The Poisson distribution has a single parameter: the mean μ. The distribution also has the property that both mean and variance are equal. That is:

$$\text{mean} \ = \ \text{variance} \ = \ \mu$$

- The proportion of instances that take the value x is given by:

$$p(x) = \frac{\mu^x e^{-\mu}}{x!}$$

- In practice, if the data are in the form of counts (or lengths, numbers and so on) then it may be possible to use a Poisson distribution to describe them. The first thing to do is to check if the mean and variance of the data are approximately equal.

Mean and Spread of the Poisson Distribution I

- The expected value of the Poisson distribution is calculated as usual by:

$$E(X) \ = \ \sum_{x=0}^{\infty} xp(x)$$

Mean and Spread of the Poisson Distribution II

- That is:

$$\begin{aligned}
\text{Mean} &= \sum_{x=0}^{\infty} x p(x) \\
&= \sum_{x=1}^{\infty} x \frac{\mu^x e^{-\mu}}{x!} \\
&= \sum_{y=0}^{\infty} (y+1) \frac{e^{-\mu} \mu^{(y+1)}}{(y+1) y!} \\
&= \mu \sum_{y=0}^{\infty} p(y) \\
&= \mu
\end{aligned}$$

- Similarly, find $E(X^2) = \mu^2 + \mu$, and Variance $= E(X^2) - (E(X))^2 = \mu$

Skewed Data: The Binomial Distribution

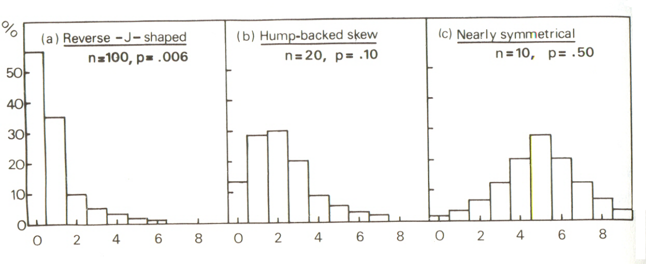
- The Poisson distribution simply counts the occurrence of a value. When observations count both occurrence and non-occurrence of a value, then we need either a Binomial (or its extension, the Multinomial distribution)
- Observed data for modelling with a Binomial distribution come in sets with a fixed number of observations n. In each such set of n-observations, we have counts of the occurrence (and therefore, the non-occurrence) of a specific characteristic. For example, the number of faulty TVs in boxes of 100 TVs of a particular manufacturer; the number of boys in families with 5 children and so on.
- From this data, we can obtain the proportion of datasets, each of size n, in which there were 0, 1, 2,. . . ,n observations with the characteristic. For example, the proportion of 5-children families with 0 boys, 1 boy, 2 boys and so on. This proportion can be modelled by the Binomial distribution.

The Binomial Distribution (contd.)

- There are 2 parameters in Binomial distribution: n, the size of data set; and p, the proportion having the specific characteristic. The proportion p_k of sets of size n that have k observations with the characteristic is then:

$$p_k = \frac{n!}{k!(n-k)!} p^k (1-p)^k$$

- Different values of n and p allow us to describe different kinds of datasets:



From: Ehrenberg (1986)

Molecules in a Cell I

- In a cell, important molecules (like an enzyme) may not be present in large numbers
- At any time instant, a molecule in a cell may: (a) stay in the cell; or (b) leave the cell. It may be possible to collect data of the following kind:

Molecule	Time				
	1	2	3	4	...
1	In	In	In	Out	...
2	Out	Out	In	In	...
3	In	In	In	In	...
...	...	...	...	...	...
...	...	...	...	...	...
10	In	In	Out	Out	Out

  From these, it will be possible to build a probabilistic model of finding a molecule inside a cell

Molecules in a Cell II

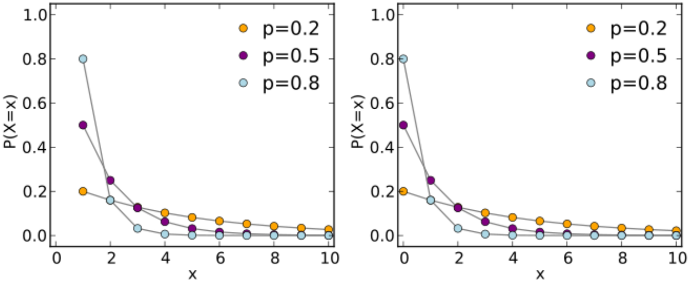
- If there are 50 molecules in a cell that randomly enter or leave the cell, and that the probability of any one molecule being within the cell is about $1/3$ in the steady state. How many molecules would you expect to be within the cell in the long run?
- Is it biologically justified to use the Binomial probability model for this?
  - All molecules are identical
  - All molecules are independent

The Geometric Distribution I

- Suppose you are interested not in the number of molecules left in a cell, but at the time instant at which a molecule leaves the cell
  - The process we want to model now is the number Bernoulli trials before a first "success" is seen
- If each independent trial has probability of success p and probability of failure q then the probability that we will see x failures before getting a success is:

$$P(X = x) \;=\; pq^x$$

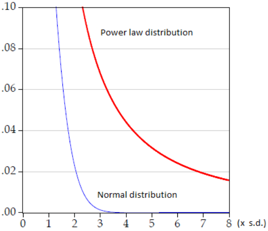
The Geometric Distribution II



The mean or expected value of number of failures before seeing a success is $E(X) = q/p$

More Skewed Data: Power Law Distributions

- For the skewed distributions described so far, the frequency of values far away from the mean is usually very small
- Many real problems are not well-described by this. For example, the cities having a certain population size; the energy of earthquakes; the frequencies of words in any language; changes in stock prices; distribution of incomes, and so on. All these distributions have "fatter tails" than the Normal distribution
- For example, oil prices jumped in 1973 to a point that was 37 × s.d. from the mean



Power Law Distributions (contd.)

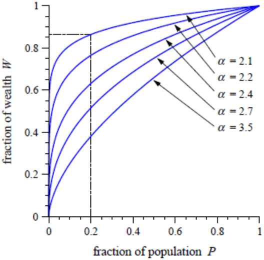
- Power-law distributions have single parameter α. The frequency with which a value x occurs is proportional to

  $$x^{-\alpha}$$

  The smaller the value of α, the "fatter" the tail.
- Usually, the mean of the distribution is far to the right. But, when $\alpha < 2$), the mean has no finite value. And when $\alpha < 3$ the variance is not finite.
- If a distribution follows a power law with parameter α then a plot of log(frequency) against log(x) will be a straight line
- So, a simple check of whether observed frequencies follow a power law is to check whether log(observed) vs. log(x) is approximately a straight line

An example: the “80/20” rule

- In 1906, the Italian engineer, sociologist and economist Vilfredo Pareto discovered that approximately 80% of the wealth in Italy was owned by 20% of the population
- This can be described by a power-law distribution with an α of about 2.1:



From: M.E.J. Newman, (2005) “Power laws, Pareto distributions and Zipf’s law”, Contemp. Phys. 46:5, pp.323–351

Modelling U-shaped and Uniform Frequency Distributions

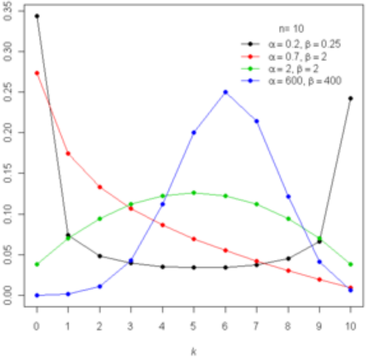
The Beta-Binomial Distribution

- The Binomial distribution requires that the proportion p having a particular characteristic is always constant. For example, the incidence of boys in families of 3 children may not be the same from one family to the next
- The Beta-Binomial Distribution allows for this proportion to change. It has three parameters: n, the size of the sample; α, related to the overall average proportion of the characteristic; and β, related to the variability of this proportion
- The formula for the proportion k having the characteristic in data-sets of size n can be calculated by a formula that is complicated, but the distribution has a mean value:

$$\frac{n\alpha}{(\alpha + \beta)}$$

The Beta-Binomial Distribution (contd.)

- For $\alpha = 1$ and $\beta = 1$, the Beta-Binomial distribution is a uniform distribution, with each value having the same frequency. Other values of α and β result in a number of other shapes:



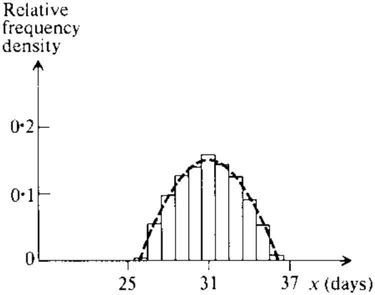
Make Your Own I

- Here is a table of frequency values observed for some event:

x days	Frequency
$25.5 \leq x < 26.5$	4
$26.5 \leq x < 27.5$	55
$27.5 \leq x < 28.5$	99
$28.5 \leq x < 29.5$	127
$29.5 \leq x < 30.5$	140
$30.5 \leq x < 31.5$	158
$31.5 \leq x < 32.5$	142
$32.5 \leq x < 33.5$	125
$33.5 \leq x < 34.5$	90
$34.5 \leq x < 35.5$	52
$35.5 \leq x < 36.5$	8

Make Your Own II

- Here is a histogram of relative frequency density and possible theoretical probability density function



- How do you fit the correct model? (The mean is $m = 39.98$ and spread is $s^2 = 4.94$.)
- Let us fit a parabolic density function $\phi(x) = b^2 - a^2x^2$ with $\mu = 31$ and $\sigma^2 = 5$

Make Your Own III

- It is easier to work with a shifted function that has the origin at $x = 31$. The new density function $\phi(x')$ will therefore have a mean of 0
- We want the total area under the function to be 1. That is

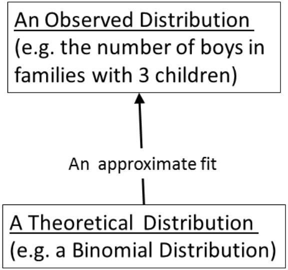
$$\int_{-b/a}^{b/a} (b^2 - a^2{x'}^2 dx' = \frac{4b^3}{3a} = 1$$

- We also want $\sigma^2 = 5$. So:

$$\sigma^2 = \int_{-b/a}^{b/a} x'2\phi(x')dx' - \mu^2 = \frac{4b^5}{15a^3} = 5$$

- Solving simulataneously, we get $b^2 = 0.15$ and $a^2 = 0.006$, giving $\phi(x') = 0.15 - 0.006{x'}^2$

Summary



An Observed Distribution
(e.g. the number of boys in families with 3 children)

An approximate fit

A Theoretical Distribution
(e.g. a Binomial Distribution)

Probability Models of Frequency Distributions

- Can we say anything more than “the theoretical distribution fits the observed values well”?
- For example, under what (mathematical) conditions is a theoretical distribution an inevitable consequence? Could these conditions provide some insight into the mechanisms that could have generated the observed data?
- Probability models or processes are mathematical explanations for theoretical distributions. They can be used to provide explanations for observations as outcomes of chance.
- This does not necessarily mean that the underlying mechanism is actually on based on chance. It simply says that it is enough to think of the underlying mechanism as operating *as if* by chance

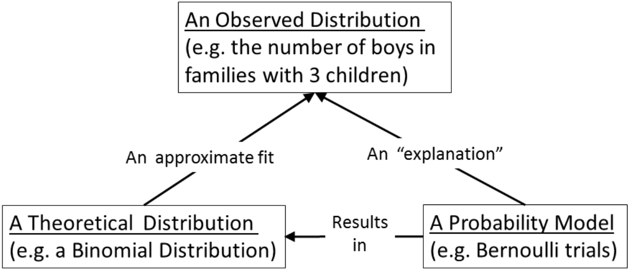
Probability Theory and Statistical Computing I

- Probability theory provides the mathematical foundations for statistics
  - One practical example of the use of probability theory in statistics is the use of probability models.
  - Another example is the derivation of results that are relevant to sampling. Two prominent examples are the ‘ ‘Law of Large Numbers” and the “Central Limit Theorem”
  - Probability theory also provides ways of calculating bounds on the relative frequency of outcomes, irrespective of the underlying distribution. An example of such a bound is given by "Chebyshev's inequality”
- What about events that do not occur repeatedly? *Subjective probabilities* are used to deal with once-off events.

> *The Miracle of the Sun was an event on 13 October 1917 in which 30,000 to 100,000 people, who were gathered near Fatima, Portugal, claimed to have an witnessed extraordinary solar event as a sign from The Virgin Mary. According to many witnesses, after a period of rain, the dark clouds broke and the sun appeared as an opaque, spinning disc in the sky. It was significantly duller than normal, and cast multicolored lights across the landscape. The sun then careened towards the earth in a zigzag pattern.*

What is the probability of such an event? Relative frequency estimates are clearly not available. Bayesian probability theory deals with answering questions like these

Summary (contd.)



An Observed Distribution (e.g. the number of boys in families with 3 children)

An approximate fit

An "explanation"

A Theoretical Distribution (e.g. a Binomial Distribution)

Results in

A Probability Model (e.g. Bernoulli trials)

Fitting Models to Data I

- When using a known theoretical distribution for the population, it is sometimes possible to find the value for a parameter that maximises the probability of obtaining the sample from that distribution. This is called the *maximum likelihood estimate*
- For example, if we were using a Poisson distribution for the population, then the proportion p_k of instances that take the value k is given by:

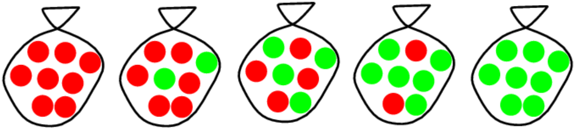
$$p_k \ = \ p(X = k) \ = \ \frac{\mu^k e^{-\mu}}{k!}$$

Fitting Models to Data II

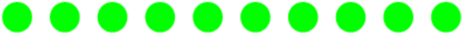
- Now suppose we have observed the values $15, 8, 13$, then we can determing the estimate of μ that maximises the probability of observing the values $15, 8, 13$. This usually means finding an expression of this probability in terms of an equation with μ as unknown; and determining the value of μ that would maximise the value of that expression

Example: The Maximum Likelihood Principle I

- Suppose there are five kinds of bags of lollies from Russell and Norvig):
  1. 10% are h_1: 100% cherry lollies
  2. 20% are h_2: 75% cherry lollies + 25% lime lollies
  3. 40% are h_3: 50% cherry lollies + 50% lime lollies
  4. 20% are h_4: 25% cherry lollies + 75% lime lollie
  5. 10% are h_5: 100% lime lollies



- Then we observe lollies drawn from some bag:



Example: The Maximum Likelihood Principle II

- What kind of bag is it? What flavour will the next lolly be?
- To answer these questions, we will first have to fit a model to the data

The Maximum Likelihood Principle (contd.) I

- Bags have a fraction θ of cherry lollies
- We are therefore dealing with binomial models (cherry vs lime lollies) in which we do not know θ. We will take this set of models to be characterised by the *parameter* θ
- Now we unwrap N lollies, and find c and $N - c$ limes. We will have to assume that these are i.i.d. (independent, identically distributed) observations
- What can we say about the probability of observed data, using the binomial distribution as our theoretical model. This is:

$$\text{Prob}(c \text{ cherries and } (N-c) \text{ limes } \propto \theta^c(1-\theta)^{(N-c)}$$

- Question: For what value of θ will this probability be highest?

The Maximum Likelihood Principle (contd.) II

- Ans: Find maximum by differentiating and setting first differential to 0. Actually easier to differentiate log(P) and set that to 0:

  $$\log(P) \; = \; L(P) \; = \; c\log\theta \; + \; (N-c)\log(1-\theta)$$

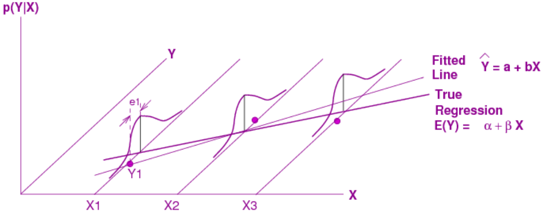
  Differentiating w.r.t. θ and setting this to zero:

  $$\frac{dL(P)}{d\theta} \; = \; \frac{c}{\theta} - \frac{N-c}{1-\theta} \; = \; 0$$

  which gives $\theta = c/N$
- This is the "Maximum Likelihood Estimate" for θ ($L(P)$ is called the likelihood function)
  (Seems sensible, but causes problems with 0 counts! But more on that later.)

Example: The linear Gaussian model I

- Recall the regression model:



- The probability model being assumed is:

$$Y_i = \alpha + \beta X_i + e_i$$

where e_i are distributed with mean 0 and variance σ^2. In addition, we are further assuming that the frequency distribution of the e_i can be approximated using a Gaussian distribution

Example: The linear Gaussian model II

- That is, we are assuming that $P(Y_i|X_i)$ is a Gaussian distribution with mean $\alpha + \beta X_i$ and variance σ^2:

$$P(Y_i|X_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - f(X_i))^2}{2\sigma^2}}$$

  (where $f(X_i) = \alpha + \beta X_i$)

- Assume we are given a set of points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$. Then the probability of obtaining these points is:

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_1{}^n (y_i - f(x_i))^2}$$

- This is the likelihood function. Maximising this, will require minimising $\sum_1{}^n (y_i - f(x_i))^2$, which is the same as finding the least squares estimate

Example: The linear Gaussian model III

- So, the least squares estimators for the regression line are the same as the maximum likelihood estimators for that linear Gaussian model (with i.i.d. data, and fixed variance)

Are Maximum Likelihood Estimators best?

- The MLE of the population mean is the sample mean. The sample mean is statistically unbiased, so the ML principle results in an unbiased estimate of the population mean
- However, the MLE of the variance is not unbiased (that is, the ML estimator is biased). So, it is not always the case that the ML principle results in an estimate with 0 bias. So, what can we say about ML estimators?
- As the sample size gets large, the variance of the MLE tends to the CR bound v_{min}. So, for all unbiased estimators (that is, all estimators that have $b = 0$), the MLE will have the lowest MSE (for large samples)

Limit Theorems and Bounds

- To understand the application of probability theory to describe the limiting behaviour of samples of data requires us to first understand random variables.
- We will look at this in forthcoming lectures. But it is nevertheless helpful to look ahead at this point to understand some of the useful theorems and bounds that can be obtained from the use of probability
  - The Central Limit Theorem
  - The Law of Large Numbers
  - Some bounds on the probability of obtaining samples with unusually high or low values

The Central Limit Theorem I

- Suppose we had n independent samples drawn from some population, which has some well-defined mean μ and variance σ^2
- Then, in the form we have seen this before, we said that, for large n, the (arithmetic) mean of the n data points is approximately symmetric and bell-shaped centred around μ and with a spread of σ^2/n. That is, we can model it with a normal (or Gaussian) distribution with mean μ and variance σ^2/n
- In general, the theorem actually applies not just to the mean, but to the sum of the data points. That is, for large n, the sum is well-modelled by a normal distribution with mean $n\mu$ and variance σ^2. The result on the mean is therefore a corollary of this more general statement of the theorem.
    - We will not look at the proof of the theorem ay this point

The Central Limit Theorem II

  - The theorem even holds if the n observations are from different populations, provided no one observation dominates the sum
  - It will not hold for power-law distributions which have infinite means and variances (see "Expectation Calculations" for these)

-

The Law of Large Numbers I

- The difference between the probability of some event and the relative frequency with which it occurs necessarily approaches 0
- This is what underlies the justification for repeated collection of observational data to measure some quantity
- Conversely, the law provides the theoretical support for using probabilities as guide to deal with randomness in the real-world
- But it does not justify the gambler's fallacy: a win (or a head) is due after a long sequence of losses (or tails)
- Here is what the theorem means:
    1. Take a sample of size n

The Law of Large Numbers II

2. See if the sample mean is "close" to the real mean. That is, for some ϵ, check if the following occurs:

$$|\mathrm{Mean}\ -\ \mu| < \epsilon$$

3. Ifyou perform this experiment many times, then you will find that this occurs nearly all the time. That is:

$$P\left(|\mathrm{Mean}\ -\ \mu|\right) < \epsilon right)\ \rightarrow\ 1$$

- An example is the relative frequency of *Heads* converging to the true probability p as the sample size n increases

Markov's Inequality I

- Suppose you are observing values X, and want to know how likely is it that X will be very large
  - If you knew the relative frequencies of the X's was well modelled by, say, a normal distribution with some μ and σ^2, you could provide a very precise answer to this
  - But what if you did not have a good theoretical distribution to model the data; or you did not know what the distribution was?
- Under some circumstances, Markov's inequality gives an upper bound on the probability of an unusually large value:
  - Suppose you did not know a theoretical distribution for the values, but you knew the mean μ of the distribution
  - Suppose all X values were known to be non-negative
- Then Markov's inequality states:

$$P(X \geq k) \leq \frac{\mu}{k} \qquad (k > 0)$$

Markov's Inequality II

- Or, in terms of expectations:

$$P(X \geq k) \leq \frac{E(X)}{k} \qquad\qquad (k > 0)$$

- Although a proof of Markov's inequality should really be done after you have had more experience with random variables, we can use what you know already:

$$\begin{aligned} E(X) &= \int_0^\infty xf(x)dx \\ &\geq \int_k^\infty xf(x)dx \\ &\geq \int_k^\infty kf(x)dx \\ &\geq kP(X \geq k) \end{aligned}$$

Markov's Inequality III

from which the result follows

Chebyshev's Inequality I

- Suppoise you want to know how likely X will be from the mean. Again, suppose we know very little about X other than it is well-modelled by a theoretical distribution with mean μ variance σ^2. We want to know the value of

$$P(|X - \mu| \geq k) \qquad\qquad (k > 0)$$

- We cannot find this value when we do not know the precise functional form of the theoretical distribution. But we can find an upper bound on the value, using Markov's inequality:

$$\begin{aligned} P(|X - \mu| \geq k) &= P((X - \mu)^2 \geq k^2) \\ &\leq \frac{E[(X - \mu)^2]}{k^2} \quad \text{Markov} \\ &\leq \frac{Var(X)}{k^2} \end{aligned}$$

Chebyshev's Inequality II

- A variant of this is:

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

  You can show that this follows in the same way as the other version
- Clearly, the bound is only useful if $k > 1$, since for $0 < k \leq 1$, the bound is trivially equal to 1

Chernoff Bounds I

- Chebyshev's inequality follows from a more general feature of Markov's inequality, namely: we can substitute any positive function f, such that:

  $$P(f(X) \geq f(k)) \leq \frac{E(f(X))}{f(k)}$$

  In order to obtain Chebyshev's inequality, we use a function $f(X) = X^2$.
- In general, it is possible to obtain tighter bounds using other kinds of functions that grow even faster than X^2. One example is the use of an exponential, which results in Chernoff bounds

Chernoff Bounds II

- Suppose X is the sum of n independent observations, each modelled by a theoretical distribution with mean $p_1, p_2, \ldots, p_n$. Then, the expected value of the sum μ is clearly $\sum_i p_i$. Using an exponential function in Markov's inequality, it is possible to show that

$$P(X \geq (1+\delta)\mu) \leq e^{\frac{-\delta^2\mu}{3}} \quad (0 < \delta < 1)$$

$$P(X \leq (1+\delta)\mu) \leq e^{\frac{-\delta^2\mu}{2}} \quad (0 < \delta < 1)$$

- More general Chernoff bounds that follow from these are:

$$P(X \geq (1+\delta)\mu) \leq e^{\frac{-\delta^2\mu}{2+\delta}} \quad (0 \leq \delta)$$

$$P(X \leq (1+\delta)\mu) \leq e^{\frac{-\delta^2\mu}{2+\delta}} \quad (0 \leq \delta)$$

Expectation Calculations I

1. The probability mass function of a discrete r.v. is as follows:

$$p(X = x) = \begin{cases} 1/3 & x = -1, 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

   What is $\mu_X = E(X)$?

Expectation Calculations II

2. You are told $Var(X) = E[(X - \mu_X)^2]$. What is $Var(X)$ for the r.v. in the above?
3. Repeat the calculations for the following mass function:

$$p(X = x) = \begin{cases} 1/3 & x = -2, 0, 2 \\ 0 & \text{otherwise} \end{cases}$$

   Why does the variance increase?

Expectation Calculations III

4. Let X be the random variable denoting the number of dots that come up on the throw of a six-sided die. What is $E(X)$? (Are store-bought dice uniform?)

Expectation Calculations IV

5. Let X be a random variable denoting the number of successes in n i.i.d. Bernoulli trials, each with probability p of success. What is $E(X)$?

Expectation Calculations V

6. Let X be an exponential random variable with pdf $f(X = x) = \lambda e^{-\lambda x}$ $(x > 0)$. What is $E(X)$? What is $E(X^2)$? Recall: integration by parts:

$$\int u dv \ = \ uv - \int v du$$

Expectation Calculations VI

7. A continuous real-valued variable has a power-law p.d.f. if $p(x) = Cx^{-\alpha}$ $(\alpha > 0)$. In fact, this function diverges as $x \to 0$: so how can it be a p.d.f. ?
8. Find an expression for C in the (modified) p.d.f. in the previous question.
9. Find the expected value for the random variable having the (modified) power-law p.d.f.
10. Power-laws with $\alpha \leq 2$ have no finite mean. This means that as we start taking more and more samples from such populations, we will start to see the mean diverge. How can this happen?
11. Similarly show that for $\alpha \leq 3$, there is no finite variance.

Maximum Likelihood Calculations I

1. You have a sample of n observations $x_1, x_2, \ldots, x_n$ from data that appear to fit a binomial distribution
with parameters N and p. Assuming N is known, derive the maximum likelihood estimate for p in terms of N, n, and the x_i.

Maximum Likelihood Calculations II

2. Let $x_1, x_2, \ldots, x_n$ be a sample of observations from a Poisson distribution with parameter λ. Find the maximum likelihood estimate of λ in terms of the x_i and n.

Maximum Likelihood Calculations III

3. Let $x_1, x_2, \ldots, x_n$ be a sample from an exponential distribution, which has a density function $f(X = x) = \lambda e^{-\lambda x}$ $(x > 0)$. Derive a maximum likelihood estimate of λ in terms of the x_i and n.

Maximum Likelihood Calculations IV

4. Let $x_1, x_2, \ldots, x_n$ be observations from a normal distribution with parameters μ and σ^2. Derive maximum likelihood estimates of μ and σ^2.

Bounds Calculations I

1. One use of the inequalities like Markov is to come up with worst-case estimates of probabilities (or relative frequencies) with very little information. An alternate statement of Markov's inequality is:

   $$P(X \geq aE(X)) \ \leq \ \frac{1}{a}$$

   Show that this inequality holds when X takes only non-negative real values and $a > 0$

Bounds Calculations II

2. Suppose you invest $1,000$ cowrie shells in some stock, that past history shows has an average annual return of about 5%. What is the probability that you double your cowrie shells in a year?

Bounds Calculations III

3. Sometimes, it may be the case that X is not always positive. There is a generalised version of Markov's inequality:

$$P(f(X) \geq a) \leq \frac{E[f(X)]}{a}$$

where $f(X)$ is non-negative. For example, if X can be negative, Markov's inequality can still be used for $|X|$. Show that Chebyshev's inequality follows from this general version of Markov's inequality.

Bounds Calculations IV

4. A biased coin has a probability $p = 0.2$ of *heads* on any one toss. What is the probability that we will get at least 80% *heads* in n tosses. Compare the probabilities that you get using: (a) Markov's inequality; and (b) the Binomial distribution.

Bounds Calculations V

5. Let X be a r.v. that assigns the number of dots that turn up when a die is thrown. Calculate the probability of the event: $X \geq 5$: (a) exactly; and (b) using Markov's inequality.

Bounds Calculations VI

6. Let X be a r.v. from a negative exponential distribution (p.d.f $= \lambda e^{-\lambda x}$ for $x > 0$ and 0 otherwise), with mean μ. Determine the probability of the event: $X \geq 3\mu$: (a) exactly; and (b) using Markov's inequality.

Bounds Calculations VII

7. One application of Chebyshev's bound is that it is sometimes used to obtain something like a confidence limit, since it is a two-sided bound. Use the Chebyshev's bound to obtain an interval around the mean, within which at least 95% will lie.

Bounds Calculations VIII

8. A factory produces bearings with an average diameter of 0.5in and a standard deviation of 0.01in. Find a lower bound on the number of bearings in a box of 400 bearings have diameters varying between 0.48 and 0.52.

Bounds Calculations IX

9. Let X be a r.v. denoting the number of successes in n independent Bernoulli trials, each with probability of success p. If p is unknown, how close is the proportion of observed successes (the r.v. X/n) to p as n increases? That is, you are asked to estimate for some fixed ϵ:

$$P\left(\left|\frac{X}{n} - p\right|\right) < \epsilon$$

as n increases (that is, the limiting value of this probability as $n \to \infty$). (You can use the fact that $Var(X/n) = Var(X)/n^2$

Bounds Calculations X

10. Let X be a r.v. with p.d.f. $f_X(x)$, which has a mean μ and variance σ^2. Let $\overline{X}_n$ be the mean of a sample n observations modelled by this p.d.f. We know $\overline{X}_n$ will have mean μ and variance σ^2/n Let ϵ and δ be numbers such that $\epsilon > 0$ and $0 < \delta < 1$. Then, if $n > \sigma^2/\epsilon^2\delta$, show:

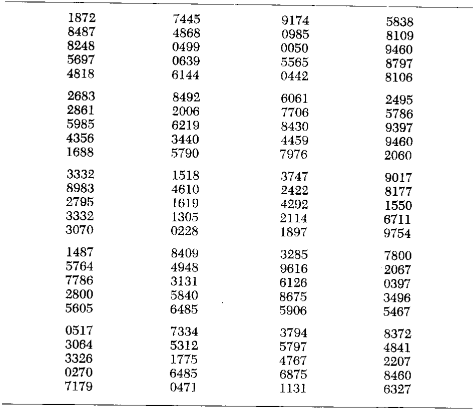
$$P\left(|\overline{X}_n - \mu| \leq \epsilon\right) \geq (1 - \delta)$$

Bounds Calculations XI

11. How large a sample must be taken in order that you are 99% certain that the sample mean $\overline{X}$ is within 0.5σ of the population mean μ?

Even Scattering I

- A random selection from The Manhattan Telephone Directory (1958-59)



1872	7445	9174	5838
8487	4868	0985	8109
8248	0499	0050	9460
5697	0639	5565	8797
4818	6144	0442	8106
2683	8492	6061	2495
2861	2006	7706	5786
5985	6219	8430	9397
4356	3440	4459	9460
1688	5790	7976	2060
3332	1518	3747	9017
8983	4610	2422	8177
2795	1619	4292	1550
3332	1305	2114	6711
3070	0228	1897	9754
1487	8409	3285	7800
5764	4948	9616	2067
7786	3131	6126	0397
2800	5840	8675	3496
5605	6485	5906	5467
0517	7334	3794	8372
3064	5312	5797	4841
3326	1775	4767	2207
0270	6485	6875	8460
7179	0471	1131	6327

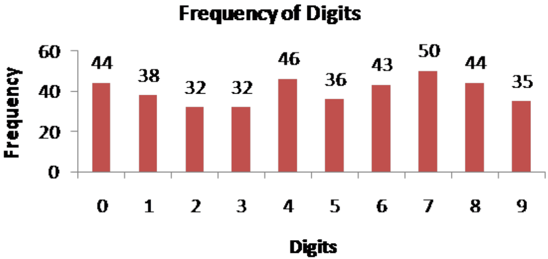
Even Scattering II

From: E.C. Berkeley, *Probability and Statistics: An Introduction Through Experiments*

- The frequency of digits

Digit	Freq.	Relative Freq.
0	44	0.110
1	38	0.095
2	32	0.080
3	32	0.080
4	46	0.115
5	36	0.090
6	43	0.108
7	50	0.125
8	44	0.110
9	35	0.087

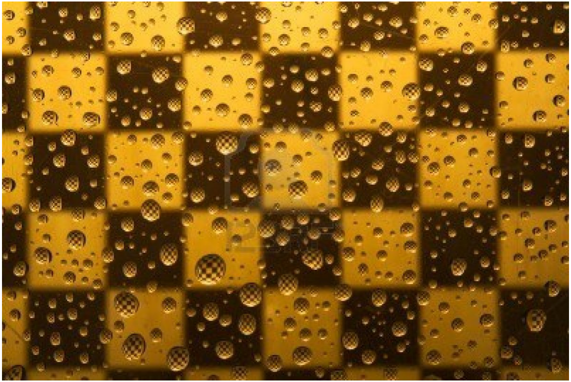
Even Scattering III



- The relative frequency of any digit ranges between 0.080 to 0.125, which is quite close to the 1/10 (the value if all digits had a uniform chance of appearing)

Even Scattering IV

- In 2-dimensions:



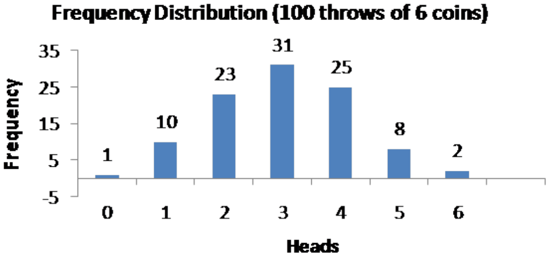
- This kind of even distribution of values is called the *discrete uniform distribution*

Uneven Scattering I

- Flipping 6 coins 100 times:

Number of Heads	Freq.	Relative Freq.
0	1	1/100
1	10	10/100
2	23	23/100
3	31	31/100
4	25	25/100
5	8	8/100
6	2	2/100

Uneven Scattering II



- Why does this shape occur? For this, you will have to examine all possible outcomes from tossing 6 cpins.
  - There will be 64 outcomes, each of which has an equal chance of occuring (why?)
  - Here is a classification of these 64 outcomes, based on the number of heads

Uneven Scattering III

Number of Heads	Number of Cases	Relative Frequency
0	1	1/64
1	6	6/64
2	15	15/64
3	20	20/64
4	15	15/64
5	6	6/64
6	1	1/64

- You can see that the actual relative frequencies (above) and the observed frequencies (earlier) are quite close
- This kind of uneven distribution of values is called the *binomial distribution*

Mathematical Thinking: Proofs

What are Mathematical Proofs

- A proof in mathematics is a demonstration that some mathematical statement is a logical consequence of a set of assertions
  - The assertions play the role of axioms
  - The statement proved is a *theorem*
  - Until a theorem is proved, it is called a *conjecture*
- Mathematical proofs usually use a combination of formal logic and careful arguments in a natural language (like English) to establish the theorems
- We will look largely at techniques that rely on using and manipulating logical formulæ, with some minimal use of natural language

Conjectures and Refutations I

- Here are two conjectures:

  Conjecture 1. If $n \in Z$ and $n > 1$ and n is prime then $2^n - 1$ is prime

  Conjecture 2. If $n \in Z$ and $n > 1$ and n is not prime then $2^n - 1$ is not prime

- A "proof" for Conjectures 1 and 2?

n	$2^n - 1$	Prime?
2	3	$\surd$
3	7	$\surd$
5	31	$\surd$
7	127	$\surd$

n	$2^n - 1$	Prime?
4	15	$\times$
6	63	$\times$
8	255	$\times$
9	511	$\times$

Conjectures and Refutations II

- No amount of correct examples can constitute a *proof* of a statement
- But a single incorrect (counter-) example is usually sufficient to show that a statement false
  - It essentially comes down to showing that a rule $P \leftarrow Q$ is *false* (here Q are the axioms). That is, the counter-example will make Q *true* and P *false*.

n	$2^n - 1$	Prime?
2	3	$\surd$
3	7	$\surd$
5	31	$\surd$
7	127	$\surd$
11	2047	$\times$

Conjectures and Refutations III

That is:

$$2^n - 1 \textit{ is prime } \leftarrow \; n \in Z \; \wedge \; n > 1 \; \wedge \; n \textit{ is prime}$$

is *false*

- Suppose we are not able to find any counter-examples for Conjecture 2
- No amount of absence of counter-examples can prove that a mathematical statement is true
- In summary: we cannot prove statements are true by demonstrating lots of examples. But we can prove statements are false (that is, we can *refute* a statement with a single counter-example. In contrast, proofs that statements are true are constructed using logical arguments

Proof Techniques: Conditional Statements I

- Proofs involving conditional statements

  To prove: A statement of the form $P \leftarrow Q$
  Strategy 1: Assume Q is *true* and then show P is *true*

- For example:
  - a and b are integers > 0 such that $a < b$. Then $a^2 < b^2$
  - In logical form:

$$(a^2 < b^2) \leftarrow a \in Z \wedge b \in Z \wedge (0 < a < b)$$

- Strategy: assume $a \in Z \wedge b \in Z \wedge (0 < a < b)$ and prove $a^2 < b^2$

  To prove: A statement of the form $P \leftarrow Q$
  Strategy 2: Assume P is *false* and then show Q is *false*

- For example:
  - a and b are integers > 0 such that $a < b$. Then $a^2 < b^2$

Proof Techniques: Conditional Statements II

- In logical form:

$$(a^2 < b^2) \leftarrow a \in Z \wedge b \in Z \wedge (0 < a < b)$$

- Strategy: assume $\neg(a^2 < b^2)$ and prove $\neg(0 < a < b)$

To prove: A statement of the form $P \leftarrow Q$

Strategy 3: Assume Q is *true* and P is *false*. Show that this leads to a contradiction.

- For example:
  - a and b are integers > 0 such that $a < b$. Then $a^2 < b^2$
  - In logical form:

$$(a^2 < b^2) \leftarrow a \in Z \wedge b \in Z \wedge (0 < a < b)$$

- Strategy: assume $a \in Z \wedge b \in Z \wedge (0 < a < b)$ and $a^2 \geq b^2$. Show that this leads to a contradiction.

Proof Techniques: Negated Statements I

- Proofs involving negated statements

  To prove: A statement of the form $\neg P$

  Strategy 1: Try to re-express P in some other form (like a conditional)

- For example:
  - Suppose $A \cap C \subseteq B$ and $a \in C$. Prove $a \notin A - B$
  - In logical form:

$$\begin{aligned} a \notin A - B \quad \equiv \quad & \neg(a \in A \wedge a \not inB) \\ & \neg(a \in A) \vee \neg(a \notin B) \\ & a \in B \leftarrow a \in A \end{aligned}$$

- That is: to prove $a \notin A - B$ is the same as proving $a \in B \leftarrow a \in A$

Proof Techniques: Negated Statements II

- Prove the conditional using one of the strategies before
- That is:

Given: $A \cap C \subset B$ and $a \in C$

To prove: $a \in B \leftarrow a \in A$

Proof. Assume $a \in A$. Since $a \in C$ (given), therefore $a \in A \cap C$. Since $A \cap C \subseteq B$, if $a \in A \cap C$, then $a \in B$.

To prove: A statement of the form $\neg P$

Strategy 2: Assume P and show that this leads to a contradiction

$$\neg P \equiv \neg P \leftarrow \mathit{true} \equiv \mathit{false} \leftarrow P$$

- For example:
  - Suppose $A \cap C \subseteq B$ and $a \in C$. Prove $a \notin A - B$
- That is:

Proof Techniques: Negated Statements III

Given: $A \cap C \subset B$ and $a \in C$

To prove: $a \notin A - B$

Proof. Assume $a \in A - B$. Then $a \in A$ and $a \notin B$. Since $a \in C$ (given) and $a \in A$, then $a \in A \cap C$. Since $A \cap C \subseteq B$, it follows that $a \in B$, which is a contradiction.

Proofs Techniques: Quantified Statements I

To prove: A statement of the form $\forall x P(x)$

Strategy: Assume the variable x can stand for *any* object from the domain and show $P(x)$ is true. Since $P(x)$ is true no matter what the of x.

- For example:
  - Suppose for sets A and B, $A \cap B = A$. Then $A \subseteq B$
- That is, we want to prove $A \subseteq B \leftarrow (A \cap B = A)$. This is of the form $P \leftarrow Q$. Let us try assuming Q and proving P.

Given: $A \cap B = A$

To prove: $A \subseteq B$. That is $\forall x (x \in A \rightarrow x \in B)$

Proofs Techniques: Quantified Statements II

Proof. Let x be any element. We now have another conditional to prove: $x \in A \rightarrow x \in B$. Suppose $x \in A$. Since $A = A \cap B$ (given) $x \in A \cap B$. That is, $x \in B$. Since x was arbitrary, it follows that $\forall x(x \in A \rightarrow x \in B)$.

To prove: A statement of the form $\exists x P(x)$

Strategy: Try to find a value x for which $P(x)$ is true.

- For example:
  - For every natural number $x > 0$, there is some number $y \geq 0$ such that $y + 1 = x$
- That is, we want to prove $\forall x(x > 0 \rightarrow \exists y(y \geq 0 \wedge y + 1 = x))$. This is a conditional, so we proceed as before.

Given: Assume x is arbitrary and $x > 0$.

Proofs Techniques: Quantified Statements III

To prove: $\exists y(y \geq 0 \wedge y+1=x)$

Proof. Let $y = x-1$ for arbitrary $x > 0$. Since $x > 0$, $y \geq 0$. Also $y+1 = (x-1)+1 = x$. That is, there is a value of y for arbitrary x such that $y \geq 0$ and $y+1=x$. That is, $\exists y(y \geq 0 \wedge y+1=x)$. Since x was arbitrary, it follows that
$\forall x(x > 0 \rightarrow \exists y(y \geq 0 \wedge y+1=x))$

Proofs Techniques: Negated Quantified Statements

To prove: A statement of the form $\neg\forall x P(x)$

Strategy: Prove $\exists x \neg P(x)$

To prove: A statement of the form $\neg\exists x P(x)$

Strategy: Prove $\forall x \neg P(x)$

- Use a combination of strategies for $\forall$, $\exists$ and $\neg$

Conjunctions and Disjunctions I

To prove: A statement of the form $P \wedge Q$

Strategy: Prove P and Q separately

To prove: A statement of the form $P \vee Q$

Strategy: First check for a proof for P. Otherwise check for a proof for Q.

Given: A statement of the form $P \wedge Q$

Then: Assume P and Q are both true when checking for a proof.

Conjunctions and Disjunctions II

Given: A statement of the form $P \vee Q$

Then: Assume P and check for a proof. Otherwise assume Q and check for a proof. Alternatively, if you know that P is false, then Q must be true. If you know Q is false then P must be true

- For example:
  - Given sets A, B and C such that $A \subseteq C$ and $B \subseteq C$, show that $A \cup B \subseteq C$.
  - Logically, given $A \subseteq C$ and $B \subseteq C$ prove $\forall x(x \in A \cup B \to x \in C)$.
  - That is, prove: $\forall x(x \in A \vee x \in B \to x \in C)$
- We can now try to prove this using strategies for conditionals, $\forall$, and disjunctions

Biconditionals (iff)

To prove: A statement of the form $P \leftrightarrow Q$
Strategy: Prove the conjunction $P \leftarrow Q \;\wedge\; Q \leftarrow P$

Given: A statement of the form $P \leftrightarrow Q$
Then: Assume the conjunction $P \leftarrow Q \wedge Q \leftarrow P$ is true

To prove: A statement of the form $\exists ! x P(x)$
Strategy: Prove $\exists x P(x)$ and $\forall y, z(P(y) \wedge P(z) \rightarrow y = z)$

Given: A statement of the form $\exists ! x P(x)$
Then: Assume you are given $\exists x P(x)$ and
$\forall y, z(P(y) \wedge P(z) \rightarrow y = z)$

Aside: What about Inference Rules

- Previously, we had looked at inference rules like *modus ponens*, *modus tollens* and *resolution*

  modus ponens From $p \leftarrow q$ and q infer p

  modus tollens From $p \leftarrow q$ and $\neg p$ infer $\neg q$

  resolution From $p \leftarrow q$ and $q \leftarrow r$ infer $p \leftarrow r$
- How are these to be used in mathematical proofs?
- The strategies that we have described only refer to the form of statements to be proved (conditionals, negated *etc.*. All rules of inference can be used to derive new statements (theorems) from what is given (axioms)
- Note: previously, we had used "strategy" to mean the choice of inference rule to be used to derive theorems from the axioms. Now, we are using the same word to mean a plan of action to establish a mathematical proof

Aside: Instantiations

- Given: $\forall x P(x)$ means that $P(a)$ can be taken to be true for any value $x = a$. This is called universal instantiation
- Given: $\exists x P(x)$ means that $P(a)$ can be taken to be true for some value $x = a$. This is called existential instantiation

Mathematical Induction I

- Special kind of proof strategy for proving statements of the form $\forall n P(n)$ for $n \in \{0, 1, 2, 3, \ldots\}$
- Strategy: prove $P(0)$ is true. Assume $P(n)$ is true and show $P(n+1)$ is true. By recursive application, it follows that $P(n)$ is true for all n.
- Logically:

$$P(0) \wedge \forall n (P(n) \rightarrow P(n+1))$$

Or, more generally:

$$P(n_0) \wedge \forall k \geq n_0 (P(k) \rightarrow P(k+1))$$

$P(n_0)$ is called the base case and the conditional $\forall k \in N (P(k) \rightarrow P(k+1))$ is called the induction step

Mathematical Induction II

- The induction step is proved using one of the strategies for conditionals we have already studied. That is, assume $P(n)$ for arbitrary n, and show $P(n+1)$ is true
- For example:

Prove: $\forall n \in N \; n^3 - n$ is divisible by 3.

Proof: Let $P(n) \;=\; 3|(n^3 - n)$. Clearly $P(0)$ is true. For the induction step, we assume that $P(n)$ is true, and try to prove $P(n+1)$. That is, assume $\exists k \in Z(3k = n^3 - n)$ and prove $\exists j(3j = (n+1)^3 - (n+1))$. Now $(n+1)^3 = n^3 + 3n^2 + 3n + 1$. So, we want to prove $\exists j 3j = n^3 + 3n^2 + 3n + 1 - n - 1$. That is, $\exists j 3j = n^3 - n + 3(n^2 + n) = 3k + 3(n^2 + n)$. Clearly, if $j = k + n^2 + n$ then $P(n+1)$ is true.

Strong Induction I

Weak. Normal induction (using 0 as base case):

$$P(0) \wedge \forall k(P(k) \rightarrow P(k+1))$$

Strong. We need the information implicit in the Weak formulation, that, for all numbers i less than $k+1$, $P(i)$ is true. That is:

$$\forall n((\forall k < nP(k)) \rightarrow P(n))$$

- Normal (or weak) induction allows us to start with the base case and show that $P(\cdot)$ is true for every successive number. By the time we get to any particular number n, we have therefore proved that $P(\cdot)$ is true for every preceding number. All that Strong induction allows us to do is to use this the fact.

Strong Induction II

- Why is this needed? Sometimes, it may not be enough in the induction step to assume just $P(n-1)$ to prove $P(n)$ (we need to know that $P(0), P(1), \ldots, P(n-1)$ are true.
- For example:

Prove: Every integer $n > 1$ is either prime or a product of primes

Proof: Let $P(n) \rightarrow \mathit{prime}(n) \vee \mathit{composite}(n)$, where $\mathit{composite}(n)$ means n is a product of primes. Suppose we want to prove this by Weak induction. The base case $P(1)$ is clearly true. Suppose $P(n-1)$ is true. We want to show that $P(n)$ is true. Now, n is either a prime number (that is, $\mathit{prime}(n)$ is true) or it is not. If it is, we are done. If not, let $n = ab$ (where $1 < a, b < n$. Now, if we can show that $P(a)$ and $P(b)$ are

Strong Induction III

true, then we are done. But with Weak induction, we can only assume $P(n-1)$ is true. . .
With Strong induction, we assume $P(k)$ is true for $1 < k < n$. Since $1 < a, b < n$, $P(a)$ is true and $P(b)$ is true. That is a and b are either prime or a product of primes. It follows that $n = ab$ is a product of primes.

- Equivalence of strong and weak forms:
  - We are interested in the following two forms:

$$W : P(0) \wedge \forall n(P(n) \rightarrow P(n+1))$$

and:

$$S : P(0) \wedge (P(0) \wedge P(1) \wedge \cdots \wedge P(n)) \rightarrow P(n+1))$$

Strong Induction IV

- Repeated application of resolution to W yields $P(1), P(2), \ldots, P(n)$. Now, if $p \leftarrow q$ and $q \leftarrow r$, it follows that $p \leftarrow r$. It also follows that $p \leftarrow q \wedge r$. Thus, it follows from W that $P(n+1) \leftarrow P(0) \wedge P(1) \wedge P(1) \wedge \cdots \wedge P(n)$. That is, the Weak form can be converted into the Strong form.
- Let $Q(n) = P(0) \wedge P(1) \wedge \cdots P(n)$. Then, $S : P(0) \wedge Q(n) \rightarrow P(n+1)$. It is clear that $Q(n) \rightarrow Q(n) \wedge P(n+1)$. That is, $Q(n) \rightarrow Q(n+1)$. That is $S : P(0) \wedge Q(n) \rightarrow Q(n+1)$. That is, the Strong form can be converted into the Weak form.

Mathematical Thinking: Approximations

Approximate Calculations I

There are 4 principal points to remember when calculating approximately:

Mathematical Thinking: Data Visualisation

An Example Table

Place	M.A. (Dev. Eco)		M.A. (I.R.)		M.A. (Soc.)		L.L.M		M.Sc. (Bio .)		M.Sc. (CS)		Total	
	Off	Acc	Off	Acc	Off	Acc	Off	Acc	Off	Acc	Off	Acc	Off	Acc
Afghan	6	5	3	2	1	–	1	–	2	2	11	8	24	17
B'desh	12	3	9	3	13	5	8	5	18	8	6	2	66	27
Bhutan	–	–	–	–	–	–	–	–	1	–	–	–	1	–
M'dives	–	–	–	–	–	–	1	1	–	–	–	–	1	1
Nepal	7	5	3	3	2	2	–	–	7	5	4	2	23	17
Pak.	4	1	7	2	1	–	5	3	3	–	3	1	23	7
S.L.	2	1	3	2	1	–	3	2	1	–	1	1	11	6
India	54	10	34	14	60	16	24	18	141	13	80	10	393	81
Total	85	25	59	26	78	23	42	20	173	29	105	24	542	156

- If you had to tell someone over the 'phone about this data, what would you tell them?

Psychology, Typography and Statistics

- We often need to remember numbers when looking at a table. It is easier for our short-term memories to remember "shorter" numbers than longer ones (140 is easier than 137; and 130,000 is easier than 131,238).
- Short-term memory finds it difficult to remember more than 2 digits if interrupted. Interpreting a table results in lots interruptions.
- It is difficult for us to sort numbers, especially if numbers are large and vary across a range.
- It is easier to read numbers down a column than across a row. It would be better if larger numbers appeared first.
- STRINGS OF CAPITALS ARE HARD TO READ

Example Table (again)

- Countries ordered by size ("large", "medium" and "small")
- Places offered grouped by broad areas, and rounded

Place	S. Sciences		Law		Sciences	
	Off	Acc	Off	Acc	Off	Acc
India	150	40	24	18	220	23
Pak.	12	3	5	3	6	1
B'desh	34	8	8	5	24	10
Approx. Tot.	200	50	35	30	250	35
Afghan	10	7	1	0	13	10
Nepal	12	10	0	0	11	7
S.L.	6	3	1	0	2	1
Approx. Tot.	30	20	2	0	25	18
Bhutan	0	0	0	0	1	0
M'dives	0	0	1	1	0	0
Approx. Tot.	0	0	1	1	1	0

Words, Tables or Graphs?

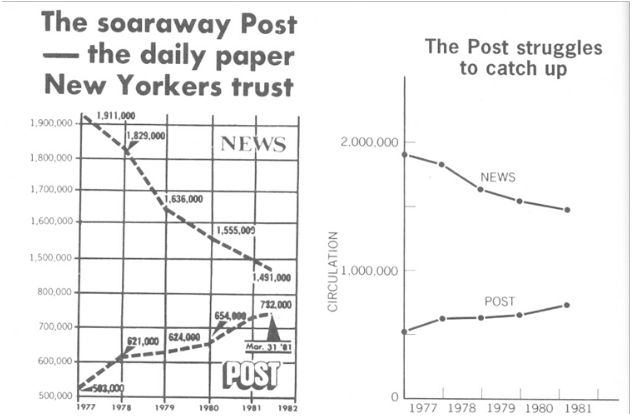
- "Is a graph worth a 1000 tables?"
  - Yes, if what is needed is a qualitative summary. (Social Sciences and Sciences are less popular in large countries than they are in medium-sized countries.)
  - Yes, if relationships in the data, or unusual features like outliers need to be identified.
  - No, if quantitative summaries are important. (Acceptance rates in the Sciences in large countries is 15% and in medium-sized countries is 70%.)
  - No, if there are too few data points or relationships are too complex.
- "Are all graphs created equal?" No: pie-charts and bar-graphs waste toner-ink and rarely convey the same information as a simple table or a graph.

Some Basic Rules for Tables and Graphs

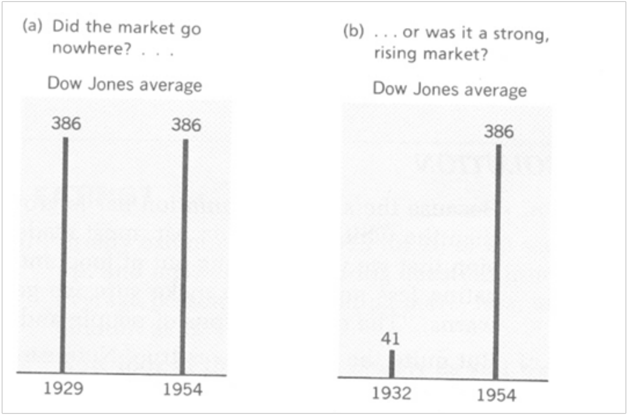
- Decide what information you want to convey. If it is qualitative, a graph is appropriate (avoid pie-charts and bar-graphs, though). If it is quantitative, then a table is usually better
- With tables:
  - Do not be afraid of rounding (to two effective digits, if possible)
  - Give row and column averages (or totals)
  - It is easier to read down a column of numbers rather than across a row of numbers
  - Order rows or columns (usually by size)
  - Avoid tablejunk: single spacing with deliberate gaps is better than double-spaced rows; avoid lots of lines across and down the table
  - Do not write out things all in CAPITALS

Bad Graphs and Good I

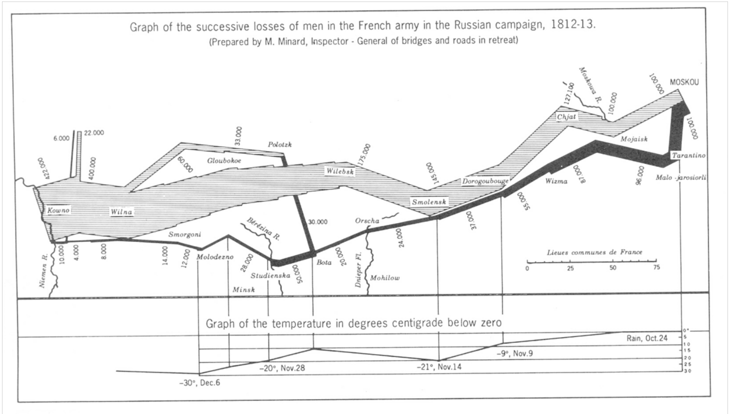
From: Wonnacott & Wonnacott, *Introductory Statistics for Business and Economics*



Bad Graphs and Good II



Bad Graphs and Good III



Summarising Data

- Ways to summarise data points $X_1, X_2, \ldots, X_n$:
  - Tables, graphs or words
  - Centre of values: mode, median and the mean
  - Scatter or spread of values: range, mean absolute deviation, mean squared deviation. variance and standard deviation.