IMPORTANT FORMULAS

I. PROBABILITY

Number of combinations of k out of n objects: $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$, where $n! = 1\cdot2\cdot3\cdot\ldots\cdot n$. ($0! = 1$.)

Probability of disjunction: $P(A \cup B) = P(A) + P(B) - P(AB)$. Odds for A: $P(A)/P(A^c)$.

Conditional probability of A given B: $P(A|B) = P(AB)/P(B)$ if $P(B) > 0$. So $P(AB) = P(A|B)P(B)$.

Theorem of total probability: $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$.

Bayes' theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$.

Independence of A and B: $P(AB) = P(A)P(B)$; equivalently (if $P(B) > 0$), $P(A|B) = P(A)$.

Variance of a random variable Y with expectation μ: $V(Y) = E(Y^2) - \mu^2 = E((Y - \mu)^2)$.

Expectation (i.e., mean) of a discrete random variable Y: $E(Y) = y_1P(Y = y_1) + y_2P(Y = y_2) + \ldots$.

$E(Y_1 + Y_2) = E(Y_1) + E(Y_2)$. $V(aY) = a^2V(Y)$. If Y_1 and Y_2 independent, $V(Y_1 + Y_2) = V(Y_1) + V(Y_2)$.

Bernoulli random variable (success/failure): $E(Y) = p$, $V(Y) = pq$, where $q = 1 - p$.

Binomial random variable (successes in n trials): $P(Y = k) = \binom{n}{k}p^kq^{n-k}$, $E(Y) = np$, $V(Y) = npq$.

Geometric random variable (trials until first success): $P(Y = n) = q^{n-1}p$, $E(Y) = 1/p$, $V(Y) = q/p^2$.

If Y is normal with parameters μ and σ^2, the *standard normal* $Z = (Y - \mu)/\sigma$ has parameters 0 and 1.

Central Limit Theorem: For any sequence $Y_1, Y_2, \ldots$ of IID random variables with expectation μ and variance σ^2, the cdf of Z is the limit, as $n \to \infty$, of the cdf of $(Y_1 + Y_2 + \ldots + Y_n - n\mu)/(\sigma\sqrt{n})$.

II. STATISTICS

Sample: n IID random variables $Y_1, \ldots, Y_n$ with $E(Y_i) = \mu$ (population mean) and $V(Y_i) = \sigma^2$ (population variance). Sample mean: $\bar{Y} = (Y_1 + \ldots + Y_n)/n$. $E(\bar{Y}) = \mu$, $V(\bar{Y}) = \sigma^2/n$. Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}(\sum_{i=1}^{n}Y_i^2 - n\bar{Y}^2)$. Measured values of $\bar{Y}$ and S: $\bar{y}$ and s.

Large sample $1 - \alpha$ confidence interval for a proportion: $(\bar{y} - z_{\alpha/2}\cdot se,\ \bar{y} + z_{\alpha/2}\cdot se)$, where $se = \sqrt{\bar{y}(1-\bar{y})/n}$ and $z_{\alpha/2}$ is the point to the right of which the area under the standard normal pdf is $\alpha/2$. For a $1 - \alpha$ confidence interval of width $\leq d$, it is enough to have $n \geq (z_{\alpha/2}/d)^2[4p(1 - p)]$.

Large sample $1 - \alpha$ confidence interval for a mean: $(\bar{y} - z_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{y} + z_{\alpha/2}\frac{s}{\sqrt{n}})$; use σ (not s) if known.

Small sample $1 - \alpha$ confidence interval for a mean: $(\bar{y} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}},\ \bar{y} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}})$, where $t_{\alpha/2,n-1}$ is the point to the right of which the area under the pdf of the t distribution with $n - 1$ degrees of freedom is $\alpha/2$. (All t-based tests assume that the population distribution is *normal*.)

Significance level α for hypothesis testing: Probability of type I error (rejecting true H_0).

Hypothesis testing for a mean: To test H_0: $\mu = \mu_0$, compute $t = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$ and see if $|t| > t_{\alpha/2,n-1}$ for a two-sided H_1 (i.e., $\mu \neq \mu_0$), or compare t to $t_{\alpha,n-1}$ for a one-sided H_1 (e.g., $\mu > \mu_0$).

Comparison of two independent means: To test H_0: $\mu_X = \mu_Y$, compute $t = \frac{\bar{x}-\bar{y}}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$ (pooled variance: $s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{i=1}^{m}(y_i-\bar{y})^2}{n+m-2}$) and see if $|t| > t_{\alpha/2,n-1}$ for a two-sided H_1 (i.e., $\mu_X \neq \mu_Y$), or compare t to $t_{\alpha,n+m-2}$ for a one-sided H_1 (e.g., $\mu_X > \mu_Y$).

Goodness of fit test: To test H_0: the distribution of $Y_1, \ldots, Y_k$ is multinomial with parameters n, $p_1, \ldots, p_k$, compute $c = \sum_{i=1}^{k}\frac{(y_i - np_i)^2}{np_i}$ and see if $c > \chi^2_{k-1}$ (check that all $np_i \geq 5$ or $n > 5k$).

Testing for independence of X and Y: If the data for X and Y are arranged in r rows and c columns, use the χ^2 test with $(r - 1)(c - 1)$ degrees of freedom.

INTRODUCTION TO LOGIC

I. LOGIC, ARGUMENTS, AND PROPOSITIONS

1. The main object of *logic* is to *evaluate arguments*: to find out *which* arguments are good (or bad), and *how good* (or how bad) they are.

2. An *argument* is an ordered pair whose first member is a set of propositions (the *premises* of the argument) and whose second member is a proposition (the *conclusion* of the argument).

3. A *proposition* is something that can be non-derivatively *true* or false, and is typically expressed by a *declarative* sentence. Different sentences can express the same proposition (e.g., "Alice is taller than Bob" and "Bob is shorter than Alice").

II. RELATIONS BETWEEN PREMISES AND CONCLUSIONS

1. An argument is (*deductively*) *valid* exactly if it is *necessary* that its conclusion is true if its premises are true (i.e., its premises *guarantee* its conclusion), and is *invalid* otherwise.

2. An argument is *(inductively) strong* exactly if (and to the extent that) it is invalid and its conclusion is *probable* given its premises (i.e., its premises render its conclusion probable but do not guarantee it), and is *weak* exactly if (and to the extent that) it is invalid and its conclusion is *improbable* given its premises.

3. An argument is *confirmatory* exactly if (and to the extent that) it is invalid and its premises *raise the probability* of its conclusion (i.e., its conclusion is *more probable* given that its premises are true than given that its premises are false), and is *disconfirmatory* exactly if (and to the extent that) it is invalid and its premises *lower the probability* of its conclusion.

4. A classification of *invalid* arguments (in the examples, "IAU" stands for "After conducting a thorough survey of celestial objects, the International Astronomical Union has declared"):

Invalid	Strong: $P(H\|E)$ high	Neither: $P(H\|E)$ medium	Weak: $P(H\|E)$ low
Confirmatory: $P(H\|E) > P(H)$	IAU: "No large asteroid will hit the Earth next year". So: No large asteroid will hit.	IAU: "50% chance a large asteroid will hit next year". So: A large asteroid will hit.	IAU: "30% chance a large asteroid will hit next year". So: A large asteroid will hit.
Neither: $P(H\|E) = P(H)$	Paris is in France. So: No large asteroid will hit the Earth next year.	Paris is in France. So: This fair coin will come up heads when tossed.	Paris is in France. So: A large asteroid will hit the Earth next year.
Disconfirmatory: $P(H\|E) < P(H)$	IAU: "30% chance a large asteroid will hit next year". So: No large asteroid will hit.	IAU: "50% chance a large asteroid will hit next year". So: No large asteroid will hit.	IAU: "No large asteroid will hit the Earth next year". So: A large asteroid will hit.

III TRUTH AND PROBABILITY OF PREMISES

1. Truth of premises: A valid argument with false premises is not good in the fullest sense. An argument is *sound* exactly if it is valid and its premises are all true, and is *unsound* exactly if it is not sound (i.e., either it is invalid or it is valid but its premises are not all true). The conclusion of a sound argument is true, but an argument with a true conclusion need not be valid or sound.

2. Probability of the premises: True premises can be improbable; e.g., any particular sequence of heads and tails in 100 tosses of a coin is improbable, but one sequence is true (i.e., will occur). To be good in the fullest sense, an argument must have premises that are not only true but also as close to *certain* (i.e., maximally probable) as possible.

3. *Deductive logic* evaluates arguments in terms of validity; *inductive logic* evaluates arguments in terms of strength and confirmation. Logic does not examine the *truth* of the premises.

COMBINATORICS

I. INTRODUCTION

1. The object of *combinatorics* is to find the number of possible outcomes of a given procedure (i.e., the number of ways in which the procedure can be performed).

2. The procedure may be complex, consisting of performing simpler procedures in successive steps. E.g., choosing a username and password consists of first choosing a username and then choosing a password.

3. Notation: $<a, b>$ is the ordered sequence whose first member is a and whose second member is b, and $\{a, b\}$ is the unordered set whose two members are a and b. So $<a, b> \neq <b, a>$ but $\{a, b\} = \{b, a\}$.

II. THE MULTIPLICATION RULE

1. The rule: Consider k procedures P_1, P_2, …, P_k. Suppose P_1 can be performed in n_1 ways, P_2 in n_2 ways, …, and P_k in n_k ways. Then the complex procedure which consists of successively performing P_1, P_2, …, and P_k can be performed in $n_1 \cdot n_2 \cdot \ldots \cdot n_k$ ways.

2. Example: If there are $n_1 = 800$ ways of choosing a username and $n_2 = 1{,}000$ ways of choosing a password, then there are $n_1 \cdot n_2 = 800{,}000$ ways of first choosing a username and then choosing a password.

III. PERMUTATIONS

1. A *permutation* of n objects is an ordered sequence of the n objects. It corresponds to a way of arranging the n objects in a sequence. E.g., there are two permutations of the two objects a and b: $<a, b>$ and $<b, a>$.

2. The *number of permutations* of n objects is $n!$ ("n factorial"), defined as $1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$ (by definition, $0! = 1$). E.g., there are $3! = 1 \cdot 2 \cdot 3 = 6$ permutations of 3 objects a, b, and c. They are: $<a, b, c>$, $<a, c, b>$, $<b, a, c>$, $<b, c, a>$, $<c, a, b>$, and $<c, b, a>$.

IV. COMBINATIONS

1. A *combination* of k out of n objects ($k \leq n$) is a collection (i.e., an unordered set) consisting of k out of the n objects. It corresponds to a way of choosing k out of the n objects without paying attention to the order of the k chosen objects. E.g., there are three combinations of 2 out of 3 objects a, b, and c: $\{a, b\}$, $\{a, c\}$, and $\{b, c\}$.

2. The *number of combinations* of k out of n objects is $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$. E.g., there are $\binom{4}{2} = \frac{4!}{2!\,(4-2)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4}{(1 \cdot 2) \cdot (1 \cdot 2)} = 6$ combinations of 2 out of 4 objects a, b, c, and d. They are: $\{a, b\}$, $\{a, c\}$, $\{a, d\}$, $\{b, c\}$, $\{b, d\}$, and $\{c, d\}$.

V. COMBINATORIAL PROBABILITY

1. If there are n possible outcomes of a procedure and a total of m of them satisfy a given condition A, then the probability that A will be satisfied is the ratio of m over n: $P(A) = m/n$. (This *assumes* that all n possible outcomes are equally probable.)

2. Example: There are $n = 6$ possible outcomes of throwing a fair die and $m = 3$ of them satisfy the condition A that a side with an even number of spots will come up, so $P(A) = 3/6 = 0.5$.

THE UNCONDITIONAL PROBABILITY CALCULUS

I. SAMPLE SPACES

1. A *sample space* is a set of possibilities (usually, possible outcomes of a procedure) that are considered to be of interest and to be mutually exclusive and collectively exhaustive.

2. Example 1: Suppose one tosses a coin. If one is interested in the probability that the coin will come up heads, one can choose the sample space {Heads, Tails}. This is a *discrete* sample space: it has a finite number of members. (A sample space with a countably infinite number of members is also discrete.) By choosing this sample space, one effectively declares that the possibility that the coin will stand on its edge is not of interest.

3. Example 2: Suppose again one tosses a coin. If one is interested in the probability that it will take longer than 15 seconds from the moment the coin is tossed until the moment the coin settles, one can choose the sample space [0, 1000]; i.e., the interval of real numbers from 0 to 1000. Each member of this sample space corresponds to a possible length of time in seconds until the coin settles. This is a *continuous* sample space: it has an uncountably infinite number of members. By choosing this sample space, one effectively declares that the possibilities in which the coin takes more than 1000 seconds to settle are not of interest.

4. One often chooses a sample space whose members are all equally probable, but this is not always possible (consider a *biased* coin in example 1). Combinatorics can be used to find out the size of the sample space.

II. THE OBJECTS OF PROBABILITY: EVENTS (OR PROPOSITIONS)

1. Not everything can have a probability. It makes no sense to talk of the probability of an object, for example of a coin (as opposed to, e.g., the probability that the coin will come up heads). Only *events* can have probabilities (e.g., the event that the coin will come up heads). One can alternatively assign probabilities to *propositions* (e.g., the proposition that the coin will come up heads).

2. *An event is (and a proposition is taken to be) a set of possibilities; i.e., a subset of a sample spa*ce. Example: Suppose a fair die is thrown. Consider the sample space {Side_1, Side_2, Side_3, Side_4, Side_5, Side_6}, where Side_i is the possibility that the side with *i* spots will come up. The proposition *Even* that a side with an even number of spots will come up is the set {Side_2, Side_4, Side_6}, a subset of the sample space. (For technical reasons, if a sample space is continuous then usually not every subset of it counts as a proposition, but this complication will be ignored.)

3. Since the possibilities in the sample space are considered to be mutually exclusive and collectively exhaustive, exactly one of them will be *actualized*. A proposition is *true* or false (and an event *occurs* or does not occur) depending on whether the actualized possibility is or not a member of the proposition. E.g., if Side_4 is actualized (i.e., the side with 4 spots comes up), then the proposition Even is true (and the event Even occurs).

4. The *contradiction* is the proposition that is *false* no matter what possibility is actualized; i.e., the *empty set* (the set that has no member, denoted by $\varnothing$). The *tautology* is the proposition that is *true* no matter what possibility is actualized; i.e., the *sample space* (denoted by Ω).

III. COMPLEX PROPOSITIONS

1. The *negation* of a proposition *A* is the *complement* of *A* (denoted by A^c), namely the set whose members are the members of the sample space that are *not* in *A*. The negation of *A* is also

denoted by $\sim A$ and is true exactly if A is false. E.g., Even^c = {Side_1, Side_3, Side_5} = Odd (the proposition that a side with an odd number of spots will come up).

2. The *conjunction* of propositions A and B is the *intersection* of A and B (denoted by $A \cap B$), namely the set whose members are the common members of A and B. The conjunction of A and B is also denoted by A & B, or just by AB, and is true exactly if both A and B are true. E.g., if Large = {Side_4, Side_5, Side_6} is the proposition that a side with a large number of spots (i.e., at least 4) will come up, then Even $\cap$ Large = {Side_4, Side_6}. For any A, $A \cap A^c = \varnothing$.

3. The *disjunction* of propositions A and B is the *union* of A and B (denoted by $A \cup B$), namely the set whose members are the members of A plus the members of B. The disjunction of A and B is also denoted by $A \vee B$ and is true exactly if A is true or B is true (or both). E.g., Even $\cup$ Large = {Side_2, Side_4, Side_5, Side_6}. For any A, $A \cup A^c = \Omega$.

4. De Morgan's Laws: $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$.

IV. RELATIONS BETWEEN PROPOSITIONS

1. Propositions A and B are *(mutually) incompatible* exactly if they cannot be both true; namely, they are *disjoint* (i.e., they have no common member: $A \cap B = \varnothing$). E.g., Even $\cap$ Odd = $\varnothing$.

2. Proposition A *entails* proposition B exactly if the truth of A guarantees the truth of B; namely, $A \subseteq B$ (i.e., every member of A is a member of B). E.g., Even $\cap$ Odd $\subseteq$ Even.

V. THE AXIOMS OF PROBABILITY

Given a sample space Ω and a set of propositions (i.e., subsets of the sample space), a *probability measure* is a function P that assigns to every proposition a real number and that satisfies the following three conditions (*axioms*):

A1. For any proposition A, $P(A) \geq 0$. (Probabilities cannot be negative.)

A2. $P(\Omega) = 1$. (The totality of possibilities, namely the tautology, has probability 1.)

A3. For any (finite or infinite) countable collection of propositions A_1, A_2, … that are pairwise incompatible (i.e., $A_i \cap A_j = \varnothing$ for $i \neq j$), $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$. (The probability of the disjunction of pairwise incompatible propositions is the sum of the probabilities of those propositions.) *Special case*: If $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$.

VI. BASIC PROBABILITY THEOREMS

1. Probability of negation: $P(A^c) = 1 - P(A)$.
2. Probability of contradiction: $P(\varnothing) = 0$.
3. If $A \subseteq B$, then $P(A) \leq P(B)$.
4. Upper bound on probabilities: for every proposition A, $P(A) \leq 1$.
5. Probability of disjunction: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

VII. INDEPENDENCE

1. Propositions A and B are *independent* exactly if $P(A \cap B) = P(A)P(B)$. This definition is intended to capture the intuitive notion that the truth of A is *unrelated* to the truth of B.

2. *Independence should not be confused with incompatibility* (i.e., disjointness). If propositions A and B are incompatible, then the truth of A entails that B is *not* true, so A and B are in general *not* independent. Formally, if $A \cap B = \varnothing$, then $P(A \cap B) = 0$, which differs from $P(A)P(B)$ except if $P(A) = 0$ or $P(B) = 0$.

THE CONDITIONAL PROBABILITY CALCULUS

I. CONDITIONAL PROBABILITIES

1. The *conditional probability of A given B* is $P(A|B) = P(A \cap B)/P(B)$ if $P(B) > 0$ (and is for our purposes undefined if $P(B) = 0$). For given B, the function $P(\bullet|B)$ satisfies the probability axioms.

2. Example: Suppose a fair die is thrown. The conditional probability of S_6 (i.e., that the side with 6 spots will come up) *given* Even (i.e., given that a side with an even number of spots will come up) is $P(S_6|\text{Even}) = 1/3 = (1/6)/(3/6) = P(S_6 \cap \text{Even})/P(\text{Even})$. The effect of the condition is to *shrink* the sample space. Similarly, $P(S_6|\text{Odd}) = 0$ because $P(S_6 \cap \text{Odd}) = 0$.

II. PROBABILITIES OF CONJUNCTIONS

1. Conjunction of two propositions: $P(AB) = P(A|B)P(B)$, from the definition above.

2. Conjunction of three propositions: $P(ABC) = P(A|BC)P(BC) = P(A|BC)P(B|C)P(C)$.

III. THE THEOREM OF TOTAL PROBABILITY

1. The theorem: $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$.

2. Proof: $P(A) = P(AB \cup AB^c) = P(AB) + P(AB^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$.

3. Example: Urn 1 contains 70 black and 30 white balls, and urn 2 contains 40 black and 60 white balls. A fair coin is tossed to select one of the urns, and a ball is randomly drawn from the selected urn. What is the probability that the ball is black? $P(\text{Black}) = P(\text{Black}|\text{Urn}_1)P(\text{Urn}_1) + P(\text{Black}|\text{Urn}_2)P(\text{Urn}_2) = 0.7 \cdot 0.5 + 0.4 \cdot 0.5 = 0.55$.

IV. BAYES' THEOREM

1. The theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

2. Proof: $P(A|B) = P(AB)/P(B) = P(BA)/P(B) = P(B|A)P(A)/P(B)$.

3. Corollary: Use the theorem of total probability to rewrite the denominator in Bayes's theorem. $P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$.

4 Example: A test for AIDS comes out positive (+) with probability 0.97 if the patient has AIDS and comes out negative (-) with probability 0.95 if the patient does not have AIDS. If 2% of people have AIDS, what is the probability that a patient has AIDS given that the result of the test was positive? $P(AIDS|+) = \frac{P(+|AIDS)P(AIDS)}{P(+|AIDS)P(AIDS)+P(+|AIDS^c)P(AIDS^c)} = \frac{0.97 \cdot 0.02}{0.97 \cdot 0.02 + 0.05 \cdot 0.98} = 0.284$.

V. INDEPENDENCE AND CONFIRMATION

1. The definition of "A and B are independent", namely $P(AB) = P(A)P(B)$, can be equivalently rewritten as $P(A|B) = P(A)$, as $P(A|B) = P(A|B^c)$, as $P(A) = P(A|B^c)$, and so on (interchanging A with B), provided that the conditional probabilities are defined. A and B are independent exactly if A^c and B^c are independent, and also exactly if A and B^c are independent.

2. The definition of "B (incrementally) confirms A", namely $P(AB) > P(A)P(B)$, can be equivalently rewritten as $P(A|B) > P(A)$, as $P(A|B) > P(A|B^c)$, as $P(A) > P(A|B^c)$, and so on (interchanging A with B), provided that the conditional probabilities are defined. So confirmation amounts to positive correlation and is symmetric: B confirms A exactly if A confirms B. Moreover, A confirms B exactly if A^c confirms B^c, and also exactly if A *dis*confirms B^c.

DISCRETE RANDOM VARIABLES

I. RANDOM VARIABLES

1. Just as a *variable* is something that can take different values, a *random variable* is something that can take different values *with different probabilities*. Example: the number of heads in two successive tosses of a fair coin is a random variable: it can take the values 0, 1, and 2, with probabilities 0.25, 0.50, and 0.25 respectively.

2. Formally, *a random variable is a function from a sample space to real numbers*. In the coin example, the sample space is {TT, TH, HT, HH}, and the random variable "number of heads in two tosses" is the function that assigns the number 0 to TT, 1 to TH, 1 to HT, and 2 to HH.

3. The above random variable is *discrete*: the set of its possible values (i.e., {0, 1, 2}) is discrete. A *continuous* random variable (e.g., temperature) has a continuous set of possible values.

II. BASIC DEFINITIONS

1. The *probability mass function* (abbreviation: *pmf*) of a discrete random variable Y is the function that gives, for every possible value of Y, the probability that Y takes that value. In the coin example, where Y is the number of heads in two tosses, the pmf of Y is the function that assigns to the value 0 the probability 0.25, to the value 1 the probability 0.50, and to the value 2 the probability 0.25. Notation: $P(Y = 0) = 0.25$, $P(Y = 1) = 0.50$, and $P(Y = 2) = 0.25$.

2. The *expectation* (or *expected value*, or *mean value*) of a discrete random variable Y whose possible values are $y_1, y_2, \ldots$ is: $E(Y) = y_1P(Y = y_1) + y_2P(Y = y_2) + \ldots$. In our example, the expectation is $0{\cdot}0.25 + 1{\cdot}0.50 + 2{\cdot}0.25 = 1$. (On average, one can "expect" one head in two tosses.)

3. The *variance* of a random variable Y with expectation μ is: $V(Y) = E(Y^2) - \mu^2 = E((Y - \mu)^2)$. In our example, the square of the number of heads can take the values 0, 1, and 4, with probabilities 0.25, 0.50, and 0.25 respectively, so $E(Y^2) = 0{\cdot}0.25 + 1{\cdot}0.50 + 4{\cdot}0.25 = 1.50$. Then $V(Y) = 1.50 - 1^2 = 0.50$. The square root of the variance of Y is called the *standard deviation* of Y.

4. Random variables X and Y are *independent* exactly if, for any sets A and B of numbers among their possible values, $P((X \in A) \cap (Y \in B)) = P(X \in A)P(Y \in B)$.

III. BERNOULLI PROCESSES

1. A *Bernoulli process* is a process that consists of *repeated independent and identically distributed (IID) trials*, with each trial having only two possible outcomes, called "success" (value 1) and "failure" (value 0). E.g., tossing a fair coin 10 times is a Bernoulli process: each toss is a trial (with heads as success and tails as failure, or the other way around) and the 10 trials are IID (they have identical probabilities of success and failure).

2. A *Bernoulli random variable* corresponds to each trial: it has two possible values, 1 and 0, with probabilities p and $q = 1 - p$ respectively. Its expectation is p, and its variance is pq.

3. A *binomial random variable* corresponds to the *number of successes in n trials* (e.g., number of heads in n coin tosses), and is the sum of n IID Bernoulli random variables. It can take values $k = 0, \ldots, n$, with probabilities: $P(Y = k) = \binom{n}{k}p^k q^{n-k}$. It has expectation np and variance npq. In general, $E(Y_1 + Y_2) = E(Y_1) + E(Y_2)$, and, for independent Y_1 and Y_2, $V(Y_1 + Y_2) = V(Y_1) + V(Y_2)$.

4. A *geometric random variable* corresponds to the *number of trials until (and including) the first success* (e.g., number of tosses until heads appears). It has infinitely many possible values ($n = 1, 2, \ldots$) with probabilities $P(Y = n) = q^{n-1}p$. Its expectation is $1/p$, and its variance is q/p^2.

CONTINUOUS RANDOM VARIABLES

I. UNIFORM RANDOM VARIABLES

1. Suppose one randomly selects a real number between 0 and 12 (e.g., by spinning a hand of a clock). Each number in the interval (0, 12) has the same probability p of being selected. But p must be 0: if it were positive, the sum of all probabilities would be infinite (since there are infinitely many numbers in the interval), but the sum must be 1. In general, *the probability that a continuous random variable takes a particular value y is zero*: $P(Y = y) = 0$ for any y.

2. For a continuous random variable, we are interested in the probability that its value falls in a *range* (or set) of possible values. In the clock example, what is the probability that the selected number is between 0 and 6? Given the randomness of the selection, $P(0 < Y < 6) = P(6 < Y < 12) = 0.5 = 6/12$. In general, the probability that Y is in an interval (y_1, y_2) is proportional to the length of the interval: $P(y_1 < Y < y_2) = (y_2 - y_1)/12$. It does not matter whether the interval is open or closed: $P(y_1 < Y \leq y_2) = P(y_1 < Y < y_2) + P(Y = y_2) = P(y_1 < Y < y_2)$, since $P(Y = y_2) = 0$.

3. The probability that Y is in a set A is the length of A times 1/12; i.e., the *area* that corresponds to A under the graph of the constant function 1/12; i.e., the *integral* of that function over A.

4. A random variable Y is *uniform* (or *uniformly distributed*) over the interval (a, b) exactly if, for every measurable subset A of (a, b), the probability that Y takes a value in A is the integral of the constant function $1/(b - a)$ over A. That constant function is the *probability density function* (abbreviation: *pdf*) of Y. The next step is to consider random variables whose pdf is *not* constant.

II. BASIC DEFINITIONS

1. The *probability density function* of a continuous random variable Y is a non-negative function $f(y)$ on all real numbers y such that, for every measurable set A of real numbers, $P(Y \in A) = \int_A f(y)dy$. In general, $f(y)$ is *not* $P(Y = y)$, which is 0. Since $P(-\infty < Y < +\infty) = 1$, the area under the *whole* graph of the function $f(y)$ must be 1. This can be so even if $f(y) > 1$ for some y. The pdf replaces the pmf (which is undefined: it would be 0 everywhere).

2. A continuous random variable can be equivalently specified by its *cumulative distribution function* (*cdf*), namely a function (also defined for discrete random variables) $F(y)$ such that, for every real number y, $F(y) = P(Y \leq y)$. So $P(a < Y \leq b) = F(b) - F(a)$. Note that $f(y) = dF(y)/dy$.

3. The *expectation* of a continuous random variable Y is: $E(Y) = \int_{-\infty}^{+\infty} y f(y) dy$. The *variance* of Y is, as in the discrete case, $V(Y) = E(Y^2) - \mu^2$, with $\mu = E(Y)$. The expectation of a random variable that is uniform over (a, b) is $(a + b)/2$, and its variance is $(b - a)^2/12$.

III. NORMAL RANDOM VARIABLES

1. A random variable Y is *normal* (or *normally distributed*) with parameters μ and σ^2 exactly if its pdf is: $f(y) = (2\pi)^{-1/2}\sigma^{-1}\exp[-(y - \mu)^2/(2\sigma^2)]$. It can be shown that $E(X) = \mu$ and $V(X) = \sigma^2$. This pdf has a "bell-shaped curve" centered around μ. With probability about 0.68, Y is within σ of μ: $P(\mu - \sigma < Y < \mu + \sigma) = 0.68$. With probabilities about 0.95 and 0.999, Y is within 2σ and 3σ of μ.

2. If Y is normal with parameters μ and σ^2, then $\alpha Y + b$ is normal with parameters $a\mu + b$ and $a^2\sigma^2$. So $Z = (Y - \mu)/\sigma$ is normal with parameters 0 and 1. Z is called the *standard normal* random variable and is very important because of the *Central Limit Theorem*: for *any* sequence $Y_1, Y_2, \ldots$ of IID random variables with expectation μ and variance σ^2, the cdf of Z is the limit, as $n \to \infty$, of the cdf of $(Y_1 + Y_2 + \ldots + Y_n - n\mu)/(\sigma\sqrt{n})$. So $Y_1 + \ldots + Y_n$ "approximates" a normal random variable. E.g., a binomial random variable approximates a normal random variable for large n.

INDUCTIVE LOGIC

I. THREE KINDS OF PROBABILITY

Mathematically, a probability measure is a function that satisfies the probability axioms. But what does it *mean* to say, e.g., that the probability of rain is 0.7 (given the presence of clouds)?

1. It can mean that the (objective) *chance* of rain is 0.7. Chances are supposed to be features of the world, not matters of opinion. They appear in scientific theories (e.g., quantum mechanics, statistical mechanics, genetics). We can infer them from relative frequencies. On a common view, the chance of a proposition can change over time: the chance that it rains at noon is low at 6am, is high at 10am, and is 1 at 2pm, assuming it did rain at noon: if A is a true proposition *about the past*, its present chance is 1. The chance of A at time t is denoted by $Ch_t(A)$.

2. Alternatively, saying that the probability of rain is 0.7 can mean that it is *rational* (i.e., rationally *required*) to have degree of belief 0.7 in the proposition that it will rain; equivalently, *every rational agent* has this degree of belief. The (subjective) *credence* of agent g at time t in A, denoted by $Cr_{gt}(A)$, is the degree of belief that g at t has in A. Credences are relative to times (like chances) but are relative to agents and thus subjective (unlike chances). To be rational, an agent must have credences that satisfy the probability axioms and some further constraints.

3. Saying that the probability of rain is 0.7 *given* the presence of clouds can mean that the *inductive probability* of the argument from "There are clouds" to "It will rain" is 0.7. Like chances, inductive probabilities are not relative to agents. Unlike chances, inductive probabilities are not relative to times either: whether an argument is (inductively) strong cannot change over time. The inductive probability of A given B is denoted by $In(A|B)$. (Unlike chances and credences, inductive probabilities are primarily *conditional*, but one can define $In(A)$ as $In(A|\Omega)$.)

II. RELATIONS BETWEEN THE THREE KINDS OF PROBABILITY

1. <u>Relation between inductive probabilities and rational credences</u>. For any number x in [0, 1], $In(A|B) = x$ exactly if $Cr_{gt}(A|B) = x$ (for any rational g and any t at which $Cr_{gt}(A|B)$ is defined): *the inductive probability of an argument is the rational credence in the conclusion given the premises of the argument* (and given no further information relevant to the conclusion). This leaves it open whether there is an independent standard for evaluating inductive probabilities to which rational agents conform, or whether inductive probabilities are just defined by agreement among rational agents. If $Cr_{gt}(A|B)$ differs among rational agents, then $In(A|B)$ is undefined.

2. <u>Relation between chances and rational credences</u>. Rational agents adjust their credences to (information about) *chances*: given only that the chance of A at t is 0.5, the rational credence at t in A is 0.5. This is a *chance-credence principle*: $Cr_{gt}(A|[Ch_t(A) = x]) = x$ (for any *rational* g). Similarly for *conditional* chances: $Cr_{gt}(A|B[Ch_t(A|B) = x]) = x$. Rational agents also adjust their credences to *frequencies*: given only that 90% of past tosses of a coin came up heads, the rational credence in heads at the next toss is 0.9. *Information about chances overrides information about frequencies*: given only *both* that 90% of past tosses came up heads *and* that the present chance of heads at the next toss is 0.5, the rational credence in heads at the next toss is 0.5 (not 0.9).

III. TRUTH VERSUS PROBABILITY ONE

1. <u>True proposistions need not have probability one</u>: A true but *unknown* proposition about the *future* can have both a present chance and a present rational credence less than 1.

2. <u>Probability-one propositions need not be true</u>: If a continuous random variable Y takes the value 0.3, the *false* proposition "$Y \neq 0.3$" had chance 1 (since $Ch_t(Y = 0.3) = 0$) and credence 1

(since rational agents adjust their credences to chances). So the argument from $Ch_t(A) = 1$ to A is invalid. But it is *maximally strong*: its inductive probability is 1, since $Cr_{gt}(A|[Ch_t(A) = 1]) = 1$.

IV. ARGUMENTS WITH PROBABILISTIC CONCLUSIONS

1. Arguments with probabilistic conclusions can be valid; e.g., $P(AB) = 0.9$ entails $P(A) \geq 0.9$. (Take the probability axioms to be implicit premises.) But do all such arguments have at least one (non-axiomatic) probabilistic premise? One might propose the principle "No Probability In, No Probability Out" (NPINPO): *no non-trivial argument with a probabilistic conclusion but no probabilistic premise is valid*. (The qualification "non-trivial" is needed to avoid, e.g., arguments with contradictory premises or arguments about the present chances of past events.)

2. Consider the argument from "This card was randomly selected from a standard deck" to "The present chance that this card is red is 0.5". Is this a counterexample to NPINPO? No: the argument is invalid. Either a red card was selected, and then the present chance of the card being red is 1, or a black card was selected, and then the chance is 0. What about the different argument from "A card *will* be randomly selected from a standard deck" to "The present chance that a red card will be selected is 0.5"? This argument is valid but is still no counterexample to NPINPO: to say that the card will be *randomly* selected is to say that each card has *the same chance* of being selected, so the premise is probabilistic after all.

3. The argument form "This card was (somehow) selected from a standard deck" to "This card is red" has inductive probability 0.5. So one might think that the argument from "This card was (somehow) selected from a standard deck" to "The present rational credence in this card being red is 0.5" is valid (and a counterexample to NPINPO). The latter argument is *not* valid, however: if it were valid, then adding *any* premise would preserve validity, but adding the premise "Everyone knows that this card is black" results in a clearly invalid argument.

V. ARGUMENTS WITH PROBABILISTIC PREMISES

1. <u>Probabilistic *Modus Ponens*</u>. The argument from "If C, then D" and C to D is valid. Analogously, the argument from $Ch_t(D|C) = 0.99$ and C to D is strong. It has degree of strength 0.99: by one of the chance-credence principles, $Ch_{gt}(D|C[Ch_t(D|C) = 0.99]) = 0.99$.

2. <u>Probabilistic *Modus Tollens*</u>. The argument from "If C, then D" and ~D to ~C is valid. But the argument from $Ch_t(D|C) = 0.99$ and ~D to ~C need not be strong. E.g., the inductive probability of the argument from Ch_t(Jim is not an American Senator|Jim is an American) = 0.99 and "Jim is an American Senator" to "Jim is not an American" is *zero*, not high.

3. <u>The Special Consequence Condition of confirmation</u>. If C *entails* D and D entails E, then C entails E. But if C *confirms* D and D entails E, C need not confirm E. For example: "This card is red" confirms "This card is the ace of hearts", and "This card is the ace of hearts" entails "This card is an ace", but "This card is red" does not confirm "This card is an ace".

VI. THE TOTAL EVIDENCE REQUIREMENT

The argument from "80% of US Senators are men" and "X is a US Senator" to "X is a man" has degree of strength 0.80. But what about the argument from "80% of US Senators are men" and "Barbara Boxer is a US Senator" to "Barbara Boxer is a man"? This also has degree of strength 0.80, although the *different* argument that one gets by adding the premise "Almost no one named 'Barbara' is a man" is *not* strong. This shows that a strong argument with premises known to be true may be useless because some further premises known to be true may be relevant to the conclusion. According to the *Total Evidence Requirement*, the credence of a rational agent in a proposition A is equal to the inductive probability of the argument whose conclusion is A and whose premises constitute the *total* evidence (relevant to A) that is available to the agent.

ESTIMATING PROPORTIONS

I. POPULATIONS, SAMPLES, AND ESTIMATORS

1. The object of *estimation* is to find out some *parameters* of a *population* (e.g., the proportion of registered Wisconsin voters who plan to vote in the next election) on the basis of data collected from a *sample* (i.e., a subset of the population; e.g., the respondents in a telephone survey).

2. Suppose the parameter to be estimated is the *proportion* (i.e., percentage) p of members of the population who have a certain feature (e.g., they plan to vote). Suppose the sample is *random* (or *randomly selected*): every member of the population has the same probability ($1/N$, where N is the population size) of being selected. Then to each member i of the sample ($i = 1, \ldots, n$, where n is the sample size) corresponds a Bernoulli random variable Y_i taking the value 1 if the member of the sample has the feature (e.g., plans to vote), with probability p, and the value 0 otherwise.

3. An *estimator* is a random variable that is a function from $Y_1, \ldots, Y_n$ to possible values of the parameter to be estimated. E.g., a simple estimator is the *sample mean* $\bar{Y} = (Y_1 + \ldots + Y_n)/n$.

II. POINT ESTIMATES

1. An *estimate* (or *point estimate*) of the parameter to be estimated is the value that an estimator takes for a given sample. E.g., if for a given sample of size $n = 3$ we have $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, then $\bar{y} = (1 + 1 + 0)/3 = 0.67$, so the estimate of the proportion (e.g., of those who plan to vote) is 0.67. (A specific value of $\bar{Y}$ is denoted by $\bar{y}$.) Different samples can result in different estimates.

2. Not knowing the population parameter, we do not know how good an *estimate* is. A good *estimator* is one that in general yields good estimates. An estimator is *unbiased* if its expected value equals the parameter, and is *consistent* if it "converges" to the parameter as n increases.

III. INTERVAL ESTIMATES (CONFIDENCE INTERVALS)

1. If the sample is random, then the Bernoulli random variables Y_i are independent, so their sum $Y_1 + \ldots + Y_n$ is a binomial random variable, and by the Central Limit Theorem it is approximately normal (with mean np and variance npq) if n is large (in practice, if $n > 20$, $np > 5$, and $nq > 5$; if the sample is without replacement, so the Y_i are not independent, the approximation can still be used if $n/N < 0.10$). Then $\bar{Y}$ is normal with mean p and variance pq/n. So $P(-1.96 < \frac{\bar{Y}-p}{\sqrt{pq/n}} < 1.96) = 0.95$. Let the standard error be $SE = \sqrt{\bar{Y}(1-\bar{Y})/n}$. Then one can show that $P(\bar{Y} - 1.96SE < p < \bar{Y} + 1.96SE) = 0.95$. For a value $\bar{y}$ of $\bar{Y}$, the interval $(\bar{y} - 1.96se, \bar{y} + 1.96se)$ is called a 95% *confidence interval* for p. E.g., if $n = 3$ and $\bar{y} = 0.67$, then $se = \sqrt{0.67(1-0.67)/3} = 0.27$, so $(0.67 - 1.96 \cdot 0.27, 0.67 + 1.96 \cdot 0.27) = (0.14, 1.20)$ is a 95% confidence interval for p. We can expect 95% of the confidence intervals constructed from many samples of size 3 to contain p.

2. The above confidence interval, (0.14, 1.20), is very wide and thus not very informative. *One way to get a narrower confidence interval is to decrease the confidence level.* In the above example, since $P(-1.645 < Z < 1.645) = 0.90$, a 90% confidence interval for p is $(\bar{y} - 1.645se, \bar{y} + 1.645se) = (0.23, 1.11)$, which is narrower than the 95% confidence interval, namely (0.14, 1.20).

3. *A better way to get a narrower confidence interval is to increase the sample size.* E.g., for the width of a 95% confidence interval to be 0.02, we need $2 \cdot 1.96se \leq 0.02$, so $n \geq \bar{y}(1 - \bar{y})(1.96/0.01)^2$. But $\bar{y}(1 - \bar{y}) \leq 0.25$ (since $0 \leq \bar{y} \leq 1$), so it is enough to take $n \geq 0.25(1.96/0.01)^2 = 9604$. In general, for a $1 - \alpha$ confidence interval of width at most d, it is enough to have $n \geq (z_{\alpha/2}/d)^2$. ($z_{\alpha/2}$ is the point to the right of which the area under the standard normal pdf is $\alpha/2$.)

ESTIMATING AND COMPARING MEANS

I. ESTIMATING MEANS

1. Suppose we want to estimate the mean IQ in the population of UW-Madison students. We randomly select a sample of n students. To each member i of the sample (i = 1, 2, …, n) corresponds a random variable Y_i whose distribution is the same as the distribution of IQ scores in our population. Call the mean of the distribution μ (this is the parameter to be estimated) and its variance σ^2. If n is large (in practice, n > 30), by the Central Limit Theorem the sample mean $\bar{Y} = (Y_1 + \ldots + Y_n)/n$ is approximately normal with mean μ and variance σ^2/n. So 0.95 = $P(-1.96 < \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} < 1.96) = P(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}})$, and a 95% confidence interval for μ is $(\bar{y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\frac{\sigma}{\sqrt{n}})$, where $\bar{y}$ is the measured value of $\bar{Y}$. Estimate σ^2 by the value s^2 of the <u>sample variance</u> $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} Y_i^{\,2} - n\bar{Y}^2)$ to get $(\bar{y} - 1.96\frac{s}{\sqrt{n}}, \bar{y} + 1.96\frac{s}{\sqrt{n}})$ as a 95% confidence interval for μ.

2. For a small sample, use this result: <u>if the population distribution is normal, then the random variable $T = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$ has the *t distribution* with $n - 1$ *degrees of freedom*</u>. So a 1 - α small-sample confidence interval for μ is $(\bar{y} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{y} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}})$, where $t_{\alpha/2,n-1}$ is the value (obtained from the table of the t distribution) such that $P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1 - \alpha$. For example, $t_{0.25,9}$ = 2.26.

II. HYPOTHESIS TESTING

1. Suppose we want to find out whether the mean IQ μ of UW-Madison students *differs* from the national average of 100. In other words, we want to *test the hypothesis* that μ = 100 (the *null hypothesis*, denoted by H_0; i.e., the hypothesis that there is no difference, that the difference is "null") *against the alternative hypothesis* (denoted by H_1) that $\mu \neq 100$. A way to perform this test is by computing a 95% confidence interval for μ and checking whether if contains 100: if it does, the null hypothesis is *accepted* (and the alternative hypothesis is *rejected*); if it does not, the alternative hypothesis is accepted (and the null hypothesis is rejected). This amounts to computing the value t of the random variable $T = \frac{\bar{Y}-100}{S/\sqrt{n}}$ (called the *test statistic*) and seeing whether its absolute value $|t|$ exceeds the *critical value* $t_{\alpha/2,n-1}$.

2. Suppose now we want to find out whether the mean IQ μ of UW-Madison students is *greater* than the national average of 100. In other words, we want to test the null hypothesis H_0 that μ = 100 against the *one-sided* alternative hypothesis H_2 that μ > 100 (because the possibility that μ < 100 is so remote that we are not interested in it). (By contrast, the previous alternative hypothesis, that $\mu \neq 100$, was *two-sided*.) Here it will not do to compute a confidence interval for μ, since confidence intervals are symmetric and thus correspond to two-sided alternative hypotheses. But we can still compute the value t of the test statistic $T = \frac{\bar{Y}-100}{S/\sqrt{n}}$ and see whether t (instead of $|t|$) exceeds the critical value $t_{\alpha,n-1}$ (instead of $t_{\alpha/2,n-1}$). The idea is that, *supposing H_0 is true*, it is improbable that t would be so far away from 0 as to exceed the critical value; so if it does exceed it, H_0 is rejected (otherwise, H_2 is rejected).

3. This way of testing hypotheses can lead to two errors. A *<u>type I error</u>* occurs when *a true null hypothesis is rejected*, and a *<u>type II error</u>* occurs when a *true alternative hypothesis is rejected.* (The other two possibilities, namely accepting a true null hypothesis or accepting a true

alternative hypothesis, are *not* errors. We are assuming that either the null or the alternative hypothesis is true.) The significance level α used to compute the critical value is the probability of a type I error: in the two-sided example, P_{H_0}(reject H_o) = $P_{\mu = 100}(\frac{|\bar{Y}-\mu|}{S/\sqrt{n}} > t_{\alpha/2,n-1}) = \alpha$.

4. *A type II error occurs when the sample size is small*: even if, e.g., $\mu \neq 100$, for a small sample the confidence internal is wide, so the interval may still include 100. We say then that the test does not have sufficient *power* to discriminate the two hypotheses. Formally, if β is the probability of a type II error, namely P_{H_1}(reject H_1), the power of the test is 1 - β, namely P_{H_1}(accept H_1). To increase power, increase the sample size.

III. COMPARING MEANS

1. Suppose we want to find out whether a new teaching method improves learning. One way to do this is by taking n pairs of identical twins and randomly assigning one twin in each pair to the new teaching method and the other twin to the old method. Then we give everyone a test to measure how much they have learned. Let the scores of those taught by the new method be X_i and the scores of those taught by the old method be Y_i (i = 1, ..., n; the same i corresponds to the two twins in the same pair). We want to test the null hypothesis H_0: $\mu_X = \mu_Y$ against the alternative hypothesis H_1: $\mu_X > \mu_Y$. Since the two samples (each of size n) are *paired* rather than independent, in effect we have a single sample of n pairs, so we can consider the *differences* $D_i = X_i - Y_i$ and test H_0: $\mu_D = 0$ (i.e., $\mu_X - \mu_Y = 0$) against H_1 : $\mu_D > 0$ (i.e., $\mu_X - \mu_Y > 0$). If X_i and Y_i are normal, then so is D_i, so we can use the t statistic to perform the test, just as in the one-sample case.

2. Identical twins are hard to come by, however, so an alternative way to find out if the new teaching method improves learning is by taking $n + m$ unrelated people and randomly assigning n of them (the *experimental group*) to the new teaching method and the remaining m of them (the *control group*) to the old teaching method. Again, we want to test H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X > \mu_Y$. Here is the crucial result: <u>if X_i and Y_i are normal and independent, then the random variable $T = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$</u> (where $S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{m}(Y_i-\bar{Y})^2}{n+m-2}$ is the *pooled variance*) <u>has a t distribution with $n + m - 2$ degrees of freedom</u>. (Strictly speaking, the result also assumes that X_i and Y_i have the same variances. There are statistical tests one can perform to check whether the assumption holds.)

3. For the purpose of finding out whether the new teaching method improved learning *in the n + m people participating in the experiment*, those people need not have been randomly selected from a large population. And even if they were randomly selected, the immediate purpose of the experiment (and of the statistical test) is to *compare the mean scores of the two independent samples*, not to make an inference about the mean score of a larger population. In this respect hypothesis testing differs importantly from estimation.

GOODNESS OF FIT

I. NULL HYPOTHESIS: MULTINOMIAL DISTRIBUTION

1. Just as a *Bernoulli* process consists of repeated IID trials each of which has *two* possible outcomes, a *multinomial* process consists of repeated IID trials each of which has *many* (say k) possible outcomes; e.g., repeatedly throwing a fair die (six possible outcomes at each trial). Just as a binomial distribution corresponds to the numbers of successes and failures in n trials of a Bernoulli process, a multinomial distribution corresponds to the numbers of occurrences of each possible outcome (e.g., each side of the die) in n trials of a multinomial process; e.g., Y_i is the number of times side i comes up in n throws (so $Y_1 + \ldots + Y_6 = n$). If p_i is the probability that side i comes up at any trial, $p_1 + \ldots + p_6 = 1$. $P(Y_1 = y_1, \ldots, Y_k = y_k) = \frac{n!}{y_1! \ldots y_k!} p_1^{y_1} \ldots p_k^{y_k}$.

2. Suppose we want to test the null hypothesis that the die is fair, namely that it corresponds to a multinomial distribution with $p_1 = \ldots = p_6 = 1/6$. We use this result: <u>if the distribution of $Y_1, \ldots, Y_k$ is multinomial with parameters $n, p_1, \ldots, p_k$, then the random variable</u> $C = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}$ <u>has approximately a χ^2 (chi square) distribution with k - 1 degrees of freedom</u>. (The approximation is good if $np_i \geq 5$ for each i or if $n > 5k$.) If the value of C exceeds the critical value obtained from the table of the χ^2 distribution for the desired level of confidence, the null hypothesis is rejected.

3. <u>Example</u>*:* We toss a die 90 times, and we get side 1, 2, 3, 4, 5, 6 respectively 16, 19, 15, 14, 12, 14 times. The null hypothesis that the die is fair gives $np_1 = \ldots = np_6 = 90 \cdot 1/6 = 15 > 5$, so we can apply the χ^2 test. The value of C is: $[(16 - 15)^2 + (19 - 15)^2 + (15 - 15)^2 + (14 - 15)^2 + (12 - 15)^2 + (14 - 15)^2]/15 = 28/15 = 1.87$. From the table of the χ^2 distribution, the critical value for $\alpha = 0.05$ and 5 (= 6 - 1) degrees of freedom is 11.1 > 1.87, so the null hypothesis is not rejected.

II. NULL HYPOTHESIS: INDEPENDENCE

1. To test the null hypothesis that men and women are equally likely to smoke (i.e., the variables of sex and smoking are independent), we select a random sample of 100 people.

	Smokers	Non-smokers	Total
Men	16	36	52
Women	11	37	48
Total	27	73	100

Table 1. Observed frequencies.

	Smokers	Non-smokers	Total
Men	0.14	0.38	0.52
Women	0.13	0.35	0.48
Total	0.27	0.73	1.00

Table 2. Expected probabilities under H_0.

From Table 1, we estimate the probabilities of being a man as 0.52 and a smoker as 0.27. If the two variables are independent, then, e.g., the probability of being *both* a man and a smoker is $0.52 \cdot 0.27 = 0.14$ (Table 2), so 14 of the 100 people are expected to be both men and smokers.

2. The null hypothesis that the two variables are independent amounts to the hypothesis that the "pair" of variables has a *multinomial* distribution with four possible outcomes (smoking man, non-smoking man, smoking woman, non-smoking woman) and probabilities given in Table 2. So we can use the χ^2 test, but we have already "used up" two degrees of freedom to estimate the probabilities of being a man and of being a smoker, so the degrees of freedom to be used in the test are 4 - 1 - 2 = 1. (In general, <u>if the data are arranged in r rows and c columns, the number of degrees of freedom is $(r - 1)(c - 1)$</u>.) For $\alpha = 0.99$, the critical value is 6.63. The value of the χ^2 statistic is $\frac{(16-14)^2}{14} + \frac{(36-38)^2}{38} + \frac{(11-13)^2}{13} + \frac{(37-35)^2}{35} = 0.813 < 6.63$, so H_0 is not rejected.

3. The χ^2 test is to be used only for *categorical* (not *numerical*) variables, namely variables whose possible values fall into non-numerical categories (e.g., man vs. woman, heads vs. tails).

BAYESIAN STATISTICAL INFERENCE

I. BAYESIAN CRITICISMS OF CONFIDENCE INTERVALS

1. Suppose one plans to randomly select a sample of size 100 from a normal population with (unknown) mean μ and (known) standard deviation 10. Then $P(\bar{Y} - 1.96 < \mu < \bar{Y} + 1.96) = 0.95$, where the probability can be either present chance or present rational credence. Suppose next one selects a sample, gets $\bar{y} = 5$, and constructs the confidence interval (3.04, 6.96). Bayesians claim that constructing this interval is pointless, since it is false that $P(3.04 < \mu < 6.96) = 0.95$. This is indeed false if the probability is present chance: the chance is 1 or 0, depending on whether μ is or not between 3.04 and 6.96. But why is it false if the probability is present rational credence?

2. Bayesians say it is fallacious to infer $P(3.04 < \mu < 6.96) = 0.95$ from the premises $P(\bar{Y} - 1.96 < \mu < \bar{Y} + 1.96) = 0.95$ and $\bar{Y} = 5$, just as it is fallacious to infer P(4 is odd) = 0.5 from the premises P(the result of throwing the die is odd) = 0.5 and "The result of throwing the die is 4". Moreover, Bayesians grant that if one randomly selects many samples one can expect about 95% of the confidence intervals one constructs to include μ, but say it is fallacious to infer from this claim about the *procedure* one uses a probability claim about the *confidence interval* one constructs.

3. Is it really fallacious, however? Just as the argument from "This card was selected according to a procedure that had a 50% chance of selecting a red card" to "This card is red" has inductive probability 0.5, the argument from "The confidence interval (3.04, 6.96) was constructed according to a procedure that had a 95% chance of constructing a confidence interval including μ" to "The confidence interval (3.04, 6.96) includes μ" has inductive probability 0.95. Proponents of confidence intervals typically do not talk about inductive probabilities, but what *they typically say* may differ from what *can be justifiably said* about confidence intervals.

4. Bayesians also note that there are multiple ways to construct confidence intervals: one could use various statistics or construct non-symmetric intervals. This observation does not undermine the practice of constructing confidence intervals; it just calls for justifying aspects of the practice.

II. BAYESIAN CRITICISMS OF HYPOTHESIS TESTING

1. <u>What does it mean to accept or reject a hypothesis?</u> Bayesians claim that no answer to this question works. (a) Is to reject a hypothesis to *regard it as definitely false*? (This interpretation is suggested by the advice of refusing to say that a hypothesis is accepted when it is not rejected, by analogy with the falsificationist advice of refusing to say that a hypothesis is verified when it is not falsified.) No: it is possible to reject a true hypothesis. (b) Is to reject a hypothesis with α = 0.05 to *regard it as less than 5% probable*? No: such an inference is unwarranted. (c) Is to reject a hypothesis to *decide to act as if it were false*? No: how one decides to act depends on further factors. (If I reject a hypothesis, I may stop investigating it, but I will not bet my entire fortune that it is false.) However, there is a plausible interpretation that Bayesians typically neglect: (d) *To reject a hypothesis is to believe that it is false* (i.e., to disbelieve it). This implies neither that one regards it as definitely false, nor that one regards it as less than 5% probable, nor that one decides to act as if it were false: binary belief is associated with a *range* of degrees of belief.

2. <u>Isn't the null hypothesis always false (so that testing is *redundant*)?</u> "Are the effects of A and B different? They are always different—for some decimal place." But some null hypotheses are true: it often happens that the mean teaching evaluations for two courses taught a given term at a given department are exactly the same. Moreover, the objection does not establish that any empirical null hypothesis is a priori false: no matter how *unlikely*, it is still *possible* that the effects of A and B are *not* different—for *any* decimal place. Maybe, more charitably, the

objection is that H_0 is always *very unlikely*, so one should instead focus on hypotheses like "μ is *approximately* 3". But if data on the basis of which $\mu = 3$ is rejected provide evidence against $\mu = 3$, they also provide evidence against "μ is approximately 3".

3. <u>Wouldn't the null hypothesis always be rejected with a large enough sample?</u> This objection relies on the claim that, if e.g. the null hypothesis is H_0: $\mu = 3$, then even if $\bar{y}$ is 3.001 (i.e., very close to 3), there is always a large enough sample size n such that $(3.001 - 3)/(s/\sqrt{n}) > 1.96$, so H_0 will be rejected. But it is fallacious to infer from this claim that every null hypothesis will be rejected if n is large enough: the claim takes $\bar{y}$ as fixed and increases n, but as n increases $\bar{y}$ may get closer to μ. E.g., Pearson tossed a coin 24,000 times and got 12,012 heads (50.05%), failing to reject the null hypothesis that the coin was fair. The reply that he would had rejected the null hypothesis if he had tossed the coin 24,000,000 times and obtained 50.05% heads misses the point: maybe he would have gotten 12,000,010 heads (50.00004%). No matter how large the sample will be, we do not know a priori that any null hypothesis will be rejected.

4. <u>Aren't some rejected null hypotheses very probably true?</u> Suppose we know that a population percentage p is either 0.4 or 0.6 and we get $\bar{y} = 0.401$ but the sample is so large that H_0: $p = 0.4$ is rejected. This seems wrong: the value of 0.401 makes it very probable that p is 0.4 (given that p is either 0.4 or 0.6). Moreover, if one had designated $p = 0.6$ as the null hypothesis, the very same data (i.e., $\bar{y} = 0.401$) would have led one to reject $p = 0.6$ (instead of rejecting $p = 0.4$). These criticisms, however, work only against the (practically nonexistent) cases in which H_0 and H_1 are *simple*, not against the (standard) cases in which H_1 is composite (e.g., $\mu > 3$ or $\mu \neq 3$).

5. <u>Isn't the logic of null hypothesis testing fallacious?</u> Let D (for "data") be the proposition that an extreme (i.e., higher than the critical value) value of the test statistic (e.g., the sample mean) was obtained. The argument behind null hypothesis testing, namely the argument from $P(D|H_0) = 0.05$ (i.e., $P(\sim D|H_0) = 0.95$) and D to $\sim H_0$ is an instance of probabilistic modus tollens, which is *not* always inductively strong. It is true that the argument from $P(H_0|D) = 0.05$ and D to $\sim H_0$ is always strong (it has inductive probability 0.95, by probabilistic modus ponens), but it is fallacious to infer $P(H_0|D) = 0.05$ from $P(D|H_0) = 0.05$. Classical statisticians grant this but reply that *in practice* $P(H_0|D)$ and $P(D|H_0)$ are *highly correlated*, so that it is unfair to focus on contrived cases in which they are very different. Compare: using Newtonian mechanics is justified for speeds low relative to the speed of light, where it gives good approximations.

III. A BAYESIAN ALTERNATIVE TO CLASSICAL STATISTICS

1. Bayesian statistical inference starts with the assignment of *prior probabilities* to hypotheses (which classical statisticians avoid). These prior probabilities are *credences*; usually they are not rationally *required*, but they must be rationally *permitted* (e.g., they must satisfy the probability axioms). Typically many assignments of prior probabilities are rationally permitted, and the arbitrariness of any such assignment is a standard objection to Bayesianism. Bayesians reply that typically *it does not matter* what prior probabilities one starts with: as evidence accumulates, people who start with different prior probabilities end up with similar posterior probabilities.

2. The main component of Bayesian statistical inference is the application of Bayes' theorem to compute the *posterior probability* of a hypothesis H given the evidence E. But to compute the posterior probability $P(H|E)$, one needs not only $P(E|H)$ and the prior probability $P(H)$, but also $P(E|H^c)$. Sometimes $P(E|H^c)$ is available (e.g., P(test positive|patient does not have AIDS)), but often it is unavailable (e.g., P(deflection of sunlight|General Theory of Relativity is false).

3. Another worry is that Bayesian statistical inference seems to make no difference *in practice*. Bayesians talk of "credible intervals" instead of "confidence intervals" and of posterior probabilities instead of accepting or rejecting hypotheses, but they still (dis)believe certain hypotheses; must these hypotheses be different from those that classical statisticians (dis)believe?

DECISION THEORY

I. DECISION PROBLEMS

1. Informally, a decision problem is an agent's problem of choosing among alternative courses of action at a given time. For example, the problem of choosing (i.e., deciding) whether to satisfy a friend's request to lend her $1,000.

2. Formally, a decision problem has three components. (1) A set of possible *actions* (e.g., lend the money vs. not lend the money). (2) A set of possible *states* of the world on which the consequences of the agent's possible actions depend (e.g., the friend returns the money vs. does not return the money if you lend it). (3) A set of *outcomes* associated with each combination of an action and a state (e.g., losing $1,000 if you lend the money but the friend does not return it).

3. It is convenient to take actions, states, and outcomes to be *propositions* (e.g., the proposition that you lend the money). The actions must be mutually exclusive and collectively exhaustive, and so must be the states. The outcomes must include *all* relevant consequences of the actions.

II. EXPECTED UTILITY MAXIMIZATION

1. The *expected utility* of an action *A*, denoted by EU(*A*), is the sum, over all states, of the product of the *probabilities* (*rational credences*) of the states with the values (or *utilities*) of the outcomes. Utilities are often understood as *monetary values*. For example, suppose you are offered to pay $1,000 for the following gamble: a fair coin will be tossed 5 times, and you will get nothing if the coin comes up heads all 5 times (with probability $0.5^5 = 0.03125$), but you will get $10,000 otherwise (with probability $1 - 0.5^5 = 0.96875$). The expected utility of refusing to play is 0, and the expected utility of playing is 0.03125·(-$1,000) + 0.96875·$9,000 = $8,678.5.

2. According to the principle of *expected utility maximization* (EUM), an agent is rationally required to choose an action that maximizes (i.e., has highest) expected utility. So it seems that, according to EUM, in the above example you should pay and play.

3. Things are not so simple, however. $1,000 may be a lot of money for you, so it need not be irrational to balk at the small chance (around 3%) that you will lose it. Moreover, the St. Petersburg game (which gives you $\$2^n$ if a coin first comes up heads at the *n*th toss and 0 otherwise) has infinite expected utility, but it need not be irrational to refuse to pay even $100 to play, even if $100 is not a lot of money for you. Also, suppose you are offered a bet that gives you $1,000,000,000,000 with probability 0.001 but requires you to pay $1,000,000 with probability 0.999; it seems clearly irrational to accept the bet, although its expected utility is $999,001,000. Finally, in many cases non-monetary values are relevant (e.g., the value of helping your friend by lending her the money); how are utilities defined in such cases?

III. CLASICAL EXPECTED UTILITY THEORY

1. To answer these objections, here is a theoretical justification for EUM: it can be shown that, if your *preferences* among propositions satisfy certain *rationality constraints*, then there is a unique probability measure on states and a utility function (unique up to the arbitrary choice of a unit and a zero point) on outcomes such that you prefer action A to A' exactly if EU(A) > EU(A'), and you are indifferent among A and A' exactly if EU(A) = EU(A'). This result is a *representation theorem*: your preferences can be *represented* by an expected utility function. If you are rational, you always choose *as if* you were maximizing expected utility.

2. So the usual introductory presentations of EUM are misleading. EUM is *not* a *decision procedure*, a way of finding out which action to choose: EUM does *not* require an agent to

consciously assign probabilities to states and utilities to outcomes and compute the expected utilities of actions with the goal of maximizing expected utility. Instead, EUM requires that an agent's preferences *be compatible with the existence* of a probability and a utility function such that the corresponding expected utility function represents those preferences. More specifically, EUM requires that an agent's preferences over actions satisfy the rationality constraints specified in a representation theorem. This does not mean that expected utility calculations are useless: often they roughly correspond to the expected utilities that represent one's preferences.

IV. THE ALLAIS PARADOX

1. An integer from 1 to 100 will be randomly selected. You are given a choice between C and D:
   C: You get \$500,000 if the integer is from 90 to 100 (prob. 0.11), otherwise you get nothing (prob. 0.89).
   D: You get \$2,500,000 if the integer is from 91 to 100 (prob. 0.10), otherwise you get nothing (prob. 0.90).

2. Now you are given a choice between F and G:
   F: You get a gift of \$500,000 (no strings attached, prob. 1).
   G: You get \$2,500,000 if the integer is from 91 to 100 (prob. 0.10), you get \$500,000 if the integer is from 1 to 89 (prob. 0.89), and you get nothing if the integer is 90 (prob. 0.01).

3. Most people prefer D to C and F to G, but the expected monetary values are \$55,000 for G, \$250,000 for D, \$500,00 for F, and \$695,000 for G. One can appeal to the following rationality constraint to argue that these preferences are irrational: if a rational agent's conditional preferences between *A* and *A′* and between *B* and *B′* given any state are the same, then the agent's unconditional preferences between *A* and *A′* and between *B* and *B′* are also the same.

Selected integer	1-89	90	91-100
Option C	0	\$500,000	\$500,000
Option D	0	0	\$2,500,000
Option F	\$500,000	\$500,000	\$500,000
Option G	\$500,000	0	\$2,500,000

V. EVIDENTIAL EXPECTED UTILITY THEORY

1. Another rationality constraint is the *dominance principle* (or *sure-thing principle*): if a rational agent prefers *A* over *A′* conditionally on some states and does not prefer *A′* over *A* conditionally on any state, then the agent unconditionally prefers *A* over *A′*. E.g., assuming option X gives you \$10 if the coin comes up heads and nothing otherwise but option Y gives you \$100 if the coin comes up heads and nothing otherwise, you should prefer Y to X.

2. *Objection*: Should you spend the night partying or studying for tomorrow's exam? Conditionally on passing the exam, you prefer partying to studying. Conditionally on failing, you prefer partying to studying. According to the dominance principle, then, you should prefer partying to studying. But this reasoning ignores the fact that you are much more likely to pass if you study than if you party. The standard reply is that the dominance principle does not apply in such cases: classical expected utility theory assumes that the probabilities of the states do not depend on the agent's actions. But then classical expected utility theory is seriously incomplete.

3. To avoid this problem, maximize not expected utility, defined as $\mathrm{EU}(A) = \Sigma_s P(S)u(O[A, S])$ (where $O[A, S]$ is the outcome of action A under state S), but rather *evidential expected utility*, defined as $\mathrm{EEU}(A) = \Sigma_s P(S|A)u(O[A, S])$. A representation theorem can be proven.

VI. NEWCOMB'S PARADOX

1. A statement of the paradox by Joyce:
   Suppose there is a brilliant (and very rich) psychologist who knows you so well that he can predict your choices with a high degree of accuracy. One Monday as you are on the way to the bank he stops you, holds out a thousand dollar bill, and says: "You may take this if you like, but I must warn you that there is a catch. This past Friday I made a prediction about what your decision would be. I deposited \$1,000,000 into your

> bank account on that day if I thought you would refuse my offer, but I deposited nothing if I thought you would accept. The money is already either in the bank or not, and nothing you now do can change this fact. Do you want the extra $1,000?" You have seen the psychologist carry out this experiment on two hundred people, one hundred of whom took the cash and one hundred of whom did not, and he correctly forecast all but one choice. There is no magic in this. He does not, for instance, have a crystal ball that allows him to "foresee" what you choose. All his predictions were made solely on the basis of knowledge of facts about the history of the world up to Friday. He may know that you have a gene that predetermines your choice ...

2. Given that whether you take the $1,000 has no causal effect on what amount is already in your bank account, it seems irrational to refuse the $1,000. But here is a standard objection (Sugden):

> Imagine two people, irrational Irene and rational Rachel, who go through the experiment. Irene [refuses the money] and wins $1 million. Rachel [takes the money] and wins $1,000. Rachel then asks Irene why she didn't [take the extra thousand]; surely Irene can see that she has just thrown away $1,000. Irene has an obvious reply: "If you're so smart why ain't you rich?" This reply deserves to be taken seriously. ... The relevant difference between Irene and Rachel is that they reason in different ways. As a result of this difference, Irene finishes up with $1 million and Rachel with $1,000. Irene's mode of reasoning has been more successful ... So, are we entitled to conclude that, nevertheless, it is Rachel who is rational?

3. Irene's reply changes the subject. Rachel could reply: "My question was why *you* didn't take the money. I know why *I* am not rich: because I am not the kind of person the psychologist thinks will refuse the money. Given that I know I am the type who takes the money, the $1,000 was the most I was going to get, so the reasonable thing for me to do was to take it." Irene might respond: "But don't you wish you were like me, Rachel?" Rachel can grant that she wishes she were like Irene (i.e., the type who refuses the money), but this is not to endorse Irene's reasoning.

VII. CAUSAL DECISION THEORY

1. Evidential decision theory gives the wrong result in Newcomb's decision problem (i.e., that one should refuse the money): EEU(Refuse) = P(Predicted refusal|Refuse)·$1,000,000 + P(Predicted acceptance|Refuse)·0 = $1,000,000 > EEU(Accept) = P(Predicted refusal|Accept)· $1,001,000 + P(Predicted acceptance|Accept)·$1,000 = $1,000.

2. To avoid this problem, maximize not evidential expected utility, but rather *causal expected utility*, defined as CEU(A) = $\Sigma_s P^*(S|A)u(O[A, S])$, where $P^*(\bullet|A)$ is a probability measure reflecting your judgments about your ability to causally influence events by doing A. $P^*(S|A)$ is high *either* when you think that A will cause S *or* when you think that S is likely to hold whether or not A does. On a common proposal, $P^*(S|A) = P$(If I were to do A, S would hold).

3. Causal decision theory gives the right result in Newcomb's decision problem (i.e., one should accept the money): CEU(Refuse) = P^*(Predicted refusal|Refuse)·$1,000,000 + P^*(Predicted acceptance|Refuse)·0 = $1,000,000 = p·$1,000,000 < CEU(Accept) = P^*(Predicted refusal| Accept)·$1,001,000 + P^*(Predicted acceptance|Accept)·$1,000 = p·$1,001,000 + (1 - p)·$1,000.

VIII. SIMPSON'S PARADOX

1. Here are success rates and numbers of cured/treated cases for two treatments of kidney stones.

Kind of stones	Treatment A	Treatment B
Small stones	93% (= 81/87)	87% (= 234/270)
Large stones	73% (= 192/263)	69% (= 55/80)
Both	78% (= 273/350)	83% (= 289/350)

Treatment A seems more effective than B on small stones, and also on large stones, but *overall* B seems more effective than A. <u>Explanation</u>: Doctors tend to give the severe cases (large stones) the better treatment (A), and the milder cases (small stones) the inferior treatment (B).

2. To choose a treatment, should one consult the aggregated or the partitioned data? If one does not know the size of the stone, should one administer treatment B? According to causal decision theory, one should look at the causal story: the conditional probabilities are not enough.

CAUSAL REASONING

I. NECESSARY AND SUFFICIENT CONDITIONS

1. Watering plants causes them to grow in the sense that watering is *necessary* (i.e., *required*) for growth: in the absence of watering, no growth occurs. But watering is *not sufficient* (i.e., *not enough*) for growth: sunlight is also necessary.

2. Decapitation causes death in the sense that decapitation is *sufficient* for death: whenever decapitation occurs, death occurs. But decapitation is *not necessary* for death: death can occur without decapitation (e.g., drowning is also sufficient).

3. The action of a force causes a body to accelerate in the sense that the action of a force is *both necessary and sufficient* for acceleration: whenever a force acts, acceleration occurs, and whenever no force acts, no acceleration occurs.

4. What counts as necessary or sufficient *may vary with the circumstances*: heating water to 100^{o}C is in normal circumstances both necessary and sufficient for boiling the water, but is *not* necessary if one is at high altitude, and is *not* sufficient if the water contains impurities.

5. To *prevent* a phenomenon, look for a *necessary* condition: to eradicate yellow fever, exterminate the anopheles mosquito, since the mosquito causes (i.e., is necessary for) the spread of the disease. To *produce* a phenomenon, look for a *sufficient* condition: to increase muscular strength, exercise regularly, since exercise causes (i.e., is sufficient for) increasing strength.

II. PROBABILISTIC CAUSATION AND CAUSAL NETWORKS

1. Smoking causes lung cancer in the sense that smoking *increases the probability* of getting lung cancer. Smoking is *not* necessary for lung cancer: one can get lung cancer even if one never smokes. Smoking is *not* sufficient for lung cancer: not everyone who smokes gets lung cancer.

2. Suppose Smith shoots Jones because Jones slept with Smith's spouse; Jones is taken to surgery and suffocates after an allergic reaction to an anesthetic. The coroner may be interested in the *proximate* cause of death, namely suffocation. The prosecutor may be interested in the *salient* cause of death, namely the shooting. The psychiatrist may be interested in a *remote* cause of death, namely Smith's miserable childhood. All these causes are parts of a *causal network*.

3. *Singular causation* is causation of a single event (e.g., Caesar's death). *General causation* is causation of a class of events (e.g., deaths by lung cancer). Assuming that nature is *uniform*, singular and general causation are related: if an event *C* caused an event *E*, there must be a *causal law* to the effect that, in similar circumstances, events like *C* cause events like *E*.

III. MILL'S METHODS OF AGREEMENT AND DIFFERENCE

1. <u>The positive method of agreement</u> (eliminating features as not necessary): *If only one among the features under consideration is present in all observed positive instances of a phenomenon, then only that feature can be necessary for the phenomenon.* (The remaining features cannot be necessary, since each of them is absent in at least one positive instance of the phenomenon.)

2. <u>Example</u>: Here are the foods eaten by three people who got sick (+: eaten; - : not eaten).

Observed instance	Feature A: Spaghetti	Feature B: Steak	Feature C: Ice cream	Feature D: Orange juice	Phenomenon: Sickness
Alice	+	+	-	-	+
Bob	+	+	-	+	+
Charlie	-	+	+	+	+

Only the steak was eaten by everyone who got sick, so only the steak can be necessary for sickness. The ice cream cannot be necessary, since it was not eaten by Bob, who got sick. Similarly, the remaining foods can be eliminated as not necessary.

3. Limitations of the method. (a) Maybe none of the features under consideration is necessary: maybe the sickness was caused not by the steak, but by the use of dirty forks, a feature not in the list. (b). Maybe there is no single common cause of all observed positive instances: maybe Alice's and Bob's sickness was caused by the spaghetti, but Charlie's was caused by the ice cream. (c) Maybe the identified single common feature is *not* present in other, unobserved positive instances: maybe Derek did not eat the steak but also got sick. (d) It is very hard to find a *unique* common feature if the list of features is reasonably comprehensive: in addition to having all eaten steak, Alice, Bob, and Charlie also all used forks, drank water, etc.

4. The negative method of agreement (eliminating features as not sufficient): *If only one among the features under consideration is absent in all observed negative instances of a phenomenon, then only that feature can be sufficient for the phenomenon*. (The remaining features cannot be sufficient, since each of them is present in at least one negative instance of the phenomenon.) Example: Given the table below, only feature D can be sufficient for the phenomenon.

Instance	Feature A	Feature B	Feature C	Feature D	Phenomenon
1	+	+	-	-	-
2	+	-	+	-	-

5. The double method of agreement (eliminating features as not both necessary and sufficient): *If only one among the features under consideration is both present in all observed positive instances and absent in all observed negative instances of a phenomenon, then only that feature can be both necessary and sufficient for the phenomenon.* Example: Given the table below, only feature B can be both necessary and sufficient for the phenomenon.

Instance	Feature A	Feature B	Feature C	Feature D	Phenomenon
1	+	+	-	-	+
2	+	-	+	-	-

6. The method of difference is a special case of the double method of agreement in which *all feature columns except one consist only of + or only of -*. This method is used in controlled experiments. Example: Given the table below, only C can be both necessary and sufficient.

Instance	Feature A	Feature B	Feature C	Feature D	Phenomenon
1	+	-	+	+	+
2	+	-	-	+	-
3	+	-	-	+	-

7. To summarize, all four methods try to find a single feature column that has *exactly the same* pattern of + and - as the phenomenon column, but (a) the *positive* method of agreement considers only phenomenon columns with all +, (b) the *negative* method of agreement considers only phenomenon columns with all -, and (c) the *double* method of agreement and the method of difference consider only phenomenon columns with both + and -.

8. A complication: complex features. Maybe only the *disjunction* (or the *conjunction*, etc.) of two or more simpler features is necessary (or sufficient, or both) for a phenomenon. This can be accounted for by expanding the list of features so as to include logical combinations of simpler features. Example: Given the table, neither A nor B can be necessary, but their disjunction can.

Observed instance	Feature A: Studying hard	Feature B: Being very smart	Feature $A \cup B$: Studying hard *or* being very smart	Phenomenon: Succeeding
Alice	-	+	+	+
Bob	+	-	+	+

IV. THE METHOD OF CORRELATION (CONCOMITANT VARIATION)

1. The methods of agreement and difference assume that features and phenomena are either present or absent. But often they come in degrees: studying hard, being smart, and having successes can be present to a greater or lesser extent. If an increase or decrease in one variable is accompanied by an increase or decrease in another (e.g., studying more or less hard is accompanied by having more or fewer successes), there is a *correlation* between the two variables, and this indicates (but does not guarantee) a causal connection.

2. The *correlation coefficient* of two random variables X and Y is: $\rho(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}$. It can be shown that ρ is between -1 and 1. If $\rho > 0$, X and Y are *positively correlated*: as X increases, Y increases, and as X decreases, Y decreases. If $\rho < 0$, X and Y are *negatively correlated*: as X increases, Y decreases, and as X decreases, Y increases. If $\rho = -1$ or $\rho = 1$, there is a perfect linear relation between X and Y: with probability 1, $Y = aX + b$. If X and Y are independent, then $\rho = 0$. But if $\rho = 0$, X and Y need not be independent: Y may be a non-linear function of X (e.g., $Y = X^2$).

3. Correlation is symmetric, but causation is not. If X is correlated with Y, does X cause Y or does Y cause X? To answer this question, look at changes over time: if increases or decreases in X are *followed* by increases or decreases in Y, this suggests that X causes Y, since causes come *before* their effects. But there is no *guarantee* that X causes Y: the correlation may be *coincidental*. Or the correlation may be due to a *common cause* of X and Y. E.g., the correlation between falling barometers and stormy weather is due to a common cause: a sharp drop in atmospheric pressure.

4. *Longitudinal* (or *diachronic*) studies that find correlations between *changes* in values of variables over a time period usually provide stronger evidence for causation than *cross-sectional* (or *synchronic*) studies that find correlations between values of variables at a particular time. *Experimental* studies (especially *randomized* ones), in which an *intervention* is made (e.g., a drug is given), usually provide stronger evidence for causation than *observational* studies.

ANALOGICAL REASONING

I. REPRESENTING ANALOGICAL ARGUMENTS

1. Analogical reasoning is used all the time: judges decide how to apply the law by making analogies with how the law was applied in the past, scientists formulate hypotheses about the effects of chemicals on humans by analogy with their effects on animals, and so on.

2. An *analogical argument* has the following form:

(1) Source is similar to Target in certain respects.
(2) Source has some further feature *Q*.
So: (3) Target also has *Q* or some feature *Q** similar to *Q*.

3. *Tabular representation* of an analogical argument:

Domains:	Earth (Source)	Mars (Target)
Known similarities:	Has a moon (*P*)	Has moons (*P**)
Known dissimilarities:	Has surface water (*A*)	Has little surface water (~*A**)
Inferred similarity:	Supports life (*Q*)	Supports microbial life (*Q**)

4. The *horizontal relations* are the relations between Source and Target: the relations of *similarity* between *P* and *P**, and the relations of *dissimilarity* between *A* and ~*A**. The *vertical relations* are the relations between features of Source: the *prior association* between *P* and *Q*.

II. CLASSIFYING ANALOGICAL ARGUMENTS

The classification is based on the *nature* (inductive vs. deductive) and on the *direction* (from *P* to *Q*, from *Q* to *P*, both, or neither) of the prior association between *P* and *Q*.

Nature of association	Direction of association			
	Predictive (from *P* to *Q*)	Explanatory (from *Q* to *P*)	Functional (both directions)	Correlative (no direction)
Deductive	Mathematical	Abductive	—	—
Inductive	Predictive/Probabilistic	Abductive/Probabilistic	Functional	Correlative

One gets then six types of analogical arguments:

1. Mathematical: The three medians of any triangle have a common intersection. By analogy, the four medians of any tetrahedron have a common intersection.

2. Predictive/Probabilistic: Microbes have been found to thrive in frozen lakes in Antarctica and glaciers in Greenland. By analogy, there may be microbial life on Mars.

3. Abductive: The absence of force inside a hollow spherical shell is a consequence of, and thus can be explained by, the fact that the gravitational force between masses follows an inverse square law. By analogy, the absence of electrical influence inside a hollow charged spherical shell suggests that charges attract and repel each other with an inverse square force.

4. Abductive/Probabilistic: The predominance of useful traits among domesticated animals is explained by artificial selection (i.e., breeding). By analogy, the predominance of useful traits among animals in the wild is explained by natural selection.

5. Functional (inferring similarities in function from similarities in form): In addition to bowl-shaped lamps, carved from rock, inside which animal fat is burned, Inuit groups occasionally use flat, uncarved slabs that allow fuel to spill over the sides as makeshift lamps when traveling and pressed for time. By analogy, flat slabs bearing traces of burned fat found by archaeologists in Southern Europe had the same function during the Ice Age.

6. Correlative: Morphine is an effective painkiller and induces an S-shaped tail curvature in mice. By analogy, the observation (in 1934) that meperidine (know also known as Demerol) induced an S-shaped tail curvature in mice suggested that meperidine has painkilling properties.

III. EVALUATING ANALOGICAL ARGUMENTS

1. Commonsense guidelines.

- The more similarities (between Source and Target), the stronger the analogy. The more differences, the weaker the analogy.
- Analogies involving casual relations are more plausible than those not involving causal relations, and structural analogies are stronger than those based on superficial similarities.
- The *relevance* of the similarities and differences to the conclusion must be taken into account.
- The weaker the conclusion, the more plausible the analogy.

These guidelines are of limited use. How to count similarities? How to determine relevance?

2. A three-step procedure to evaluate analogical arguments.

Preliminary step: Represent the argument in tabular form (identify P, P^*, Q, Q^*).

First step: Formulate explicitly the prior association between P and Q and *evaluate* it. Is it valid (if deductive), is it strong (if inductive), is it a good explanation (if abductive), is there a high or at least a statistically significant correlation (if correlative)? If the prior association fails to satisfy these standards, then the analogical argument cannot be strong.

Second step: Determine which features are *relevant* to the evaluation of the argument, in the sense of playing an *essential* role in the prior association between P and Q. If the association is deductive, which premises are indispensable and which ones are redundant? If the association is predictive, which causal factors are important?

Third step: Assess the potential for *generalizing* the prior association. Do the essential features identified in the second step have analogues in Target that are known to hold, or at least not known not to hold? Are there reasons to believe that generalization might be *blocked*?

Source: P. Bartha, *By parallel reasoning: The construction and evaluation of analogical arguments* (Oxford University Press, 2010).