Notes on Probability Theory: The Logic of Science by Jaynes

March 26, 2018

Contents

1 Notation and Equations

1.1 Notation

- Greek letters (α, β, etc.) denote continuously variable parameters and Latin letters (a, b, etc) denote discrete indices or data values, unless otherwise stated
- Probabilities are denoted by capital P's, which signifies that the arguments are propositions

  $P(A|B)$
- Probabilities whose arguments are numerical values are denoted by other function symbols, such as

  $f(r|np)$
- Small p functions means the arguments can be propositions or numerical values

  $p(x|y)$ or $p(A|B)$ or $p(x|B)$
- Should be noted that this book only applies to finite sets of propositions
- *Kernel*: This means something different in probability theory. Form of the PDF in which all factors which are not functions of any of the variables in the domain are omitted.

1.2 General Equations

- Binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Fourier transform:

$$\mathscr{F}_i(\alpha) = \int_{-\infty}^{\infty} \mathrm{d}x\, f_i(x) e^{i\alpha x} \qquad f_i(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}\alpha\, \mathscr{F}_i(\alpha) e^{-i\alpha x}$$

- Power series:

$$\sum_{n=0}^{\infty} a_n (x-c)^n = a_0 + a_1 (x-c)^1 + a_2 (x-c)^2 + \dots$$

  with special cases, such as for exponentials:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

- Stirling's approximation:

$$\log(n!) \sim n\log(n) - n + \log\sqrt{2\pi n} + O\left(\frac{1}{n}\right)$$

- Gaussian error function:

$$\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} \mathrm{d}t = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} \mathrm{d}t$$

- Poisson distribution (probability that n counts will appear in one timestep):

$$p(n|l) = \frac{l^n}{n!} e^{-l} \qquad n = 1, 2, \dots$$

  where l is the sampling expectation value of n, $\langle n \rangle = l$
- Cauchy–Schwarz Inequality:

$$|\langle \hat{u}, \hat{v} \rangle|^2 \le \langle \hat{u}, \hat{u} \rangle \cdot \langle \hat{v}, \hat{v} \rangle$$

  where $\langle \dots, \dots \rangle$ denotes an inner product. So, it states that the square of the inner product of two vectors, $\hat{u}$ and $\hat{v}$, is less than or equal to the dot product of their inner products.
  a more common notation for probability theory is

$$|\langle XY \rangle|^2 \le \langle X^2 \rangle \langle Y^2 \rangle$$

  where X and Y are random variables

- Laplace transform (takes a function of a real variable t, usually time, to a function of a complex variable s, usually frequency):

$$F(s) = \int_0^\infty \mathrm{d}t\, f(t) e^{-st}$$

  where s is a complex number frequency parameter, $s = \sigma + i\omega$, with σ and ω being real numbers

2 Boolean Algebra

- AB – logical product or conjunction
  - both A and B are true, order does not matter
- A + B – logical sum or disjunction
  - at least one of A or B is true, order does not matter
- if A or B is true iff the other is true, they both have the same truth value
  - does not matter how it is established that they have the same truth value
  - leads to the most primative axiom of plausible reasoning: two propositions with the same truth value are equally plausible

2.1 Trivial identities of Boolean algebra

Idempotence: $AA = A$ $A + A = A$

Commutativity: $AB = BA$ $A + B = B + A$

Associativity: $A(BC) = B(AC) = ABC$ $A + (B + C) = (A + B) + C = A + B + C$

Distributivity: $A(B + C) = AB + BC$ $A + (BC) = (A + B)(A + C)$

Duality: If $C + AB$, then $\bar{C} = \bar{A} + \bar{B}$ If $D = A + B$, then $\bar{D} = \bar{A}\bar{B}$

- these trivial identities can be used to prove more important relations
- $A \Rightarrow B$
  - A implies B, does not assert that either A or B is true
  - same thing as $A\bar{B}$ is false
  - if A is false, it says nothing about B, and vice versa
  - means A and AB have the same truth value
- the three operations, conjunction, disjunction, and negation (the bar over the letter) are adequate to generate all logic functions of a single proposition
- Conditional probability: $A|B$
  - the conditional probabiltiy that A is true given that B is true
  - can use all the above identities

3 The quantitative rules

This chapter focuses on the deduction of quantitative rules for inference which follow these desiderata:

1. Representation of degrees of plausibility by real numbers
2. Qualtitative correspondence with common sense
3. Consistency

- $F(x, y)$, i.e. a Boolean function of propositions x and y (i'm fuzzy on this), must be a continuous monotonic increasing function of both x and y.
  - monotonic increasing function: a function which never decreases
  - $F_1(x, y) \equiv \frac{\delta F}{\delta x} \geq 0$
  - $F_2(x, y) \equiv \frac{\delta F}{\delta y} \geq 0$
  - equivalency only when x represents an impossibility
- Consistency requires that propositions be true regardless of association:
  - $F[F(x, y), z] = F[x, F(x, y)]$
  - This functional equation is a big deal in mathematics. Called "The Associativity Equation"
- Product rule:

  $w(AB|C) = w(A|BC)w(B|C) = w(B|AC)w(A|C)$
  - $w(x)$ must be continuous monotonic
- Sum rule (super long derivation which I didn't quite follow):

  $p(A|B) + p(\bar{A}|B) = 1$, or more generally,
  $p(A + B|C) = p(A|C) + p(B|C) - p(AB|C)$
- given several propositions A_i which are mutually exclusive, it can be shown that,

  $p(A_1 + \ldots + A_m|B) = \sum_{i=1}^{m} p(A_i|B)$ where $1 \leq m \leq n$
- if these propositions A_i are not only mutually exclusive but also exhaustive,

  $\sum_{i=1}^{m} p(A_i|B) = 1$
- we are given two sets of mutually exclusive propositions, $\{A_1, \ldots, A_n\}$ and $\{A'_1, \ldots, A'_n\}$, with the only differences between the two sets being the subscripts 1 and 2 being swapped in the prime set
  - If the given information B is the same between the two sets of propositions,

    $p(A_1|B)_I = p(A'_2|B)_{II}$
    and
    $p(A_2|B)_I = p(A'_1|B)_{II}$
  - If the information B is indifferent between propositions A_1 and A_2, and since we know that equivalent states of knowledge must be represented by equivalent plausibility, we can say

    $p(A_i|B)_I = p(A'_i|B)_{II}$
  - Therefore, $P(A_1|B)_I = p(A_2|B)_I$
  - this is a ''baby" version of the group invariance principle for assigning plausibilites
  - If the information B is indifferent between all propositions A_i, it can be shown that

    $p(A_i|B)_I = \frac{1}{n}$ where $1 \leq i \leq n$)
    * this is called the ''principle of indifference"
- the information given can determine numerical values of the quantities $p(x) = p(A_i|B)$, not the numerical values of the plausibilities $x = A_i|B$
- the plausibility $x \equiv A|B$ is an arbitrary monotonic function of p, defined in $(0 \leq p \leq 1)$
  - these functions p are called ''probabilities"

3.1 EXAMPLE: Bernoulli urn

- Prior information:
  - Ten balls of identical size and weight are in an urn
  - three balls (4, 5, and 6) are black, the rest are white
  - what is the probability that we draw a black one?
- Propositions:
  - $A_i \equiv$ the i^{th} ball drawn
- the probability to choose a particular ball is,

  $$p(A_i|B) = \frac{1}{10}$$
- the probability to choose a black ball is,

  $$p(\text{black}|B) = p(A_4 + A_5 + A_6|B)$$
- and since A_4, A_5, and A_6 are mutually exclusive,

  $$p(\text{black}|B) = \frac{3}{10}$$

4 Elementary sampling theory

Recall the basic rules:

- Product: $P(AB|C) = P(A|BC)P(B|C) = P(B|AC)P(A|C)$
- Sum: $P(AB) + P(\bar{A}|B) = 1$
- Extended sum: $P(A + B|C) = P(A|C) + P(B|C) - P(AB|C)$
- Principle of indifference: $P(H_i|B) = 1/N, 1 \leq i \leq N$ - iff the set H_i is exhaustive and mutually exclusive
- Bernoulli urn rule: $P(A|B) = M/N$ - iff B specifies that A is true on some subset of H_i and false on the remaining $N - M$

4.1 Sampling without replacement

EXAMPLE: Bernoulli urn reexamined

- Propositions
  1. $B \equiv$ An urn contains N balls, identical except that they are labeled sequentially and M of them are coloured red, with the remaining $N - M$ coloured white. We draw a ball, observe and record its colour, and do not replace it. This is done until n balls are drawn, $0 \leq n \leq N$
  2. $R_i \equiv$ Red ball on the ith draw
  3. $W_i \equiv$ White ball on the ith draw
- Since only red or white can be drawn, we know that $P(R_i|B) + P(W_i|B) = 1$
- The propositions are related by the negations $\bar{R}_i = W_i$ and $R_i = \bar{W}_i$
- So, the first draw is then defined by the probabilities: $P(R_1|B) = M/N$ and $P(W_1|B) = 1 - M/N$
- Subsequent draws can be derived from the product rule: $P(R_1R_2|B) = P(R_1|B)P(R_2|B)$, but need to take into account that the ball being drawn is not replaced.
  - Therefore, $P(R_1R_2|B) = \frac{M}{N}\frac{M-1}{N-1}$
  - can be extended to r draws: $P(R_1R_2 \ldots R_r|B) = \frac{M!(N-r)!}{(M-r)!N!}$, where $r \leq M$
- What is the probabiltiy of drawing exactly r red balls in n draws, regardless of order?

- must multiply by the bionomial coefficient: $\binom{n}{r} = \frac{n!}{r!(n-r)!}$, which represents the number of possible orders of drawing r red balls in n draws, called the multiplicity of the event r.
  - for example, to get three red in three draws, $\binom{3}{3} = 1$, can only happen in one way, $R_1R_2R_3$
  - However, to get two red in three draws, $\binom{3}{2} = 3$, can happen in three ways, $R_1R_2W_3$, $R_1W_2R_3$, and $W_1R_2R_3$.
- We can then derive an expression for drawing exactly R red balls in n draws, defined by the function $h(r|N,M,n) \equiv P(A|B)$

  $$h(r|N,M,n) = \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}}$$

  - This is called the hypergeometric distribution, often abbreviated as $h(r)$
- it can be demonstrated that the hypergeometric distribution is symmetric on exchange of M and n, i.e. $h(r|N,M,n) = h(r|N,n,M)$. So, the probability of drawing ten ball balls from an urn containing 50 red ones is the same as drawing 50 balls from an urn containing 10 red ones.
- generalized hypergeometric distribution:

  $$h(r_1 \dots r_k|N_1 \dots N_k) = \frac{\binom{N_1}{r_1} \dots \binom{N_k}{r_k}}{\binom{\sum N_i}{\sum r_i}}$$

  where there are k different colours of N balls, drawn $n = \sum r_i$ times

- we can find the most probable value of r by setting $h(r') = h(r'-1)$, and solving for r'. (this makes sense if you think about it like a distribution with a peak)
- the width of the distribution $h(r)$ gives an indication of the accuracy with which we can predict r
- Cumulative probability distribution: $H(R) \equiv \sum_{r=0}^{R} h(r)$, which is the probability of finding R or fewer red balls
  - $H(R)$ is a step function (think the first term in your first project)
- the median of the probability function is defined to be a numer m which has equal probabilities associated with $(r < m)$ and $(r > m)$ (think a transition state)
- What is the probability of drawing a red ball on the second draw, $P(R_2|B)$, without knowledge of the first draw's result?
  - we know that either R_1 or W_1 is true, so $R_2 = (R_1 + W_1)R_2 = R_1R_2 + W_1R_2$
  - applying the product rule, we get

    $$P(R_2|B) = P(R_1R_2|B) + P(W_1R_2|B) = P(R_2|R_1B)P(R_1|B) + P(R_2|W_1B)P(W_1|B)$$

  - but $P(R_2|R_1B) = \frac{M-1}{N-1}$ and $P(R_2|W_1B) = \frac{M}{N-1}$
  - so $P(R_2|B) = \frac{M-1}{N-1}\frac{M}{N} + \frac{M}{N-1}\frac{N-M}{N} = \frac{M}{N}$
  - notice that everything cancels out such that we have the same probability for red on the first and second draws.
  - this holds true generally, so the probability to draw red at any draw is the same, iff we do not know the result of any other draw

4.2 Logic vs Propensity

- since we know that knowledge of a earlier drawn ball's colour can change the probability of the current draw, can knowledge of a future ball's colour change the probability of the current ball?
- usually a fundamental axiom that future events can't change the probability of a current event
- but lets say we have an urn with two balls, one red and one white. the probability to draw a white ball with priors B is $P(W_1|B) = 1/2$. But what if we knew that the second draw was red? Then the probability becomes unity ($P(W_1|B) = 1$).
- So, while information about a later draw does not change the physical nature of the system, i.e. it does not change the number of balls of each colour in the system, it does change the state of knowledge of the system in the same way that knowledge of a past draw would

- this suggests that logical inference is fundamentally different from physical causation, e.g. physical influences propagate only forward in time while logical inferences propagate equally in either direction
- if the probability of an event is invariant under an permutation of the events (e.g. when the event took place), the probability distribution is called exchangeable. the hypergeometric distribution discussed above in the Bernoulli's urn example is exchangeable

4.3 Expectations

- definition: if a variable x can take on the set of values $(x_1, \ldots, x_n)$, in n mutually exclusive and exhaustive situations, and we assign probabilities $(p_1, \ldots, p_n)$, the quantity $\langle x \rangle = E(x) = \sum_{i=1}^{n} p_i x_i$ is the expectation balue of x
  - (you know this from quantum mechanics)
  - it is a weighted average of the possible values of x, weighted by their corresponding probabilities
- this brings us to a an easier way to discuss probabilities with prior knowledge of an event at some later time not known
  - if $F = M/N$ of red balls is known, then $P(R_1|B) = F$
  - if F is unknown, $P(R_1|B) = \langle F \rangle$

4.4 Binomial Distribution

- the hypergeometric distribution takes into account the changing nature of the urn, i.e. drawing a ball changes the contents to N-1.
- but what if $N \gg n$? the probability changes very little, and in the limit $N \to \infty$, this becomes negligible
- thus the hypergeometric distribution simplifies to the binomial distribution (through a derivation i don't care to type):

$$h(r|N, M, n) \to b(r|n, f) \equiv \binom{n}{r} f^r (1-f)^{n-r} \text{ where } M/N \to f$$

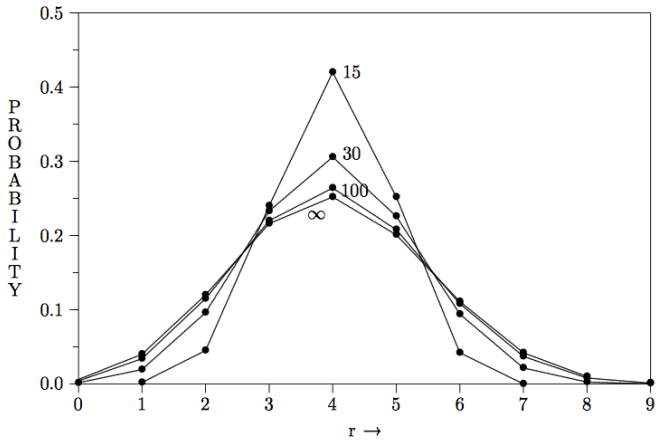

Fig. 3.1. The hypergeometric distribution for $N = 15, 30, 100, \infty$.

Figure 1: Comparing the hypergeometric distribution to the binomial distribution (the hypergeometric distribution in the $N \to \infty$ limit

- another limiting case, where $N_i \to \infty$, in such a way that $f_i \equiv \frac{N_i}{\sum N_j}$ = constant,

$$m(r_1 \ldots r_k | f_1 \ldots f_k) = \frac{r!}{r_1! \ldots r_k!} f_1^{r_1} \ldots f_k^{r_k} \text{ where } r \equiv \sum r_i$$

  - this is called the multinomial distribution

4.5 Sampling with replacement

- suppose instead of placing a drawn ball to the side, we instead put it back into the urn?
- with background information B', the probability of drawing two red balls in succession is $P(R_1R_2|B') = P(R_1|B')P(R_2|R_1B')$
  - clearly the first factor, $P(R_1|B')$ is still M/N, but what about the second factor?
  - (there is an interesting digression on reality vs models wrt randomization, on pp. 73-75. randomization is not truly something found in nature, what we mean by randomization is that no human is able to implicitly or explicitly influence the results)
  - if we assume the urn to be randomized upon replacement, information of R_1 is irrelevant to draw 2, so $P(R_2|R_1B') = P(R_2|B') = M/N$. this holds true generally for any draw.
  - the probability for drawing exactly r balls in n trials is simply $\binom{n}{r}\left(\frac{M}{N}\right)^r\left(\frac{N-M}{N}\right)^{n-r}$, which is just the binomial distribution
  - thus, the probability of drawing r red balls with replacement is the same as drawing r red balls without replacement in the limit $N \to \infty$
  - this approximation, however, can accumulate errors for large n
    * suppose that drawing and replacing a red ball increases the probability of drawing a red ball on the next draw by some small amount $\epsilon > 0$, while drawing a white ball decreases it by some small amount $\delta > 0$ (think of this as nonoptimal shaking of the urn)
    * letting C be the background information described above, we have the following probabilities:

$$P(R_k|R_{k-1}C) = p + \epsilon \qquad P(R_k|W_{k-1}C) = p - \delta$$
$$P(W_k|R_{k-1}C) = 1 - p - \epsilon \qquad P(W_k|W_{k-1}C) = 1 - p + \delta$$

      where $p \equiv M/N$ (this is referenced below)
    * from this, the probability of drawing r red balls and $(n-r)$ white balls in any order is $p(p+\epsilon)^c(p-\delta)^{c'}(1-p+\delta)^w(1-p-\epsilon)^{w'}$, where c is the number of red draws preceded by red ones, c' is the number of red preceded by white, w is the number of white preceded by white, and w' is the number of white preceded by red.
    * we can additionally see that $c + c' = \left[\begin{smallmatrix} r-1 \\ r \end{smallmatrix}\right]$ and $w + w' = \left[\begin{smallmatrix} n-r \\ n-r-1 \end{smallmatrix}\right]$, where the upper and lower cases are when the first draw is red or white, respectively.
      · when r and $(n-r)$ are small, ϵ and δ are negligible, and the expression simplifies to $p^r(1-p)^{n-r}$, as in the binomial distribution above
      · but as $r, n \to \infty$, we can use the relation $\left(1+\frac{\epsilon}{p}\right)^c \approx \exp\left(\frac{\epsilon c}{p}\right)$, so that the probability goes to $p^r(1-p)^{n-r}exp\left(\frac{\epsilon c - \delta c'}{p} + \frac{\delta w - \epsilon w'}{1-p}\right)$
      · thus, the probability now depends on the order, and depending on ϵ and δ, the deviation from the binomial distribution can be large
  - Let's see how this affects previous calculations
    * for the first draw we still have $p = P(R_1|C) = M/N$ and $q = 1 - p = P(W_1|C) = \frac{N-M}{N}$
    * but for the second trial we have $P(R_2|C) = p + (p\epsilon - q\delta)$, and $P(R_3|C) = p + (1 + \epsilon + \delta)(p\epsilon - q\delta)$ for the third. does $P(R_k|C)$ approach some limit as $k \to \infty$?
    * if we write the probabilities for the kth trial as a vector $V_k \equiv \left[\begin{smallmatrix} P(R_k|C) \\ P(W_k|C) \end{smallmatrix}\right]$, the set of the four probabilites above can be written in matrix form $V_k = MV_{k-1}$ where $M = \left[\begin{smallmatrix} p+\epsilon & p-\delta \\ q-\epsilon & q+\delta \end{smallmatrix}\right]$
      · This is a Markov chain of probabilities (!), and M is called the transition matrix.
      · So draws after the first can be expressed like $V_k = M^{k-1}V_1$.
      · To find a general solution, we need to find the eigenvectors and eigenvalues of M: $C(\lambda) \equiv \det(M_{ij} - \lambda\delta_{ij}) = \lambda^2 - \lambda(1 + \epsilon + \delta) + (\epsilon + \delta)$, so $\lambda_1 = 1$ and $\lambda_2 = \epsilon + \delta$
      · the eigenvectors are $x_1 = \binom{p-\delta}{q-\epsilon}$ and $x_2 = \binom{1}{-1}$, which are not orthogonal.
      · using these to define the transition matrix, $S = \left(\begin{smallmatrix} p-\delta & 1 \\ q-\epsilon & -1 \end{smallmatrix}\right)$, we can diagonalize M, $(S^{-1}MS)$, to eventually get to the general solution:

$$P(R_k|C) = \frac{(p-\delta) - (\epsilon+\delta)^{k-1}(p\epsilon - q\delta)}{1 - \epsilon - \delta}$$

- when $\epsilon = \delta = 0$, this simplifies to $P(R_k|C) = p$

* interesting to note that $\epsilon + \delta = 1$ iff $\epsilon = q$ and $\delta = p$ (since $0 \leq p \leq 1$, ϵ and δ must be bounded to the interval [-1,1]), the transition matrix simplifies to $M = \begin{bmatrix} p+\epsilon & p-\delta \\ q-\epsilon & q+\delta \end{bmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and no transitions occur
* likewise, if $\epsilon + \delta = -1$, $M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, and nothing but transitions occur, i.e. colours alternate after the first draw
* in the realistic case where $0 < |\epsilon + \delta| < 1$, the general solution attenuates exponentially with k to give the limit $P(R_k|C) \to \frac{p-\delta}{1-\epsilon-\delta}$
  - it is clear that this limiting distribution is not exchangeable, since the conditional probabilities depend on the separation $|k - j|$ of each draw.
  - let us consider $P(R_k|R_jC)$. the general solution is

$$P(R_k|R_jC) = \frac{(p-\delta) + (\epsilon+\delta)^{k-j}(q-\epsilon)}{1-\epsilon-\delta}$$

    where $j < k$
  - this approaches the limit above. since we have seen that $P(R_k|C) \neq P(R_j|C)$, it follows that $P(R_j|R_k|C) \neq P(R_k|R_jC)$, due to the product rule.
  - this is the ‘‘baby” form of the irreversible Markov process, where the current state depends only on the previous state and is asymmetric around the current state.

5 Elementary hypothesis testing

Reminder: the fundmental principle underlying all probabilistic inference:

> To form a judgement about the likely truth of falsity of any proposition A, the correct procedure is to calculate the probability that A is true, $P(A|E_1E_2\ldots)$, conditional on all the evidence, E_i, at hand.

5.1 Prior probabilities

- When we are given a problem, we have some set of data D, but we almost always have some other information X, which is all our past experience.
  - so all probabilities are conditional, at least in some part, on X, so the probability of proposition A is $P(A|DX)$
  - any probability that is conditional on X alone, $P(A|X)$ is called a *prior probability*
  - note that *prior* does not necessarily mean ‘‘earlier in time’’ it merely denotes information outside the scope of D
- four general principles for assigning priors:
  1. group invariance
  2. maximum entropy
  3. marginalization
  4. coding theory
- in sampling theory, the probabilities that arise presuppose the contents of the population, and seek to predict the data we would get from drawing from that population
  - however, most scientific inference requires us to look at problems the other way, alreading knowing the data and needing to know the population
  - for more on sampling theory, see Feller 1950, 1966 and Kendall & Stuart 1977
- generally in hypothesis testing, with a given data D, we want to know which set of hypotheses $\{H_1, H_2, \ldots\}$ is most likely true
  - begin with the notation:

    X = prior information
    H = some hypothesis to be tested
    D = the data

- using the product rule we can write the probability as:

$$P(DH|X) = P(D|HX)P(H|X) = P(H|DX)P(D|X)$$

- $P(H|DX)$ is simply a sampling distribution as seen earlier
- since we are looking for probabilities that are not conditional on H, but are still conditional on X, we need separate notations for them:

$$P(H|DX) = P(H|X)\frac{P(D|HX)}{P(D|X)} \tag{1}$$

  * $P(H|DX)$ is called a *posterior probability*, which again means logically later, not temporally later
  * $\frac{P(D|HX)}{P(D|X)}$ is called the *likelihood*, $L(H)$
  * the likelihood $L(H)$ is not itself a probability for $H_{\text{¡}}$ it is a dimensionless factor which may become a probability when multiplied by a prior probability and a normalization factor
- the above equation is an essential principle underlying all scientific inference. IF $P(H|DX)$ is close to 1 or 0, we know that the hypothesis H is likely true or false. If close to 1/2, we know that we need more information

5.2 Testing binary hypotheses with binary data

- the simplest nontrivial case of hypothesis testing is one where we have two hypotheses to test using two data values
- adapting eq. 1 to this, we can also write it as the probability that H is false:

$$P(\bar{H}|DX) = P(\bar{H}|X)\frac{P(D|\bar{H}X)}{P(D|X)}$$

- combining the two equations and getting the ratio, we get

$$\frac{P(H|DX)}{P(\bar{H}|DX)} = \frac{P(H|X)}{P(\bar{H}|X)}\frac{P(D|HX)}{P(D|\bar{H}X)}$$

  and the $P(D|X)$ term cancels
- the ratio of the probability that H is true to the probability that H is false is called the *odds*, $O(H|X)$ on the proposition H

$$O(H|DX) = O(H|X)\frac{P(D|HX)}{P(D|\bar{H}X)}$$

- so the postieror odds on H are equal to the prior odds multiplied by the likelihood ratio. this is a strictly monotonic function of the probability
- convenient to write the odds in terms of logarithms, because then we can add the terms. $\log_{10}$ is used due to historical purposes
- define a new function, calle the *evidence* for H given D and X: $e(H|DX) \equiv 10\log_{10} O(H|DX)$
  - still a monotonic function of the probability
  - base 10 multiplied by 10 puts it in terms of decibels (db)
- the evidence for H given D is the prior evidence plus the db provided by taking the log of the likelihood:

$$e(H|DX) = e(H|X) + 10\log_{10}\left[\frac{P(D|HX)}{P(D|\bar{H}X)}\right]$$

- if $D = D_1D_2\ldots$, the above equation is additive wrt the log of the likelihood. if $P(D_j|D_iHX) = P(D_j|HX)$, D_i and D_j are called logically *independent*. important to note that logical independence does not necessitate causal independence, and visa versa

e	O	p
0	1:1	1/2
3	2:1	2/3
6	4:1	4/5
10	10:1	10/11
20	100:1	100/101
30	1000:1	0.999
40	10^4:1	0.9999
$-e$	$1/O$	$1-p$

Figure 2: Table comparing the depiction of plausibilities in evidence, odds, and probabilities. Note that evidence provides a better intuitive depiction at $p \to 0, 1$

- if all D_i are logically independent given both (HX) and $(\bar{H}X)$, the evidence becomes

$$e(H|DX) = e(H|X) + 10\log_{10}\sum_i \log_{10}\left[\frac{P(D_i|HX)}{P(D_i|\bar{H}X)}\right] \tag{2}$$

- Example: industrial control

  $X \equiv$ 11 automatic machines turn out widgets which pour into 11 boxes. ten of the machines have a fail rate of one in six. the eleventh machine has a fail rate of one in three. the output of each machine is collected in an unlabeled box and stored

  - we choose one of the boxes and test a few of the widgets and classify them as ‘‘good’’ or ‘‘bad’’ based on the fail rate
  - Propositions

    $A \equiv$ bad batch (1/3 fail)

    $B \equiv$ good batch (1/6 fail)
  - prior information X told us there are only two possibilities, therefore A and B are related by negation: $\bar{A} = B$ and $\bar{B} = A$
  - by the principle of indifference (as we know that there are 11 machines and we do not know which made the batch), we know that $P(A|X) = 1/11$, so

$$e(A|X) = 10\log_{10}\frac{P(A|X)}{P(\bar{A}|X)} = 10\log_{10}\frac{(1/11)}{(10/11)} = -10 \text{ db}$$

    and by necessity, $e(A|X) = +10$ db
  - now lets pull out a widget and test for failure. if we pull a fail, what will that due to the evidence that this is a bad patch? Since we know that likelihoods are additive when the log is taken, this adds $10\log_{10}\frac{P(\text{bad}|AX)}{P(\text{bad}|\bar{A}X)}$ db to the evidence, where $P(\text{result}|AX)$ is the sampling distribution of the result given A
    - this results in the following probabilities:

$$P(\text{bad}|AX) = \frac{1}{3} \qquad P(\text{good}|AX) = \frac{2}{3}$$

$$P(\text{bad}|BX) = \frac{1}{6} \qquad P(\text{bad}|BX) = \frac{5}{6}$$

    - thus a bad draw will increase the evidence for A by $10\log_{10}\frac{(1/3)}{(1/6)} = 10\log_{10} 2 = 3$ db
    - pulling a second widget would update the probability according to the hypergeometric distribution (as this is sampling without replacement). but if we assume that N is very large, we can use the binomial distribution to sample and say that every bad draw will provide +3 db of evidence for hypothesis A
    - a good draw will provide $10\log_{10}\frac{P(\text{good}|AX)}{P(\text{good}|BX)} = 10\log_{10}\frac{(2/3)}{(5/6)} = -0.97 \text{ db} \approx -1$ db of evidence for hypothesis A

* so over all, if we draw n widgets, of which n_b are bad and n_g are good, the evidence that the batch is bad is

$$e(A|DX) = e(A|X) + 3n_b - 1n_g$$

* how do we eventually make a decision to reject or accept the batch?
  - if we say reject the batch if the evidence for A is $< +0$ db and accept if > -13 db, this is rejecting or accepting the hypothesis based on the posterior probability
  - decision making on the posterior probability is called *sequential inference* and denotes that the number of tests is not determined in advance and that the decision is based on the sequence on data values that are found

5.3 Noextensibility beyond the binary case

- unfortunately, the independent additivity over data (eq. 2) and linearity are not general rules
- one could always break up the analysis of a set of hypotheses into a set of binary comparisons to some null hypothesis, but this requires $n - 1$ calculations
- lets examine this reason for nonextensibility in the form of an exercise:
  - suppose we have a set of hypotheses $\{H_1, \ldots, H_n\}$ which are mutually exclusive and exhaustive on prior information X

$$P(H_iH_j|X) = P(H_i|X)\delta_{ij} \qquad \sum_{i=1}^{n} P(H_i|X) = 1$$

  - we have acquired m data sets $\{D_1, \ldots, D_m\}$, so we can write the probabilities as odds,

$$O(H_i|D_1, \ldots, D_mX) = O(H_i|X)\frac{P(D_1, \ldots, D_m|H_iX)}{P(D_1, \ldots, D_m|\bar{H}_iX)}$$

  - commonly, the numerator will factor due to the logical independence of D_j given H_i

$$P(D_1, \ldots, D_m|H_iX) = \prod_j P(D_j|H_iX), \qquad 1 \le i \le n$$

  - if the denominator can also factor in the same way, into

$$P(D_1, \ldots, D_m|\bar{H}_iX) = \prod_j P(D_j|\bar{H}_iX), \qquad 1 \le i \le n$$

    then the log form of the odds would again become independently additive over D_j
  - however true this is at $n = 2$, at $n > 2$, the factorization of the denominator reduces the problem into triviality, because at most one of the factors $\frac{P(D_m|H_iX)}{P(D_m|\bar{H}_iX)} \neq 1$

5.4 Multiple hypothesis testing

- remembering back to the industrial industrial control example, suppose we test 50 widgets and all are fails? according to the formula for $e(A|E)$ we would have 150 db of evidence for the proposition of it being a bad batch.
- common sense rejects this conclusion, as a bad batch has 1/3 being failures, not all of them
- how do we rectify this and make the testing more ‘‘skeptical’’ given an outcome which defies the logic of the original propositions?
- let’s reconsider the example:
  - let’s add a proposition C which denotes something going completely wrong with the machine and it gives us a 99% fail rate
  - let’s give this hypotheses a very low prior probability $P(C|X)$ of 10^{-6}, or -60 db.

– supposing we start out with these initial probabilities

$$P(A|X) = \frac{1}{11}(1-10^{-6}) \qquad P(B|X) = \frac{10}{11}(1-10^{-6}) \qquad P(C|X) = 10^{-6}$$

where A is a box with 1/3 defective, B is a box with 1/6 defective, and C is a box with 99/100 defective

– with $(1-10^{-6})$ being negligible, we start with the initial evidence values $A = -10$ db, $B = +10$ db, and $C = -60$ db

– with the data proposition D denoting that m widgets were drawn and were all fails, tehe posterior evidence for proposition C is

$$e(C|DX) = e(C|X) + 10\log_{10}\frac{P(D|CX)}{P(D|\bar{C}X)}$$

– if $N \gg m$, $P(D|CX) = \left(\frac{99}{100}\right)^m$ is the probability that the first m drawn are all bad, if C is true

– we also need to know the probability $P(D|\bar{C}X)$:

  * since there are only three possibilities, A, B, and C, $\bar{C} \Rightarrow A + B$,

$$P(\bar{C}|DX) = P(A+B|DX) = P(A|DX) + P(B|DX)$$

  * we also know that

$$P(D|\bar{C}X) = P(D|X)\frac{P(\bar{C}|DX)}{P(\bar{C}|X)}$$

  * combining these two equations we get

$$P(D|\bar{C}X) = \frac{P(D|AX)P(A|X) + P(D|BX)P(B|X)}{P(A|X) + P(B|X)} = \left(\frac{1}{3}\right)^m\left(\frac{1}{11}\right) + \left(\frac{1}{6}\right)^m\frac{10}{11}$$

– putting everything together, we have the evidence for proposition C,

$$e(C|DX) = -60 + 10\log_{10}\left[\frac{\left(\frac{99}{100}\right)^m}{\frac{1}{11}\left(\frac{1}{3}\right)^m + \frac{10}{11}\left(\frac{1}{6}\right)^m}\right]$$

– if $m > 5$ is a good approximation, $e(C|DX) \approx -49.6 + 4.73m$ and if $m < 3$ a cruder approximation is $e(C|DX) \approx -60 + 7.73m$

– so 10 consecutive bad widgets would raise the evidence for C by ~ 58 db, while 11 would make us consider it more likely true than false

– what is happening to A and B during this? the equations below denote the evidence for the two propositions and their approximate forms

$$e(A|DX) = -10 + 10\log_{10}\left[\frac{\left(\frac{1}{3}\right)^m}{\left(\frac{1}{6}\right)^m + \frac{11}{10}\times 10^{-6}\left(\frac{99}{100}\right)^m}\right] \approx \begin{bmatrix} -10+3m & \text{for } m<7 \\ +49.6-4.73m & \text{for } m>8 \end{bmatrix}$$

$$e(B|DX) = +10 + 10\log_{10}\left[\frac{\left(\frac{1}{6}\right)^m}{\left(\frac{1}{3}\right)^m + \frac{11}{10}\times 10^{-6}\left(\frac{99}{100}\right)^m}\right] \approx \begin{bmatrix} 10-3m & \text{for } m<10 \\ +59.6-7.73m & \text{for } m>11 \end{bmatrix}$$

– this is interesting, because initially A and C are so much more plausible than C that we are functionally comparing A and C in binary

– once enough $e(C|DX) \approx e(B|DX)$, we are functionally testing A against C, instead of against B

– all of the changes in slopes can be interpreted in this way, and this leads us to a general principle:

> As long as we have a discrete set of hypotheses, a change in plausibility for any one of them will be approximately the result of a test of this hypothesis against a single alternative, being one of the remaining hypotheses which is most plausible at that time.

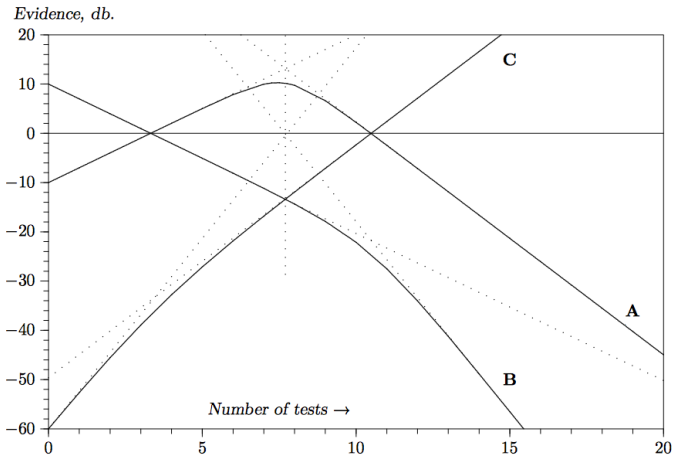

Figure 3: Evidence for propositions A, B, and C as a function of draws m. Note the peak in proposition A at $\sim m = 7$.

5.5 Continuous probability distribution functions

- we could continue the above example by adding in more and more 'discrete' hypotheses. however a more interesting task would be to introduce a continuous range of hypotheses such that $H_f \equiv$ the machine putting out a fraction f of fails
- instead of a discrete prior probability distribution, we would have a continous distribution in the interval $[0,1]$, and we could calculate posterior probabilities for various values of f
- suppose f is any continuously variable parameter. the propositions

$$F' \equiv (f \leq q)$$
$$F'' \equiv (f > q)$$

  are discrete, mutually exclusive, and exhaustive
- with some prior information Y, the probability for F' will depend on q: $G(q) \equiv P(F'|Y)$
- what is the probability of finding f in the interval $(a < f \leq b)$?
  - Propositions:
$$A \equiv (f \leq a) \qquad B \equiv (f \leq b) \qquad W \equiv (a < f \leq b)$$
  - since $B = A + W$, and A and W are mutually exclusive, the sum rule is $P(B|Y) = P(A|Y) + P(W|Y)$
  - and since $P(B|Y) = G(b)$ and $P(A|Y) = G(a)$, $P(W|Y) = P(a < f \leq b|Y) = G(b) - G(a)$
  - if $G(q)$ is continuous and differentiable, we can also write $P(a < f \leq b|Y) = \int_a^b \mathrm{d}f\, g(f)$, where $g(f) = G'(f) \geq 0$ is the derivative of G
    * G' is the *probability distribution function* (PDF) for f, given Y
    * should be noted that this is should not be described as a posterior probability **of** f, as that implies that f is the thing being distributed. it is not, the *probability* of f is
- it should be noticed that the same rules as we used in the treatment of discrete, finite sets of propositions were used here with a continuous, infinite set of propositions
  - generally true if the continuous set of propositions is defined from a basis of finite sets of propositions

5.6 Testing an infinite set of hypotheses

- say we are testing an uncountable, infinite set of hypotheses. because log forms now become awkward (for some complicated reasons), we go back to using the original form for the probability of $P(A|DX)$ (the first equation in section 5.2)
- if A is now the proposition that the fraction of fails is in teh range $(f, f + \mathrm{d}f)$, then the prior PDF is $P(A|X) = g(f|X)\mathrm{d}f$

- if D is the data N widgets were tested such that we found n fails and $N-n$ good ones, the posterior PDF is

$$P(A|DX) = P(A|X)\frac{P(D|AX)}{P(D|X)} = g(f|DX)\mathrm{d}f$$

- furthermore, the prior and posterior PDFs are related by

$$g(f|DX) = g(f|X)\frac{P(D|AX)}{P(D|X)}$$

- this problem is usually easier if we require that $P(0 \leq f \leq 1|DX) = \int_0^1 \mathrm{d}f\, g(f|DX) = 1$
- it should be clear that the evidence of the data is entirely within the f dependence of $P(D|AX)$
- note that the proposition A uses an interval $\mathrm{d}f$, not a point f
  - if f is the only variable, we can usually use the limit $\mathrm{d}f \to 0$ and replace $P(D|AX)$ with $P(D|H_f X)$
  - this only works reliably if f is not dependent on any other variable, otherwise we don't know if $\mathrm{d}f \to 0$ in the same way for all values of the other variable
  - see the Borel-Kolmogorov paradox
- if f is the fraction of fails, then $(1-f)$ is the probability of getting a good one. the probabilities at different trials are logically independent given f, so we follow the binomial distribution and write

$$g(f|DX) = \frac{f^n(1-f)^{N-n}g(f|X)}{\int_0^1 \mathrm{d}f f^n(1-f)^{N-n}g(f|X)} \tag{3}$$

  - aside: this should absolutely evoke your first project derivation which has

$$P(\lambda, t) = \frac{\int \mathrm{d}x_0 \rho(x_0) h_\mathrm{R}(x_0)\delta[\lambda - \lambda(x_t)]}{\int \mathrm{d}x_0 \rho(x_0) h_\mathrm{R}(x_0)}$$

- the entirety of the above section of multiple hypothesis testing can be dealt with using eq. 3, using three delta functions for $g(f|X)$, $\delta(f - f_H)$, to denote the three hypotheses A, B, and C
- going back to the machine, what if the machine's starting information was simply that it was *possible* for a machine to make a fail, and *possible* for a machine to make a good one?
  - we have no definite knowledge of a prior for f, and we haven't yet determined how to assign such priors
  - best thing to do is definte a uniform prior PDF, $g(f|X) = 1$, where $0 \leq f \leq 1$
  - eq. 3 then reduces to the *complete beta-function*,

$$g(f|DX) = \frac{(N+1)!}{n!(N-n)!}f^n(1-f)^{N-n}$$

  - this has a single peak in the interval $(0 \leq f \leq 1)$, $f = \hat{f} \equiv \frac{n}{N}$, or the observed fraction of fails to good widgets
  - the sharpness, however, is found by

$$L(f) \equiv \log g(f|DX) = n\log(f) + (N-n)\log(1-f) + \text{const}$$

  - expanding $L(f)$ in a power series around $\hat{f}$,

$$L(f) = L(\hat{f}) - \frac{(f-\hat{f})^2}{2\sigma^2}$$

  where

$$\sigma \equiv \frac{\hat{f}(1-\hat{f})}{N}$$

- and so the complete beta function in this approximation is

$$g(f|DX) \approx K \exp\left[-\frac{(f-\hat{f})^2}{2\sigma^2}\right] \tag{4}$$

  which is the *Gaussian distribution*
- so the most likely value of f is the observed fraction $\hat{f}$, and the accuracy of this estimate is such that the interval $\hat{f} \pm \sigma$ is likely to contain the tru value
- σ is the *standard deviation* and σ^2 is the *variance* of the PDF

  50% probability that the true value of f is contained in the interval $\hat{f} \pm 0.68\sigma$

  90% probability it is contained in $\hat{f} \pm 1.65\sigma$

  99% probability it is contained in $\hat{f} \pm 2.57\sigma$
- as the number of tests N increases, the intervals shrink proportional to $1/\sqrt{N}$

5.7 Simple and compound hypotheses

- so far the hypotheses we have considered refer to a single parameter, or value, $f = M/M$, and are sharply defined for this value f
- if we consider an abstract ''parameter space'' Ω, consisting of all possible values of all possible parameters, these hypotheses are considered ''simple'' hypotheses as they refer to a single point within Ω
- testing all simple hypotheses within Ω would be too much. could we test subsets, like $\Omega_1 \in \Omega$? can we proceed directly to this question instead of testing all possible simple hypotheses within Ω_1?
  - let's start from eq. 3. if we want to take some action with the machine, say readjust if $f > 0.1$, we can take some subset of $\Omega = [0,1]$, i.e. Ω_1 comprising all f in the interval $[0.1,1]$, and H as the hypothesis that f is in Ω_1
  - the actual value of f is not needed anymore and is considered a *nuisance parameter*, and we need to take action to remove it
  - if there are no parameters other than f and the intervals $\mathrm{d}f$ are mutually exclusive, then f is removed by integrating it out of eq 3:

$$P(\Omega_1|DX) = \frac{\int_{\Omega_1} \mathrm{d}f f^n (1-f)^{N-n} g(f|X)}{\int_{\Omega} \mathrm{d}f f^n (1-f)^{N-n} g(f|X)}$$

    and if we have a uniform prior PDF for f,

$$P(a < f < b|DX) = \frac{(N+1)!}{n!(N-n)!} \int_a^b \mathrm{d}f f^n (1-f)^{N-n}$$

- thus if we have any compound hypothesis to test, the proper procedure is to sum or integrate out the nuisance parameters the PDF contains, with respect to appropriate priors.

6 Elementary parameter estimation

- When we have a small number of discrete hypotheses $\{H_1, \ldots, H_n\}$, we want to choose a specific one which is most likely, in light of priors and data (this procedure was shown in Ch.5)
- if n is very large, however, we must use a different approach.
- a set of discrete hypotheses can be classified by one or more numerical indices, such that H_t $(1 \le t \le n)$. In this case, deciding between the hypotheses and estimating the index t become practically the same thing, i.e. we can set the index as the quantity of interest rather than the hypothesis itself
- this is called *parameter estimation*

6.1 Inversion of the urn distributions

- in this case, the index is discrete
- recall the urn sampling examples we considered in ch. 4, wherein we knew the number of balls, N, of which R were red and $N - R$ were white. we were able to make inferences about what the numbers of red, r, and white, $n - r$, balls were likely to be drawn in n draws
- now lets invert this example: we know the data, $D \equiv (n, r)$, but not the composition of the population of the urn, (N, R)
- recall that the sampling distribution for this situation is the hypergeometric distribution:

$$p(D|NRI) = h(r|NR, n) = \binom{N}{n}^{-1}\binom{R}{r}\binom{N-R}{n-r}$$

where I is the prior information.

6.1.1 Both N and R are unknown

- we know intuitively that if we draw n balls from the urn, then $N \geq n$. is the number of red balls drawn, r, or the order, be relevant to N?
  - the joint posterior probability distribution for N and R, after using the product rule $p(NR|I) = p(N|I)p(R|NI)$, is

$$p(NR|DI) = p(N|I)p(R|NI)\frac{p(D|NRI)}{p(D|I)}$$

where

$$p(D|I) = \sum_{N=0}^{\infty}\sum_{R=0}^{N} p(N|I)p(R|NI)p(D|NRI)$$

    - it should be obvious that $p(D|NRI)$ is zero in the cases $N < n$, $R < r$, or $N - R < n - r$
  - it follows that the marginal posterior probability for N is thus

$$p(N|DI) = \sum_{R=0}^{N} p(NR|DI) = p(N|I)\frac{\sum_R p(R|NI)p(D|NRI)}{p(D|I)}$$

    - could alternatively apply Bayes' theorem directly, and it must agree with the above posterior probability by way of sum and product rules
  - whatever prior $p(N|I)$ we assigned, data can only truncate impossible values:

$$P(N|DI) = \begin{Bmatrix} Ap(N|I) & \text{if } N \geq n \\ 0 & \text{if } 0 \leq N \leq n \end{Bmatrix}$$

where A is a normalization constant

    - it should be noted that this only applies if we have no information linking N and R. if we know, for example, that $R < 0.06N$, we know that, having the observed data $(n, r) = (10, 6)$, that $N \geq 10$ *and* that $N > 100$, making datum r relevant to the estimation of N
    - this is usually not known, however, so it is irrelevant
  - generally, this is a condition on the prior probability $p(R|NI)$

$$p(D|NI) = \sum_{R=0}^{N} p(D|NRI)p(R|NI) = \begin{Bmatrix} f(n,r) & \text{if } N \geq n \\ 0 & \text{if } N < n \end{Bmatrix}$$

where $f(n, r)$ may depend on the D, but is independent of N

  - using the hypergeometric distribution, this is explicitly

$$\sum_{R=0}^{N}\binom{R}{r}\binom{N-R}{n-r}p(R|NI) = f(n,r)\binom{N}{n}, \qquad N \geq n \tag{5}$$

- so, if the constraint on $p(R|NI)$ is that $f(n,r)$ must be independent of N if $N \geq n$, estimation of N is uninteresting and irrelevant
- by factoring the joint posterior distribution found earlier, we know that $p(NR|DI) = p(N|DI)p(R|NDI)$, and we are mainly concerned with the distribution R:

$$p(R|DNI) = p(R|NI)\frac{p(D|NRI)}{p(D|NI)} \tag{6}$$

- different choices for the prior probability $p(R|NI)$ will yield different results
- **Uniform prior**
  - Consider the prior knowledge, I_0, to be us absolutely ignorant about R while knowing N exactly, i.e. the *uniform distribution*

$$p(R|NI_0) = \left\{ \begin{array}{cc} \frac{1}{N+1} & \text{if } 0 \leq R \leq N \\ 0 & \text{if } R > N \end{array} \right\}$$

  - some terms cancel and eq 6 reduces to

$$p(R|DNI_0) = S^{-1}\binom{R}{r}\binom{N-R}{n-r}$$

    where S^{-1} is a normalization constant, $S \equiv \sum_{R=0}^{N}\binom{R}{r}\binom{N-R}{n-r} = \binom{N+1}{n+1}$
  - combining this and the prior, we get

$$\sum_{R=0}^{N}\frac{1}{N+1}\binom{R}{r}\binom{N-R}{n-r} = \frac{1}{N+1}\binom{N+1}{n+1} = \frac{1}{n+1}\binom{N}{n}$$

    - this satisfies the integral equation (eq 5), as it can tell us nothing about N beyond $N \geq n$ (because $\binom{N}{n}$ where $n > N$ is zero)
    - goes to zero when $R < r$ or $R > N - n + r$
    - goes to $\delta(R,r)$ if $n = N$
    - all these in accordance to intuition
    - if we obtain no data, $n = r = 0$, this reduces to the prior distribution
  - the most probable value of R is found within one unit by setting $p(R') = p(R'-1)$ and solving for R', yielding $R' = (N+1)\frac{r}{n}$, or the peak of the sampling distribution discussed in the Bernoulli urn section
  - can find the expectation value of R

$$\langle R \rangle = \sum_{R=0}^{N} Rp(R|DNI_0) \Rightarrow \frac{(N+2)(r+1)}{(n+2)}$$

    - when (n, r, N) are large, $\langle R \rangle \approx R'$, indicating either a sharply peaked or symmetric posterior distribution
    - can further extend this to ask what the expected fraction F of red balls left in the urn: $\langle F \rangle = \frac{\langle R \rangle - r}{N-n} = \frac{r+1}{n+2}$
  - instead of estimating unobserved contents, we can instead predict future observations, asking ''after drawing r red balls in n draws, what is the probability that the next will be red?'', with the proposition $R_i \equiv$ red on the ith draw, where $1 \leq i \leq N$
  - this gives us the posterior probability

$$p(R_{n+1}|DNI_0) = \sum_{R=0}^{N} p(R_{n+1}R|DNI_0) = \sum_{R} p(R_{n+1}|RDNI_0)p(R|DNI_0)$$

    or

$$p(R_{n+1}|DNI_0) = \sum_{R=0}^{N}\frac{R-r}{N-n}\binom{N+1}{n+1}^{-1}\binom{R}{r}\binom{N-R}{n-r} = \frac{r+1}{n+2}$$

    - this result is the same as the expected fraction of red balls left in the urn, above, $\langle F \rangle$

    * brings us to a rule: a probability is not the same thing as a frequency; but, under general conditions, the *predictive probability* of an event at a single trial is numerically equal to the *expectation* of its frequency in a specified class of trials
    * also called *Laplace's rule of succession*
  - however, this only gives a point estimate. what accuracy is claimed? find the variance, $\langle R^2\rangle - \langle R\rangle^2$, resulting in

$$\mathrm{var}(R) = \frac{p(1-p)}{n+3}(N+2)(N-n)$$

    after a derivation, where $p = \langle F\rangle = (r+1)/(n+2)$
  - therefore, our (mean) ± (standard deviation) is

$$(R)_{\mathrm{est}} = r + (N-n)p \pm \sqrt{\frac{p(1-p)}{n+3}(N+2)(N-n)}$$

    with the $N-n$ term inside the square root indicating, as expected, that the estimate becomes more accurate as more of the urn is sampled
  - we can also find the (mean) ± (standard deviation) of F

$$(F)_{\mathrm{est}} = p \pm \sqrt{\frac{p(1-p)}{n+3}\frac{N+2}{N-n}}$$

    and as $N \to \infty$,

$$(F)_{\mathrm{est}} = p \pm \sqrt{\frac{p(1-p)}{n+3}}$$

  - EXAMPLE: a poll of 1600 voters finds that 41% ±3% of the population favours a candidate. Is this consistent?
    * to obtain $(F)_{\mathrm{est}} = \langle F\rangle(1 \pm 0.03)$, we need an n of,

$$n+3 = \frac{1-p}{p}\frac{1}{(0.03)^2} = \frac{1-0.41}{0.41} \times 1111 \approx 1596$$

- **Truncated uniform priors**
  - say we know from the start that $0 < R < N$, i.e. that there is at least one red ball and one white ball. the new prior is

$$p(R|NI_1) = \left\{\begin{matrix} \frac{1}{N-1} & \text{if } 1 \le R \le N-1 \\ 0 & \text{otherwise} \end{matrix}\right\}$$

  - the summation formula is the same as for a uniform prior (S, above), corrected by subtracting an $R = 0$ and $R = N$ term, $\binom{R}{r} = \binom{R+1}{r+1} = \delta(r,0)$ and $\binom{N-R}{n-r} = \delta(r,n)$,respectively. the new summation expressions are

$$S = \sum_{R=1}^{N-1}\binom{R}{r}\binom{N-R}{n-r} = \binom{N+1}{n+1} - \binom{N}{n}\delta(r,n) - \binom{N}{n}\delta(r,0)$$

$$S = \sum_{R=1}^{N-1}\binom{R+1}{r+1}\binom{N-R}{n-r} = \binom{N+2}{n+2} - \binom{N+1}{n+1}\delta(r,n) - \binom{N}{n}\delta(r,0)$$

    * note that the new terms vanish if $0 < r < n$, and the posterior distribution is unchanged, $p(R|DNI_1) = p(R|DNI_0)$

- **Binomial monkey prior**
  - suppose our prior information, I_2 is that the urn is filled with monkeys who toss balls up at random such that each ball had a probability g of being red. then our prior for R will be the binomial distribution,

$$p(R|NI_2) = \binom{N}{R}g^R(1-g)^{N-R} \qquad 0 \le R \le N$$

    with the prior estimate of the number of red balls being $(R)_{\mathrm{est}} = Ng \pm \sqrt{Ng(1-g)}$

- the sum for this prior is thus
$$p(D|NI) = \binom{n}{r} g^r (1-g)^{n-r} \qquad N \geq n$$
- note that this sum is independent of N and thus satifies the integral equation (eq 5), $p(NR|DI_2) = p(N|DI_2)p(R|NDI_2)$ and $p(NR|DI_2) = p(R|DI_2)p(N|RDI_2)$
- we are interested in $p(R|NDI_2)$, where N is known, and $P(R|DI)$, which tells us what we know about R regardless of N
- using the prior and the hypergeometric distribution we find
$$p(R|DNI_2) = A\binom{N}{R} g^R (1-g)^{N-R} \binom{R}{r}\binom{N-R}{n-r}$$
with A being the normalization,
$$1 = \sum_R p(R|DNI_2) = A\binom{N}{n}\binom{n}{r} g^r (1-g)^{n-r}$$
leading to a normalized posterior distribution for R of
$$p(R|DNI_2) = \binom{N-n}{R-r} g^{R-r} (1-g)^{N-R-n+r} \tag{7}$$
- this has a (mean) ± (standard deviation) of $(R)_{\text{est.}} = r + (N-n)g \pm \sqrt{g(1-g)(N-n)}$
- we can also estimate the fraction of red balls left in the urn:
$$\frac{(R-r)_{\text{est.}}}{N-n} = g \pm \sqrt{\frac{g(1-g)}{N-n}}$$
- though these two estimates seem similar to those derived for the uniform priors, note that these do not depend on p at all, only on g, i.e. the binomial prior leads us to estimates which are exactly the same as prior estimates with no data whatsoever.
  * ''More precisely, with that prior the data can tell us nothing at all about the unsampled balls.''
  * if the prior information about the population is described accurately by the binomial prior, sampling is futile unless you sample the entire population
  * the binomial prior is more informative about the population than the uniform prior, as it is moderately peaked as opposed to flat
  * however, extra data does not inform us further as each draw has an *independent* probability g to be red (again, this is *logical independence* of the prior and it is preserved in the posterior distribution)

6.2 Continuous parameter estimation

- if our hypotheses become so dense, i.e./ the intervals between their indices $t \to 0$, have increasingly similar observables, and as such have similar posterior probabilities
- thus one particular hypothesis is likely not favoured over all others and it is more appropriate to think not of discrete indices t, but instead of a parameter θ
  - we have changed the hypothesis testing problem into a parameter estimation problem
  - can be changed back if, say, for a hypothesis that a parameter θ lies in an interval $a < \theta < b$, this is an interval estimation procedure for a compound hypothesis (see section 5)

6.2.1 Estimation with a binomial sampling distribution

- experiments in which there is a binary result, either yes or no, are called *Bernoulli trials*, after Bernoulli's urn
- the conditions of one of these experiments will tell us if order is known, but probability theory tells us whether it is relevant. for example, flipping a coin 100 times: the order is known but not meaningful if determining if it is rigged. sampling a population for disease before and after the introduction of a new drug: order is known and is relevant to whether or not the drug worked

- Let's set up a simple binomial sampling problem, first defining a variable

$$x_i \equiv \begin{Bmatrix} 1 & \text{success on ith trial} \\ 0 & \text{otherwise} \end{Bmatrix}$$

- our data is $D \equiv \{x_1, \ldots, x_n\}$
- our prior information I specifies that there is a parameter θ such that for each logically independent trial we have a probability θ for success and $(1-\theta)$ for failure
- our sampling distribution is thus

$$p(D|\theta I) = \prod_{i=1}^{n} p(x_i|\theta I) = \theta^r (1-\theta)^{n-r}$$

  where r is successes and $n-r$ is failures
- we thus have the posterior PDF

$$p(\theta|DI) = \frac{p(\theta|I)p(D|\theta I)}{\int d\theta\ p(\theta|I)p(D|\theta I)} = Ap(\theta|I)\theta^r(1-\theta)^{n-r}$$

  where A is a normalizing constant, and with a uniform prior such that $p(\theta|I) = 1$,

$$A^{-1} = \int_0^1 d\theta\ \theta^r(1-\theta)^{n-r} = \frac{r!(n-r)!}{(n+1)!}$$

  providing the normalized PDF

$$p(\theta|DI) = \frac{(n+1)!}{r!(n-r)!}\theta^r(1-\theta)^{n-r} \tag{8}$$

- this is identical to the complete-beta function described in section 5
- we have a predictive probability for success at the next trial of

$$p \equiv \langle\theta\rangle = \int_0^1 d\theta\ \theta p(\theta|DI) = \frac{r+1}{n+2}$$

  which is Laplace's rule of succession
- and a (mean) ± (standard deviation) of

$$(\theta)_{\text{est.}} = \langle\theta\rangle \pm \sqrt{\langle\theta^2\rangle - \langle\theta\rangle^2} = p \pm \sqrt{\frac{p(1-p)}{n+3}}$$

  which is identical to the results for the uniform prior for a discrete set of hypotheses.
- so the continuous results must be derivable from the discrete results, in the limit $N \to \infty$, and since the discrete results are independent of N, they are identical
- **A digression on stopping**
  - we did not include n or r in the conditioning statements in $p(D|\theta I)$, as both are learned from the data D
  - what if we decided to stop after n trials? we could write $p(D|n\theta I)$
  - or if we stopped after r successes, $p(D|r\theta I)$
  - but does this affect our conclusions about θ?
  - in deductive logic $AA = A$, saying A is true twice is the same as saying it once. additionally, when something is known already from the priors, no matter what the data says, the likelihood is the same: $p(nr \text{ order}|n\theta I) = p(r \text{ order}|n\theta I)p(n|n\theta I) = p(r \text{ order}|n\theta I)$
  - likewise, $p(\theta n|DI) = p(\theta|nDI)p(n|DI) = p(n|\theta DI)p(\theta|DI) \Rightarrow p(\theta|nDI) = p(\theta|DI)$, since $P(n|\theta DI) = p(n|DI) = 1$
  - thus if any part of the data is included in the priors, then that part is redundant and cannot affect the conclusions
  - this is general to any function $f(D)$; if f is known beforehand it can have a major effect on sampling space and sampling distributions, but it cannot have any effect on rational inferences from the data
  - furthermore, inference must depend on the data that was actually observed, since possible unobserved data sets gives us no information beyond our priors

6.3 Effects of qualitative prior information

- Two robots A and B have different prior information as to the source of particles hitting a detector. the particles hit the detector and create counts with efficiency $\phi = 0.1$
  - A has no knowleodge about the source
  - B knows that the source is a radioactive sample of long lifetime, in a fixed position
- if during the first second, $c_1 = 10$ counts are registered, what can A and B say about the number n_1 of particles?

$$p(n_1|\phi c_1 I_A) = p(n_1|I_A)\frac{p(c_1|\phi n_1 I_A)}{p(c_1|\phi I_A)}$$

- we now are stuck with determining $p(n_1|I_A)$. how can we assign prior probabilities based on purely qualitative evidence?

6.3.1 Choice of a prior

- The prior for A should avoid all structure which would cause great variations in $p(n_1|I_A)$, as variations such as oscillations or sudden jumps/falls would imply prior information that A does not have
  - Jeffreys (1939) states that almost any prior which is smooth in the region of high likelihood will lead to substantially the same final conclusions
- Let's assign a uniform prior probability out to some large, finite N,

$$p(n_1|I_A) = \begin{Bmatrix} 1/N & \text{if } 0 \le n_1 \le N \\ 0 & \text{if } N \le n_1 \end{Bmatrix}$$

- choice of N is important as the final conclusions depend strongly on it. need to analyze exact prior information to determine a valid choice and if $n_1 = N$ needs to be smoothed
- plugging this into the above probability, we get

$$p(n_1|\phi c_1 I_A) = \begin{Bmatrix} Ap(c_1|\phi n_1) & \text{if } 0 \le n_q < N \\ 0 & \text{if } N \le n_1 \end{Bmatrix}$$

  where A is a normalization factor

$$A^{-1} = \sum_{n_1=0}^{N-1} p(c_1|\phi N - 1)$$

- the normalization factor A converges so rapidly that its sum is not appreciably different than the sum to infinity, allowing us to use the simplifaction

$$\sum_{m=0}^{\infty}\binom{m+a}{m}m^n x^m = \left(x\frac{d}{dx}\right)^n \frac{1}{(1-x)^{a+1}}, \qquad |x| < 1$$

  yielding,

$$A^{-1} \approx \sum_{m=0}^{\infty} p(c_1|\phi n_1) = \phi^c \sum_{m=0}^{\infty}\binom{m+c}{m}(1-\phi)^m = \phi^c\left\{\frac{1}{[1-(1-\phi)]^{c+1}}\right\} = \frac{1}{\phi}$$

- putting this together, we obtain the result,

$$p(n_1|\phi c_1 I_A) = \phi p(c_1|\phi n_1) = \binom{n_1}{c_1}\phi^{c_1+1}(1-\phi)^{n_1-c_1}$$

  yielding a most probable value of $(\hat{n}_1) = c_1/\phi$, or the same as the maximum likelihood estimate earlier

6.3.2 The Jeffreys prior

- Jeffreys suggests that the correct way to express 'complete ignorance' of a continuous variable known to be positive is to assign uniform prior probability to its logarithm,

$$p(s|I_J) \propto \frac{1}{s}, \qquad (0 \leq s \leq \infty) \tag{9}$$

- should be noted that this cannot be normalized (why?)
- using this prior gives the following results for the above subsection:

$$p(n_1|I_J) = \frac{1}{n_1} \qquad p(c_1|I_J) = \frac{1}{c_1} \qquad p(n_1|\phi c_1 I_J) = \frac{c_1}{n_1} p(c_1|\phi n_1)$$

with most probable and mean value estimates of

$$\left(\hat{n}_1\right)_J = \frac{c_1 - 1 + \phi}{\phi} = 91 \qquad \langle n_1 \rangle_J \frac{c}{\phi} = 100$$

- Jeffreys prior probability rule reduces the most probable and posterior mean value estimates by nine, bringing the mean value back to the maximum likelihood estimate. shows us that different prior probabilities which are not sharply peaked give numerically very similar answers
- General rule of thumb: changing the prior probability $p(\alpha|I)$ for a parameter by one power of α has about the same effect as does having one more data point, as the likelihood function has a width of $1/\sqrt{n}$, and one more power of α adds an extra small amount of slope in the neighborhood of the maximum, shifting it slightly

7 The central, Gaussian, or normal distribution

- we can only understand why Guassian distributions work for error estimation when ''we learn to think of probability distributions in terms of their demonstrable *information content* instea dof their imagined (and, as we shall see, irrelevant) frequency connections.'' (p 198)
- General properties of Gaussians:
  1. Any smooth function with a single rounded maximum, if raised to higher and higher powers, goes into a Gaussian function.
  2. The product of two Gaussian functions is another Gaussian function.
  3. The convolution of two Gaussian functions is another Gaussian function.
  4. The Fourier transform of a Gaussian function is another Gaussian function.
  5. A Gaussian probability distribution has higher entropy than any other with the same variance; therefore any operation on a probability distribution which discards information but conserves variance takes us close to a Gaussian. (see the central limit theorem for example)

7.1 The gravitating phenomenon

- in probability theory, there is a central, universal distribution

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \tag{10}$$

toward which all other distributions gravitate under various conditions (ex. binomial distribution goes asymptotically to a Gaussian when number of trials increases, and posterior distributions for parameters go into Gaussian when number of data points increases)

- cumulative Gaussian:

$$\begin{aligned}\Phi(x) &\equiv \int_{-\infty}^{x} \mathrm{d}t\phi(t) \\ &= \int_{-\infty}^{0} \mathrm{d}t\phi(t) + \int_{0}^{x} \mathrm{d}t\phi(t) \\ &= \frac{1}{2}[1 + \ \mathrm{erf}(x)]\end{aligned}$$

- Jaynes presents three derivations of the Gaussian distribution here (Herschel—Maxwell, Gauss), which I will not replicate.
- Gauss showed that (maximum likelihood estimate) = (arithmetic mean) determines the Gaussian error law, not the uniform error law.

7.2 Why the ubiquitous use of Gaussian distributions?

- If a physicist takes some measurements electronically, the mean square error of those measurements follow the Nyquist thermal fluctuation law. If he takes two moments of a noise probability distribution and has no further information about the noise, a Gaussian distribution fits to those moments and will represent the state of knowledge of the noise.
- However, it must be clear that the physicist's state of knowledge is about the specific samples of noise for which he had taken data, not about measurements he is about to make. *Past noise samples are relevant for predicting future noise only through those aspects which are reproducible in the future.*
- ''. . .if the Gaussian law is nearly always a good representation of our state of knowledge about the errors *in our specific data set*, it follows that inferences made from it are nearly always the best ones that could have been made from the information that we actually have.''
- Because Bayes' theorem takes into account whatever can be inferred about the noise from the data, Bayesian inferences using a Gaussian sampling distribution can only be approved upon by knowing additional information about the actual errors in a specific data set, beyond its first two moments and what is known from the data.

7.2.1 Non–Gaussian estimators?

- Since Gauss showed that (maximum likelihood estimator) = (arithmetic mean of the observations) uniquely determines the Gaussian distribution, if our sampling distribution is not Gaussian, these two estimators should be different.
- Most sampling distributions which are used are of the independent, identically distributed form,

$$p(x_1 \ldots x_n|\mu I) = \sum_{i=1}^{n} f(x_i - \mu)$$

- if the sampling distribution is not Gaussian, the estimator is usually a linear combination of the observations $(\mu)_{\mathrm{est}} = \sum w_i y_i$, with variable weighting coefficients w_i based on the data configuration $(y_i - y_j), 1 \le i, j \le n$.
- Let's consider a case in which we have no prior probabilities or loss functions.
  - trying to estimate a location parameter μ, with data D of n observations: $D = \{y_1, \ldots, y_n\}$
  - the data points have errors which vary in an uncontrolled and unpredictable manner
  - with the unknown true value being μ_0, μ being the running variable, and e_i being the error in the ith measurement, our model is

$$y_i = \mu_0 + e_i \qquad (1 \le i \le n)$$

  - if we assign an independent Gaussian sampling distribution for the errors $e_i = y_i - \mu_0$,

$$p(D|\mu_0 \sigma I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{\sum (y_i - \mu_0)^2}{2\sigma^2}\right]$$

    where

$$\sum_{i=1}^{n} (y_i - \mu_0)^2 = n\left[(\mu_0 - \bar{y})^2 + s^2\right]$$

with

$$\bar{y} \equiv \frac{1}{n}\sum y_i = \mu_0 + \bar{e} \qquad s^2 \equiv \bar{y^2} - \bar{y}^2 = \bar{e^2} - \bar{e}^2$$

being the only two properties of the data appearing in the likelihood function.

- Thus, as a consequence of applying the Gaussian error distribution to the data, only the first two moments of the data are used for inferences about μ_0
- These two moments are called *sufficient statistics*.
- Note that $\bar{e}$ is the average of the actual errors, not an average over a probability distribution, and this holds for however the errors e_i are distributioned (be it Gaussian or not)

7.2.2 Error cancellation

- Error cancellation is an important component of the success of the Gaussian sampling distribution
- Suppose we estimate μ through a linear combination of data values,

$$(\mu)_{\text{est}} = \sum_{i=1}^{n} w_i y_i$$

where w_i are weighting coefficients, real numbers satisfying $\sum w_i = 1$

- with this model, the square of the error is

$$\Delta^2 = \left[(\mu)_{\text{est}} - (\mu)_0\right]^2 = \left(\sum_i w_i e_i\right)^2 = \sum_{i,j=1}^{n} w_i w_j e_i e_j$$

with the expectation value of

$$\langle \Delta^2 \rangle = \sum_{i,j} w_i w_j \langle e_i e_j \rangle$$

- if we assign identical and independent probabilities to each e_i, as is always assumed, then $\langle e_i e_j \rangle = \sigma^2 \delta_{ij}$ and

$$\langle \Delta^2 \rangle = \sigma^2 \sum_i {w_i}^2$$

- if we set $e_i = n^{-1} + q_i$ (why??), where $\{q_i\}$ are real numbers constrained by $\sum w_i = 1$ and $\sum q_i = 0$, the square of the error is then

$$\langle \Delta^2 \rangle = \sigma^2 \sum_i \left(\frac{1}{n^2} + \frac{2q_i}{n} + {q_i}^2\right) = \sigma^2 \left(\frac{1}{n} + \sum_i {q_i}^2\right)$$

which reaches a minimum,

$$\langle \Delta^2 \rangle_{\min} = \frac{\sigma^2}{n}$$

iff all $q_i = 0$.

- what this shows is that uniform weighting, $w_i = 1/n$, leads to a smaller expected square error, σ^2/n, than any other weighting, leading to square error $\sigma^2 \sum_i {w_i}^2$.
- important point:

> If we have no important prior information, use of the Gaussian sampling distribution automatically leads us to estimate μ by the arithmetic mean of the observations; and Gauss proved that the Gaussian distribution is the only one which does this. Therefore, among all sampling distributions which estimate μ by the arithmetic mean of the observations, the Gaussian distribution is uniquely deterined as the one that gives maximun error cancellation. (p. 214)

- another important point, about the irrelevancy of frequency distributions:

> If we assign the independent Gaussian error distribution, then the error in our estimate is always the arithmetic mean of the true errors in our data set; and whether the frequency distribution of those errors is or is not Gaussian is totally irrelevant. Any error vector $\{e_1, \ldots, e_n\}$ with the same first moment $\bar{e}$ will lead us to the same estimate of μ; and any error vector with the same first two moments will lead us to the same estimates of both μ and σ and the same accuracy claims, *whatever the frequency distributions of the individual errors.* (p. 216)

7.3 Nuisance parameters as safety devices

- if we do not have actual knowledge about the magnitude of σ, it would be harmful to assume an arbitrary value; we should adopt a model which acknowledges ignorance by allowing various values of σ, with a prior $p(\sigma|I)$ in accordance to prior information (recall that σ is the variance)
- the joint posterior PDF for μ and σ is,

$$p(\mu\sigma|DI) = p(\mu\sigma|I)\frac{p(D|\mu\sigma I)}{p(D|I)}$$

- using the product rule, we obtain,

$$p(\mu\sigma|DI) = p(\sigma|I)p(\mu|\sigma I)\frac{p(D|\sigma I)p(\mu|\sigma DI)}{p(D|I)p(\mu|\sigma I)} = p(\mu|\sigma DI)p(\sigma|DI)$$

- integrating out σ, we get the marginal posterior PDF for μ,

$$p(\mu|DI) = \int \mathrm{d}\sigma\; p(\mu|\sigma DI)p(\sigma|DI)$$

- this is a weighted average of the PDFs $p(\mu|\sigma DI)$ for all possible values of σ, weighted by the marginal posterior PDF $p(\sigma|DI)$, which is everything we know about σ. Thus, by integrating out σ as a nuisance parameter, we don't lose information relevant to the parameters we keep
- important point:

> . . .whatever question we ask, probability theory as logic automatically takes into account all possibilities *allowed by our model* and information. (p. 219)

7.4 Convolution of Gaussians

- NOTE: A convolution is an integral which denotes the amount of overlap of two functions as one is shifted over the other
- for an excellent description of convolutions, see: http://colah.github.io/posts/2014-07-Understanding-Convolutions
- expanding eq 10, we obtain,

$$\phi(x-\mu|\sigma) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] = \sqrt{\frac{w}{2\pi}}\exp\left[-\frac{w}{2}(x-\mu)^2\right]$$

  where $w \equiv 1/\sigma^2$ is the 'weight.'
- the product of two such functions is,

$$\phi(x-\mu_1|\sigma_1)\phi(y-x-\mu_2|\sigma_2) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-x-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

- we can arrange the quadratic as

$$\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-x-\mu_2}{\sigma_2}\right)^2 = (w_1+w_2)(x-\hat{x})^2 + \frac{w_1w_2}{w_1+w_2}(y-\mu_1-\mu_2)^2$$

  where $\hat{x} \equiv (w_1\mu_1 + w_2y - w_2\mu_2)/(w_1+w_2)$
- this product is still a Gaussian *wrt* x, and upon integrating out x, we obtain the convolution law,

$$\int_{-\infty}^{\infty}\mathrm{d}x\;\phi(x-\mu_1|\sigma_1)\phi(y-x-\mu_2|\sigma_2) = \phi(y-\mu|\sigma) \tag{11}$$

  where $\mu \equiv \mu_1+\mu_2$ and $\sigma = \sigma_1^2+\sigma_2^2$
- Therefore, two Gaussians convolve to make another Gaussians, with the means μ and variances σ^2 being additive.

7.5 An aside: cumulants

- Jaynes states that whether non-Gaussian distributions also have parameters additive under convolution leads to the notion of *cumulants*, and that this is critical to understand before going into the central limit theorem.
- Given a set of real functions $\{f_1(x), f_2(x), \ldots, f_n(x)\}$, defined on the real line and not necessarily non-negative, their integrals (zeroth moment) and their first, second, and third moments are:

$$Z_i \equiv \int_{-\infty}^{\infty} \mathrm{d}x\ f_i(x) < \infty \qquad S_i \equiv \int_{-\infty}^{\infty} \mathrm{d}x\ x^2 f_i(x) < \infty$$
$$F_i \equiv \int_{infty}^{\infty} \mathrm{d}x\ x f_i(x) < \infty \qquad T_i \equiv \int_{-\infty}^{\infty} \mathrm{d}x\ x^3 f_i(x) < \infty$$

- the convolution of f_1 and f_2 is defined by

$$h(x) \equiv \int_{-\infty}^{\infty} \mathrm{d}y\ f_1(y) f_2(x-y)$$

  or, condensed, $h = f_1 * f_2$, and it is associative: $(f_1 * f_2) * f_3 = f_1 * (f_2 * f_3)$
- what happens to the zeroth moment under convolution?

$$Z_h = \int_{-\infty}^{\infty} \mathrm{d}x\ \int_{-\infty}^{\infty} \mathrm{d}y\ f_1(y) f_2(x-y) = \int \mathrm{d}y\ f_1 Z_2 = Z_1 Z_2$$

- therefore, if $Z_i \neq 0$, we can multiply $f_i(x)$ by some constant which makes $Z_i = 1$
- the first moment under convolution is

$$\begin{aligned} F_h &= \int_{-\infty}^{\infty} \mathrm{d}x\ \int_{-\infty}^{\infty} \mathrm{d}y\ f_1(y) x f_2(x-y) = int \mathrm{d}y\ f_1(y)\ \int_{-\infty}^{\infty} \mathrm{d}q\ (y+q) f_2(q) \\ &= \int_{-\infty}^{\infty} \mathrm{d}y\ f_1(y)[y Z_2 + F_2] = F_1 Z_2 + Z_1 F_2 \end{aligned}$$

- therefore, the first moments are additive under convolution: $F_h = F_1 + F_2$
- doing the same for the second moment, we get, $[S_h - (F_h)^2] = [S_1 - (F_1)^2] + [S_2 - (F_2)^2]$
- and for the third moment, $T_h = T_1 Z_2 + 3 S_1 F_2 + 3 F_1 S_2 + Z_1 T_2$
- So, we can generalize the moments under convolution as:

$$F_h = \sum_{i=1}^{n} F_i$$
$$S_h = F_h^2 = \sum_{i=1}^{n} \left(S_i - F_i^2\right)$$
$$T_h - 3 S_h F_h + 2 F_h^3 = \sum_{i=1}^{n} \left(T_i - 3 S_i F_i + 2 F_i^3\right)$$

- these quantities accumulate additively under convolution, and are called *cumulants*.
- we define the *n*th cumulant as the *n*th moment, with correction terms.
  - begs two questions:
    1. do these correction terms always exist?
    2. is there a general algorithm to develop them?
  - recall that $\log \mathscr{F}(\alpha)$ is additive under convolution, and the power series expansions of $\mathscr{F}\alpha$ and $\log \mathscr{F}(\alpha)$:

$$\mathscr{F}(\alpha) = M_0 + M_1 (i\alpha) + M_2 \frac{(i\alpha)^2}{2!} + M_3 \frac{(i\alpha)^3}{3!} + \ldots$$
$$\log \mathscr{F}(\alpha) = C_0 + C_1 (i\alpha) + C_2 \frac{(i\alpha)^2}{2!} + C_3 \frac{(i\alpha)^3}{3!} + \ldots$$

with the coefficients

$$M_n = \frac{1}{i^n}\frac{\mathrm{d}^n \mathscr{F}(\alpha)}{\mathrm{d}\alpha^n} = \int_{-\infty}^{\infty} \mathrm{d}x\, x^n f(x)$$
$$C_n = \frac{1}{i^n}\frac{\mathrm{d}^n}{\mathrm{d}\alpha^n} \log \mathscr{F}(\alpha)\Big|_{\alpha=0}$$

which are additive under convolution and therefore cumulants.

7.5.1 Relation of cumulants and moments

- lets use a more general notation for the nth moment of a function:

$$M_n \equiv \int_{-\infty}^{\infty} \mathrm{d}x\, x^n f(x) = \frac{\mathrm{d}^n}{\mathrm{d}(i\alpha)^n} int\mathrm{d}x\, f(x) e^{i\alpha x}\Big|_{\alpha=0} = i^{-n}\mathscr{F}^{(n)}(0) \qquad n = 0, 1, 2, \ldots$$

  or, as a notation, $M_n = \overline{x^n}$, indicating the average of x_n wrt $f(x)$

- cumulants → moments
  - if a function $f(x)$ has moments of all orders, then its Fourier transform has a power series expansion,

$$\mathscr{F}(\alpha) = \sum_{n=0}^{\infty} M_n (i\alpha)^n$$

  - the first cumulant is the same as the first moment, $C_1 = M_1 = \overline{x}$
  - the second cumulant is $C_2 = M_2 - M_1^2$, which is

$$C_2 = \int \mathrm{d}x\, \left[x - M_1\right]^2 f(x) = \overline{(x - \overline{x})^2} = \overline{x^2} - \overline{x}^2$$

    which is the second moment of x about its mean value, called the *second central moment* of $f(x)$.
  - the third central moment is,

$$\int \mathrm{d}x\, (x - \overline{x})^3 f(x) = \int \mathrm{d}x\, \left[x^3 - 3\overline{x}x^2 + 3\overline{x}^2 x - \overline{x}^3\right] f(x)$$

    which is the same as the third cumulant, $C_3 = M_3 - 3M_1M_2 + 2M_1^3$
  - this (nth moment) = (nth cumulant) relation does not hold past the third moment

- example: Gaussian distribution
  - recall the Gaussian distribution,

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

    with the Fourier transform

$$\mathscr{F}(\alpha) = \exp\left[i\alpha\mu - \frac{\alpha^2\sigma^2}{2}\right]$$

  - such that

$$\log \mathscr{F}(\alpha) = i\alpha\mu - \frac{\alpha^2\sigma^2}{2}$$

    and

$$C_0 = 0 \qquad C_1 = \mu \qquad C_2 = \sigma^2$$

  - all higher C_n are zero, meaning the Gaussian distribution only has two nontrivial cumulants, the mean and the variance

7.6 The central limit theorem

- Definition: in most situations, when random independent variables are summed, their normalized sum trends towards a normal distribution
- if we have a set of functions $f_i(x)$ which are probability distributions, then they are by definition non-negative and normalized: $f_i(x) \geq 0$, $\int \mathrm{d}x\ f_i(x) = 1$.
- the zeroth moments are all $Z_i = 1$ and the Fourier transforms are absolutely convergent for all real α (means that the sum of the absolute integrands is finite):

$$\mathscr{F}_i(\alpha) \equiv \int_{-\infty}^{\infty} \mathrm{d}x\ f_i(x) e^{i\alpha x}$$

- consider two variables, x_1 and x_2, which are assigned probability distributions conditional on some information I: $f_1(x) = p(x_1|I)$, $f_2(x) = p(x_2|I)$
- we want the cumulative probability distribution $f(y)$ for the sum $y = x_1 + x_2$,

$$P(y' \leq y|I) = \int_{-\infty}^{\infty} \mathrm{d}x_1\ f_1(x_1) \int_{-\infty}^{y-x_1} \mathrm{d}x_2\ f_2(x_2)$$

  (Note: I don't understand why the upper limit of the second integral is $y - x_1$ and the first integral is ∞)
- so, the probability distribution for y is

$$f(y) = \left[\frac{\mathrm{d}}{\mathrm{d}y} P(y' \leq y|I)\right]_{y=y'} = \int \mathrm{d}x_1 f_1(x_1) f_2(y - x_1) = (f_1 * f_2)$$

- the probability distribution of a third variable $z = y + x_3$ is

$$g(x) = \int \mathrm{d}y\ f(y) f_3(z - y) = (f_1 * f_2 * f_3)$$

  and so on. so through induction, we can say that the probability distribution for the sum $y = x_1 + \cdots + x_n$ is the convolution $h_n(y) = (f_1 * \cdots * f_n)$
- in the cumulants section, we saw that convolution in x space corresponds to multiplication in Fourier transform space
- with the *characteristic function* (a function which completely defines its probability distribution, the Fourier transform of the probability distribution) for $f_k(x)$,

$$\phi_k(\alpha) \equiv \langle e^{i\alpha x} \rangle = \int_{-\infty}^{\infty} \mathrm{d}x\ f_k(x) e^{i\alpha x}$$

  with the inverse Fourier transform,

$$f_k(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}\alpha\ \phi_k(\alpha) e^{-i\alpha x}$$

- the probability distribution for the sum of n variables x_i is then,

$$h_n(q) = \frac{1}{2\pi} \int \mathrm{d}\alpha\ \phi_1(\alpha) \cdots \phi_n(\alpha) e^{-i\alpha q}$$

  of, if all probability distributions $f_i(x)$ are identical,

$$h_n(q) = \frac{1}{x\pi} \int \mathrm{d}\alpha\ [\phi(\alpha)]^n e^{-i\alpha q}$$

  with the probability density of the arithmetic mean, $\overline{x} = q/n$,

$$p(\overline{x}) = n h_n(n\overline{x}) = \frac{n}{2\pi} \int \mathrm{d}\alpha \left[\phi(\alpha) e^{-i\alpha\overline{x}}\right]^n$$

- there is only *one* probability distribution with this property.

– if the probability distribution $p(x|I)$ for a single observation x has the characteristic function,

$$\phi(\alpha) = \int \mathrm{d}x\; p(x|I)e^{i\alpha x}$$

then the one for the average of n observations, $\overline{x} = n^{-1}\sum x_i$, has the characteristic function of a form $\phi^n(n^{-1}\alpha)$

– it is a necessary and sufficient condition that for x and $\overline{x}$ to have the same probability distribution is for $\phi(\alpha)$ to satisfy the funtional equation $\phi^n(n^{-1}\alpha) = \phi(\alpha) \Rightarrow n\log\phi(\alpha) = \log\phi(n\alpha)$ for $-\infty < \alpha < \infty$, $n = 1, 2, 3, \ldots$

– this requires a linear relation, $\log\phi(\alpha) = C\alpha$, $0 \leq \alpha < \infty$, where C is some complex number $C = -k + i\theta$

– the most general solution satisfying $\phi(-\alpha) = \phi^*(\alpha)$ is

$$\phi(\alpha) = e^{i\alpha\theta - k|\alpha|} \qquad -\infty < \theta < \infty \qquad 0 < k < \infty$$

– which ultimately gives us the probability distribution,

$$p(x|I) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}x\; e^{-k|\alpha|}e^{i\alpha\theta - x} = \frac{1}{\pi}\left[\frac{k}{k^2 + (x-\theta)^2}\right]$$

which is the *Cauchy distribution* with median θ and quartiles $\theta \pm k$.

- there are a few interesting applications of this in pp. 224–240, which I will not replicate (see pg. 229 in particular)
- important note from these applications: any sufficently smooth non–Gaussian distribution can be generated from many different superpositions of different Gaussians of varying widths

8 Sufficiency, ancillarity, and all that

- these are features of probability theory that ''take care of themselves as long as we obey the rules.''
- help us better understand the inner workings of probability theory

8.1 Sufficiency

- Probability theory sometimes disregards information that we provide it. As an example, when we estimated a parameter θ of a binomial distribution from n trials, the posterior PDF for θ only depended on the n number of trials and r number of successes, and not at all on order of those successes/failures (several other examples are presented).
- it seems that some information provided must be *irrelevant* to the question at hand. can we confirm this?
  - if certain information is known but not used, it would follow that it would not matter to the conclusion if the information was instead completely unknown
  - so, if the posterior PDF for parameter θ is found to depend on the data $D = \{x_1, \ldots, x_n\}$ only through a function $r(x_1, \ldots, x_n)$, then it would follow that we would reach the same posterior PDF if given r alone
  - with a sampling density function $p(x_1 \ldots x_n|\theta)$ and prior $p(\theta|I) = f(\theta)$, the posterior PDF is,

$$p(\theta|DI) = h(\theta|D) = \frac{f(\theta)p(x_1 \ldots x_n|\theta)}{\int \mathrm{d}\theta' f(\theta')p(x_1 \ldots x_n|\theta')}$$

  - now let's change variables $(x_1, \ldots, x_n) \to (y_1, \ldots, y_n)$ in sample space S_x, such that $y_1 = r(x_1, \ldots, x_n)$ and choose $(y_2, \ldots, y_n)$ so that the Jacobian

$$J = \frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)}$$

is bounded and nonvanishing over all S_x.

  - thus, the change of variables is a 1:1 map of S_x onto S_y, and the sampling density,

$$g(y_1, \ldots, y_n|\theta) = J^{-1}p(x_1 \ldots x_n|\theta)$$

may be used in place of $p(x_1 \ldots x_n|\theta)$ (with the Jacobian, being independent of θ, cancelling out),

$$h(\theta|D) = \frac{f(\theta)g(y_1 \ldots y_n|\theta)}{\int \mathrm{d}\theta' f(\theta')g(y_1 \ldots y_n|\theta')}$$

- the property R, described by $r(x_1 \dots x_n)$, is the statement that for all $\theta \in S_\theta$ is independent of $(y_2 \dots y_n)$.
- this defines $n-1$ simultaneous integral equations that the prior $f(\theta)$ must satisfy,

$$\int_{S_\theta} d\theta' \, K_i(\theta, \theta') f(\theta') = 0 \qquad \left\{ \begin{matrix} \theta \in S_\theta \\ 2 \le i \le n \end{matrix} \right\} \tag{12}$$

where the ith kernel is,

$$K_i(\theta, \theta') \equiv g(y|\theta) \frac{\partial g(y|\theta')}{\partial y_i} - g(y|\theta') \frac{\partial g(y|\theta)}{\partial y_i}$$

where $y \equiv (y_1, \dots, y_n)$. $K_i(\theta, \theta')$ is antisymmetric.

8.1.1 Generalized sufficiency

- unlike in Fisher sufficiency (no notes here, see the book), the property R may hold under weak conditions that depend on the prior chosen.
- since the integral equations in 12 are linear, we may think of them in terms of linear vector spaces.
- Let all priors span a Hilbert space H of functions on the parameter space S_θ. If property R holds only for some priors $f(\theta) \in H'$ that span a subspace $H' \subset H$, then it is only required that the projection of $K_i(\theta, \theta')$ onto that subspace vanishes.
- so $K_i(\theta, \theta')$ is an arbitrary function on the Hilbert space $(H - H')$ of functions orthogonal to H' (need to think about this some more).
- given any specified function $r(x_1, \dots, x_n)$ and sampling density $p(x_1 \dots x_n|\theta)$, this determines a kernel $K_i(\theta, \theta')$. If this kernel is not complete (i.e. if as (θ, θ', i) vary, the kernel does not span all of Hilbert space H), then there are a set of nonvanishing solutions $f(\theta)$
- however, if there are non-negative solutions, then they determine a subclass of priors $f(\theta)$ for which r is a sufficient statistic.
- this leaves open the possibility that for different priors, different functions $r(x_1, \dots, x_n)$ may be sufficient statistics.
  - VERY IMPORTANT NOTE: ''This means that *use of a particular prior may make certain particular aspects of the data irrelevant. Then a different prior may make different aspects of the data irrelevant.*'' (pg. 249)

8.1.2 The likelihood principle

- According to Bayes' theorem, since the posterior PDF of θ is always a product of a prior, $p(\theta|I)$, and a likelihood function, $L(\theta) \propto p(D|\theta I)$, the only place the data appears is in the likelihood function.
- Therefore, we can make the following statement, called the *likelihood principle*,

  > *Within the context of the specified model*, the likelihood function $L(\theta)$ from data D contains all the information about θ tht is contained in D. (pg. 250)

- Think of it like this: given two data sets D and D' that lead to the same likelihood function $L(\theta) = aL'(\theta)$, where a is some constant independent of θ, both $L(\theta)$ and $L'(\theta)$ have the same importance for any inferences we make about θ
- Early argument supporting this principle, by George Barnard:
  - Any irrelevant data should cancel out of the inferences
  - suppose in obtaining D, we flip a coin and record the result $Z = H$ or T.
  - the sampling probability is then $p(DZ|\theta) = p(D|\theta)p(Z)$
  - the result of a coin flip can tell us nothing more about θ than the data D can, so inferences about θ based on D and DZ should be the same
  - so, constant factors in the likelihood must be irrelevant to inferences, i.e. inferences about θ may depend only on the ratios of likelihoods:

$$\frac{L_1}{L_2} = \frac{p(DZ|\theta_1 T)}{p(DZ|\theta_2 T)} = \frac{p(D|\theta_1 I)}{p(D|\theta_2 I)}$$

- Further arguments by Alan Birnbaum gave rise to the *conditionality principle*:

> Suppose we can estimate θ from either of two experiments, E_1 and E_2. If we flip a coin to decide which to do, then the information we get about θ should depend only on the experiment that was actually performed. That is, recognition of an experiment that might have been performed, but was not, cannot tell us anything about θ. (pg. 251)

- orthodox probability theory violates the likelihood principle for three reasons:
  1. the central statement of orthodox probability theory, that the merit of an estimator is solely determined by its sampling properties over time, makes no reference to the likelihood function.
  2. its second central statement, that the accuracy of an estimate is determined by the width of the sampling distribution, again makes no reference to the likelihood function.
  3. ''procedures in which 'randomization' is held to generate the probability distribution *used in the inference*!'' (not sure what this means yet)
- for example, in the case of the coin flip, if there is any connection between θ and the coin flip such that knowing θ would tell us something about the coin flip, then conversely, knowing the result of the coin flip must tell us something about θ
  - if the coin was tossed in some gravitational field, then there is a logical connection between the coin flip and knowledge of the gravitational field.
  - both Barnard and Birnbaum's arguments disregard this

8.2 Ancillarity

- consider the estimation of a parameter θ from a sampling distribution $p(x|\theta I) = f(x - \theta|I)$ (if the mean of a set of observations is used as the estimator, then the observed variation of the mean is the sampling distribution of the mean)
  - if we take some function $\theta^*(x_1, \ldots, x_n)$ as the estimator, two different data sets may give the same estimate for θ but have different configurations, e.g. range, central moments, etc.
  - so, a broad distribution and sharply peaked distribution could give the same estimated θ, but tell us different conclusions about the accuracy
  - but we previously held that the accuracy of the estimate was determined by the width of the *sampling distribution*. from this we would have to conclude that both estimates are equally accurate. how do we rectify this?
- Fisher's remedy was to propose using sampling distributions conditional on some 'ancillary' statistic, $z(x_1, \ldots, x)n)$ that gives some information about the configuration of the data
  - this could require as many as $(n-1)$ ancillary statistics, which Fisher often could not provide, as he demanded that $p(z|\theta I) = p(z|I)$
- what can we make of this from a Bayesian viewpoint?
  - Fisher's sampling distribution would be,

$$p(D|z\theta I) = \frac{p(zD|\theta I)}{p(z|\theta I)} = p(D|\theta I)\frac{p(z|D\theta I)}{p(z|\theta I)}$$

  - if $z = z(D)$ is a function of the data alone, then $p(z|D\theta I) = \delta[z - z(d)]$
  - so, if $p(z|\theta I)$ is independent of θ, then the conditional sampling distribution $p(D|z\theta I)$ also is independent of θ
  - another way, from a Bayesian standpoint, Fisher's procedure does nothing, the likelihood is unchanged
  - can also thought of in terms of previous discussions; if z is a function of only the data, then we know the value of z from the data already. i.e. $AA = A$

8.2.1 Generalized ancillary information

- Let's take a broad view of ancillary information, referring to any additional quantity Z that is not part of prior information or data
- Let

$$\theta = \text{parameters} \tag{13}$$
$$E = e_1, \ldots, e_n \quad \text{noise} \tag{14}$$
$$D = d_1, \ldots, d_n \quad \text{data} \tag{15}$$
$$d_i = f(t_i\theta) + e_i \quad \text{model} \tag{16}$$
$$Z = z_1, \ldots, z_n \quad \text{ancillary data} \tag{17}$$

- we want to estimate θ from the posterior PDF

$$p(\theta|DZI) = p(\theta|I)\frac{p(DZ|\theta I)}{p(DZ|I)}$$

  - assume Z is independent of θ, i.e. $p(\theta|ZI) = p(\theta|I)$ and $p(Z|\theta I) = p(Z|I)$ (these expressions are the same, as per the product rule)
  - expanding the likelihood ratio with this assumption, we can rewrite the posterior PDF as

$$p(\theta|DZI) = p(\theta|I)\frac{p(DZ|\theta I)}{p(DZ|I)} = p(\theta|ZI)\frac{p(D|\theta ZI)}{p(D|ZI)}$$

    so that the generalized ancillary information appears to be part of the prior information

- one important property of generalized ancillary information is that the relationship between Z and θ is reciprocal, i.e. if we are interested in estimating Z, θ is then the generalized ancillary statistic

8.2.2 Asymptotic likelihood: Fisher information

- Given a data set $D \equiv \{x_1, \ldots, x_n\}$, the log likelihood is

$$\frac{1}{n}\log L(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log[p(x_i|\theta)]$$

- what happens as continue to accumulate data?
  - usual assumption is that as $n \to \infty$, the sampling distribution $p(x|\theta)$ is equal to the limiting relative frequencies of the various data values x_i
  - with this assumption, as $n \to \infty$,

$$\frac{1}{n}\log L(\theta) \to \int \mathrm{d}x\; p(x|\theta_0)\log[p(x|\theta)]$$

    where θ_0 is the 'true' value
  - the entropy of the 'true' value is

$$H_0 = -\int \mathrm{d}x\; p(x|\theta_0)\log[p(x|\theta_0)]$$

    we have the asymptotic likelihood function

$$\frac{1}{n}\log[L(\theta)] + H_0 = \int \mathrm{d}x\; p(x|\theta_0)\log\left[\frac{p(x|\theta)}{p(x|\theta_0)}\right] \leq 0$$

  - this is an equality iff $p(x|\theta) = p(x|\theta_0)$ for all x for which $p(x|\theta_0 > 0$
  - however, if two values θ and θ_0 lead to the same sampling distributions, then the data cannot distinguish between them

- if different values of θ always leads to different sampling distributions, then the asymptotic likelihood function reaches equality iff $\theta = \theta_0$, so $L(\theta)$ reaches its maximum at $\theta = \theta_0$
  - given a multidimensional parameter $\theta \equiv \{\theta_1, \ldots, \theta_m\}$, and expanding around $\theta = \theta_0$, we get

$$\log[p(x|\theta)] = \log[p(x|\theta_0) - \frac{1}{2}\sum_{i,j=1}^{m} \frac{\partial^2 \log[p(x|\theta)]}{\partial\theta_i \partial\theta_j}\delta\theta_i \delta\theta_j \Rightarrow$$

$$\frac{1}{n}\log\left[\frac{L(\theta)}{L(\theta_0}\right] = -\frac{1}{2}\sum_{ij} I_{ij}\delta\theta_i\delta\theta_j$$

    where

$$I_{ij} \equiv \int \mathrm{d}^n x \; p(x|\theta_0)\frac{\partial^2 \log[p(x|\theta)]}{\partial\theta_i\partial\theta_j}$$

    is called the *Fisher information matrix*, which tells us how large the separation $|\theta - \theta'|$ between two close values θ and θ' must be before we can distinguish them in an experiment

8.3 Combining evidence from different sources

- Starts with a quote that I like:

  > We all know that there are good and bad experiments. The latter accumulate in vain. Whether there are a hundred or a thousand, one single piece of work by a real master — Pasteur, for example — will be sufficient to sweep them into oblivion. (Henri Poincaré, 1904, pg. 141)

- it is the naïve assumption that, for example, if we have 25 experiments with accuracy of ±10%, then by averaging them we would get an accuracy of $\pm 10/\sqrt{25} = \pm 2\%$. we can use probability theory to determine when this is a safe assumption
- this is called *meta-analysis*
- an example of how this assumption can lead us awry
  - suppose that each person in china knows the height of the emperor to within ±1 meter.
  - if we poll $N = 1 \times 10^{10}$ inhabitants, then, by this logic, we would get an accuracy of

$$\frac{1}{\sqrt{1 \times 10^{10}}} = 3 \times 10^{-5}\mathrm{m} = 0.03\mathrm{mm}$$

    which is obviously absurd
  - therefore, the $\sqrt{N}$ rule is not always valid
  - the data values must not be merely causally independent, they must be *logically* independent for this to hold true
  - most inhabitants of china have never seen the emperor, so their estimates are built on folklore, which surely introduces some systematic error
  - can put this as

$$\text{error in estimate} = S \pm \frac{R}{\sqrt{N}}$$

    where S is the systematic error in each datum, R is the RMS random error in the individual data points

- so we know that we must use some kind of probabilistic model in scientific inference to be truly rigorous
- unless we know that the systematic error is less than rougly one–third of the random error, we can't be sure that the average of one million data points is more accurate than the average of ten
  - Jaynes throws some shade at 'soft sciences' here:

    > Indeed, this has been well recognized by experimental physicists for generations: but warnings about it are conspicuously missing from textbooks written by statisticians, and so it is not sufficiently recognized in the 'soft' sciences whose practitioners are educated from those textbooks. (pp. 258–259)

- let's investigate this from a Bayesian perspective:

- suppose we want to know about hypothesis H using two experiments which yield data sets A and B
- with prior information I, we first know

$$p(H|AI) = p(H|I)\frac{p(A|HI)}{p(A|I)}$$

- with this as the prior probability, adding in data set B gives us

$$p(H|ABI) = p(H|AI)\frac{p(B|AHI)}{p(B|AI)} = p(H|I)\frac{p(A|HI)p(B|AHI)}{p(A|I)p(B|AI)}$$

- knowing that $p(A|HI)p(B|AHI) = p(AB|HI)$ and $p(A|I)p(B|AI) = p(AB|I)$ (the product rule), this further reduces to

$$p(H|ABI) = p(H|I)\frac{p(AB|HI)}{p(AB|I)}$$

- this is the same probability as if we used $C = AB$ in a single Bayes theorem. This is called the *chain consistency property*
- therefore, it is valid to combine *evidence* from several experiments iff:
  1. the prior information I is the same
  2. the prior for each experiment includes the results of the earlier experiments
- condition 2 can apparently be violated properly if A and B are totally logically independent

- his conclusion is that, although meta-analysis is not always incorrect, if it is applied without careful thought, it is wildly misleading

8.4 A folk theorem

- suppose we are given a number of unknowns $\{x_i, \ldots, x_n\}$ in some domain X to be determined, and are given m functions of them,

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_n) \\ y_2 &= f_2(x_1, \ldots, x_n) \\ &\ldots \\ y_m &= f_m(x_1, \ldots, x_n) \end{aligned}$$

- if $m = n$ and the Jacobian $\partial(y_1, \ldots, y_n)/\partial(x_1, \ldots, x_n)$ is non–zero, then we can solve for individual x_i
- but if $m < n$, then the system is *underdetermined* and we cannot find individual x_i as we have insufficient information
- gives rise to the folk theorem, that from m observations one cannot estimate more than m parameters.
- nothing in probability theory places this limited. as our data tend to zero, it is not that we cannot estimate many parameters, its just that those estimates must relax back to the prior estimates
- a grain of truth is found in this theorem if adding parameters correlates to the posterior PDF, i.e. if the posterior PDFs for different parameters are not independent

9 Repetitive experiments: probability and frequency

- traditionally, probability theory has focused on drawing inferences from experiments which can be repeated indefinitely under identical conditions, but give different results on different trials
- applications–oriented definitions of probability focus on 'limiting frequency in independent repetitions of random experiments' rather than on its logic
- many examples of how frequency–based probability theory fails

9.1 Physical experiments

- inferences from experiments must obviously take into account knowledge of the physical laws concerning the situation; e.g. inferences from medicatal experiments must take into account physiology and biochemistry
- this knowledge will determine the model used to help solve the problem
- in repeatable experiments, we can keep some relevant factors the same for all trials, but some are not under our control
  - the factors that are teh same are called *systematic*
  - the factors which vary uncontrollably are called *random*, or as Jaynes likes to say, *irreproducible by the experimental technique used*
- let our experiment consist of n trials with m possible results at each trial (so if tossing a coin, $m = 2$, if rolling a die, $m = 6$)
  - m is a description of the state of knowledge within which we conduct probabilistic analysis, but it may or may not correspond to the number of *real* possibilities. we typically cannot know in advance the true value of m
  - generally, we should take m as the number of results per trial that we will take into account *in the present calculation*
    - by specifing m, we are defining a tentative working hypothesis
  - we are concerned with two sample spaces: S for a single trial, consisting of m points, and an extension space, $S^n = S \otimes S \otimes \cdots \otimes S$ ($\otimes$ is a tensor product of vector spaces, yielding another vector space)
    - *result* refers to S space, *outcome* refers to S^n
    - the number of results being considered is then m, the number of outcomes is then $N = m^n$
  - let the ith trial be denoted by r_i $(1 \le r_i \le m, 1 \le i \le n)$
  - any outcome of the experiment denoted by the set of results $D = \{r_1, \ldots, r_n\}$
  - outcomes are mutually exclusive and exhausive, so, given prior information I, we can assign a probability $P(D|I) = p(r_1 \ldots r_n)$, satisfying the sum over all data sets,

$$\sum_{r_1=1}^{m} \sum_{r_2=1}^{m} \cdots \sum_{r_n=1}^{m} p(r_1 \ldots r_n) = 1$$

  - regard r_i as digits (% m) in a number R in base m: $(0 \le R \le N-1)$ (oh god why)

> More generally, for an experiment with m possible results at each trial, repeated n times, we communicate with the robot in the base m system, whereupon each number displayed with have exactly n digits, and for us the ith digit will represent, % m, the result of the ith trial. (pg. 273)

9.1.1 Poorly defined prior

- suppose we know only that there are N possibilities, and nothing else; we are completely ignorant not only of the physical laws of the system, but also that the full experiment consist of n repetitions of a simpler experiment
- let's first see what predictions we can make while this poorly informed, and see how this changes as we add more and more information
  - denote the initial state of ignorance by I_0
  - the principle of indifference then applies, with the sample space consisting of $N = m^n$ discrete points, each with N^{-1} probability
  - a proposition A defined to be true in the subspace $S' \subset S^n$ and false in the subspace $S^n - S'$ will be assigned the probability

$$P(A|I_0) = \frac{M(n, A)}{N}$$

    where $M(n, A)$ is the multiplicity of A, the number of points in S^n where A is true
  - trivial to show that if $m = 6$ (rolling a die), $p(r_1 = k|I_0) = 1/6$ and $p(r_1 = k, r_2 = j|I_0) = 1/36$ where $(1 \le k \le 6)$ and $(1 \le i \le n)$, and so on
- adding more information: out of the possible N outcomes, the correct one belongs to the subclass in which the first digit is $r_1 = 3$

- what probability can we now assign to the proposition $r_2 = j$?
- probability determined by product rule,

$$p(r_1|r_1 I_0) = \frac{p(r_1 r_2|I_0)}{p(r_1|I_0)}$$

- which is then

$$p(r_2|r_1 I) = \frac{1/36}{1/6} = 1/6$$

- so the probability is then unchanged

- defining further die rolls will give similar results: this type of information clearly doesn't help our inferences

9.2 Induction

- a human, noticing a regular pattern in results, would expect the pattern to continue. this is called *induction*.
- our inferences made with the poorly defined prior does not have the ability to reason inductively
- want to show that, if we provide a prior which provides logical connections between different trials, that induction can be done logically
- for example, by what logic can we say that, because 672 people of 1000 people polled support some proposition, 67% of the entire population support the proposition?

9.2.1 Are there general inductive rules?

- we showed that with prior information I_0, the results of each toss are logically independent propositions
  - this is because the prior does not include the fact that each successive toss $\{r_1, r_2, \ldots\}$ are successive repetitions of the same experiment
- what, then, is the 'extra, hidden' information that we use to inductively reason that we could include in this prior?
- suppose the data we have say that the first three tosses of the coin are all heads: $D = H_1 H_2 H_3$. what is our *intuitive* probability $p(H_4|DI)$
  - this depends on the prior used
  - if we use I_0, the answer is *always* $p(H_4|DI) = 1/2$
  - other priors could be:

    $I_1 \equiv$ we have examined the coin and know that each side is symmetrical, with 'no funny business' $I_2 \equiv$ we did not examine the coin and cannot know the honesty of the coin or the person tossing it
  - with prior I_1, we have no reason to believe $p(H_4|DI)$ is anything other than $1/2$
  - but with prior I_2, we would feel that three heads in a row would increase the probability of a fourth head, $p(H_4|DI) > 1/2$
  - interesting paradox that I_1 has more information than I_2, but it agrees more with I_0
- there are a great variety of conclusions based on what prior you use, so there is no universal inductive rule; different inductive rules correspond to different prior information

9.3 Multiplicity

- recall the multiplicity factor, $M(n, A)$
- this may seem trivial, but could in fact be extremely computationally intensive. for example, what if we toss the die twelve times? then the number of conveivable outcomes is $6^{12} = 2.18 \times 10^9$
- so while it is true that we are dealing in finite sets, we are dealing in very large finite sets. how do we calculate on these sets?
- most problems can fit into the following scheme:

- let $\{g_1, g_2, \ldots, g_m\}$ be any set of m finite, real numbers
- the individual g_j are independent, but additive
- lets also say we are interested in predicting linear functions of the n_j
- the total amount of G generated is then

$$G = \sum_{i=1}^{n} g(r_i) = \sum_{j=1}^{m} n_j g_j$$

where the sample number n_j is the number of times the jth result happened

- the probability of obtaining G is

$$p(G|n, I_0) = f(G|n, I_0) = \frac{M(n, G)}{N}$$

where $N = m^n$ and $M(n, G)$ is the multiplicity of G

9.3.1 Partition function algorithms

- expanding $M(n, G)$ around the nth result gives

$$M(n, G) = \sum_{j=1}^{m} M(n-1, G - g_j)$$

- cannot solve this directly for large n. it is a linear difference equation, so the solutions must be of the form $\exp\{\alpha n + \lambda G\}$
- this is a solution if α and λ are related by

$$e^{\alpha} = Z(\lambda) \equiv \sum_{j=1}^{m} e^{-\lambda g_j}$$

where $Z(\lambda)$ is a *partition function*

- an arbitrary superposition of the elementary solutions is

$$H(n, G) = \int \mathrm{d}\lambda \; Z^n(\lambda) e^{\lambda G} h(\lambda)$$

which is a form of an *inverse Laplace transformation*

- the true multiplicity $M(n, G)$ must satisfy the boundary condition $M(0, G) = \delta(G, 0)$ and is only defined at certain discrete values
- to find the discrete Laplace transform of $M(n, G)$, we multiply it by $\exp\{-\lambda G\}$ and sum over all possible values,

$$\sum_{G} e^{-\lambda G} M(n, G) = \sum_{n_j \in U} W(n_1, \ldots, n_m) \exp\left\{-\lambda \sum n_j g_j\right\} \tag{18}$$

where

$$W(n_1, \ldots, n_m) \equiv \frac{n!}{n_1! \cdots n_m!}$$

is the multinomial coefficient, containing the number of outcomes leading to sample numbers n_j

- this contains contributions from every possible outcome in the experiment.
- if we let $N_j^{n_j} = \exp\{-n_j g_j\}$, then $\exp\left\{-\sum_j^m n_j g_j\right\} = x_1^{n_1} x_2^{n_2} \ldots x_m^{n_m}$ and the multinomial expansion is

$$(x_1 + \cdots + x_m)^n = \sum_{n_j \in U} W(n_1, \ldots, n_m) X_1^{n_1} \ldots X_m^{n_m} \tag{19}$$

- the universal set which we sum over contains all possible sample numbers in n trials, and is defined by

$$\left\{ U : n_j \geq 0, \qquad \sum_{j=1}^{m} n_j = n \right\}$$

- comparing eqs. 18 and 19, this is just
$$\sum_G e^{-\lambda G} M(n, G) = Z^n(\lambda)$$
(I think this is because eq. 19 is equivalent to $Z^n(\lambda)$)
- this shows the number of ways $M(n, G)$ in which G can be realized is a coefficient of $\exp\{-\lambda G\}$ in $Z^n(\lambda)$
- so $Z(\lambda)$ to the nth power is the exact way in which all possible outcomes in n trials are partitioned among the possible values of G
- there follows (pp. 282–285) a very interesting discussion and calculation of the partition functions in simple cases, wherein he derives the binomial distribution from this information and the 'poor prior', I_0
- he says that the usual probability distributions, Poisson, gamma, Gaussian, chi–squared, etc., are all limiting forms of the binomial distribution, and that ''*frequentist probability theory is, functionally, just the reasoning of the poorly informed robot.* (pg. 284)
- so, the poor prior is unable to lead to inductive reasoning

9.3.2 Entropy algorithms

- Consider a proposition $A(n_1, \ldots, n_m)$ which is a function of sample numbesr n_j and is defined to be true when $(n_1, \ldots, n_m)$ are in some subset $R \in U$ and false in a complementary set $\overline{R} = U - R$
- if A is linear on n_j, then A acts like G, and the multiplicity of A is
$$M(n, G) = \sum_{n_j \in U} W(n_1, \ldots, n_m)$$
- how many terms $T(n, m)$ are in this sum?
$$T(n, m) = \binom{n+m-1}{n} = \frac{(n+m-1)!}{n!(m-1)!}$$
this is the same as the 'Bose–Einstein multiplicity factor' from statistical mechanics, and a $n \to \infty$,
$$T(n, m) \sim \frac{n^{m-1}}{(m-1)!}$$
- we find, after a few steps I won't replicate, that
$$\frac{1}{n} \log M(n, A) \to \frac{1}{n} \log(W_{\max})$$
where $n \to \infty$ and $W_{\max} \equiv \max_R [W(n_1, \ldots, n_m)]$, or the greatest term in region R
- so W grows so rapidly with n that a single maximum term in the sum dominates it
- how does $\log(W/n)$ behave in the limit in which the sample frequencies $f_j = n_j/n$ tend to constants, i.e. the limit of $n \to \infty$ of
$$\frac{1}{n} \log \left[\frac{n!}{(nf_1)! \cdots (nf_m)!} \right]$$
as f_j are held constant
- using Stirling's approximation, we find that $\log(W/n)$ tends to a finite constant independent of n,
$$\frac{1}{n} \to H \equiv \sum_{j=1}^{m} f_j \log(f_j) \tag{20}$$
which is called the *entropy* of the frequency distribution $\{f_1, \ldots, f_m\}$
- so, for large n, if the f_j tend to constants, the multiplicity of A tends to $M(n, A) \sim \exp\{nH\}$

- IMPORTANT:

  > We now see what was not evident before; that this multiplicity is to be found by determining the *frequency* distribution $\{f_1, \ldots, f_m\}$ which has maximum entropy subject to whatever constraints define R. (pg. 287)

- this is, along with eq. 20, the *maximum entropy principle*
- how does acquiring this information change the results of the poor prior?
  - first note that if A is linear in n_j, then the multiplicity is asymptotically equal to $M(n, G) = \exp\{nH\}$
    - the probability of getting the total G is then

      $$p(G|n, I_0) = m^{-n} e^{nH} = e^{-n(H_0 - H)}$$

      where $H_0 = \log(m)$ is the absolutely maximum of the entropy (derived below)
    - the difference between the maxium entropy and the observed entropy is a measure of how strong the constraints R are
  - we now know that a specified trial yields some amount g_j, which changes the multiplicity of A, since now the remaining $(n-1)$ trials must yield $(G - g_j)$, with a multiplicity of $M(n-1, G-g_j)$
  - the frequencies also change, as one trial yielding g_j is omitted,

    $$f_k = \frac{n_k}{n} \Rightarrow f'_k = \frac{n_k - \delta_{jk}}{n-1} \qquad 1 \leq k \leq m$$

  - these changes lead to a small change in the entropy, $H' = H + \delta H$, where

    $$\delta H = \sum_k \frac{\partial H}{\partial f_k} + O\left(\frac{1}{n^2}\right) = \left[\frac{H + \log(f_j)}{n-1}\right] + O\left(\frac{1}{n^2}\right)$$

    so,

    $$H' = \frac{nH + \log(f_j)}{n-1} + O\left(\frac{1}{n^2}\right)$$

  - leading to a multiplicity of

    $$M(n-1, G-g_j) = e^{(n-1)H'} = f_j e^{nH}\left[1 + O\left(\frac{1}{n}\right)\right]$$

  - in large n limit, the set S^n disappears, and all we need to do is calculate the f_k which maximizes the entropy over the domain R
  - with this new information, the probability to get total G is

    $$p(G|r_i = j, nI_0) = \frac{M(n-1, G-g_j)}{m^{n-1}}$$

  - given I_0, the prior probability for $r_i = j$ is

    $$p(r_i = j|nI_0) = \frac{1}{m}$$

  - and applying Bayes' theorem,

    $$\begin{aligned} p(r_i = j|GnI_0) &= p(r_i = j|nI_0)\frac{p(G|r_i = j, nI_)}{p(G|nI_0)} \\ &= \frac{1}{m}\frac{[M(n-1, G-g_j)/m^{n-1}]}{[M(n,G)/m^n]} = \frac{M(n-1, G-g_j)}{M(n,G)} = f_k \end{aligned}$$

  - therefore, knowing G changes the probability of the jth result from uniform $1/m$ to the observed frequency f_k of that result

9.3.3 Maximum entropy

- in the situation where $G = A$, and we are concerned with a linear function $G = \sum n_j g_j$, the domain R is defined by the average of G over n trials:

$$\overline{G} = \frac{G}{n} = \sum_{j=1}^{m} f_j g_j$$

- the problem of maximization thus has a solution, provided by Gibbs (!). a full description of maxiumum entropy will come later, but this focuses on the situation at hand
  - Let $\{f_1, \ldots, f_m\}$ be any possible frequency distribution over m points, satisfying $(f_j \geq 0)$ and $\sum_j f_j = 1$
  - let $\{u_1, \ldots, u_m\}$ be any other frequency distribution satisfying the same critera
  - then using the fact that on the positive real line $\log(x) \leq (x-1)$ iff $x = 1$ (need to think more about this), we get

$$\sum_{j=1}^{m} f_j \log\left(\frac{u_j}{f_k}\right) \leq 0$$

    with equality iff $f_j = u_j$ for all j
  - in this, we see eq. 20, so this inequality becomes

$$H(f_1, \ldots, f_m) \leq -\sum_{j=1}^{m} f_j \log(u_j)$$

  - for example, let's choose

$$u_j = \frac{e^{-\lambda g_j}}{Z(\lambda)}$$

    where we choose some λ such that an average $\overline{G} = \sum u_j g_j$ is attained
  - the Gibb's inequality then becomes

$$H \leq \sum f_j g_j + \log\left[Z(\lambda)\right]$$

  - varying f_j over all frequency distributions that yield the wanted $\overline{G}$, and we get

$$H_{\max} = \overline{G} + \log\left[Z(\lambda)\right]$$

    iff $f_j = u_j$
  - it is further evident, from the definition we chose for u_j, that

$$\overline{G} = -\frac{\partial}{\log(Z)\partial\lambda}$$

    and the solution must be a decreasing monotonic function
  - this is the canonical ensemble formalism from statistical mechanics

9.4 Significance tests

- Significance tests demonstrate the subtle differences between frequency and probability
- recall that how some evidence E affects our view of a certain hypothesis depends entirely on which hypothesis it is being tested against
- suppose we wish to consider two hypotheses, H and H', and with data D and prior information I, we must have $P(H|DI) + P(H'|DI) = 1$
- in terms of logarithmic measure of plausibility (in decibels), we have

$$e(H|DI) = e(H|I) + 10\log_{10}\left[\frac{P(D|H)}{P(D|H')}\right]$$

  which is a precise way of saying ''Data D supports hypothesis H relative to H' by $10\log_{10}[P(D|H)/P(D|H')]$ decibels''

- the world *relative* is critical in this sentence; an different hypothesis H'' could change the evidence the evidence in a different manner
- no matter what H' is, we *must* have $p(D|H') \leq 1$
- so a statement independent of any alternative hypothesis is

$$e(H|DI) \geq p(H|I) + 10 \log_{10} p(D|H) = e(H|I) - \psi_\infty$$

  where

$$\psi_\infty \equiv -10 \log_{10} p(D|H) \geq 0$$

- so there is no possible alternative hypothesis which the data could support, relative to H, by more than ψ_∞ decibels
- we can answer the question ''Are there any alternatives H' which data D would support relative to H, and how much support is possible?''
- we are not concerned with considering all possible alternatives, only those in some class Ω which we consider to be 'reasonable'. here's an example
  - consider tossing a die with m possible results $\{A_1, \ldots, A_m\}$ at each trial
  - further more, let $x_i \equiv k$, if A_k is true at the ith trial. so each x_i can take on the values $(1, 2, \ldots, m)$
  - let our class of hypotheses of interest be a 'Bernoulli class' B_m, where there are m possible results at each trial with the probabilities of A_k on successive repetitions being independent and stationary
  - so when H is in B_m, the conditional probability of any specific sequence of observations has the form

$$p(x_1 \ldots x_n|H) = p_1^{n_1} \ldots p_m^{n_m}$$

    where n_k is the sample number
  - to every hypothesis in B_m there is a set of numbers $\{p_1 \ldots p_m\}$ where $p_k \geq 0$ and $\sum_k p_k = 1$, and, for our purposes, these numbers characterize the hypothesis completely
  - An important lemma from Gibbs
    * let $x = n_k/np_k$, and knowing that on the positive real line $\log(x) \geq (1 - x^{-1})$ with equality iff $x = 1$, we know that

$$\sum_{k=1}^{m} n_k \log\left(\frac{n_k}{np_k}\right) \geq 0$$

      with equality iff $p_k = n_k/n$ for every k
    * this is the same as

$$\log p(x_1 \ldots x_n|H) \leq n \sum_{k=1}^{m} f_k \log(f_k)$$

      where $f_k = n_k/n$ is the observed frequency of result A_k
    * the right hand side of the ψ_∞ inequality depends only on D, so the closer to equality each hypothesis brings this value to, the better the fit to the data
  - for convenience, lets express ψ_∞ in decibels

$$\psi_B \equiv 10 \sum_{k=1}^{m} n_k \log_{10}\left(\frac{n_k}{np_k}\right)$$

  - consider two hypotheses $H = \{p_1, \ldots, p_m\}$ and $H' = \{p'_1, \ldots, p'_m\}$ with respective ψ_B and ψ'_B
  - Bayes' theorem is then

$$e(H|x_1 \ldots x_n) = e(H|I) + 10 \log_{10}\left[\frac{p(x_1 \ldots x_n|H)}{p(x_1 \ldots x_n|H')}\right] = e(H|I) + \psi'_B - \psi_B$$

  - we can always find a hypothesis H' in B_m for which $p'_k = n_k/n$ and so $\psi'_B = 0$. so ψ_B then means:

> Given an hypothesis H and the observed data $D \equiv \{x_1, \ldots, x_n\}$, compute ψ_B … Then, given any ψ in the range $0 \leq \psi \leq \psi_B$, it is possible to find an alternative hypothesis H' in B_m such that the data support H' relative to H by ψ decibels. There is no H' in B_m which is supported relative to H by more than ψ_B decibels. (pg. 298)

  - $(-\psi_B/n)$ is the entropy per symbol $H(f;p)$ of the observed distribution relative to the 'expected' distribution

- he does a comparison of the χ^2 and ψ tests, and shows that χ^2, because of the $(1/p_i)$ term, concentrates on unlikely possibilities of the hypothesis and over-penalizes hypotheses which have slight discrepancies between the expected and observed sample numbers

- concludes:

> For testing hypotheses involving moderately large probabilities, which agree moderately well with observation, it will not make much difference whether we use ψ or χ^2. But for testing hypotheses involving extremely unlikely events, we had better use ψ; or life might become too exciting for us. (pg. 302)

10 Discrete prior probabilities: the entropy principle

- in order to consistently provide 'correct' probabilities, we must first know how to recognize the relevance of prior information

- if we can break the situation up into mutually exclusive, exhaustive probabilities, we can use the principle of indifference. but there is often prior information that doesn't change the number of possibilities, but does give reason to prefer some over the others

10.1 A new kind of prior information

- imagine a problem in which the prior information consists of average values of certain things, like the average window is broken into $\overline{m} = 9.76$ pieces. what we are not told is that out of 100 windows, 976 pieces of glass were found.

- given this information, what is the probability that a window will break into m pieces? or more generally, how do we assign informative prior probabilities?

- this is actually two problems: estimating a frequency distribution and assigning a probability distribution. in an exchangeable sequence, these are almost identical mathematically

- a uniform prior represents a state of knowledge absolutely noncommittal with respect to all possibilities. what we would like to do is assign a probability that is as uniform as possible, while still agreeing with the available information

- this is a *variational* problem (think principle of least action or variational method in QM)

10.1.1 Minimum $\sum p_i^2$

- a measure of how spread out a probability distribution is is the sum of the squares of the probabilities assigned to each possibility

- the distribution which minimizes this expression, with relevant constraints, may be a reasonable way to solve the above stated problem

- mathematically, we want to minimize

$$\sum_m p_m^2 \tag{21}$$

  subject to the constraint that the sum of all p_m is unity and the average over the distribution is $\overline{m}$

  - applying the variational method, we get the solution

$$\delta\left[\sum_m p_m^2 - \lambda \sum_m m p_m - \mu \sum_m p_m\right] = \sum_m (2p_m - \lambda m - \mu)\delta p_m = 0$$

    where λ and μ are *Lagrange multipliers*.

- Lagrange multipliers are used to find local maxima and minima of functions subject to constraints. It seems that there is one multiplier for each constraint. for a case of a n choice variables and M constraints, the Lagrangian takes the form
$$\mathscr{L}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_M) = f(x_1, \ldots, x_n) - \sum_{k=1}^{M} \lambda_k g_k(x_1, \ldots, x_n)$$
- so p_m will be a linear function of m, $2pm - \lambda m - \mu = 0$
- μ and λ can be found from
$$\sum_m p_m = 1$$
and
$$\sum_m m p_m = \overline{m}$$
- if m is a positive, real integer, then we get the solutions
$$p_1 = \frac{4}{3} - \frac{\overline{m}}{2} \qquad p_2 = \frac{1}{3} \qquad p_3 = \frac{\overline{m}}{2} - \frac{2}{3}$$
- but then the probability for p_1 becomes negative when $\overline{m} > 2.667$ and p_3 becomes negative when $\overline{m} < 1.33$. so the formal solution for the min $[\sum p_i]$ lacks the property of non-negativity

- so while not exactly what we want, variational methods do have promise

10.1.2 Entropy: Shannon's theorem

- comes from Shannon's work on information theory
- if there exists a consistent measure of 'uncertainty' represented by a probability distribution, there are certain conditions which it must satisfy (this is taken verbatim)
  1. We assume that some numerical measure $H_n(p_1, \ldots, p_n)$ exists; i.e. that it is possible to set up some kind of association between 'uncertainty' and real numbers
  2. We assume a continuity principle: H_n is a continuous function of the p_i. Otherwise an arbitrarily small change in the probability distribution would lead to a large change in the uncertainty.
  3. We require that this measure should correspond qualitatively to common sense in that, when there are many possibilities, we are more uncertain than when there are few. This condition takes the form that in the case that all p_i are equal, the quantity:
$$h(n) = H_n\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$$
is a monotonic increasing function of n. This establishes the 'sense of direction'
  4. We require that the measure H_n be consistent in the same sense as before; i.e. if there is more than one way of working out its value, we must get the same answer for every possible way
- suppose we perceive two possibilities, which we assign p_1 and $q \equiv 1 - p_1$. the amount of uncertainty is then represented by the distribution $H_2(p_1, q)$
- what if we learn that the second possibility is actually two possibilities such that $p_2 + p_3 = q$? what then is the full uncertainty, $H_3(p_1, p_2, p_3)$
  - this process can be broken into two steps
  - first decide if the first possibility is true. this removes uncertainty $H(p_1, q)$
  - then we encounter the additional uncertainty of events 2 and 3, with probability q, leading to
$$H_3(p_1, p_2, p_3) = h_2(p_1, q) + qH_2\left(\frac{p_2}{q}, \frac{p_3}{q}\right)$$
as the condition that we obtain the same net uncertainty for either method
  - generally H_n can be broken down in many different ways

- he shows a proof of Shannon's theorem on pp. 348–350, which I will not replicate (also, I really have a hard time following it, as a not math person)
- Shannon's theorem: the only function $H(p_1, \ldots, p_n)$ satisfying the conditions we have imposed is

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{22}$$

- it follows that the distribution $(p_1, \ldots, p_n)$ which maximizes eq. 22, subject to constraints, will represent the 'most honest' description about what we know about propositions $(A_1, \ldots, A_n)$
- the function H is the *information entropy* of the distribution $\{p_i\}$
- he provides another derivation, the Wallis derivation, which he says is more satisfying conceptually
  - We are given information I, which is to be used in assigning probabilities $\{p_1, \ldots, p_m\}$ to m possibilities, with the total probability $\sum_{i=1}^{m} p_i = 1$
  - choose some integer $n \gg m$, where we have n quanta of probability, each with the magnitude $\delta = n^{-1}$ to distribute
  - suppose we were to distribute the n quanta randomly among the m possibilities; the first m receives n_1 quanta, the second n_2 quanta, and so on, until we have the probability assignment

    $$p_i = n_i \delta = \frac{n_i}{n} \qquad i = 1, 2, \ldots, m$$

    with the probability that this will happen being the multinomial distribution

    $$m^{-n} = \frac{n!}{n_1! \cdots n_m!}$$

  - suppose we do this procedure repeatedly, rejecting each probability assignment if it does not conform to I. what is the most likely probability distribution result?
  - it is the one that maximizes the probability

    $$W = \frac{n!}{n_1! \cdots n_m!}$$

  - we can refine this by using smaller and smaller quanta, and in this limit we have, by Stirling's approximation

    $$\log(n!) = n \log(n) - n + \sqrt{(2\pi n)} + \frac{1}{12n} + O\left(\frac{1}{n^2}\right)$$

    where O denotes terms that tend to zero as $n \to \infty$
  - using this and writing $n_i = np_i$, we find that as $n \to \infty$, $n_i \to \infty$ in such a way that $n_i/n \to p_i = \text{const.}$,

    $$\frac{1}{n} \log(W) \to -\sum_{i=1}^{m} p_i \log(p_i) = H(p_1, \ldots, p_m)$$

  - so the most likely probability assignment is the one that has maximum entropy subject to the given information I

10.1.3 Formal properties of maximum entropy distributions

- want to list the formal properties of the canonical distribution (look at pg. 357)

$$u_i \equiv \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \exp\left\{-\sum_{j=1}^{m} \lambda_j f_j x_i\right\} \tag{23}$$

- the maximum H attainable by holding averages fixed depends on the averages we specify,

$$H_{\max} = S(F_1, \ldots, F_m) = \log Z(\lambda_1, \ldots, \lambda_m) + \sum_{k=1}^{m} \lambda_k F_k$$

- H is the measure of the 'amount of uncertainty' in a probability distribution and, once maximized, it becomes a function of the definite data of the problem $\{F_i\}$, which we'll call $S(F_1, \ldots, F_m)$
- $S(F_1, \ldots, F_m)$ is still a measure of uncertainty, just uncertainty *when all the information we have consists of just these numbers*
- if S is a function only of $(F_1, \ldots, F_m)$, then the partition function $Z(\lambda_1, \ldots, \lambda_m)$ is also a function of $(F_1, \ldots, F_m)$
- different Lagrange multipliers, λ_i are different canonical probability distributions, in which the averages over these distributions agree with the given averages F_k if

$$F_k = \langle f_k \rangle = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k} \qquad k = 1, 2, \ldots, m$$

  which is a set of m simultaneous nonlinear equations which must be solved for the multipliers in terms of F_k
- small changes in F_k changes the maxium attainable H by

$$\lambda_k = \frac{\partial S(F_1, \ldots, F_m)}{\partial F_k} \tag{24}$$

- differentiating either eqs. 23 or 24, with respect to λ_j, we get a general reciprocity law

$$\frac{\partial F_k}{\partial \lambda_j} = \frac{\partial^2 \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_j \partial \lambda_k} = \frac{\partial F_j}{\partial \lambda_k}$$

- differentiating again will give a second reciprocity law dependent on the first derivative
- lets now consider the possibility that one of the functions $f_k(x)$ contains a parameter α which can vary (e.g. say $f_k(x_i; \alpha)$ stands for the ith energy level of a system and α is the volume of the system), where we want to predict how quantities change as we change α
  - the best estimate of the derivative would be the mean value over the probability distribution

$$\left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = \frac{1}{Z} \sum_i \exp\{-\lambda_i f_1(x_i) - \cdots - \lambda_k f_k(x_i; \alpha) - \cdots - \lambda_m f_m(x_i)\} \frac{\partial f_k(x_i; \alpha)}{\partial \alpha}$$

$$\Rightarrow \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{1}{\lambda_k} \frac{\partial \log Z(\lambda_1, \ldots, \lambda_m; \alpha)}{\partial \alpha}$$

    which assumes α appears in only one function. if the same parameter appears in multiple functions, we can generalize it to

$$\sum_{k=1}^{m} \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m; \alpha)}{\partial \alpha}$$

- now lets note some *fluctuation laws*, or moment laws
  - Note: F_k and $\langle f_k \rangle$ stand for the same number, since we specified that the expectation values $\{\langle f_1 \rangle, \ldots, \langle f_m \rangle\}$ are set to be equal to the given data $\{F_1, \ldots, F_m\}$
    - when we want to emphasize that the quantities are averages over the distribution, we will use $\langle f_k \rangle$, and when we want to emphasize that they are the given data, we use F_k
  - the reciprocity law can be written as

$$\frac{\partial \langle f_k \rangle}{\partial \lambda_j} = \frac{\partial \langle f_j \rangle}{\partial \lambda_k} = \frac{\partial^2 \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_j \partial \lambda_k}$$

  - varying λ's changes from one canonical distribution to another, where their averages are slightly different.
  - the new distribution corresponds to $(\lambda_k + \delta\lambda_k)$, and since it is of canonical form, its maximum entropy corresponds to slightly different data $(F_k + dF_k)$
  - how are the different quantities $f_k(x)$ correlated? measure of covariance of the distribution

$$\big\langle (f_j - \langle f_j \rangle)(f_k - \langle f_k \rangle) \big\rangle = \big\langle f_j f_k - f_j \langle f_k \rangle - \langle f_j \rangle f_k + \langle f_j \rangle \langle f_k \rangle \big\rangle = \langle f_j f_k \rangle - \langle f_j \rangle \langle f_k \rangle$$

    if f_k is greater than $\langle f_k \rangle$, it is likely that the f_j is also larger than $\langle f_j \rangle$ and the covariance is positive. if the oppositive, it is negative. if their variations are uncorrelated, the covariance is zero

  - if $j = k$, this reduces to the variance
$$\left\langle (f_k - \langle f_k \rangle)^2 \right\rangle = \langle f_k^2 \rangle - \langle f_k \rangle^2 \geq 0$$
  - he goes on to calculate these for the canonical distribution explicitly

- we now have a new class of problems we can solve wholesale: first evaluate the partition function, then by differentiating this with respect to all arguments, we can obtain predictions in the form of mean values over the maximum entropy distribution
- I need to spend more time looking at this to truly grok it

10.1.4 Conceptual problems — frequency correspondence

- Maximum entropy is conceptually difficult, especially from a frequentist's perspective
- some common objections:
  1. The only justification for the canonical entropy is 'maximum uncertainty', which is a negative; you can't get results out of ignorance
  2. no reason to assume that distributions observed experimentally would correspond to those found from maximum entropy, as the distributions have nothing to do with frequencies
  3. the principle is restricted to cases where constraints are average values, but data F_k are almost never averages
  4. the principle can't lead to definite physical results because if different people had different information, they would come up with different distributions
- to which he responds:
  1. the uncertainty was always there; maximizing the entropy does not *create* uncertainty
  2. maximum entropy fundamentally has nothing to do with the frequencies of 'random experiments', but this does not mean that it cannot be applied to such cases
     - see the dice example — maximum entropy probabilities do have precise connections to frequencies
     - this relation is usually not needed
     - maximum entropy is most useful when observed frequencies do not agree with maximum entropy probabilities
  3. if given information does consist of mean values, then the math is neat and gives us a partition function. but for given information which places any type of constraint, we can conclude that the *probability* distribution which maximizes the entropy is identical to the *frequency* distribution which can be realized in the greatest number of ways
  4. goes into quite some detail here, saying that this misses the point of maximum entropy
     - if we have two people with different prior information, B having more than A, the measure of the ratio of maximum probabilities is
$$\frac{W_A}{W_B} \sim \exp\{N(H_A - H_B)\}$$
     - even for large N, a slight decrease in the entropy leads to a large decrease in the number of possibilities
     - justifies the weak statement of frequency correspondence:

       > If the information incorporated into the maximum entropy analysis includes all the constraints actually operating in the random experiment, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally. Indeed, most frequency distributions observed in Nature are maximum entropy distributions, simply because they can be realized in so many more ways than can any other. (pg. 370)

- he says, to summarize: ''…the principle of maximum entropy is not an oracle telling which predictions *must* be right; it is a rule for inductive reasoning that tells us which predictions *are most strongly indicated by our present information*. (pg. 370)

11 Ignorance priors and transformation groups

- how do we translate prior information into prior probability assignments?. maximum entropy provides one powerful tool
- if prior probabilities just represent prior opinions, they are useless. problems of inference are ill–posed until we recognize three things (this is taken verbatim):
  1. The prior probabilities represent our prior *information*, and are to be determined, not only by introspection, but by *logical analysis* of that information.
  2. Since the final conclusions depend necessarily on both the prior information and the data, it follows that, in formulating a problem, one must specify the prior information to be used just as fully as one specifies the data
  3. Our goal is that inferences are to be completely 'objective' in the sense that two persons with the same prior information must assign the same prior probabilities. (pg. 373)
- the natural starting point in translating prior information is the state of complete ignorance
- maximum entropy tells us that for discrete probabilities complete ignorance is represented by a uniform prior probability assignment. for continuous probabilities, this problem is more complicated
- this chapter talks about the method of transformation groups, but first goes into some detail about using maximum entropy for continuous distributions

11.1 Continuous distributions

- Shannon's theorem holds true only for discrete distributions, and the corresponding expression for continuous distributions must pass to the limit from a discrete distribution
  - taking the discrete entropy expression
    $$H_I^{\mathrm{d}} = -\sum_{i=1}^{n} p_i \log[p_i]$$
    suppose that the points x_i become more and more numbers such that in the limit $n \to \infty$
    $$\lim_{n\to\infty} \frac{1}{n}(\text{number of points in } a < x < b) = \int_a^b \mathrm{d}x\; m(x)$$
  - if this limit is well–behaved, then it is also true that the difference between discrete points, $(x_{i+1} - x_i)$, tends to zero such that
    $$\lim_{n\to\infty} [n(x_{i+1} - x_i)] = [m(x_i)]^{-1}$$
  - the discrete probability distribution p_i will go over into a continuous probability $p(x|I)$ according to
    $$p_i = p(x_i|I)(x_{i+1} - x_i)$$
    $$p_i \to p(x_i|I)[nm(x_i)]^{-1}$$
  - so the discrete entropy goes over into an integral
    $$H_I^{\mathrm{d}} \to \int \mathrm{d}x\; p(x|I) \log\left[\frac{p(x|I)}{nm(x)}\right]$$
    where $m(x)$ is some 'invariant measure' function, which is proportional to the limiting density of discrete points
  - contains an infinite term $\log(n)$, and subtracting this causes the integral to tend towards a definitely limit
    $$H_I^{\mathrm{c}} \equiv \lim_{n\to\infty} [H_I^{\mathrm{d}} - \log(n)] = -\int \mathrm{d}x\; p(x|I) \log\left[\frac{p(x|I)}{m(x)}\right]$$
  - H_I^{c} is invariant, since $p(x|I)$ and $m(x)$ transform in the same way under change of variables
  - we seek a normalized probability distribution $p(x|I)$

* constrained by information fixing the mean values of m different functions $f_k(x)$

$$F_k = \int \mathrm{d}x\ p(x|I) f_k(x)$$

* the solution is

$$p(x|I) = Z^{-1} m(x) \exp\{\lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)\}$$

with partition function

$$Z(\lambda_1, \ldots, \lambda_m) \equiv \int \mathrm{d}x\ m(x) \exp\{\lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)\}$$

and Lagrange multipliers

$$F_k = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k} \qquad k = 1, \ldots, m$$

– what is $m(x)$?

* consider a one–dimensional case, supposing that we know $a < x < b$ but no other prior information.
* there are then no Lagrange multipliers and the distribution reduces to

$$p(x|I) = \left[\int_a^b \mathrm{d}x\ m(x)\right]^{-1} m(x)$$

* this tells us that the measure $m(x)$ is also the prior distribution describing complete ignorance of x. but how do we find this?

11.2 Aside on assigning priors

- Bayes suggests we express ignorance by assigning a uniform prior
  - not invariant under change of parameters, and no criterion to tell us which parameterization to use
- Jeffreys suggested assigning a prior $\mathrm{d}\sigma/\sigma$ to a continuous, positive parameter σ, since this is invariant whether we use σ or σ^m
  - but we don't want it to be invariant under *all* parameter changes
- the real problem is then, ''For which choice of parameters does a given form, such as that of Bayes of Jeffreys, apply?''
- can apply groups of transformations

11.3 Transformation groups

- best illustrated by example

11.3.1 Location and scale parameters

- *location parameter*: parameter which translates a PDF along the x–axis
- *scale parameter*: parameter which widens or sharpens the PDF
- Sample from a continuous two–parameter distribution $p(x|v\sigma) = \phi(x, v, \sigma)\ \mathrm{d}x$
- Given a sample $\{x_1, \ldots, x_n\}$, estimate v and σ.
  - cannot do this until we define some prior distribution

$$p(v\sigma|I)\ \mathrm{d}v\ \mathrm{d}\sigma = f(v, \sigma)\ \mathrm{d}v\ \mathrm{d}\sigma$$

  - if in complete initial ignorance, doesn't tell us which function $f(v, \sigma)$ to use

- suppose we change variables to $\{x', \nu', \sigma'\}$ such that

$$\begin{aligned} \nu' &= \nu + b \\ \sigma' &= a\sigma \\ x' - \nu' &= a(x - \nu) \end{aligned}$$

  where $(0 < a < \infty)$ and $(-\infty < b < \infty)$
- the new distribution is then

$$p(x'|\nu'\sigma') = \psi(x', \nu', \sigma') = \phi(x, \nu, \sigma)\, \mathrm{d}x$$

  or, $\psi(x', \nu', \sigma') = a^{-1}\phi(x, \nu, \sigma)$
- the prior has changed to $g(\nu', \sigma') = a^{-1}f(\nu, \sigma)$, according to the Jacobian of the transformation
- suppose that the distribution is invariant under transformation, i.e. $\psi(x, \nu, \sigma) = \phi(x, \nu, \sigma)$, which means that $\phi(x, \nu, \sigma)$ must satisfy

$$\phi(x, \nu, \sigma) = a\phi(ax - a\nu + \nu + b, \nu + b, a\sigma)$$

- differentiating with respect to a, b and solving the differential equation, we get

$$\phi(x, \nu, \sigma) = \frac{1}{\sigma} h\left(\frac{x - \nu}{\sigma}\right)$$

  where $h(q)$ is an arbitrary function
- ''Thus, the usual definition of a location parameter ν and a scale parameter σ is equivalent to specifying that the distribution shall be invariant under the group of transformations.'' (pg. 379)
- If a change of scale makes the problem different, then we are not completely ignorant in the prior. complete ignorance of a location and scale parameter is a state of knowledge such that changing either parameter does not change the state of knowledge

- lets look at some consequences of this
  - given a sample $\{x'_1, \ldots, x'_n\}$, estimate ν' and σ'
  - if in complete ignorance, this has the same sampling distribution as before, and our state of knowledge is also the same as before
  - since we have formulated two problems which have the same prior information, we would then, to be consistent, assign the same prior, $f(\nu, \sigma) = g(\nu, \sigma)$
  - now the form of the prior is uniquely determined, and, combining this with above, we get

$$f(\nu, \sigma) = af(\nu + b, a\sigma)$$

    with a general solution of

$$f(\nu, \sigma) = \frac{\text{const.}}{\sigma}$$

    which is Jeffreys rule, described above.
  - we can see that this result is uniquely determined by the *transformation group*, not the form of the distribution (he goes on to show that for a different transformation group, the end result is different)
  - so, it is not enough to say that a change of scale and location does not change the state of knowledge, we must specify a definite group of transformations

- to summarize this, specifying complete initial ignorance precludes us from obtaining any definite prior distribution, but defining a set of operations which transforms the problems into an equivalent one places nontrivial restrictions on the form of the prior, allowing us to obtain it

- goes on to present several other examples of complete ignorance and its effects on priors

12 Decision theory, historical background

12.1 Inference vs. decision

- Nothing inherent in probability theory which would tell us where to make decisions, like rejecting or accepting a batch of widgets, or when to make another test
- need to develop a defined criterion for making estimates, not just intuitive *ad hockeries*
- Probability theory alone can only solve inference, giving us the probability distribution which represents our final state of knowledge with all available prior information and data taken into account
- missing a concrete way to turn those probabilities into definite action
- two opposing ways to develop this

12.1.1 Bernoulli's suggestion

- consider possibilities $i = 1, 2, \ldots, n$, with probabilities p_i, and numbers M_i representing 'profit' we would obtain if the ith possibility is true
- the expectation of profit is then
$$E(M) = \langle M \rangle = \sum_{i=1}^{n} p_i M_i$$
- Bernoulli discovered some paradoxes that led him to believe that simple expectation of profit does not always lend itself as a sensible criterion for action
  - for example, if you can bet any amount of money with the probability $(1 - 10^{-6})$ that wou will lose all your money, but with the probability 10^{-6} that you will win 1000001 times the amount bet, the criterion of maximizing profit would tell you to bet, regardless of how stupid it would be
  - he proposed that the true value to a person, of receiving a certain amount of money, is not measured simply by the amount received, but also by how much he already has
- he proposed that the 'moral value', or 'utility', of an amount of money M should be taken proportional to $\log(M)$
- he goes on to show that intuition indicates that the utility of money must increase less rapidly than the log for extremely large values

12.1.2 Example: The honest weatherman

- a weatherman's prior and data yield a probability $\rho = P(\text{rain}|\text{data}, I)$ that it will rain
- what is the probability q that he will announce this forecast? this depends on his perceived utility function
- we would imagine that he would overstate the probability for bad weather, i.e. announce a value $q > p$, so as to not incur criticism
- but is there a utility environment which would induce him to always tell the truth?
- suppose he won't ever be fired for making a wrong prediction, but his pay for the day will be $B\log(2q)$ if it actually rains, and $B\log(2[1-q])$ if it does not, where B is the base pay. his expected pay for today, if announcing a probability q is then
$$B[p\log(2q) + (1-p)\log(2[1-q])] = B[\log(2) + p\log(q) + (1-p)\log(1-q)]$$
which is maximum at $q = p$
- any continuous function appears linear if we look at a small enough portion of it, so if the weatherman considers a single day's pay small enough so that his utility for it is linear, it will always be advantagous to tell the truth
- what if there was a reward?
  - let there be n possible events $(A_1, \ldots, A_n)$, for which the priors and data indicate probabilities $(p_1, \ldots, p_n)$. but the weatherman announces probabilities $(q_1, \ldots, q_n)$

- let him be paid $B\log(nq_i)$ if the event A_i occurs, puting a reward for placing a high probability on a true event
- his expectation of pay is then

$$B[\log(n) - I(q;p)]$$

where $I(q;p) = \sum p_i \log(q_i)$, or the relative entropy of the distributions

- it will then be to the weatherman's advantage to always announce $q_i = p_i$, with his maximum expectation of pay being

$$B[\log(n) - H(p_1, \ldots, p_n)]$$

where $H(p_i) = -\sum p_i \log(p_i)$ is the entropy that measures uncertainty about A_i. it is in his advantage to acquire maximum amount of data so as to minimize the entropy

12.2 Wald's decision theory

- begin by enumerating a set of possible 'states of nature', $\{\theta_1, \ldots, \theta_N\}$
  - important to note that enumerating these states of nature is not describing any real property, rather it is describing a state of knowledge about the range of possibilities
- next enumerate a set of decisions, $\{D_1, \ldots, D_k\}$, which might be made. for example, in quality control there would be three decisions:

$$D_1 \equiv \text{accept the batch}$$
$$D_2 \equiv \text{reject the batch}$$
$$D_3 \equiv \text{make another test}$$

- the enumeration of these decisions is a means of describing our knowledge as to what kind of actions are *feasible*, as it is wasteful to enumerate a decision that we would never act on
  - two points
    1. There is a continuous gradient — the consequences of an action might be serious without being intolerable
    2. the consequences of an action will depend on the true state of nature
- this idea of consequence brings up a third concept, the loss function $L(D_i, \theta_j)$, which is a set of numbers representing our judgement as to the loss incurred by making some decision D_i should θ_j turn out to be the true state of nature
  - if both D_i and θ_j are discrete, the loss function → a loss matrix, L_{ij}
- a few possible criteria:
  - minimax criterion: for each D_i find the maximum possible loss $M_i = \max_j(L_{ij})$, then choose the D_i for which M_i is minimum. represents an adversarial Nature
  - minimin criterion: for each D_i find the minimum possible loss $m_i = \min_j(L_{ij})$, then choose the D_i for which m_i is minimum. represents a benevolent Nature
  - science would choose an intermediate between the two, that Nature is neutral
- full decision theory must also take into account evidence E
- if we knew the true state of nature, making the correct decision would be trivial; if θ_3 was the true state of nature, then the best decision D_i is the one that minimizes L_{i3}
  - so once the loss function is specified, the uncertainty as to the best decision comes entirely from the uncertainty as to the true state of nature
  - so, rephrased, what we should ask is 'Conditional on all the available evidence, what is the *probability* P_3 that θ_3 is the true state of nature?'
- goes on to show (pp. 410–417) how a purely sampling theory point of view, relegating parameter estimation and hypothesis testing to a separate field, misses some detail and information which introducing the loss function rectifies

– determines the 'best' estimator, β of some variable α from n observations $(x_1, \ldots, x_n)$ is

$$\langle R \rangle = \int \mathrm{d}\alpha \; g(\alpha) R_\alpha$$

or, a weighted average of R_α, the 'risk'

– the variation, $\delta\langle R \rangle$, due to an arbitrary variation $\delta\beta(x_1, \ldots, x_n)$ in the estimator is

$$\delta\langle R \rangle = \int \cdots \int \mathrm{d}x_1 \cdots \mathrm{d}x_n \left\{ \int \mathrm{d}\alpha \; g(\alpha) \frac{\partial L(\alpha, \beta)}{\partial \beta} f(x_1, \ldots, x_n | \alpha) \right\} \delta\beta(x_1, \ldots, x_n)$$

which vanishes, independently of $\delta\beta$ if

$$\int \mathrm{d}\alpha \; g(\alpha) \frac{\partial L(\alpha, \beta)}{\partial \beta} f(x_1, \ldots, x_n | \alpha) = 0$$

– eventually leads to the conclusion that estimation of parameters requires estimators which end up being medians/modes over the Bayesian posterior PDF

– concludes:

> …if the θ_j are discrete and we agree not to include in our enumeration of states of nature any θ_j that is known to be impossible, then the class of admissible strategies is just the class of Bayes strategies (i.e. those that minimize expected loss over a posterior pdf). If the possible θ_j form a continuum, the admissible rules are the proper Bayesian ones; i.e. Bayes rules from proper (normalizable) prior probabilities. …For a given sampling distribution and loss function, we are content to say simply that the defensible decision rules are the Bayes rules character by different proper priors, and their well–behaved limits. (pg. 415)

– leads to a 'fundamental integral equation' for minimization of loss,

$$\frac{\partial}{\partial \beta} \int \mathrm{d}\alpha \; g(\alpha) L(\alpha, \beta) f(x_1, \ldots, x_n | \alpha) = 0 \tag{25}$$

12.3 General decision theory

- Eq. 25 is generalized outside of just parameter estimation, and we can use it as a criterion to find the optimal decision in any case
- to solve the problem of inference, we have four steps (taken verbatim):
  1. Enumerate the possible states of nature θ_j, discrete or continuous, as the case may be.
  2. Assign prior probabilities $p(\theta_j | I)$ which represent whatever prior information I you have about them
  3. Assign sampling probabilities $p(E_i | \theta_j)$ which represent your prior knowledge about the mechanism of the measurement process yielding the possible data sets E_i
  4. Digest any additional evidence $E = E_1 E_2 \cdots$ by application of Bayes' theorem, thus obtaining posterior probabilities $p(\theta_j | EI)$
- $p(\theta_j | EI)$ expresses all information about θ_j contained in the prior and data. to solve the problem of decision, we have three more steps:
  5. Enumerate the possible decisions D_i
  6. Assign the loss function $L(D_i, \theta_j)$ that tells what you want to accomplish
  7. Make that decision D_i which minimizes the expected loss over the posterior probabilities for θ_j

13 Simple applications of decision theory

- statisticians would call these procedures ''significance tests''
- gist of what we need to do is apply the Bayesian inference rules and hypothesis testing, supplemented by the loss function

13.1 Definitions and preliminaries

- Employ the following notation for this chapter (this is a summary of the rules of Bayesian statistics)
  - $p(A|B)$ = conditional probability for A, given B
  - $p(AB|CD)$ = joint conditional probability for A and B, given C and D
  - everything follows from the product rule, $p(AB|C) = p(A|BC)p(B|C) = p(B|AC)p(A|C)$
  - if the propositions B and C are not mutually contradictory, this can be rearranged into Bayes' theorem,

$$p(A|BC) = p(A|C)\frac{p(B|AC)}{p(B|C)} = p(A|B)\frac{p(C|AB)}{p(C|B)}$$

  - if there are several mutually exclusive and exhaustive propositions B_i, then by summing over them, we obtain the chain rule,

$$p(A|C) = \sum_i p(A|B_iC)p(B_i|C)$$

  - also, let the following definitions

$$\begin{aligned} X &= \text{prior knowledge, of any kind} \\ S &= \text{signal} \\ N &= \text{noise} \\ V &= V(S,N) = \text{observed voltage} \\ D &= \text{decision about the nature of the signal} \end{aligned}$$

  - so, we have

    $p(S|X)$ = prior probability for a particular signal, S
    $p(N|X) = W(N)$ = prior probability for the particular sample of noise N
  - prior information X is always built into the right–hand side of probability symbols, whether explicitly written or not. thus, in a linear system $V = S + N$

$$p(V|S) \equiv p(V|SX) = W(V - S)$$

  - a *decision rule*, $p(D_i|V_j)$, or $p(D|V)$ for simplicity, is the process of drawing inferences about the signal from observed voltage
    - any decision must be made on the basis of V alone, as it, by definition, contains all information actually used (including X) in arriving to that decision
    - so, if $Y \neq D$ is another proposition, $p(D|V) = p(D|VY)$
    - or, equivalently, we can say that the probabilty for reaching decision D depends on any proposition Y only through the influence Y has on X,

$$p(D|Y) = \sum_V p(D|V)p(V|Y)$$

13.2 Sufficiency and information

- that $p(D|V) = p(D|VY)$ leads to an interesting consequence
  - suppose we wish to know the probability for Y, on the basis of knowing D and V. using the product rule,

$$p(DY|V) = p(Y|VD)p(D|V) = p(D|VY)p(Y|V) \Rightarrow p(Y|VD) = p(Y|V)$$

  - note that the reduction (indicated by $\Rightarrow$) comes from using $p(D|V) = p(D|VY)$
  - this is interesting as it indicates that if V is known, D is redundant and does not help us in estimating any other quantity. the reverse, however, is not true
- this leads to the following theorem:

Theorem 13.1 *Let D be a possible decision, given V. then $p(V|D) \neq 0$ and*

$$p(Y|V) = p(Y|D) \text{ iff } p(V|D) = p(V|YD)$$

- in other words, knowledge of D is as good as knowledge of V for judgements about Y iff Y is irrelevant for judgements about V, given D.
- another rewording: in the 'environment' produced by knowledge of D, the probabilities for Y and V are independent, i.e. $p(YV|D) = p(Y|D)p(V|D)$
- so, D is a *sufficient statistic* for judgements about Y
- sufficiency is closely related to the concept of information. the theorem could be state equally as well as: D is a sufficient statistic for judgements about Y if it contains all the information about Y that V contains
- he goes on to describe this loose definition of 'information' and how it connects to sufficiency using Shannon's measure of information, concluding that

  > …if by 'information' we mean minus the expectation of the entropy Y over the prior distribution of D or V, zero information loss in going from V to D is equivalent to sufficiency of D. (pg. 430)

- also, he notes that acquisition of new data can never increase $\overline{H}$ (the expectation value of the entropy)

13.3 Loss functions and criteria of optimum performance

- need some criterion to determine which decision rule to use over another. criterion will change with application, no one criterion to rule them all (lol)
- general criterion is obtained by assigning a *loss function*, $L(D, S)$, representing the judgement of how serious it is to make decision D when signal S is present
- example:
  - suppose there are only two possible signals, $S_0 = 0$ and $S_1 > 0$, which in turn lead to two decisions, D_0 and D_1
  - there are two types of error, a false alarm $A = (D_1, S_0)$ and false rest $R = (D_0, S_1)$, and we consider a false rest ten times more serious than a false alarm, while the correct decision of either type represents 'no loss'
  - so our loss functions are then $L(D_0, S_0) = L(D_1, S_1) = 0$, $L(D_0, S_1) = 10$, and $L(D_1, S_0) = 1$, leading to a *loss matrix* (in the case of discrete sets of decisions),
    $$Lij = \begin{pmatrix} 0 & 10 \\ 1 & 0 \end{pmatrix}$$
  - we can also consider *information loss* by assigning $L(D, S) = -\log[p(S|D)]$, instead of arbitrary loss values
    * more difficult, as the loss function now depends on the decision rule
    * minimization of the information loss leads to a decision which is as close as possible to being a sufficient statistic for judgements about the signal
  - the *conditional loss*, $L(S)$, is the expected loss when a signal S is present,
    $$L(S) = \sum_D L(D, S)p(D|S)$$
  - *average loss* is indicated by the expectation of conditional loss over all possible signals,
    $$\langle L \rangle = \sum_S L(S)p(S|X)$$
- now can see two types of criteria for optimal performance (taken verbatim):
  - **The minimax criterion**: For a given decision rule $p(D|V)$, consider the conditional loss $L(S)$ for all possible signals, and let $[L(S)]_{\max}$ be the maximum value attained by $L(S)$. We seek that decision rule for which $[L(S)]_{\max}$ is as small as possible.…this criterion concentrates attention on the worst possible case, regardless of the probability for occurrence of this case, and it is thus in a sense too conservative. However, it gives some psychological comfort that it does not involve the prior probabilities for the different signals $p(S|X)$, and therefore can be applied by persons who, under the handicap of orthodox training, have a mental hangup against prior probabilities.
  - **The Bayes criterion**: We seek the decision rule for which the expected loss $\langle L \rangle$ is minimized. In order to apply this, a prior distribution $p(S|X)$ must be available. (pg. 431)

- note that Wikipedia shows the formal definition of Bayes' criterion as

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

  where

  - $\hat{L}$ = maximum value of the likelihood function of model M, $\hat{L} = p(D|\hat{\theta}, M)$, where $\hat{\theta}$ is the parameter values that maximize the likelihood function
  - D = observed data
  - n = number of data points in D
  - k = number of parameters estimated by the model

- let's find the Bayes solution
  - substitution leads to the expression for expected loss

$$\langle L \rangle = \sum_{DV} \left[\sum_{S} L(D, S) p(VS|X) \right] p(D|V)$$

  - if $L(D, S)$ is independent of $p(D|V)$, there is no $p(D|V)$ for which this expression is stationary
  - then minimize $\langle L \rangle$ by choosing each possible V leading to decision $D_1(V)$ for which the coefficient

$$K(D, V) \equiv \sum_{S} L(D_1, S) p(VS|X)$$

    is minimum
  - therefore, we adopt the decision rule $p(D|V) = \delta(D, D_1)$

- goes on to provide a detailed example (pp. 432–437)

13.4 The widget problem

- interesting problem, as we have no occasion to use Bayes' theorem, since no 'new' information is acquired; it is a pure use of maximum entropy
- Formulation of the problem:

  …Mr *A* is in charge of a widget factor, which proudly advertises that it can make delivery in 24 hours on any size order. This, of course, is not really true, and Mr *A*'s job is to protect, as best he can, the advertising manager's reputation for veracity. This means that each morning he must decide whether the day's run of 200 widgets will be painted red, yellow, or green. (For complex technological reasons, not relevant to the present problem, only one color can be produced per day.) We follow his problem of decision through several stages of increasing knowledge. (pg. 441)

13.4.1 Stage 1

- Mr *A* checks the stock at the beginning of the day to find 100 red widgets, 150 yellow widgets, and 50 green widgets.
- with complete ignorance of the day's orders, common sense would say to build up the stock of green widgets

13.4.2 Stage 2

- Calling the front desk leads him to learn that average orders per day break down to 50 red, 100 yellow, and 10 green widgets per day.
- likely would change decision to yellow widgets

13.4.3 Stage 3

- he gets further information on the total number of orders processed, leading to a break down of average widgets per order of 75 red, 10 yellow, and 20 green widgets
- likely would change decision to red widgets

Table 1: Summarized data of the 4 stages

Stage	R	Y	G	Decision
1. In stock	100	150	50	G
2. Avg. daily order total	50	100	10	Y
3. Avg. individual order	75	10	20	R
4. Specific order	NA	NA	40	?

13.4.4 Stage 4

- gets word of an emergency order for 40 green widgets
- no longer can make a qualitative decision, must make a quantitative decision

13.4.5 Mathematical solution for Stage 2

- only will solve truncated problem of today not affecting tomorrow's manufacturing
- begin by enumerating states of nature θ_j corresponding to all possible order situations
- let $n_1 = 1, 2, 3, \ldots$ be the number of red widgets that will be sold today, with n_2 and n_3 indicating the yellow and green widgets, respectively. therefore, any conceivable order can be represnted by a set of three non–negative integers, $\{n_1, n_2, n_3\}$
- next we assign prior probabilities $p(\theta_j|X) = p(n_1 n_2 n_3|X)$ to the states of nature, maximizing the entropy of the distribution subject to constraints (see §10 for solution to this)
- the index i in x_i indicates the three integers n_1, n_2, and n_3, the function $f_k(x_i)$ corresponds to the n_i, since the expectation values $\langle n_1 \rangle$, $\langle n_2 \rangle$, and $\langle n_3 \rangle$ are fixed by information given in Stage 2
- leads to the partition function

$$\begin{aligned} Z(\lambda_1, \lambda_2, \lambda_3) &= \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \sum_{n_3=0}^{\infty} \exp\{-\lambda_1 n_1 - \lambda_2 n_2 - \lambda_3 n_3\} \\ &= \prod_{i=1}^{3} \left(1 - \exp\{-\lambda_i\}\right)^{-1} \end{aligned}$$

  with λ_i being the Lagrange multipliers for the three expectation value constraints, determined by

$$\langle n_i \rangle = -\frac{\partial \log(Z)}{\partial \lambda_i} = \frac{1}{\exp\{\lambda_i\} - 1}$$

- the maximum entropy probability assignment factors as $p(n_1 n_2 n_3) = p_1(n_1) p_2(n_2) p_3(n_3)$ where

$$\begin{aligned} p_i(n_i) &= (1 - \exp\{-\lambda_i\}) \exp\{-\lambda_i n_i\} \qquad n_i = 1, 2, 3, \ldots \\ &= \frac{1}{\langle n_i \rangle + 1} \left[\frac{\langle n_i \rangle}{\langle n_i \rangle + 1} \right]^{n_i} \end{aligned}$$

- so Mr A's initial state of knowledge about today's order is given by the three exponential distributions

$$p_1(n_1) = \frac{1}{51}\left(\frac{50}{51}\right)^{n_1} \qquad p_2(n_2) = \frac{1}{101}\left(\frac{100}{101}\right)^{n_2} \qquad p_3(n_3) = \frac{1}{11}\left(\frac{10}{11}\right)^{n_3}$$

- since there is no new evidence E, we cannot use Bayes' theorem and must make decisions based solely on the prior distributions
- enumerate the decisions $D_1 \equiv$ make red ones, $D_2 \equiv$ make yellow ones, $D_3 \equiv$ make green ones, with the loss function $L(D_i, \theta)$ (we take the loss as no loss for all orders filled, and loss proportional to the number of unfilled orders)
- present stock is $S_1 = 100$, $S_2 = 150$, $S_3 = 50$

- loss on each decision is

$$L(D_1; n_1, n_2, n_3) = R(n_1 - S_1 - 200) + R(n_2 - S_2) + R(n_3 - S_3)$$
$$L(D_2; n_1, n_2, n_3) = R(n_1 - S_1) + R(n_2 - S_2 - 200) + R(n_3 - S_3)$$
$$L(D_3; n_1, n_2, n_3) = R(n_1 - S_1) + R(n_2 - S_2) + R(n_3 - S_3 - 200)$$

  where

$$R(x) \equiv \begin{cases} x & x \leq 0 \\ 0 & x \geq 0 \end{cases}$$

- expected loss is then

$$\langle L \rangle_1 = \langle n_1 \rangle \exp\{-\lambda_1(S_1 + 200)\} + \langle n_2 \rangle \exp\{-\lambda_2 S_2\} + \langle n_3 \rangle \exp\{-\lambda_3 S_3\} \Rightarrow 22.70$$
$$\langle L \rangle_2 = \langle n_1 \rangle \exp\{-\lambda_1 S_1\} + \langle n_2 \rangle \exp\{-\lambda_2(S_2 + 200)\} + \langle n_3 \rangle \exp\{-\lambda_3 S_3\} \Rightarrow 10.6$$
$$\langle L \rangle_3 = \langle n_1 \rangle \exp\{-\lambda_1 S_1\} + \langle n_2 \rangle \exp\{-\lambda_2 S_2\} + \langle n_3 \rangle \exp\{-\lambda_3(S_3 + 200)\} \Rightarrow 29.38$$

  leading to the expected decision to make yellow widgets

- he goes on to show a mathematical solution for Stage 3, which I will not replicate
  - because we know average of individual orders, we need to define new states of nature $\theta = \{u_1 \ldots; v_1 \ldots; w_1 \ldots\}$ where u_r, v_y, and w_g are orders for r red widgets, y yellow widgets, and g green widgets, respectively
  - the expectation values end up giving the same expectation value form as in Bose–Einstein statistics,

$$\langle u_r \rangle = \frac{1}{\exp\{r\lambda_1 + \mu_1\} - 1}$$

    which is *fucking cool*.

14 Paradoxes of probability theory

- Defines a paradox ‘‘something which is absurd or logically contradictory, but which appears at first glance to be the result of sound reasoning.’’ (pg. 451)
- says they often arise from misuse of infinite sets and infinite/infinitesimal quantities
- if bad reasoning always lead to ridiculous conclusions, it would be easy to identify; but once that reasoning has led to a few correct conclusions, it is hard to safeguard against it
- removing a paradox from probability theory will require
  1. the result is indeed absurd
  2. the reasoning leading to it violates the rules of inference developed earlier
  3. when one obeys those rules, the paradox disappears and we have a reasonable result

14.1 Summing a series the easy way

- Paradox: can prove that any infinite series $S = \sum_i a_i$ converges to any number x you choose
- the sum of the first n terms is $s_n = a_1 + a_2 + \cdots + a_n$
- defining $s_0 \equiv 0$, we have $a_n = (s_n - x) - (s_{n-1} - x)$, where $1 \leq n < \infty$
- the series then becomes

$$S = (s_1 - x) + (s_2 - x) + (s_3 - x) + \cdots$$
$$- (s_0 - x) - (s_1 - x) - (s_2 - x) - \cdots$$

  so all terms $(s_n - x)$ cancel out, leading to $S = (s_0 - x) = x$

- we can avoid such fallacious arguments by following the advice, ‘‘passage to a limit should always be the last operation, not the first.’’ (pg. 452)
- the correct line of reasoning leads to cancelling all the x terms, leading to the final s value, which is the correct summation

14.2 Nonconglomerability

- another example of misusing infinite sets
- if $(C_1, \ldots, C_n)$ denote a finite set of mutually exclusive, exhaustive propositions with prior information I, the probability for a proposition A is

$$P(A|I) = \sum_{i=1}^{n} P(AC_i|I) = \sum_{i=1}^{n} P(A|C_iI)P(C_i|I)$$

  which shows that the prior probability $P(A|I)$ is a weighted average of conditional probabilities $P(A|C_iI)$
- a weighted average of a set of real numbers cannot lie outside the range spanned by those numbers; i.e. if $L \leq P(A|C_iI) \leq U$, thence $L \leq P(A|I) \leq U$, a property which is called *conglomerability*, or, more precisely, conglomerability in the partition $\{C_i\}$
- he says that obviously nonconglomerability cannot arise from a correct application of the rules of probability theory, and goes on to analyze some examples where nonconglomerability has been claimed, which I'm not going to replicate (pp. 452–464)
- why does nonconglomerability matter?
  - in and of itself, it does not, but it is a symptom of a larger issue
  - recall that if $A \equiv A_1 + A_2 + \cdots + A_n$ is a disjunction of a finite number of mutually exclusive propositions, then

$$p(A|C) = \sum_{i=1}^{n} p(A_i|C)$$

  - Definition—*disjunction*: Given two propositions, A and B, A **or** B is true if A is true, if B is true, and if A and B are true. Denoted by $\vee$ or $+$, as in $A \vee B$ or $A + B$
  - these probabilities have 'finite additivity'. as $n \to \infty$ we would suppose that the sum goes in the limit into a sum over a countable number of terms, thusly converging. if the sum does not converge, we would refuse to pass to the limit at all
  - however, suppose we pass to the infinite limit before considering additivity (as shown to be the cause of the nonconglomerability paradox described in the examples). we are then concerned with additivity over propositions about intervals on infinite sets

14.3 The Borel–Kolmogorov paradox

- the transition from discrete to continuous probabilities is typically uneventful, but one tricky point can lead to erroneous results
- suppose that I is prior information according to which (x, y) are assigned a bivariate normal PDF with variance unity and correlation coefficient ρ,

$$p(\mathrm{d}x\mathrm{d}y|I) = \frac{\sqrt{1-\rho^2}}{2\pi} \exp\left\{\frac{1}{2}(x^2 + y^2 - 2\rho xy)\right\} \mathrm{d}x\mathrm{d}y$$

- integrate out either x or y to get marginal PDFs,

$$p(\mathrm{d}x|I) = \sqrt{\left(\frac{1-\rho^2}{2\pi}\right)} \exp\left\{-\frac{1}{2}(1-\rho^2)x^2\right\} \mathrm{d}x$$
$$p(\mathrm{d}y|I) = \sqrt{\left(\frac{1-\rho^2}{2\pi}\right)} \exp\left\{-\frac{1}{2}(1-\rho^2)y^2\right\} \mathrm{d}y$$

- this is all rote so far. but what if we wanted the conditional PDF for x, given $y = y_0$? we might assume that we need to just set $y = y_0$ and renormalize with A,

$$p(\mathrm{d}x|y = y_0 I) = A \exp\left\{-\frac{1}{2}(x^2 + {y_0}^2 - 2\rho x y_0)\right\} \mathrm{d}x$$

- this is, however, an *ad hoc* device and not derived from the prior PDF

- working through the proper derivation (described in previous chapters) we get

$$p(A|BI) = p(\mathrm{d}|x\mathrm{d}yI) = \frac{p(\mathrm{d}x\mathrm{d}y|I)}{p(\mathrm{d}y|I)} = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(x-\rho y_0)^2\right\}\mathrm{d}x$$

  where $A \equiv x$ in $\mathrm{d}x$ and $B \equiv y$ in $(y_0 < y < y_0 + \mathrm{d}y)$.

- we can see that the $\mathrm{d}y$ term cancels out, so $\mathrm{d}y \to 0$ does nothing. so this gives the same result as our intuitive *ad hoc* result, after working out the normalization constant. so why not just always make these intuitive leaps?

- we can see that it all depends on how we choose to write the *ad hoc* equation. suppose we instead used variables (x, u), where $u \equiv y/f(x)$

  - the Jacobian is then

$$\frac{\partial(x,u)}{\partial(x,y)} = \left(\frac{\partial u}{\partial y}\right)_x = \frac{1}{f(x)}$$

  - which leads to the conditional PDF, expressed in the new variables and after integrating out u or x (same procedure as above),

$$p(\mathrm{d}x|u = 0I) = A\exp\left\{-\frac{1}{2}x^2\right\}f(x)\mathrm{d}x$$

  - from the definition of u, we can see that $u = 0$ is the same as $y = 0$, meaning the above conditional PDF differs from the previous conditional PDF by a factor of $f(x)$

- he goes on to show that this arises from the ambiguity of $y = 0$, without defining which of any number of limits is intended by passing to a measure–zero limit (as the limit $y = 0$ and be defined in any number of sequences, such as $A_\epsilon \equiv |y| < \epsilon$ or $B \equiv |y| < \epsilon|x|$)

- in other words, the final result will depend on the limiting operation specified

14.4 Discussion

- he goes through a few more paradoxes (specifically the marginalization paradox, which he states is still incomplete and much more complex than the previous ones), but I'm tired of writing them and I don't really care about them that much.

15 Orthodox methods: historical background

- as much of this chapter is historical ramblings, I will only reproduce some insights I find interesting

- this chapter is concerned with the history of orthodox statistics circa 1900–1970, with the express purpose of informing us about the failings of the orthodoxy of statistics

- orthodox and Bayesian statistics differ in how they relate to data

  - orthodox statistics is limited at the outset to pre–data considerations, giving correct answers to questions of the form (verbatim)

    1. Before you have seen the data, what data do you expect to get?
    2. If the as yet unknown data are used to estimate parameters by some known algorithm, how accurate do you expect the estimates to be?
    3. If the hypothesis being tested is in fact true, what is the probability that we shall get data indicating that it is true?

  - however, essentially all real scientific inferences problems are concerned with post–data questions (verbatim),

    1. After we have seen the data, do we have any reason to be surprised by them?
    2. After we have seen the data, what parameter estimates can we now make, and what accuracy are we entitled to claim?
    3. What is the probability *conditional on the data*, that the hypothesis is true?

15.1 Sampling distribution for an estimator

- a major part of orthodoxy is devoted to calculating, approximating, and comparing sampling PDFs for estimators, as it is the only criterion orthodoxy has for judging estimators
- on the other hand, in Bayesian analysis, we do not need to do this, as we know that an estimator derived from Bayes' theorem with a specific loss function is the optimal estimator for the problem as defined
- suppose we are estimating some parameter, α,
  - in orthodoxy, the width of the sampling PDF for the estimator would answer the pre–data question, ''How much would the estimate of α vary over the class of all data sets that we might conceivably get?''
  - as this is not quite relevant to scientists, we should be concerned with the post–data question, ''How accurately is the value of α determined by the one data set D that we actually have?''
- this difference is a source of contention between orthodox and Bayesian statistics, so let's examine how they can often be the same
  - scientific inference has been, historically, dominated by Gaussian sampling distributions.
  - suppose we have a data set $D = \{y_1, \ldots, y_n\}$, with the sampling distribution,

    $$p(D|\mu\sigma I) \propto \exp\left\{-\sum_i \frac{(y_i-\mu)^2}{w\sigma^2}\right\}$$

    where we know the value of sigma
  - the Bayesian posterior PDF for μ, with a uniform prior, is

    $$p(\mu|D\sigma I) \propto \exp\left\{-\frac{n(\mu-\overline{y})^2}{2\sigma^2}\right\}$$

    from which the post–data estimate of μ is

    $$(\mu)_{\text{est}} = \overline{y} \pm \frac{\sigma}{\sqrt{n}}$$
  - we can see that the sample mean, $\overline{y} \equiv n^{-1}\sum y_i$ is a sufficient statistic
  - an orthodox satistician using $\overline{y}$ as an estimator of μ would find the sampling distribution to be

    $$p(\overline{y}|\mu\sigma I) \propto \exp\left\{-\frac{n(\overline{y}-\mu)^2}{2\sigma^2}\right\}$$

    leading to a pre–data estimate of

    $$(\overline{y})_{\text{est}} = \mu \pm \frac{\sigma}{\sqrt{n}}$$
  - though differing in conceptual meaning, the pre– and post–data estimates are nearly idential mathematically that the Bayesian and orthodox statisticians would make the same numerical estimate of μ with the same claimed accuracy
  - **Important**: in problems where we have sufficient statistics but no nuisance parameters, there is mathematical symmetry which can make pre– and post–data questions closely related

16 Principles and pathology of orthodox statistics

- analyze the consequences of failure to use Bayesian methods in simple models
- orthodox objections to Bayesian statistics are usually ideological in nature
- want to answer the questions: ''In what circumstances, and in what ways, do the orthodox results differ from the Bayesian results? What are the pragmatic consequences of this in real applications?''
- this chapter reallllyyy starts to lose me early on

- talks a lot about how orthodoxy puts a lot of emphasis on using 'unbiased' estimators, but that biased estimators are sometimes better in arriving at the correct value with faster convergence
  - the *bias* of an estimator is the difference between the estimator's expectation value and the true value of the parameter being estimated
  - *unbiased estimator*: an estimator with no bias, i.e. its expectation value is the true value
  - he (and Fisher) argue that bias is not meaningful, as it is not invariant under change of parameters, e.g. the square of an unbiased estimate of a parameter α is not the same as an unbiased estimate of α^2

16.1 Information loss

- orthodox methods must waste some of the information contained in the data
- consider estimation of a parameter θ from some data set $D \equiv \{x_1, \ldots, x_n\}$, represented by a point R^n
  - orthodoxy requires us to define some estimator $b(x) \equiv b(x_1, \ldots, x_n)$, before seeing any data, and then use this estimator, and this estimator only, for estimation of θ
  - specifying a numerical value of $b(x)$ locates the sample on a subspace of R^n of dimension $(n-1)$
  - specifying the data D tells us both that and where on the subspace we are
  - if position on the subspace is independent of θ, then $b(x)$ is a sufficient statistic, and orthodox and Bayesian methods will give similar results
  - if position is relevant, then the data contains information that orthodox methods do not take into account
  - put differently, given the actual data set, all estimators that might be chosen, $\{b_1, b_2, \ldots\}$, are known. Therefore, Bayes' theorem contains all the information contained in the class of all estimators
- so, Bayesian methods can continue if a single estimator is not a sufficient statistic, but orthodox methods must continue to create estimators until a sufficient statistic is found, or else it will produce inaccurate results
- shows that the notion of sufficiency can be definable in terms of Shannon's information measure of entropy, in addition to in the terms of information solely
- the condition for a sufficient statistic to exist is that the sampling distribution be of the functional form

$$\log p(x|a) = -l(\alpha)\beta(x) + \int \mathrm{d}l\, \langle\beta\rangle + \text{const.}$$

  where α and β are vectors of any dimensionality and l is the sampling expectation value of some observable x
- ''…if we think of a maximum entropy distribution as a sampling distribution paramterized by the Lagrange multipliers l_j, we find that the sufficient statistics are precisely the data images of the constraints that were used in defining that distribution.'' (pg. 520)
  - so, if we have a maximum entropy distribution generated from a set of constraints $\{\langle\beta_1(x)\rangle, \langle\beta_2(x)\rangle, \ldots, \langle\beta_k(x)\rangle\}$ as expectations over that probability distribution, it has k sufficient statistics which are just $\{\beta_1(x), \ldots, \beta_k(x)\}$, where x is the data set observed
  - Note: I presume this means that $\beta = 1/kT$ is a sufficient statistic for the Maxwell–Boltzmann distribution?

16.2 Bayesian spectrum analysis

- compares the orthodox and Bayesian methods for determining periodicity in data.
- says that the orthodox method throws away information relevant to periodicity, as well as further pointing out the folly of comparing solely against a ''null hypothesis'' (see section 5 of this document for more)
- will only replicate his discussion on the Bayesian method, examining the periodicity of temperature in New York City
- first, we consider it possible that temperature data has some periodic element due to a systematic, physical influence, $A\cos\omega t + B\sin\omega t$, where $|\omega| \leq \pi$
- additionally, presume that the data are contaminated with variables e_t, which we cannot control or predict

  - almost always a good idea to set a Gaussian prior, in this case with parameters (μ, σ) to the e_t variables
  - μ is the 'nominal true mean temperature' not known in advance, σ is a nuisance parameter to be integrated out (how did he determine this?)

- model equation for the data is then

$$y_t = \mu + A\cos\omega t + B\sin\omega t + e_i \qquad 1 \le t \le n$$

with sampling distribution for e_t,

$$p(e_i \cdots e_n|\mu\sigma I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_t e_t^2\right\}$$

and sampling distribution for the data,

$$p(y_1 \cdots y_n|\mu\sigma I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{Q}{2\sigma^2}\right\}$$

where

$$\begin{aligned} Q(A,B,\omega) &= \sum (y_t - \mu - A\cos\omega t - B\sin\omega t)^2 \\ &\Rightarrow n\left[\overline{y^2} - 2\overline{y}\mu + \mu^2 - 2A\overline{y_t\cos\omega t} - 2B\overline{y_t\sin\omega t} + 2\mu A\overline{\cos\omega t}\right. \\ &\left. + 2\mu B\overline{\sin\omega t} + 2AB\overline{\cos\omega t\sin\omega t} + A^2\overline{\cos^2\omega t} + B^2\overline{\sin^2\omega t}\right] \end{aligned}$$

- in this problem, A, B, and ω are the parameters of interest, with μ and σ being nuisance parameters
- so in the definition of Q, the four sums involving y_t are the sufficient statistics for all five parameters, with the others being evaluated analytically before incorporating the data
- what about the priors?
  - the sheer fact that New York exists is relevant, as we know then that A and B must be less than 200° F
  - have no information about the periodicity phase, $\theta = \tan^{-1}(B/A)$, so we assign a uniform prior to θ
  - assign a joint prior to (A, B),

$$p(AB|I) = \frac{1}{2\pi\delta^2}\exp\left\{-\frac{A^2+B^2}{2\delta^2}\right\}$$

    where δ is on the order of magnitude of 100°F

- applying Bayes' theorem and integrating out the nuisance parameters gives us

$$p(AB\omega|DI) = Cp(AB\omega|I)L^*(A,B,\omega)$$

where C is a normalization constant and L^* is the quasi–likelihood,

$$L^*(A,B,\omega) = \int \mathrm{d}\mu \int \mathrm{d}\sigma\; p(\mu\sigma|AB\omega I)p(D|AB\omega\mu\sigma I)$$

and, to be on the safe side,

$$p(\mu\sigma|I) \propto \frac{1}{\sigma\sqrt{2\pi\alpha^2}}\exp\left\{-\frac{\mu^2}{2\alpha^2}\right\} \qquad a \le \sigma \le b$$

in which α and b (where the fuck is this b?) are of the order of 100°F, and $a \backsimeq 10^{-6}$, making L^* then

$$\begin{aligned} L^*(A,B,\omega) &= \int_{-\infty}^{\infty}\mathrm{d}\mu\;\exp\left\{-\frac{\mu^2}{2\alpha^2}\right\}\int_b^a \frac{\mathrm{d}\sigma}{\sigma^{n+1}}\exp\left\{-\frac{Q}{2\sigma^2}\right\} \\ &\Rightarrow \frac{1}{2}\frac{(n/2-1)!}{(Q/2)^{n/2}} \end{aligned}$$

if $n > 0$

16.2.1 The folly of randomization

- ''randomization'' is often introduced by way of Monte Carlo integration
- let a function $y = f(x)$ exist in a unit square $0 \leq x, y \leq 1$, and we wish to compute its integral

$$\theta = \int_0^1 \mathrm{d}x\, f(x)$$

- let's assume that we cannot solve this analytically. we can just choose n points at random (x, y) within the unit square and determine whether or not $y \leq f(x)$
- for r points, we estimate the integral as $(\theta)_{\text{est.}} = r/n$, and, as $n \to \infty$, this estimate might approach the Riemannian integral
- how accurate is the estimate?
  - suppose we have an intependent binomial sampling distribution on r

    $$p(r|n\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}$$

    with (mean) ± (standard deviation) of

    $$\theta \pm \sqrt{\frac{\theta(1-\theta)}{n}}$$

  - the width of the sampling distribution would indicate the accuracy of the estimate, so we could determine the probable error of $(\theta)_{\text{est.}}$ as

    $$(\theta)_{\text{est.}} = \frac{r}{n} \pm \sqrt{\frac{r(n-r)}{n^3}}$$

  - you will notice that the accuracy only improves as $1/\sqrt{n}$
  - this leads to him showing that the maximum possible error per step is

    $$[\text{error in determining } f(x)] \times [\text{width of step}] = \frac{1}{2\sqrt{n}} \times \frac{1}{\sqrt{n}} = \frac{1}{2n}$$

- he goes on to show that the error in choosing nonrandom points is significantly lower than choosing random points

16.3 Continuing on

- proposes a general principle: ''*Whenever there is a randomized way of doing something, there is a nonrandomized way that yields better results from the same data, but requires more thinking.*'' (pg. 532)
- Bayesian methods have built in safety devices
  - in parameter estimation, the log–likelihood function is

    $$\log L(\alpha) = \sum_{i=1}^{n} \log p(x_i|\alpha) = n\overline{\log p(x_i|\alpha)}$$

  - this is spread out over the full range of variability of all data, so if we get bad data, no good estimate is possible and Bayes' theorem returns to us a wide posterior distribution
  - however, orthodox methods claim accuracy which is essentially the width of the sampling distribution for whatever estimator you choose, ignoring the range of data while taking into account all data sets which might have been obtained but were not
- fundamental differences in analyzing trends in data
  - Orthodox methods attempt to 'detrend' data before analysis, usually irreversibly removing data which may or may not be relevant
  - Bayesian methods remove the contamination in *final conclusion*, taking into account *all relevant information*

17 The A_p distribution and rule of succession

- want to give our probability calculation machine a definite mechanism to store conclusions, not just isolated facts
- one can produce identical 'external' probabilities for two events, while having very different 'internal' states of knowledge
- up to this point, all propositions have been 'Aristotelian' in that they are binary choices, either true or false.
- what about propositions that are not so black and white, but to which we still want to assign real number probabilities?
  - introduce a new proposition A_p, defined by
    $$p(A|A_pE) \equiv p$$
    where E is any additional evidence
  - A_p, in words, is something to the effect of ''regardless of anything else you may have been told, the probability of A is p''
  - this is a weird proposition, and nothing stops us from using Bayes' theorem to get its probabilities, $p(A_p|E)$. this is weird, because it is essentially a probability of probabilities
  - not claiming that $p(A_p|E)$ is a real probability, only that it is a number which obeys the rules of probability theory
    * to end confusion, use the notation $(A_p|E)$, and call it 'the density of A_p, given E'
  - if some information X tells us nothing other than it is possible for A to be true or false, in a completely ignorant population (Ch. 11),
    $$(A_p|X) = 1 \qquad 0 \leq p \leq 1$$
  - use Bayes' theorem to compute the density of A_p conditional on other things
    $$(A_p|EX) = (A_p|X)\frac{P(E|A_pX)}{P(E|X)} = \frac{P(E|A_p)}{P(E|X)}$$
  - Now,
    $$P(A|E) = \int_0^1 \mathrm{d}p\ (AA_p|E)$$
    because every A_p are mutually exclusive and exhaustive
  - use the product rule to get
    $$P(A|E) = \int_0^1 \mathrm{d}p\ P(A|A_pE)(A_p|E) \Rightarrow P(A|E) = \int_0^1 \mathrm{d}p\ p(A_p|E)$$
    which states that the probability of A is the first moment of the density of A_p
  - therefore, the density for A_p should contain more information about the robot's state of mind concerning A than just the probability for A

17.1 Relevance

- note some lemmas about relevance to see why we proposed that the density of A_p should be more informative
- suppose we have evidence $E = E_aE_b$, with only E_a being relevant to A,
  $$P(A|E) = P(A|E_aE_b) = P(A|E_a)$$
- according to Bayes' theorem, A, given E_a, must also be irrelevant to E_b,
  $$P(E_b|AE_a) = P(E_b|E_a)\frac{P(A|E_bE_a)}{P(A|E_a)} = P(E_b|E_a)$$
- this is called 'weak irrelevance'. this does not imply that E_b is irrelevant to A_p, it only says that the first moments of $(A_p|E_b)$ and $(A_p|E_aE_b)$ are the same
- suppose that for a given E_b, the relationship $P(A|E_aE_b) = P(A|E_a)$ holds no matter what E_a is

- this is called 'strong irrelevance'
- we then have

$$P(A|E) = \int_0^1 \mathrm{d}p\ p(A_p|E_aE_b) = \int_0^1 \mathrm{d}p\ p(A_p|E_a)$$
$$\Rightarrow P(A_b|A_pE_a) = P(E_b|E_a)$$

since the integrands must be identical if it is to hold for all $(A_p|E_a)$

- how does new evidence F change the state of knowledge about A
  - expand by Bayes' theorem using A_p,

$$P(A|EF) \Rightarrow \int_0^1 \mathrm{d}p\ (A_p|E)\frac{P(F|A_pE)}{P(F|E)} \tag{26}$$

  - since any evidence irrelevant to A can be ignored, the likelihood ratio goes to

$$\frac{P(F|A_pE_aE_b)}{P(F|E_aE_b)} \Rightarrow \frac{P(F|A_pE_a)}{P(F|E_a)}$$

  - if we continue doing this for all parts of E_a that are irrelevant to A, we get some E_{aa} which is relevant only to A
  - the above expression for $P(A|EF)$ then reduces to

$$P(A|EF) = \frac{1}{P(F|E_{aa})}\int_0^1 \mathrm{d}p\ p(A_p|E)P(F|A_p) \tag{27}$$

  - we can eliminate the normalization factor, $1/(F|E_{aa})$ by calculating the odds on A instead of the probability,

$$O(A|EF) = \frac{P(A|EF)}{P(\overline{A}|EF)} = \frac{\int_0^1 \mathrm{d}p\ p(A_p|E)P(F|A_p)}{\int_0^1 \mathrm{d}p\ (A_p|E)P(F|A_p)(1-p)}$$

    - the significant thing here is that the prior information, E, only appears in the density $(A_p|E)$, which indicates that the only property of E that we need in order to understand the effect of new information is the density itself
    - in other words, everything we need in order to reason about A from past experience is contained in the $(A_p|E)$
  - we can then store a density function, $(A_p|E)$ for each proposition A. when new evidence F is introduced, we can calculate $(A_p|EF)$ and store this as the new $(A_p|E)$
  - looking back to eq. 27, we can see that the state of mind when we have evidence E is determined by the width of the density $(A_p|E)$; when the density is already peaked sharply around one value of p, then $P(F|A_p)$ will be even more sharply peaked around some value p', and if the density is very broad, then a small change in $P(F|A_p)$ will make a large chang ein the probability of A
- can think of conventional probability formulas as an 'outer' process, which reasons about the external world, and the density A_p as an 'inner' process which observes and analyzes the 'outer' process

17.2 An application

- imagine a 'random' experiment is being performed, and we want to predict the results in the future based upon results from the past
- make the problem definite by introducing the propositions,

  $X \equiv$ For each trial we admit two prior hypotheses, A true, and A false
- assume two things
  1. probability assigned to A at the nth trial is independent of n
  2. the evidence concerning past trials retains full relevance for all times, i.e trial $n = 1$ is as relevant as trial $n = 99$ is to the prediction of trial $n = 100$

- no other prior information. additional definitions,

  $N_n \equiv A$ true n times in N trials in the past
  $M_m \equiv A$ true m times in M trials in the future

- the precise statement of information X is $(A_p|X) = 1, 0 \leq p \leq 1$
  - note that the *same* A_p is used for calculations pertaining to *all* trials
  - we are after $P(M_m|N_n)$
- through repetitions of the sum and product rules, we come to the binomial distributions

$$P(N_n|A_p) = \binom{N}{n} p^n (1-p)^{N-n}$$
$$P(M_m|A_p) = \binom{M}{n} p^m (1-p)^{M-m}$$

- find the prior probability $P(N_n|X)$, recalling the trick for resolving a proposition into mutually exclusive alternatives,

$$P(N_n|X) = \int_0^1 \mathrm{d}p\ (N_n A_p|X) \Rightarrow \binom{N}{n} \int_0^1 \mathrm{d}p\ p^n (1-p)^{N-n} \Rightarrow P(N_n|X) = \begin{cases} \frac{1}{N+1} & 0 \leq n \leq N \\ 0 & N < n \end{cases}$$

  the result of which is the uniform distribution of maximum entropy. $P(M_m|X)$ is found in the same manner

- expanding with Bayes' theorem,

$$(A_p|N_n) = (A_p|X)\frac{P(N_n|A_p)}{P(N_n|X)} = (N+1)P(N_n|A_p)$$

  giving us the desired probability

$$P(M_m|N_n) = \int_0^1 \mathrm{d}p\ (M_m A_p|N_n) = \int_0^1 \mathrm{d}p\ P(M_m|A_p N_n)(A_p|N_n) \tag{28}$$
$$\Rightarrow P(M_m|N_n) = \frac{\binom{n+m}{n}\binom{N+M-n-m}{N-n}}{\binom{N+M+1}{M}} \tag{29}$$

  since $P(M_m|A_p N_n) = P(M_m|A_p)$ (see the previous section). it should be noted that although this looks like it, this is not the hypergeometric distribution discussed in earlier chapters

- in the special case of $M = m = 1$, this collapses to the probability of A being true in the next trial, given it has been true n times in the previous N trials,

$$P(A|N_n) = \frac{n+1}{N+2}$$

  which is Laplace's rule of succession.

17.3 Laplace's rule of succession

- though identified several times before, we need to go into more detail
- converts raw information into numerical values of probabilities, and gives a connection between probabilities and frequencies
- must be stated that the rule of succession gives the probability based *only* on the information that the event occurred n times in N trials, without additional information
- goes through some objections to the rule of succession, pointing out how they consider problems which are not under the umbrella of the rule. Concludes,

  > Probability theory, like any other mathematical theory, cannot give a definite answer unless we ask it a definite question. We should always start a problem with an explicit enumeration of the 'hypothesis space' consisting of the different propositions that we are going to consider in that problem. That is part of the 'boundary conditions' which must be specified before we have a well–posed mathematical problem. (pg. 566)

- concludes the section with

  > …Laplace's rule of succession provides a definite, useful solution to a definite, real problem. …The case where the problem can be reasonably idealized to one with only two hypotheses to be considered, a belief in a constant 'causal mechanism', *and no other prior information*, is the only case where it applies. (pg. 568)

17.4 An example of the rule of succession: bass or carp?

- suppose a certain lake consists of only bass and carp. we catch ten fish and they all turn out to be carp — what, then, is our state of belief about the percentage of bass in the lake?
- common sense would tell us that the bass population is likely somewhere in the range $(0\%, 15\%)$, but does not tell us quantitatively how likely this is
- what does the rule of succession say?
  - with the bass fraction denoted by f, its posterior cumulative PDF is $P(f < f_0|DX) = 1 - (1 - f_0)^{11} \Rightarrow 0.833$, meaning we have $5:1$ odds that the bass population is below 15%
  - the posterior median value is
    $$f_{1/2} = 1 - \left(\frac{1}{2}\right)^{11} = 0.061$$
    or even odds that the bass population is below 6.1%
  - the 'best' estimate, by minimum mean–square error, is $\langle f \rangle = 8.3\%$
- how does a bass catch on the 11th draw change our view?
  - common sense says that if one out of eleven draws is a bass, we would find it hard to believe that the bass population is below 5%
  - Laplace agrees, with mean–square estimate being $\langle f \rangle = 2/13 = 15\%$
- this is the kind of problem Laplace's rule is good at dealing with: only two possibilities at each trial, and our prior knowledge was nothing more than that either of those two events were possible

17.5 Generalization of the rule

- provides a derivation of the mathematical technique of Laplace's rule of succession that is generalized (pp. 568–571), which I will not replicate
  - the problem set up is that we have K hypotheses, $\{A_1, A_2, \ldots, A_K\}$, a belief that the 'causal mechanism' is constant, and no other prior information
  - a random experiment is performed N times, with us observing A_1 to be true n_1 times, A_2 to be true n_2 times, and so on
  - based on this, what is the probability that in the next $M = \sum_i m_i$ repetitions of the experiment, A_i will be true exactly m_i times?
- so, the generalization of eq. 29 is
  $$P(m_1 \cdots m_K | n_1 \cdots n_K) = \frac{\binom{n_1+m_1}{n_1} \cdots \binom{n_K+m_K}{n_k}}{\binom{N+M+K-1}{M}} \tag{30}$$
- again, the special case where we just want the probability that A_1 will be true on the next trial, so $M = m_1 = 1$ with all other $m_i = 0$,
  $$P(A_1|n_1 NK) = \frac{n_1 + 1}{N + K}$$
- using this rule for small N is rather dumb, as, having no other prior information about A, the numerical probabilities will be very 'soft', since A_p will be very broad

17.6 Weight of new evidence

- the stability of a probability assignment in the face of new evidence is determined by the width of the A_p distribution
- with E being prior evidence, and F being new evidence,

$$P(A|EF) = \int_0^1 \mathrm{d}p\; p(A_p|EF) = \frac{\int_0^1 \mathrm{d}p\; p(A_p|F)(A_p|E)}{\int_0^1 \mathrm{d}p\; (A_p|F)(A_p|E)}$$

- F is compatible with E if $P(A|EF) = P(A|E)$, i.e. F makes no change in the probability of A
- F can make a large change in the distribution A_p without changing the first moment, i.e. F can broaden or sharpen the distribution, but unless it changes the maximum, we still will be likely to assign the same probability to A as without F
  - this is called *confirmation*, as the new evidence sharpens the A_p distribution, making us more confident in the previously assigned probability of A
- consider A_p given two different pieces of evidence, E and F,

$$(A_p|EF) = (\text{constant}) \times (A_p|E)(A_p|F)$$

  - if the distribution $A_p|F$ is significantly sharper than the distribution $A_p|E$, then the resulting distribution's width and peak will be determined by F.
  - we would then say that F carries more *weight* than E
  - having information F makes information E almost irrelevant to our final probability assignment

17.7 Indifference through knowledge or ignorance

- before we can use the principle of indifference to assign numerical probabilities, we must satisfy two conditions:
  1. we must be able to analyze the situation into mutually exclusive, exhaustive possibilities
  2. we must then find the available information gives us no reason to prefer any of the possibilities to any other
- these conditions are almost never met unless there is some symmetry in the problem
- condition (2) might be satisfied either through ignorance or through some positive knowledge of the situation
  - the difference is that the extra knowledge makes no change in the probability for A, but does change the density for A_p

17.8 Carnap's inductive methods

- ''…an infinite family of possible 'inductive methods' by which one can convert prior information and frequency data into a probability assignment and an estimate of frequencies for the future.'' (pg. 574)
- the principle is *ad hoc*, and states that the final probability assignment $P(A|N_nX)$ should be a weighted average of the prior probability $P(A|X)$ and the observed frequency, $f = n/N$
  - assigning a weight N to the prior 'empirical factor', f, and a weight λ to the prior 'logical factor', $P(A|X)$, leads to the method $c_\lambda(h, e)$
  - λ can be interpreted as the weight of prior evidence
  - so, with two hypotheses, Carnap's λ method is what you can calculate from the prior density $(A_p|X) = \text{constant} \cdot [p(1-p)]^r$ with $2r = \lambda - 2$, resulting in

$$P(A|N_nX) = \frac{2n + \lambda}{2N + 2\lambda} = \frac{(n - r) + 1}{(N + 2r) + 2}$$

  - so, a greater λ results in a more sharply peaked $(A_p|X)$ density
  - one should notice that prior evidence is weighted as λ = (number of prior trials + 2), while new evidence N_p is weighted only as (number of new trials) = N

    * the $(+2)$ in the weighting of prior evidence comes from the fact that prior knowledge that it is *possible* for A to be true or false is equivalent to knowledge that A has been true at least once and false at least once
    * our 'pre–prior' distribution for A_p is

$$(A_p|X_0)\mathrm{d}p = (\text{constant})\frac{\mathrm{d}p}{p(1-p)}$$

    * if we have prior knowledge as defined above, then it would be appropriate to use Laplace's rule $(A_p|X) = 1$
    * if not, we should use the above stated pre–prior distribution, which is the quasi–distribution representing 'complete ignorance'

17.9 More on probability and frequency connections

- two types of connections:
    1. given an observed frequency in a random experiment, convert this information into a probability assignment
    2. given a probability assignment, predict the frequency with which some condition will be realized
- problem (2) can be often solved by maximum entropy and transformation groups
- problem (1) is solved mostly by rule of succession,

    > If we have observed whether A was true in a very large number of trials, *and the only knowledge we have about A is the result of this random experiment, and the consistency of the 'causal mechanism'*, then it says that the probability we should assign to A at the next trial becomes practically equal to the observed frequency. (pg. 576)

- let's consider how to do this
    - this is just parameter estimation (frequency being the parameter)
    - suppose instead of determining the probability that A will be true in the next trial, we wish to infer the relative frequency of A in an indefinitely large number of trials, based on some evidence N_n
    - this is akin to taking the limit of eq. 29 as $M \to \infty$ and $m \to \infty$, such that $(m/M) \to f$
    - introduce the proposition,

        $A_f \equiv$ the frequency of A true in the indefinitely large number of trials is f

    - we find that the mean–value estimate of the frequency is equal to Laplace's rule of succession, and can interpret it as, ''*the probability which Laplace's theory assigns to A at a single trial is numerically equal to the estimate of frequency which minimizes the expected square of the error.*'' (pg. 577)
    - these results also hold for the generalized Laplace rule (eq. 30)
- this demonstrates the confusion between probability and frequency. when the available information consists of observed frequencies within very large samples and constancy of the 'causal mechanism', Laplace's theory becomes mathematically equivalent to frequency theory
    - most classical problems of statistics are of this type
- goes on to, through the example of one–dimensional neutron multiplication, and how the frequency–probability connection applies to real–world problems (pp. 579–586)
- notes an important distinction between frequentist and Bayesian statistics: the definition of independence
    - frequentist: only recognizes *causal independence*, i.e. the fact that one event occurs does not in itself extert a *physical* influence on the occurence of another.
    - Bayesian: independence means that *knowledge* of some event B does not affect the probability that we assign to an event A, i.e. $P(A|BC) = P(A|C)$. this is not just causal independence, but *logical* independence

17.10 The de Finetti theorem

- only considered A_p under the restriction that the underlying 'causal mechanism' is constant, though unknown, i.e. that we use the same A_p distribution for all trials
- how general is this?
  - define,
$$x_n \equiv \begin{cases} 1 & \text{if } A \text{ is true on the } n \text{ th trial} \\ 0 & \text{if } A \text{ is false on the } n \text{ th trial} \end{cases}$$
  - the state of knowledge about N trials is described generally by the probability function $P(x_1 \dots x_n|N)$
  - what is the necessary and sufficient condition on this distribution for it to be derivable from an A_p distribution?
  - any distribution obtained from A_p has the property that the probability of A is true in n specified trials and false in the remaining $(N-n)$ trials, depending only on the numbers n and N
  - the distribution $P(x_1 \dots x_n)$ defines an *exchangeable sequence* (the probability of an event is invariant under the permutation of the events)
- the de Finetti theorem argues that the converse is also true,

  Theorem 17.1 *Any exchangeable probability function $P(x_1 \dots x_n)$ can be generated by an A_p distribution.*

- ergo, there is a function $(A_p|X) = g(p)$ such that $g(p) \geq 0$, where $\int_0^1 \mathrm{d}p\, g(p) = 1$, and the probability that A is true in n trials and false in the remaining $(N-n)$ trials, is given by
$$P(n|N) = \int_0^1 \mathrm{d}p\, p^n (1-p)^{N-n} g(p)$$
- this theorem is important, since it shows that the connection between probability and frequency holds for a large class of probability functions (i.e. all exchangeable sequences)

18 Physical measurements

- in this section, he develops a way to estimate two parameters from three observations

18.1 Reduction of equations of condition

- suppose we wish to determine the charge e and mass m of an electron
  - Milikan oil–drop experiment measures e (I despise doing this experiment)
    - measures e with ±2% accuracy
  - deflection of an electron beam in a known elecromagnetic field measures e/m
    - measures (e/m) with ±1% accuracy
  - deflection of an electron toward a metal plate due to attraction of image charges measures e^2/m
    - measures (e^2/m) with ±5% accuracy
- these three experiments all measure e and m differently, and thus different experiments will obtain values which might not agree. wish to answer three questions:
  1. how do we process the data to make use of all information available to obtain the best estimates of e and m?
  2. what is the probable error remaining?
  3. how much would our estimates be improved by including another experiment?
- suppose the values of e and m known in advance to be $e \approx e_0$ and $m \approx m_0$, the measures are linear functions of the corrections

- writing the values of e and m as

$$e = e_0(1 + x_1)$$
$$m = m_0(1 + x_2)$$

  where x_1 and x_2 are dimensionless corrections, the results of the measurements are three numbers,

$$M_1 = e_0(1 + y_1)$$
$$M_2 = \frac{e_0}{m_0}(1 + y_2)$$
$$M_3 = \frac{e^2}{m_0}(1 + y_3)$$

  where y_n are also dimensionless numbers known in terms of e_0, m_0, and M_n

- the 'true' values are expressible in terms of x_i,

$$e = e_0(1 + x_1)$$
$$\frac{e}{m} = \frac{e_0(1 + x_1)}{m_0(1 + x_2)} = \frac{e_0}{m_0}(1 + x_1 - x_2 + \cdots)$$
$$\frac{e^2}{m} = \frac{e_0^2(1 + x_1)^2}{m_0(1 + x_2)} = \frac{e_0}{m_0}(1 + 2x_1 - x_2 + \cdots)$$

- taking into account all the errors, the general expression for the y_i coefficients is

$$y_i = \sum_{j=1}^{n} a_{ij}x_j + \delta_i \qquad i = 1, 2, \ldots, N$$

  or

$$y = Ax + \delta$$

  where δ_i are fractional errors and

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \\ 2 & -1 \end{pmatrix}$$

- old reasoning is as follows
  - plausible that x_j will be a linear combination of all y
  - if $N > n$, cannot just solve the matrix equation, since A will not be a square matrix, and thus will have no inverse
  - can multiply the left side by some $(n \times N)$ matrix B. the product BA exists, and will have an inverse if we choose B as such
  - the linear combinations are the n rows of

$$By = BAx + b\delta$$

    with the unique solution

$$x = (BA)^{-1}B(y - \delta)$$

  - the best estimate of x_j will be the jth row of

$$\hat{x} = (BA)^{-1}By$$

    if the fractional errors are symmetric, i.e. $p(\delta_i) = p(-\delta_i) \Rightarrow \langle \delta_i \rangle = 0$
  - different choices of B will give us different estimates. so which is the best choice of B?
  - this is the *reduction of equations of condition*

- this problem is actually solved in section 12 (Decision Theory),

> …where we have seen in generality that the best estimate of any parameter, by the criterion of any loss function, is found by applying Bayes' theorem to find the probability, conditional on the data, that the parameter lies in various intervals, then making the estimate which minimizes the expected loss taken over the posterior probabilities. (pg. 592)

- goes on to do this calculation for the case in which the errors have independent Gaussian probabilities (pp. 592–599). not replicated, but some important notes are below
  - ''…acquisition of new information does not affect our inferences if that new information is only what we would have predicted from our old information.'' (pg. 592)
  - find that no matter how accurately we know (e^2/m), if (e) and (e/m) measurements have the same accuracy, no matter how poor, we should ignore (e^2/m) and base our measurement of (m) solely on the (e) and (e/m) measurements

19 Model comparison

- how do we decide between two models which both account for the facts?
- Ockham said, ''Reality exists solely in individual things, and universals are merely abstract signs.'' i.e. the abstract creations of the mind are not realities in the external world
- really just a generalization of compound hypothesis testing
- working within one model, the normalization constants usually cancel. however, when multiple hypotheses are under consideration, we need to take into account all normalizations

19.1 Formulation of the problem

- let's first see why the normalization constants do not cancel. recall parameter estimation (section 6)
  - a model M contains parameters given by θ
  - given data D and prior information I, we first start with Bayes' theorem,

    $$p(\theta|DMI) = p(\theta|MI)\frac{p(D|\theta MI)}{p(D|MI)}$$

    where M on the right–hand side assumes the model's correctness
  - the denominator is the normalization constant,

    $$p(D|MI) = \int \mathrm{d}\theta\; p(D\theta|MI) = \int \mathrm{d}\theta\; p(D|\theta MI)p(\theta|MI)$$

    which is the prior expectation of the likelihood $L(\theta) = p(D|\theta MI)$
  - to judge which model of a given set $\{M_1, \ldots, M_r\}$ should be chosen, Bayes' theorem gives us the posterior probability of the jth model as

    $$p(M_j|DI) = p(M_j|I)\frac{p(D|M_jI)}{p(D|I)} \qquad 1 \le j \le r$$

  - eliminating the denominator by calculating the odds ratios, the posterior odds ratio for model M_j over M_k is

    $$\frac{p(M_j|DI)}{p(M_k|DI)} = \frac{p(M_j|I)}{p(M_k|I)}\,\frac{p(D|M_jI)}{p(D|M_kI)}$$

- so the probability $p(D|M_jI)$ which appears in single parameter estimation as a normalization constant is now key to determining the status of the model M_j over others
- the measure of what the data tell us about this is the prior expectation of its likelihood function over the prior probability $p(\theta_j|M_jI)$
- intuitively, the model which assigns the highest probability to the observed data is the correct one
  - this is a colloquial restatement of the likelihood principle for parameter estimation within a model

19.2 Fair judge vs. cruel realist

- two ways to approach model comparison
  1. fair judge: compare models by giving each the best possible prior probability for its parameters, as to compare the best case scenarios
  2. cruel realist: compare models by assigning prior probabilities based on the prior information we actually have pertaining to them, so as to show the most realistic performance of the models
- when real results are at stake, best to be a cruel realist

19.2.1 Known parameters

- assume there is no internal parameter space, i.e. the parameters of the model are known in advance ($\theta = \theta'$)
- problem reduces to simple hypothesis testing
- assign prior $p(\theta_j|M_jI) = \delta(\theta_j - \theta'_j)$, which reduces the normalization constant $p(D|MI)$, discussed earlier, to

$$p(D|M_jI) = p(D|\theta'_jM_jI) = L_j(\theta'_j)$$

  or the likelihood of θ' in the jth model
- the fair judge would note that this is maximum if θ_j is equal to the maximum–likelihood estimate $\hat{\theta}_j$ for that model, reducing the posterior odds ratio to

$$\frac{p(M_j|DI)}{p(M_k|DI)} = \frac{p(M_j|I)}{p(M_k|I)} \frac{(L_j)_{\max}}{(L_k)_{\max}} \tag{31}$$

- we saw previously that, given enough data, most models will produce such sharpley peaked likelihood functions that the prior ends up making no important inferences about the parameters
  - they are still important, however, to inferences about the *models*

19.2.2 Unknown parameters

- let model M have parameters $\theta \equiv \{\theta_1, \ldots, \theta_m\}$
- comparing the two above definition of the posterior odds, we write, $p(D|MI) = L_{\max}W$, where W is the Ockham factor, the amount by which model M is penalized by our nonoptimal prior information,

$$W \equiv \int \mathrm{d}\theta \, \frac{L(\theta)}{L_{\max}} p(\theta|MI)$$

- if the data are much more informative about the parameters than the prior information, we can define a 'high–likelihood' subspace, Ω', the smallest subspace of the parameter space Ω, which contains some large, specified amount of integrated likelihood
- therefore, most of the contribution to the Ockham factor W comes from this subspace
- the subspace Ω' is defined as the region of volume $V(\Omega')$ that contains the maximum possible amount of integrated likelihood,

$$\int \mathrm{d}\theta \, L(\theta) = L_{\max}V(\Omega')$$

- if the prior density $p(\theta|MI)$ is sufficiently broad that it is near uniform over this subspace, the expression for W reduces to

$$W \approxeq V(\Omega')p(\hat{\theta}|MI)$$

  meaning the Ockham factor is the *amount of prior probability contained in the high–likelihood subspace* Ω', as defined by the data

- the posterior odds ratio then becomes to

$$\frac{p(M_j|DI)}{p(M_k|DI)} = \frac{p(M_j|I)}{p(M_k|I)} \frac{(L_j)_{\max}}{(L_k)_{\max}} \frac{W_j}{W_k} \tag{32}$$

- this odds ratio depends only on the data and models. if two models achieve the same maximum–likelihood, they can account for the data equally well, and orthodox probability theory cannot distinguish between them. Bayesian statistics takes into account prior information, giving us a basis for choosing between the two models
- if the data are highly informative compared to the prior information, then the relative merit of two models is determined by two factors,
  1. how high a likelihood can be attained on their respective parameter spaces, Ω_j and Ω_k
  2. how much prior probability is concentrated in their respective high–likelihood subspaces, Ω'_j and Ω'_k
- how does this relate to simplicity?
  - suppose we have two explanations, A and B which account for some fact equally well. if A makes four highly plausible assumptions, while B makes only two, but unlikely, assumptions, intuition would tell us to choose explanation A, regardless of its higher complexity
  - so our intuition would choose not the simplest hypotheses, but the most plausible hypotheses
    * these two ideas are connected, though, since the more complicated a set of hypotheses is, the larger the set of conceivable alternatives is, and the smaller the prior probability of any particular hypothesis in the set is
  - while most statisticians state that simpler hypotheses are more plausible, Bayesians state that more plausible hypotheses tend to be simpler

20 Some quotes

Probability theory gives us the results of consistent plausible reasoning from the information *that was actually used* in our calculation. (p. 123)

But if our prior probability for S is lower than our prior probability that we are being deceived, hearing this claim has the opposite effect on our state of belief from what the claimant intended. The same is true in science and politics; the new information a scientist gets is not that an experiment did in fact yield this result, with adequate protection against error. It is that some colleague has claimed that it did. The information we get from the TV evening news is not that a certain event actually happened in a certain way; it is that some news reporter has claimed that it did. (p. 128)

Seeing is not a direct apprehension of reality, as we often like to pretend. Quite the contrary: *seeing is inference from incomplete information*, no different in nature from the inference that we are studying here. (p. 133)

For example, if you ask a scientist, 'How well did the Zilch experiment support the Wilson theory?' you may get an answer like this: 'Well, if you had asked me last week I would have said that it supports the Wilson theory very handsomely; Zilch's experimental points lie much closer to Wilsonfhs predictions than to Watson's. But, just yesterday, I learned that this fellow Woffson has a new theory based on more plausible assumptions, and his curve goes right through the experimental points. So now I'm afraid I have to say that the Zilch experiment pretty well demolishes the Wilson theory.' (p. 135)

…there is not the slightest use in rejecting any hypothesis H_0 unless we can do in it favor of some definite alternative H_1 which better fits the facts. (p. 135)

…Euler concentrated his attention entirely on the worst possible thing that could happen, as if it were certain to happen – which makes him perhaps the first really devout believer in Murphy's Law. (p. 203)

For example, the philosopher Karl Popper (1974) has gone so far as to flatly deny the possibility of induction. He asked the rhetorical question: 'Are we rationally justified in reasoning from repeated instances of which we have experience to instances of which we have no experience?' (pg. 276)

The fundamental, inescapable distinction between probability and frequency lies within this relativity principle: probabilities change when we change our state of knowledge; frequencies do not. (pg. 292)

The First Commandment of scientific data analysis publication ought to be: 'Thou shalt reveal thy full original data, unmutilated by any processing whatsoever.' (pg. 309)

In denying the possibility of induction, Popper holds that theories can never attain a high probability. But this presupposes that the theory is being tested against an infinite number of alternatives. We would observe that the number of atoms in the known universe is finite; so also, therefore, is the amount of paper and ink available to write alternative theories. It is not the absolute status of an hypothesis embedded in the universe of all conceivable theories, but the plausibility of an hypothesis *relative to a definite set of specified alternatives*, that Bayesian inference determines. (pg. 310)

What is done in quantum theory today is just the opposite; when no cause is apparent one simply postulates that no cause exists — ergo, the laws of physics are indeterministic and can be expressed only in probability form. The central dogma is that the light determines not whether a photoelectron will appear, but only the probability that it will appear. The mathematical formalism of present quantum theory — incomplete in the same way that our present knowledge is incomplete — does not even provide the vocabulary in which one could ask a question about the real cause of an event. (pg. 328)

Indeed, quite apart from probability theory, no scientist ever has sure knowledge of what is 'really true'; the only thing we can ever know with certainty is: *what is our state of knowledge*? (pg. 411)

In any field, the most reliable and instantly recognizable sign of a fanatic is a lack of any sense of humor. (pg. 497)

Inside every Non-Bayesian, there is a Bayesian struggling to get out.
– Dennis V. Lindley

Note in passing a simple counter–example to a principle sometimes stated by philosophers, that theories cannot be proved true, only false. We seem to have just the opposite situation for the theory that there was once life on Mars. To prove it false, it would not suffice to dig up every square foot of the surface of Mars; to prove it true one needs only to find a single fossil. (pg. 554)

…we have to remember that probability theory never solves problems of actual practice, because all such problems are infinitely complicated. We solve only idealizations of the real problem, and the solution is useful to the extent that the idealization is a good one. (pg. 568)