Systems and Distributions
MANUSCRIPT

William F. Barnes
[1]

September 15, 2024

Contents

Chapter 1

Systems and Distributions

1 Large-n Systems

For a stochastic process that produces events A_k in a continuous range instead of a discrete set, the normalization condition

$$\sum_{k=1}^{n} P(A_k) = 1$$

becomes an infinite sum. When confronted with this, the sum becomes an integral according to

$$\sum_{k=1}^{n} P(A_k) \rightarrow \int dP(A_k) = 1\,.$$

Probability Distribution Function

At this point, we abbreviate $A_k \rightarrow k$, and then use the chain rule to write

$$1 = \int_n \frac{dP(k)}{dk}\,dk = \int_n w(k)\,dk\,.$$

The continuous function $w(k)$ is called the *probability density*, or *probability distribution function* (although 'w' stands for *weight*). Specifically, $\omega(k)$ is the probability of an event occurring within a window $[k, k+dk]$.

In the continuous limit, the equations for the statistical average, general expectation value, and standard deviation generalize to

$$\langle k \rangle = \int_n k \cdot w(k)\,dk \tag{1.1}$$

$$\langle f \rangle = \int_n f(k)\,w(k)\,dk \tag{1.2}$$

$$\sigma_f = \sqrt{\int_n \left(f(k) - \langle f \rangle\right)^2\,w(k)\,dk} \tag{1.3}$$

in accordance with the definitions:

$$\sigma_f = \sqrt{\langle f^2 \rangle - 2\langle f \rangle \langle f \rangle + \langle f \rangle^2}$$

$$\sigma_f = \sqrt{\langle f^2 \rangle - \langle f \rangle^2}$$

Problem 1

What is the expected area of a right triangle with a hypotenuse of k whose non-right angles are uniformly distributed over the interval $(0, \pi/2)$? Answer:

$$\langle A \rangle = \frac{k^2/2}{\pi/2}\int_0^{\pi/2} \cos(\theta)\sin(\theta)\,d\theta$$
$$= \frac{k^2/2}{\pi/2}\int_0^{1} x\,dx = \frac{k^2}{2\pi}$$

Problem 2

Divide a given line segment into two other line segments. Then, cut each of these new line segments into two more line segments. What is the probability that the resulting four line segments are the sides of a quadrilateral?

Answer: Let the total length be L, and require that no one side be longer than $L/2$. After the initial cut, let the longer segment have length x, and the shorter segment $L - x$. Diving the longer segment at point z (from the start of x), it is required that $z < L/2$ and simultaneously $x - z < L/2$. Therefore, the window of allowed z has width $L/2 - (x - L/2) = L - x$. The normalized probability of an allowed z along x is:

$$\begin{aligned} P &= N \int_{L/2}^{L} \frac{L - x}{x}\, dx \\ &= \frac{\left(L \ln x - x\right)\big|_{L/2}^{L}}{L/2} = 2\ln 2 - 1 \approx 38.6\% \end{aligned}$$

1.1 Random Variables

Consider a set $\{A_k\}$ of random (not necessarily independent) variables.

Sum of Random Variables

Suppose that the sum of random variables comes to A:

$$A = \sum_{k=1}^{n} A_k$$

In the continuous large-n limit, the average value of A can be written as an n-dimensional integral

$$\langle A \rangle = \int A \cdot w\left(A_1, \ldots, A_n\right) dA_1 \ldots dA_n \, .$$

Replace A in the above with its sum representation:

$$\langle A \rangle = \sum_{k=1}^{n} \int A_k \cdot w\left(A_1, \ldots, A_n\right) dA_1 \ldots dA_n \, ,$$

where the 'sum' symbol has been harmlessly pulled outside all n of the integrals.

Simplifying the above is a straightforward exercise, with the majority of integrals satisfying the normalization condition and resolving to one. After the dust settles, one finds

$$\langle A \rangle = \langle A_1 \rangle + \langle A_2 \rangle + \cdots + \langle A_n \rangle \, ,$$

which, strictly translated, means *the average of the sum is the sum of the averages*:

$$\langle A \rangle = \sum_{k=1}^{n} \langle A_k \rangle \tag{1.4}$$

Independent Random Variables

More can be said about the weight function $w(k)$ in the regime of independent random variables. In the same sense that $P(A \cap B) = P(A)\, P(B)$ applies to independent events, we write

$$w\left(A_1, A_n, \ldots, A_n\right) = w\left(A_1\right) w\left(A_2\right) \cdots w\left(A_n\right)$$

when all probability distribution values $w(A_k)$ are independent.

Product of Independent Random Variables

Suppose that the product of random variables $\{B_k\}$ of n comes to

$$B = \prod_{k=1}^{n} B_k = B_1 \cdot B_2 \cdots B_n \, ,$$

and let us calculate the average value $\langle B \rangle$. Going by definition, this amounts to

$$\langle B \rangle = \prod_{k=1}^{n} \int B_k \cdot w\left(B_1\right) \cdots w\left(B_n\right) dB_1 \ldots dB_n \, ,$$

the 'product' symbol has been pulled outside all n of the integrals, and the probability distribution is factored to accommodate independent B_k.

From this, we see the right side is the product of n independent integrals, and conclude

$$\langle B \rangle = \langle B_1 \rangle \cdot \langle B_2 \rangle \cdots \langle B_n \rangle \, ,$$

which, strictly translated, means *the average of the product is the product of the averages*:

$$\langle B \rangle = \prod_{k=1}^{n} \langle B_k \rangle \tag{1.5}$$

1.2 Variance

Starting from the sum

$$A = \sum_{k=1}^{n} A_k \, ,$$

square both sides and convince yourself that

$$A^2 = \left(\sum_{i=1}^{n} A_i\right)\left(\sum_{j=1}^{n} A_j\right) = \sum_{k=1}^{n} A_k^2 + \sum_{i \neq j} c_{ij} A_i A_j \, ,$$

where c_{ij} are the binomial coefficients to represent all cross terms.

Meanwhile, the square of the average $\langle A \rangle$ comes out to

$$\langle A \rangle^2 = \sum_{k} \langle A_k \rangle^2 + \sum_{i \neq j} c_{ij} \langle A_i \rangle \langle A_j \rangle \, .$$

We can also calculate $\left\langle A^2\right\rangle$ by exploiting the the independence among A_k, resulting in

$$\left\langle A^2\right\rangle = \sum_{k=1}^{n}\left\langle A_k^2\right\rangle + \sum_{i\neq j} c_{ij}\left\langle A_i\right\rangle\left\langle A_j\right\rangle \ .$$

Taking the difference $\left\langle A^2\right\rangle - \left\langle A\right\rangle^2$, the cross terms cancel and we arrive at a simple relation connecting A to its members:

$$\begin{aligned}\left\langle A^2\right\rangle - \left\langle A\right\rangle^2 &= \sum_{k=1}^{n}\left\langle A_k^2\right\rangle - \left\langle A_k\right\rangle^2 \\ &+ \cancel{\sum_{i\neq j} c_{ij}\left\langle A_i\right\rangle\left\langle A_j\right\rangle} - \cancel{\sum_{i\neq j} c_{ij}\left\langle A_i\right\rangle\left\langle A_j\right\rangle}\end{aligned}$$

The square root of $\left\langle A^2\right\rangle - \left\langle A\right\rangle^2$ is defined as the *variance* in A:

$$\mathrm{Var}\left(A\right) = \sqrt{\left\langle A^2\right\rangle - \left\langle A\right\rangle^2} \tag{1.6}$$

As we've built it, the variance has some more handy expressions:

$$\mathrm{Var}\left(A\right) = \sqrt{\sum_{k=1}^{n}\left\langle A_k^2\right\rangle - \left\langle A_k\right\rangle^2} = \sqrt{\sum_{k=1}^{n}\left(\mathrm{Var}\left(A_k\right)\right)^2}$$

1.3 Dispersion

A variation in the sum A of independent variables, denoted ΔA, is also known as *dispersion*, defined as:

$$\Delta A = A - \left\langle A\right\rangle = \sum_{k=1}^{n}\left(A_k - \left\langle A_k\right\rangle\right) \tag{1.7}$$

From this, it's easy to show that the average dispersion is zero:

$$\left\langle \Delta A\right\rangle = \left\langle A\right\rangle - \left\langle A\right\rangle = 0$$

The expectation value $\left\langle \Delta A^2\right\rangle$, however, is more telling. By brute force, first write

$$\begin{aligned}\Delta A^2 &= \left(\left(A_1 - \left\langle A_1\right\rangle\right) + \left(A_2 - \left\langle A_2\right\rangle\right) + \cdots\right)^2 \\ &= \sum_{k=1}^{n}\left(A_k - \left\langle A_k\right\rangle\right)^2 + \sum_{i\neq j} c_{ij}\Delta A_i \Delta Aj \ ,\end{aligned}$$

so then:

$$\left\langle \Delta A^2\right\rangle = \sum_{k=1}^{n}\left\langle\left(A_k - \left\langle A_k\right\rangle\right)^2\right\rangle + \cancel{\sum_{i\neq j} c_{ij}\left\langle \Delta A_i\right\rangle\left\langle \Delta Aj\right\rangle}$$

This is result is perhaps not surprising, telling us the total ΔA^2 is the sum of its constituents:

$$\left\langle \Delta A^2\right\rangle = \sum_{k=1}^{n}\left\langle \Delta A_k^2\right\rangle \tag{1.8}$$

In the large-n limit, the average $\left\langle A\right\rangle$ scales with n, and meanwhile we see $\left\langle \Delta A^2\right\rangle$ also scales with n. The ratio of the RMS dispersion over the average thus tends to zero, as

$$\frac{\sqrt{\left\langle \Delta A^2\right\rangle}}{\left\langle A\right\rangle} \approx \frac{1}{\sqrt{n}} \rightarrow 0 \ ,$$

telling us that fluctuations in A become negligibly small.

2 Discrete Distributions

2.1 Two-State System

Consider a balanced coin that is tossed to generate n random events resulting in either H(eads) or T(ails). If we are interested in the portion m 'heads' events that occur without the order of events being important, the combination number

$$C_n^m = \frac{n!}{m!\left(n-m\right)!}$$

summarizes the system. Said another way, the multiplicity of the system Ω is 'n choose m':

$$C_n^m = \Omega\left(m,n\right) = \binom{n}{m}$$

The sum of all C_n^m across the whole range of m, namely from 0 to n, must resolve to the total multiplicity of events, namely 2^n for a coin tossing game:

$$2^n = \sum_{m=0}^{n}\frac{n!}{m!\left(n-m\right)!}$$

Normalized Probability Distribution

Knowing the total states available to the two-state system, we can write the probability of attaining m events among n trials in any two-state system as:

$$P\left(m,n\right) = \frac{1}{2^n}\frac{n!}{m!\left(n-m\right)!} \tag{1.9}$$

In the above definition, we divide by the factor 2^n so that the sum of all probabilities - accounting for all outcomes - sums to one.

The combination number C_n^m can be interpreted nicely by spotting the pattern that emerges in trivial cases. A single toss can result in T or H, which we denote

$$\omega_1 = \left(T,H\right) \ .$$

Denoting m as the number of H-events, we write

$$C_1^0 = 1 \qquad C_1^1 = 1$$

For a game of $n = 2$ tosses, the list of possible events is

$$\omega_2 = (TT, TH, HT, HH) \ .$$

Again denoting m as the number of H-events, we write

$$C_2^0 = 1 \qquad C_2^1 = 2 \qquad C_2^2 = 1$$

Similarly, a game of three tosses has

$$\omega_3 = (TTT, TTH, THT, THH, HTT, HTH, HHH)$$

with combinations

$$C_3^0 = 1 \qquad C_3^1 = 3 \qquad C_3^2 = 3 \qquad C_3^3 = 1 \ .$$

The pattern in C_n^m (stand back and look at the page) matches the rows of Pascal's triangle.

Heuristic Derivation

There is a (perhaps) intuitive way to derive C_n^m. Take n coins and lay them all down showing T, represented by $C_n^0 = 1$. Turn any one of the coins to H and find $C_n^1 = n$. Turn any two of the coins to H and find

$$C_n^2 = n\frac{(n-1)}{2} \ ,$$

and for three,

$$C_n^3 = n\frac{(n-1)}{2}\frac{(n-2)}{3} \ ,$$

and so on. Building this up for m total H-faces, we find

$$C_n^m = \frac{n!}{m!\,(n-m)!} \ ,$$

the familiar combination number.

2.2 Multi-State System

A generalization of the two-state system is the *multi-state* system. Going for a modest example, consider a three-sided coin with faces A, B, C. Flipping such a coin to generate n total events, let:

- $n_A \rightarrow$ Number of outcomes A
- $n_B \rightarrow$ Number of outcomes B
- $n_C \rightarrow$ Number of outcomes C
- $p_A \rightarrow$ Probability of of outcome A
- $p_B \rightarrow$ Probability of outcome B
- $p_C \rightarrow$ Probability of of outcome C

With this, we can write the multiplicity of the system exhibiting the state (n_A, n_B, n_C, n):

$$\Omega\left(n_A, n_B, n_C, n\right) = \frac{n!}{n_A! n_B! n_C!}\, p_A^{n_A} p_B^{n_B} p_C^{n_C}$$

In the special case $C = 0$, the above reduces to the non-normalized binomial distribution.

3 Binomial Distribution

Derivation

Consider an *unbalanced* coin having inherent probability p to land on H(eads), and correspondingly $1-p$ to land on T(ails). As a generalized two-state system, a game of n tosses generates the same potential outcomes:

$$\begin{aligned}
\Omega_1 &= T, H \\
\Omega_2 &= TT, TH, HT, HH \\
\Omega_3 &= TTT, TTH, THT, THH, HTT, HTH, HHH
\end{aligned}$$

Of course, the probability P of generating m Heads-events requires an extra argument to account for the imbalance p. Denoting the modified combination symbol $P_n^m(p)$, the two-state analysis generalizes by:

$$\begin{aligned}
P_1^0(p) &= 1 - p \\
P_1^1(p) &= p
\end{aligned}$$

$$\begin{aligned}
P_2^0(p) &= (1-p)^2 \\
P_2^1(p) &= 2 \cdot p\,(1-p) \\
P_2^2(p) &= p^2
\end{aligned}$$

$$\begin{aligned}
P_3^0(p) &= (1-p)^3 \\
P_3^1(p) &= 3 \cdot p\,(1-p)^2 \\
P_3^2(p) &= 3 \cdot p^2\,(1-p) \\
P_3^3(p) &= p^3
\end{aligned}$$

Evidently, the factors of p and $1-p$ compound into the terms p^m and $(1-p)^{n-m}$, but otherwise this analysis traces that of the two-state system exactly. Scanning for a pattern in the above, we evidently have

$$P_n^m(p) = \binom{n}{m}(1-p)^{n-m}\, p^m \ .$$

This result is known as the *binomial distribution*, and gives the probability of attaining, in general, m events of weight p among n trials:

$$P(m,n,p) = \frac{n!}{m!\,(n-m)!}\,(1-p)^{n-m}\,p^m \qquad (1.10)$$

Note there is no need to divide by 2^n. The binomial distribution as written is unit-normalized already.

Average Values

Define a random variable z_k that is equal to one if the event H with weight p occurs in the k-th trial, and is equal to zero otherwise. The average value of z_k is then

$$\begin{aligned}\langle z_k\rangle &= P(H)\,z(H) + P(T)\,z(T)\\ &= p\cdot 1 + (1-p)\cdot 0 = p\,,\end{aligned}$$

and, simply enough, the average of z_k^2 reads

$$\langle z_k^2\rangle = p\cdot 1^2 + (1-p)\cdot 0^2 = p\,.$$

Standard Deviation

The standard deviation in z, denoted σ_z, is evidently

$$\sigma_z = \sqrt{\langle z_k^2\rangle - \langle z_k\rangle^2} = \sqrt{p-p^2} = \sqrt{p\,(1-p)}\,.$$

Next, note that the number m of H-events among the n independent trials is the sum

$$m = \sum_{k=1}^{n} z_k\,,$$

implying

$$\langle \Delta m^2\rangle = \sum_{k=1}^{n}\langle \Delta z_k^2\rangle = n\,\langle \Delta z^2\rangle\,,$$

or, in tighter notation for large-n systems,

$$\sigma_m = \sqrt{n\sigma_z^2} = \sqrt{np\,(1-p)}\,.$$

Mode

The mode of the binomial distribution corresponds to the value m that maximizes $P(m,n,p)$. To gain on this question, consider the ratio

$$R = \frac{P(m+1,n,p)}{P(m,n,p)}\,,$$

readily simplifying to

$$R = \left(\frac{n-m}{m+1}\right)\frac{p}{1-p}\,.$$

The special value $R=1$ indicates where the derivative of $P(m,n,p)$ is zero, i.e. the probability transitions from increasing with m to decreasing with m. Setting $R=1$ at the critical m^*, we further find

$$m^* = np - (1-p)\,.$$

In the large-n limit, the mode m^* approximates to:

$$m^* \approx np$$

Problem 3

Haldor the Viking has slain sixteen ooze creatures in the swamp. After a thorough forensic analysis, Haldor finds a single gold cup among the corpses. He remembers from swamp lore that a slain ooze has a 1/3 chance to drop a gold cup. What are the chances he found just one cup after slaying sixteen oozes? Repeat the calculation for finding two cups, three cups, etc., up to sixteen cups. Also account for zero cups.

Answer: Model a slain ooze as a weighted coin with a Heads probability of 1/3, and a Tails probability of 2/3, which calls for a straightforward application of the binomial distribution. For finding one gold cup, we have

$$\begin{aligned}P(16,1,1/3) &= \frac{16!}{1!\,(16-1)!}\,(2/3)^{16-1}\,(1/3)^1\\ &= \frac{16}{3}\left(\frac{2}{3}\right)^{15} \approx 0.01218\,,\end{aligned}$$

and then for other numbers of gold cups:

$$\begin{aligned}P(16,2,1/3) &\approx 0.04567\\ P(16,3,1/3) &\approx 0.1066\\ P(16,4,1/3) &\approx 0.1732\\ P(16,5,1/3) &\approx 0.2078\end{aligned}$$

$$\begin{aligned}P(16,6,1/3) &\approx 0.1905\\ P(16,7,1/3) &\approx 0.1361\\ P(16,8,1/3) &\approx 0.07654\\ P(16,9,1/3) &\approx 0.03402\end{aligned}$$

$$\begin{aligned}P(16,10,1/3) &\approx 0.01191\\ P(16,11,1/3) &\approx 0.003247\\ P(16,12,1/3) &\approx 0.0006765\\ P(16,13,1/3) &\approx 0.0001041\end{aligned}$$

$$\begin{aligned}P(16,14,1/3) &\approx 0.00001115\\ P(16,15,1/3) &\approx 0.0000007434\\ P(16,16,1/3) &\approx 0.00000002323\\ P(16,0,1/3) &\approx 0.001522\end{aligned}$$

Problem 4

Monique is practicing netball. She knows from experience that the probability of her making any one shot is 70%. Her coach has asked her to keep practicing until she scores 50 goals. How many shots would she need to attempt to ensure that the probability of making at least 50 shots is more than 99%?

Answer: This problem is analogous to flipping a weighted coin with bias p. The probability of scoring k shots in N tosses is

$$P(k, N, p) = \frac{N!}{k!\,(N-k)!}\,(1-p)^{N-k}\,p^k\,,$$

where summing over k gives the cumulative distribution:

$$99\% = \sum_{k=50}^{N} \frac{N!}{k!\,(N-k)!}\,(0.3)^{N-k}\,(0.7)^k$$

This is best solved by a computer, where one should find $N = 86$.

3.1 Plumbing System

Introductory Remarks

Plumbers in working Massachusetts must follow the State *Plumbing Code* to ensure the safe and adequate operation of finished plumbing systems. Formally called *248 CMR 10.00*, the *Uniform State Plumbing Code* is available via *mass.gov* as PDF. Edition 12/8/23 is used for the majority of this study.

Fixture Unit - Definition

A term used heavily throughout the plumbing code is the *fixture unit*, which has a definition tailored for plumbers appearing in *Section 10.03: Definitions*:

Fixture Unit. One cubic foot of water drained in a 1.25 inch pipe over a period of one minute. One cubic foot of water is equal to 7.5 gallons.

The definition provided is alarmingly vague. Whether (or not) one cubic foot of water *actually* drains through a 1.25 inch diameter pipe in precisely one minute is questionable. Needless to mention, one cubic foot is closer to 7.4813 gallons of liquid. As 7.5 has two digits of precision, the writer may as well have stated the pipe is 1.3 inches in diameter.

Despite the definition, there is enough information to crudely solve for the velocity v, if uniform, of the water in the 1.25 inch diameter pipe. Assuming the pipe is cylindrical and the water fills the cross section, a straightforward exercise in kinematics tells us the velocity is:

$$v = 23.5\,\frac{\text{in}}{\text{sec}} = 117\,\frac{\text{ft}}{\text{min}}$$

Once again, one should be wondering if a ≈ 7.5 gallon slug of water is actually moving through a pipe at about two feet per second. One can readily imagine factors that could offset this number, i.e. the effects of gravity and the roughness of the interior.

One can also imagine the preceding observations as being pedantic, or perhaps unfairly picky, especially when plumbers are compensated for their labors rather than theories. On the other hand, students of plumbing should expect - and deserve - something better than the State-provided definition of the fixture unit. (The story gets worse before it gets better.)

Fixture Unit - Dimensionality

Given the definition, one may expect that a fixture unit is a flow rate, perhaps measured in gallons per minute, and then perhaps multiplied (or divided?) by 1.25 somewhere, because surely the pipe diameter factors in. It turns out none of this is correct.

Browsing CMR 10.00 or comparable resources, one instead finds that fixture units are *not measured in physical units*, but are instead always *dimensionless* scale factors like 3, π, or 1000. This is also troublesome for the definition, because we lack three numbers to balance the units of gallons, minutes, and inches.

Also found in *Section 10.03: Definitions* is the definition of *load factor*, which reads:

Load Factor. The percentage of the total connected fixture unit flow which is likely to occur at any point in the drainage system. It varies with the type of occupancy, the total flow unit above this point being considered, and with the probability factor of simultaneous use.

This definition is a slight improvement over the previous one by alluding to likelihood of occurrence and type of occupancy. Sadly, the terms ‘probability’ and ‘load factor’ appear nowhere else in CMR 10.00. It will turn out, however, that probability theory is at the heart of what a fixture unit *is*.

Fixture Unit - Alternate Definitions

Reaching for the *Uniform Plumbing Code, IAMPO/ANSI 1 - 2021*, one finds a more honest yet similarly vague word salad explaining the fixture unit:

Fixture Unit. A quantity in terms of which the load-producing effects on the plumbing system of different kinds of plumbing fixtures are expressed on some arbitrarily chosen scale.

We finally encounter a sensible definition of the fixture unit in the *International Plumbing Code, ICC A117.1-2017*, which reads:

Drainage Fixture Unit. A measure of the probable discharge into the drainage system by various types of plumbing fixtures. The drainage fixture-unit value for a particular fixture depends on its volume rate of drainage discharge, on the time duration of a single drainage operation and on the average time between successive operations.

One can speculate as to why the definition of the fixture unit varies so vividly per authority. Regardless of this, any literal understanding of the matter will remain elusive if plumbing codes are the only resource.

The Plan

We've racked up plenty of technical debt by skewering the definition of the fixture unit, and the intent here is not to solve the problems with the definition stated in CMR 10.00 or other publications. Instead, the plan is to take a first-principles approach using probability theory in the same way as done by the inventor of the fixture unit, Roy B. Hunter, in the years leading to 1940.

We start with a homemade warm-up exercise called the **Valve and Tank Problem**. This problem is informed by a mixture of real data and more-or-less made up numbers corresponding to likeliness of occurrence, type of occupancy, etc., of certain plumbing fixtures.

Next, we deploy a **Deconvoluted Calculation** to once-and-for-all force an understanding of the fixture unit. Such an effort also yields the so-called *Hunter's curve*, which is the source of the various 'gallons per minute versus fixture unit' tables occurring in the myriad of plumbing resources, including CMR 10.00.

Fixture unit values tend to change over time and location, however the work that follows will readily generalize to suit any fixture parameters. Data contemporary to the 1940s is used to ensure we're on track with the original work of R. B. Hunter.

Valve and Tank Problem

Two devices, or *fixtures* in American plumbing systems are the (i) *flush valve*, and (ii) the *flush tank*. The flush valve conveys 4 gallons of water over a period of 9 seconds.[1] The flush tank conveys 4 gallons in 60 seconds. Assume that all flush valves and flush tanks are used six times per hour on average.

If a certain building has $V = 20$ flush valves and $T = 30$ flush tanks, use only the information provided to (i) estimate the combined number of devices j being used simultaneously. (ii) Calculate the probability $P(j)$ of any j occurring and check that the estimate for j is correct. (iii) Calculate the probability that zero devices are in use at a given moment. (iv) Determine the number k at which $P(k) \approx 1\%$. (v) Calculate the final water demand using k.

Per hour (3600 seconds), a flush valve is operating for an average of 54 seconds. Similarly, a flush tank operates for 360 seconds. At a given moment, there is a

$$p_v = \frac{54}{3600} = 0.015$$

probability that a given flush valve is operating, and a

$$p_t = \frac{360}{3600} = 0.1$$

probability that a flush tank is operating.

Modal Analysis

Let x_v denote the number of flush valves in use simultaneously, and let x_t be the number of flush tanks being used such that

$$j = x_v + x_t$$

at a given moment.

Modeling each fixture as a weighted coin, we know from the binomial distribution that the most probable values for x_v, x_t are approximately the respective modes

$$x_v^* = Vp_v - (1 - p_v) \approx Vp_v$$
$$x_t^* = Tp_t - (1 - p_t) \approx Tp_t \ ,$$

[1] Figures gathered from PDH Course M126, *Sizing Plumbing Water System.* A. Bhatia. 2020. *www.PDHonline.prg*

or

$$x_v^* = (20)(0.015) = 0.3$$
$$x_t^* = (30)(0.1) = 3 \,.$$

The convoluted mode j^* ought to be the sum of the individual modes

$$j^* \approx x_v^* + x_t^* \,,$$

and using the numbers on hand, one finds

$$j^* \approx 3.3 \,.$$

That is, we expect about three devices to be conveying water at a given moment. The total mode is essentially dominated by x_t^*.

In terms of x_v, x_t, the water demand rate $D(x_v, x_t)$ is given by

$$D(x_v, x_t) = x_v \left(\frac{4\text{ gal}}{9\text{ s}}\right) + x_t \left(\frac{4\text{ gal}}{60\text{ s}}\right) \,.$$

Using $j = 3$, it follows that x_v can take on any value 0, 1, 2, 3. Thus:

$$D(0,3) = 0.2\text{ gal/s}$$
$$D(1,2) = \left(0.44\overline{4} + 0.13\overline{3}\right)\text{ gal/s} = 0.57\overline{7}\text{ gal/s}$$
$$D(2,1) = \left(0.88\overline{8} + 0.066\overline{6}\right)\text{ gal/s} = 0.95\overline{5}\text{ gal/s}$$
$$D(3,0) = 1.3\overline{3}\text{ gal/s}$$

To make use of the above information, recall from the values of x_v^*, x_t^* that $(0,3)$ (zero flush valves in use, three flush tanks in use) is the most likely configuration at a given moment, and the typical water demand is estimated at 0.2 gal/s, or 12 gallons per minute.

Of course, we don't want to design based on typical use. It's much better to anticipate the worst-probable case, which is not the worst *possible* case. (To design based on the worst possible case scenario is inefficient, costly, or worse.)

Probabilistic Analysis

The probability that there are x_v flush valves in use is given by the binomial distribution:

$$P_v(x_v) = \binom{V}{x_v}(1-p_v)^{V-x_v} p_v^{x_v}$$

Also in terms of x_v, we write for the flush tank:

$$P_t(j - x_v) = \binom{T}{j - x_v}(1-p_t)^{T-(j-x_v)} p_t^{j-x_v}$$

The total probability is the sum of convolutions of the two above distributions:

$$P(j) = \sum_{x_v=\max(0,j-T)}^{\min(j,V)} P_v(x_v) \cdot P_t(j - x_v)$$

Using $j = 3$, the above probability is

$$P(3) = \sum_{x_v=0}^{3} P_v(x_v) \cdot P_t(j - x_v) \,,$$

where:

$$P_v(x_v) = \binom{20}{x_v}(1-0.015)^{20-x_v}(0.015)^{x_v}$$
$$P_t(j - x_v) = \binom{30}{3 - x_v}(1-0.1)^{30-(3-x_v)}(0.1)^{3-x_v}$$

Evaluating $P(3)$ is quite a chore. For $x_v = 0$, we have:

$$P_v(0) = \binom{20}{0}(1-0.015)^{20}(0.015)^0 \approx 0.7391$$
$$P_t(3) = \binom{30}{3}(1-0.1)^{27}(0.1)^3 \approx 0.2361$$
$$P_v(0) \cdot P_t(3) \approx 0.1745$$

Continuing for $x_v = 1$:

$$P_v(1) = \binom{20}{1}(1-0.015)^{19}(0.015)^1 \approx 0.2251$$
$$P_t(2) = \binom{30}{2}(1-0.1)^{28}(0.1)^2 \approx 0.2277$$
$$P_v(1) \cdot P_t(2) \approx 0.05126$$

Continuing for $x_v = 2$:

$$P_v(2) = \binom{20}{2}(1-0.015)^{18}(0.015)^2 \approx 0.03257$$
$$P_t(1) = \binom{30}{1}(1-0.1)^{29}(0.1)^1 \approx 0.1413$$
$$P_v(2) \cdot P_t(1) \approx 0.004602$$

Continuing for $x_v = 3$:

$$P_v(3) = \binom{20}{3}(1-0.015)^{17}(0.015)^3 \approx 0.002976$$
$$P_t(0) = \binom{30}{0}(1-0.1)^{30}(0.1)^0 \approx 0.04239$$
$$P_v(3) \cdot P_t(0) \approx 0.0001262$$

The total probability that any three of the $20 + 30$ flush valves and flush tanks are in simultaneous use is the sum of the above convolutions:

$$P(3) \approx 0.2394 \approx 23.94\%$$

Returning briefly to the issue of water demand, compare each convolution $P_v(x_v) \cdot P_t(j - x_v)$ to see the figure 0.1745 dominating its siblings, thus $(0, 3)$ is the most likely configuration in accordance with our estimate of $j^* = 3$.

We still need to establish that $P(3)$ is greater than all other $P(j)$. To study the $j = 2$ case, we need

$$P(2) = \sum_{x_v=0}^{2} P_v(x_v) \cdot P_t(j - x_v) \ ,$$

or:

$$\begin{aligned} P(2) = & P_v(0) \cdot P_t(2) + P_v(1) \cdot P_t(1) \\ & + P_v(2) \cdot P_t(0) \end{aligned}$$

Using the figures calculated above, we learn

$$\begin{aligned} P(2) &\approx 0.1683 + 0.03181 + 0.001381 \\ &\approx 0.2015 = 20.15\% \ . \end{aligned}$$

The $j = 1$ case is also done easily, as

$$P(1) = P_v(0) \cdot P_t(1) + P_v(1) \cdot P_t(0)$$

readily computes to

$$\begin{aligned} P(1) &\approx 0.1044 + 0.009542 \\ &\approx 0.1139 \approx 11.39\% \ . \end{aligned}$$

The $j = 0$ case is trivial from the information on hand, coming out to

$$P(0) \approx (0.7391)(0.04239) \approx 0.03133 \approx 3.133\% \ .$$

Reading this result backward, we see there is a 96.87% chance that at least one device is operating at a given moment.

Technically, we need to also check all additional $P(j)$ up to $j = 50$. This is best left to a machine, so we'll do one more case by hand, namely $j = 4$. For this, we need

$$\begin{aligned} P(4) = & P_v(0) \cdot P_t(4) + P_v(1) \cdot P_t(3) \\ & + P_v(2) \cdot P_t(2) + P_v(3) \cdot P_t(1) \\ & + P_v(4) \cdot P_t(0) \ . \end{aligned}$$

Most of these figures were calculated above, with the new members being:

$$\begin{aligned} P_v(4) &= \binom{20}{4} (1 - 0.015)^{16} (0.015)^4 \approx 0.0001926 \\ P_t(4) &= \binom{30}{4} (1 - 0.1)^{26} (0.1)^4 \approx 0.1771 \end{aligned}$$

Turning the crank, one finds

$$\begin{aligned} P(4) = & 0.1309 + 0.05315 \\ & + 0.007416 + 0.0004205 \\ & + .000008164 \ , \end{aligned}$$

or

$$P(4) \approx 0.1919 \approx 19.19\% \ .$$

To summarize, we found

$$\begin{aligned} P(0) &= 3.133\% \\ P(1) &= 11.39\% \\ P(2) &= 20.15\% \\ P(3) &= 23.94\% \\ P(4) &= 19.19\% \ , \end{aligned}$$

which is maximal at $j = 3$. Given that $P(4)$ begins a downward trend, we can be sure that all subsequent $P(j > 4)$ are all less than $P(3)$.

Summing each percentage above, we conclude that there is a 77.80% chance that any number from zero to four devices are in use simultaneously. This means there is a 22.20% chance that any number $5 \le j \le 50$ devices are in use simultaneously.

Reading the trend in the $P(j)$, we estimate that the probability should be less than 1% by say, $j = 10$, thus we define a variable

$$k = 10 \approx 3j^* \approx 1 + 9 \ ,$$

which corresponds to one flush valve and nine flush tanks.

For a final flow rate we find

$$D(1, 9) = 1.04\bar{4} \, \frac{\text{gal}}{\text{s}} \ ,$$

which is about 63 gallons per minute.

To reiterate the last step, one reasons that $P(k) = 1\%$ corresponds to the worst probable use case. The special value k is also called a *design factor*. For this problem, there is about a 1% chance that more than 10 of the 50 fixtures are operating simultaneously.

Deconvoluted Calculation

While adequate, the above calculation is admittedly too detailed for application in the field, especially when there are multiple types of plumbing devices in the system.

To work toward something simpler, separately consider a (i) flush valve, (ii) flush tank, (iii) bathtub having the following characteristics[2]:

[2] National Bureau of Standards Report: BMS 65 by Late Dr. R. B. Hunter (1940)

- The flush valve conveys 4 gallons over an interval of 9 seconds per use, and is used once every 5 minutes (300 s, twelve uses per hour).
- The flush tank conveys 4 gallons over an interval of 60 seconds per use, and is used once every 5 minutes (300 s, twelve uses per hour).
- A bathtub requires 16 gallons over an interval of 120 seconds, and is used once every 30 minutes (1800 s).

From these, we find that the flush valve operates a total of 108 seconds for every 3600. Thus the probability of any given flush valve being in use is

$$p_v = \frac{108}{3600} = 0.03\,.$$

Similarly, the flush tank has

$$p_f = \frac{720}{3600} = 0.2\,,$$

and finally for the bathtub:

$$p_b = \frac{120}{1800} = 0.06667$$

Then, one can immediately write the probability that x_v flush valves are in use out of V total valves:

$$P_v\left(x_v, V\right) = \binom{V}{x_v}\left(1 - p_v\right)^{V - x_v} p_v^{x_v}$$

From each value V we can derive a most-likely number of valves x_v^* in simultaneous use, along with a design factor k_v such that $P\left(k_v, V\right) = 1\%$. The very same can be said for flush tanks by switching indices $v \to t$, $V \to T$, leading to

$$P_t\left(x_t, T\right) = \binom{T}{x_t}\left(1 - p_t\right)^{V - x_t} p_t^{x_t}\,,$$

and then switching indices to b, B for bathtubs, we have

$$P_b\left(x_b, B\right) = \binom{B}{x_b}\left(1 - p_b\right)^{V - x_b} p_b^{x_b}\,.$$

Now we must find a design factor for each probability considered. To proceed, choose $V = T = B = 25$ and use a computer to find:

- $P_v\left(x_v, 25\right)$ equals 1% at $k_v = 3.671$
- $P_t\left(x_t, 25\right)$ equals 1% at $k_t = 0.4622$
- $P_b\left(x_b, 25\right)$ equals 1% at $k_b = 5.412$

Using each k-value, calculate the total water demand for each case of 25 fixtures:

$$D_v\left(k_v\right) = 3.671\left(\frac{4\text{ gal}}{9\text{ s}}\right) = 97.89\frac{\text{gal}}{\text{min}}$$
$$D_t\left(k_v\right) = 10.16\left(\frac{4\text{ gal}}{60\text{ s}}\right) = 40.64\frac{\text{gal}}{\text{min}}$$
$$D_b\left(k_b\right) = 5.412\left(\frac{8\text{ gal}}{60\text{ s}}\right) = 43.30\frac{\text{gal}}{\text{min}}$$

Repeating for $V = T = B = 50$, find:

- $P_v\left(x_v, 50\right)$ equals 1% at $k_v = 5.194$
- $P_t\left(x_t, 50\right)$ equals 1% at $k_t = 16.72$
- $P_b\left(x_b, 50\right)$ equals 1% at $k_b = 8.139$

Then, for 50 fixtures:

$$D_v\left(k_v\right) = 5.194\left(\frac{4\text{ gal}}{9\text{ s}}\right) = 138.5\frac{\text{gal}}{\text{min}}$$
$$D_t\left(k_v\right) = 16.72\left(\frac{4\text{ gal}}{60\text{ s}}\right) = 66.88\frac{\text{gal}}{\text{min}}$$
$$D_b\left(k_b\right) = 8.139\left(\frac{8\text{ gal}}{60\text{ s}}\right) = 65.11\frac{\text{gal}}{\text{min}}$$

Repeating again for $V = T = B = 75$, find

- $P_v\left(x_v, 75\right)$ equals 1% at $k_v = 6.509$
- $P_v\left(x_t, 75\right)$ equals 1% at $k_t = 22.84$
- $P_b\left(x_b, 75\right)$ equals 1% at $k_b = 10.58$

Then, for 75 fixtures:

$$D_v\left(k_v\right) = 6.509\left(\frac{4\text{ gal}}{9\text{ s}}\right) = 173.6\frac{\text{gal}}{\text{min}}$$
$$D_t\left(k_v\right) = 22.84\left(\frac{4\text{ gal}}{60\text{ s}}\right) = 91.36\frac{\text{gal}}{\text{min}}$$
$$D_b\left(k_b\right) = 10.58\left(\frac{8\text{ gal}}{60\text{ s}}\right) = 84.64\frac{\text{gal}}{\text{min}}$$

Repeating once more for $V = T = B = 100$, find

- $P_v\left(x_v, 100\right)$ equals 1% at $k_v = 7.720$
- $P_v\left(x_t, 100\right)$ equals 1% at $k_t = 28.74$
- $P_b\left(x_b, 100\right)$ equals 1% at $k_b = 12.87$

Then, for 100 fixtures:

$$D_v\left(k_v\right) = 7.720\left(\frac{4\text{ gal}}{9\text{ s}}\right) = 205.9\frac{\text{gal}}{\text{min}}$$
$$D_t\left(k_v\right) = 28.74\left(\frac{4\text{ gal}}{60\text{ s}}\right) = 115.0\frac{\text{gal}}{\text{min}}$$
$$D_b\left(k_b\right) = 12.87\left(\frac{8\text{ gal}}{60\text{ s}}\right) = 103.0\frac{\text{gal}}{\text{min}}$$

It would be more efficient to state the above results and all subsequent calculations in table form. In the following, the first column holds the number of fixtures, and the remaining three columns hold the gallons-per-minute (gpm) flow rates through the respective fixtures:

Fixture (count)	**Valve** (gpm)	**Tank** (gpm)	**Bath** (gpm)
5	51.65	15.25	20.15
10	66.29	22.58	27.20
15	78.06	29.00	33.08
20	88.41	34.97	38.37
25	97.89	40.64	43.30
50	138.5	68.88	65.11
75	173.6	91.36	84.64
100	205.9	114.9	103.0
125	236.4	138.0	120.5
150	265.8	160.6	137.6
175	294.2	183.0	154.4
200	322.0	205.1	170.8
250	375.8	248.8	202.9
300	428.0	292.1	234.4
350	489.0	334.9	265.4

Fixture Units

One can construct a table similar to the above with a fixed gpm rate and a variable number of fixtures. Aiming for 50 gal/m, we find (i) $V = 5$ yields $D_v = 51.65\,\text{gal/m}$, (ii) $T = 34$ yields $D_t = 50.42\,\text{gal/m}$, (iii) $B = 33$ yields $D_v = 50.68$ gal/m. Together, we jot the ratio $5 : 34 : 33$ for flush valves, flush tanks, and bathtubs, respectively. Repeating this for incrementing gpm rates yields the following:

Demand (gpm)	**Valve** (count)	**Tank** (count)	**Bath** (count)
50	5	34	33
100	27	85	96
150	58	139	169
200	96	195	246
250	137	252	326
300	181	310	407

The relationship between the flow rate in gallons per minute and total fixture count is more-or-less linear for each fixture type in the domain 150 gpm to 300 gpm. In ratio form, a subset of the above table reads:

Demand (gpm)	**Valve** (ratio)	**Tank** (ratio)	**Bath** (ratio)
150	1	2.397	2.914
200	1	2.031	2.563
250	1	1.839	2.380
300	1	1.713	2.249
(Average:)	(1)	(1.995)	(2.523)

Reading the average, observe that the water demand of one flush valve is like ≈ 2 flush tanks, or like ≈ 2.5 bathtubs. Stated in integer form, we have that ten flush valves is like five flush tanks, or like four bathtubs, implying the ratio

$$10 : 5 : 4 \, .$$

The numbers 10, 5, 4 are the respective *fixture units*, abbreviated FU, for the flush valve, flush tank, and bathtub, respectively.

The fixture unit is born from comparing the likely use of flush valves, flush tanks, and bathtubs at a given flow rate such that the ratio of the respective fixture counts numerically rounds to a set of integers. Hunter didn't *need* to choose integers, all that matters is that the average fixture count ratio $1 : 2 : 2.5$ is respected.

Generalizing the above, we may consider any water-conveying device, differentiating hot water demand, cold water demand, and drainage demand when applicable, as equivalent to some number $N/10$ flush valves. N is measured in fixture units.

For a minimal example, we have that 10 fixture units corresponds to one flush valve, which to remind, conveys 4 gallons over an interval of 9 seconds of water per use, and is used once every 5 minutes. Note these numbers are decades old and have been refined since the 1940s.

The modern flush valve, according to CMR 10.00, is classified as *Toilet, Valve Operated*, assigned to have 6 fixture units. Accordingly, such a modern valve conveys ≈ 1.5 gallons over 6 seconds. Thus we deduce that one fixture unit corresponds to 0.25 gallons conveyed over one second, or 15 gallons per minute. Up to a factor of ≈ 2, we've essentially recovered the definition listed in CMR 10.00.

Note, finally, that the 1.25 inch diameter pipe never entered the analysis, and is essentially a red herring in the definition. This is for good reason, because we are meant to choose the pipe diameter *after* knowing the approximate flow rate through the pipe.

Hunter's Curve

Fixture unit calculations were originally carried out by Roy B. Hunter in the years leading to 1940 and published by the *National Bureau of Standards Report: BMS 65*.

Using 10, 5, 4 as scale factors, recast the 'gpm vs. count' table by multiplying all V by 10, all T by 5, and all B by 4. This produces an equivalent table with fixture counts replaced by fixture units:

Demand	**Valve**	**Tank**	**Bath**
(gpm)	(FU)	(FU)	(FU)
150	580	695	676
200	960	975	984
250	1370	1260	1304
300	1810	1550	1628

Extending the table above and plotting the information on a graph, much as Hunter did, leads to Figure 1.1. Using the Figure, Hunter reasoned:

'...the error made by using curve 2 for both flush tanks and bathtubs for any number of either up to 300 would be small. Also, the demand load relative to the number of fixture units may be approximately represented in this range by a smoother curve drawn above the two probability curves and merged with curve 1 as shown by the broken line in [the] Figure...'

That is, curve 2 and curve 3 are essentially interchangeable before the broken line, and then the broken line takes over for curves 2 and 3 until joining curve 1, giving rise to Figure 1.2. The result is called called *Hunter's Curve*.

Demand vs Fixture Units

Finally, we summarize the information in Hunter's curve using the tables that follow.

Demand (Load)	**Valve**	**Tank**
(FU)	(gpm)	(gpm)
10	32.82	9.63
20	39.09	13.56
30	43.92	18.49*
40	48.02	27.20
50	51.65	30.24
60	54.97	33.08
70	58.04	35.78
80	60.93	38.37
90	63.67	40.87

Before the asterisk we derive values from curve 2. At the asterisk we use the average of curves 2 and 3. After the asterisk we use curve 3 and throughout the next table.

Demand (Load)	**Valve**	**Tank**
(FU)	(gpm)	(gpm)
100	66.29	43.30
140	75.84	52.45
180	84.40	61.00
200	88.41	65.11
250	97.89	75.05
300	106.8	84.60
400	123.2	102.9
500	138.5	120.5
750	173.6	162.6

After $\approx$ 1000 fixture units, all values are represented by curve 1.

Demand (Load)	**Valve**	**Tank**
(FU)	(gpm)	(gpm)
1000	205.9	205.1
1250	236.4	236.4
1500	265.8	265.8
1750	294.2	294.2
2000	322.0	322.0
2500	375.8	375.8
3000	428.0	428.0
4000	529.0	529.0
5000	626.8	626.8

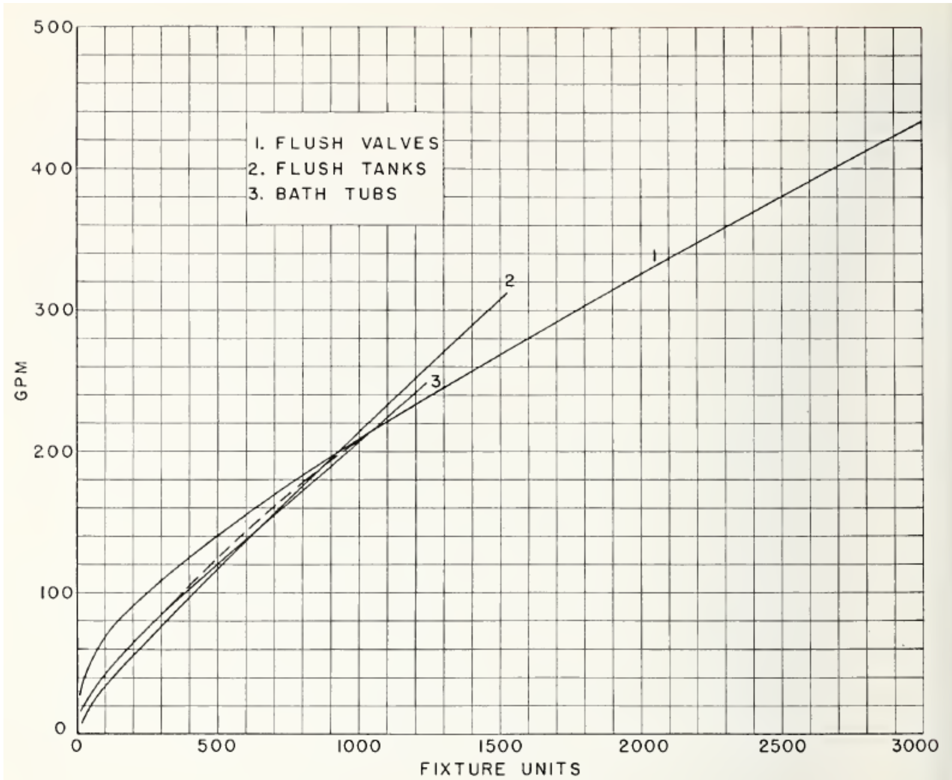

Figure 1.1: Gallons per minute versus fixture unit count for flush valves, flush tanks, and bathtubs. (*National Bureau of Standards Report: BMS 65*, 1940)

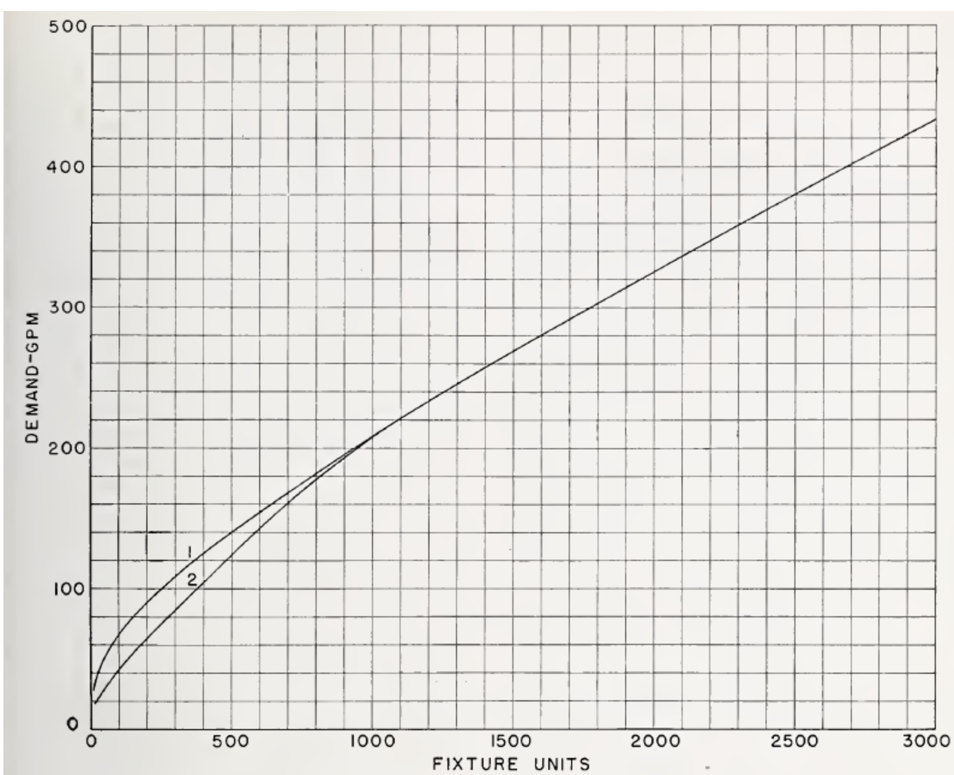

Figure 1.2: Hunter's Curve. (*National Bureau of Standards Report: BMS 65*, 1940)

4 Continuous Distributions

4.1 Gaussian Distribution

Recall that the probability of generating k results among n total trials in a two-state system is given by

$$P(k,n) = \frac{1}{2^n}\frac{n!}{k!\,(n-k)!}\ .$$

Introduce the shift

$$k \to k + \frac{n}{2}\ ,$$

which modifies the above:

$$P(k,n) = \frac{1}{2^n}\frac{n!}{\left(\frac{n}{2}+k\right)!\left(\frac{n}{2}-k\right)!}$$

In the large-k limit, making k a continuous variable, it makes sense to describe the system solely in terms of expectation values and their deviations, a notion formally called the *central limiting theorem*. Here we develop this idea on a two-state system to derive a central equation in probability theory called the *Gaussian distribution*.

To proceed in the large n-limit, we deploy Stirling's approximation for large numbers

$$\ln(n!) \approx n\ln(n) - n + \ln\left(\sqrt{2\pi n}\right)$$
$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\ ,$$

and the probability density reduces to

$$w(k) = e^{-2k^2/n}\sqrt{\frac{2}{\pi n}}\ . \tag{1.11}$$

The result $w(k)$ is the famed normalized Gaussian distribution centered at $k=0$. Introducing a nonzero shift of base-point value a, the generalized equation is

$$w(k) = e^{-2(k-a)^2/n}\sqrt{\frac{2}{\pi n}}\ .$$

Using Gaussian integrals, the average values and standard deviation are readily calculated:

$$\langle k\rangle = \int_n k\cdot w(k)\ dk = a$$
$$\left\langle k^2\right\rangle = \int_n k^2\cdot w(k)\ dk = \frac{n}{4}+a^2$$
$$\sigma_k = \sqrt{\left\langle k^2\right\rangle - \langle k\rangle^2} = \sqrt{\frac{n}{4}}$$

4.2 Poisson Distribution

Imagine trying to count the number of water molecules that pass a point in a river flowing at average speed v. Over time interval t, the average molecule count is directly proportional to vt. To reduce notation clutter, let us ignore the proportionality constant and take vt as a dimensionless quantity. Due to local random fluctuations in the river, an actual measurement would never precisely land on vt, but instead on an interval surrounding vt. Naturally we wonder, what is the time-varying probability $P_k(t)$ that k molecules are measured over the interval t?

To begin, partition the elapsed time t into n identical bins of width Δt such that $\Delta t \to 0$, and observe that each $P_k(\Delta t)$ relates to its $k-1$ and $k+1$ neighbors as:

$$\lim_{\Delta t\to 0} P_0(\Delta t) \gg P_1(\Delta t) \gg P_2(\Delta t) \gg P_3(\Delta t) \gg \cdots$$

This means it's more likely to measure few molecules in a small Δt-interval as opposed to many. We may proceed using weighted two-state analysis, wherein a Δt-interval may either be unfilled with zero molecules, or filled with one or more molecules. Borrowing the apparatus developed previously, we write

$$P(k,n,v\Delta t) = \frac{n!}{k!\,(n-k)!}\left(1-v\Delta t\right)^{n-k}\left(v\Delta t\right)^k\ ,$$

where n and k are integers. Substituting $t = n\Delta t$, we have

$$P(k,n,vt) = \frac{(vt)^k}{k!}\left(\frac{n!}{(n-k)!}\frac{1}{n^k}\right)\left(1-\frac{vt}{n}\right)^{n-k}\ .$$

In the large-n limit, the approximations

$$\frac{n!}{(n-k)!} \approx n^k$$
$$\left(1-\frac{vt}{n}\right)^{n-k} \approx e^{-vt}$$

are valid, and re-casting vt as a dimensionless variable q lands us at the anticipated *Poisson distribution*:

$$P_k(q) = \frac{q^k}{k!}e^{-q} \tag{1.12}$$

Summing over the variable k tells us $P_k(t)$ is already normalized:

$$\sum_{k=0}^{\infty}\frac{q^k}{k!}e^{-q} = e^{-q}\left(\sum_{k=0}^{\infty}\frac{q^k}{k!}\right) = e^{-q}e^{q} = 1$$

With $P_k(t)$ on hand, we may calculate $\langle k\rangle$, $\langle k^2\rangle$, and the standard deviation:

$$\begin{aligned}\langle k\rangle &= \sum_{k=0}^{\infty} k\frac{q^k}{k!}e^{-q} = e^{-q}\sum_{k=1}^{\infty}\frac{q^k}{(k-1)!}\\ &= e^{-q}\sum_{p=0}^{\infty}\frac{q^{(p+1)}}{p!} = e^{-q}qe^{q} = q\\ \langle k^2\rangle &= \sum_{k=0}^{\infty} k^2\frac{q^k}{k!}e^{-q} = e^{-q}q\sum_{p=0}^{\infty}(p+1)\frac{q^p}{p!}\\ &= q+q^2\\ \sigma_k &= \sqrt{q^2+q-q^2} = \sqrt{q}\end{aligned}$$

4.3 Random Product Problem

Consider the real numbers in the interval $(0:2)$. Let $\tilde{x}_1$ be a random number chosen from this interval, let $\tilde{x}_2$ be a second random number, and so on up to $\tilde{x}_n$. (Repeats are allowed but unlikely.)

Expectation

With this setup, suppose we are interested in the product of numbers in the list:

$$X_n = \prod_{j=1}^{n}\tilde{x}_j = \tilde{x}_1\cdot\tilde{x}_2\cdot\tilde{x}_3\cdots\tilde{x}_n$$

Sampling from $(0:2)$, it is true that the average random value is one:

$$\langle\tilde{x}_j\rangle = 1\,.$$

This should mean right away that the average product is also one:

$$\langle X_n\rangle = 1\cdot 1\cdot 1\cdots = 1$$

Disaster

All seems well until we try to verify $\langle X_n\rangle = 1$ on a calculator. To illustrate, take the contrived list with five members

$$\{\tilde{x}_j\} = \{0.8, .9, 1.0, 1.1, 1.2\}\ ,$$

so the product is

$$X_5 = (0.8)\,(0.9)\,(1.0)\,(1.1)\,(1.2) = 0.9504\,,$$

which is less than one.

The effect gets worse for increasing n, for if we continue the pattern so the list includes 0.7, 1.3, the product is

$$X_7 \approx 0.8648\,.$$

The members $\tilde{x}_j < 1$ weigh down the product X_n more than members $\tilde{x}_j > 1$ weigh the product up. After many trials, the net result is $X_n \to 0$, in contradiction to $\langle X_n\rangle = 1$.

You're encouraged to verify this on a computer with a variety of $\tilde{x}_j$ and a variety of n-values to see there is clearly something wrong with the way X_n is expected to behave. It seems that X_n reliably *decreases* for increasing n, so we inevitably conclude $X_n \to 0$.

Modified Interval

Going back to the beginning, adjust the interval to $(0:2.5)$ so that

$$\langle\tilde{x}_j\rangle = 1.25\,,$$

and run similar experiments. Now we're multiplying a list of numbers who average is clearly greater than one. However, much like the previous setup, the product X_n still goes to zero.

Adjust the interval once more to $(0:3)$ and start over. This time, we have

$$\langle\tilde{x}_j\rangle = 1.5\,,$$

and pattern finally breaks. One can check that product X_n tends to grow for increasing n, and for large n, the trend $X_n \to \infty$ occurs.

Tuning the Interval

Given the evidence on hand, there should be some interval $(0:p)$, where p is some number between 2.5 and 3 such that X_n does not tend to zero and does not tend to infinity:

$$X_n \propto \langle X_n\rangle$$

To estimate p, one may write a simple trial-and-error program that allows p to vary:

1. Choose an initial value for p.
2. Choose a sufficiently large sample of n values from the interval $(0:p)$ and calculate the corresponding X_n.
3. If X_n goes to zero, increase p.
4. If X_n goes to infinity, decrease p.
5. Goto step 2.

Doing this, one finds, after many trials:

$$p \approx 2.718\ldots$$

This answer is tantalizingly close to Euler's constant. Who saw that coming?

Proper Analysis

To reconcile the random product problem, begin with the natural logarithm of the product X_n:

$$\ln(X_n) = \ln(\tilde{x}_1) + \ln(\tilde{x}_2) + \ln(\tilde{x}_3) + \cdots$$

In the limit $n \to \infty$, it stands to reason that *every* real number in the interval $(0:p)$ is represented by some $\tilde{x}_j$ or another. Rearranging the sum to write these in order, we have

$$\ln(X) = \lim_{n\to\infty} \sum_{j=1}^{n} \ln\left(p\frac{j}{n}\right) .$$

The total interval $(0:p)$ can be made from n copies of a small interval Δx, which means $p/n = \Delta x$. Also substituting $x = j/n$, the above becomes

$$\ln(X) = \lim_{\Delta x\to 0} \frac{1}{p} \sum_{x>0}^{1} \ln(px)\,\Delta x .$$

The sum becomes an integral in the continuous limit

$$\ln(X) = \frac{1}{p}\int_0^1 (\ln(p) + \ln(x))\,dx ,$$

and the solution is straightforward:

$$p\ln(X) = (x\ln(p) + x\ln(x) - x)\Big|_0^1$$
$$p\ln(X) = \ln(p) - 1$$

Now comes the final argument. By avoiding $\ln(0) \to -\infty$ and also $\ln(\infty) \to \infty$, we're asking for X to be a finite number. In the infinite limit, it can only be that $X \to 1$:

$$p\ln(1) = 0 = \ln(p) - 1$$

The only solution to $\ln(p) = 1$ is $p = e$ and we're done.

Problem 5

Consider the real numbers in the interval $(0:1)$, and let $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$, etc. represent random samples from this interval. How many times n must a random $\tilde{x}_j$ be multiplied into a very large number $A \gg 1$ until the product is approximately one? In other words, solve for n in the following:

$$1 \approx A \cdot \tilde{x}_1 \cdot \tilde{x}_2 \cdots \tilde{x}_n$$

Hint:

$$0 \approx \ln(A) + \sum_{j=1}^{n} (\ln(x) + 1) - \sum_{j=1}^{n} (1)$$

The answer is $n \approx \ln(A)$.

4.4 Random Sums Problem

Accumulating random values $0 < r_k < 1$ in a sum, how many iterations $\langle n \rangle$ until the total is greater than one, on average?

Geometric Analysis

Begin by interpreting each interval $0 \le r_k \le 1$ as an independent 'number line' for each of the n variables needed. For $n = 2$, r_1, r_2 lie on orthogonal axes of a two-dimensional plane. For $n = 3$, r_1, r_2, r_3 lie on orthogonal axes of a three-dimensional volume, and so on.

Geometrically, the criteria

$$\sum_{j=1}^{n} r_j > 1$$

thus defines a triangular area in two dimensions, a pyramid-like volume in three dimensions, a hypervolume in four-dimensions, and so on. The space enclosed by each 'volume' is defined by

$$\sum_{j=1}^{n} r_j \le 1 .$$

For convenience, let us label $r_1 \to z$, $r_2 \to y$, $r_3 \to x$, $r_4 \to t$, $r_5 \to u$.

Examining $n = 2$, the line $z + y = 1$ encloses half of the unit square, formally shown via

$$V_2 = \int_0^1 \int_0^{1-z} dy\,dz$$
$$= \int_0^1 (1-z)\,dz = \left(z - \frac{z^2}{2}\right)\Big|_0^1 = \frac{1}{2} .$$

For $n = 3$, the plane $z + y + x = 1$ encloses one sixth of the unit cube:

$$V_3 = \int_0^1 \int_0^{1-z} \int_0^{1-z-y} dx\,dy\,dz = \frac{1}{6}$$

Jumping to $n = 4$ is impossible to visualize, however the required integral is easy enough to write and solve:

$$V_4 = \int_0^1 \int_0^{1-z} \int_0^{1-z-y} \int_0^{1-z-y-x} dt\,dx\,dy\,dz = \frac{1}{24}$$

Evidently, the enclosed volume is always the inverse of the factorial of the number of dimensions,

$$V_n = \frac{1}{n!} .$$

Probabilistic Calculation

We ultimately seek the expectation value $\langle n \rangle$, given by

$$\langle n \rangle = \sum_{n=2}^{\infty} n \cdot P(n) \ ,$$

where $P(n)$ is the probability of satisfying

$$\sum_{j=1}^{n} r_j < 1 \ .$$

By the geometric analysis, observe that $P(n)$ corresponds to the 'window' of volume bounded between V_n and V_{n-1}:

$$P(n) = \frac{1}{(n-1)!} - \frac{1}{n!} = \frac{n-1}{n!}$$

With this, we can calculate the expectation value

$$\langle n \rangle = \sum_{n=2}^{\infty} n \cdot \frac{n-1}{n!} = \sum_{n=2}^{\infty} \frac{1}{(n-2)!} = \sum_{0}^{\infty} \frac{1}{n!} \ ,$$

which indeed converges to Euler's constant:

$$e = \sum_{0}^{\infty} \frac{1}{n!}$$

Amazingly, we conclude:

$$\langle n \rangle = e$$

Index