Appendix G: Random Variables, Probability Theory and Pulse Train Amplitudes 2
(a) What is a Random Variable ? Part I 2
(b) What is a Random Variable ? Part II: the Capital Letter Notation 4
(c) Basic Probability Theory 6
(d) Ensemble Experiments 15
(e) Experiments rolling two dice at the same time 17
(f) Experimental determination of discrete distribution functions 20
(g) Experiments with Sequences of Pulse Train Amplitudes 22

Appendix G: Random Variables, Probability Theory and Pulse Train Amplitudes

Appendix D discusses random variables X and Y, and then, for a pulse train, the random variables Y_n and Y_m . The narrow purpose of that Appendix is to establish the relation among parameters α,β,μ,σ since these relations are needed to verify certain results of our pulse train analysis.

In this Appendix we take a more general view of random variables which then provides a context in which Appendix D is a particular special case. This subject always seems "slippery". One can find a discussion like the one presented below in every book on probability theory, but here we wish perhaps to put our own "spin" on the subject with comments not always appearing in texts.

(a) What is a Random Variable ? Part I

Our **opening definition** is that a random variable is a real parameter for which there exists some probability distribution. The parameter is "random" in the sense that it does not take a fixed value, but can take a range of values. This definition does not require that the probability distribution be a "flat" distribution in which all possible values of the parameter are equally likely, though that is possible, Sometimes one associates the word "random" with this flat distribution sense, such as the face up number of a rolled unloaded die having probability P = 1/6. But the random in "random variable" just means there is variability in the value of the parameter that can be connected with *some* distribution. The parameter can of course be viewed as a "variable", since it varies, and in particular the parameter will be the variable which is the argument of the probability distribution function. If that distribution happens to be a Gaussian (normal) distribution, then one might refer to the parameter as a "Gaussian random variable".

The above definition is rather abstract and it helps to put it into a context which involves other concepts which also have to be defined. Perhaps the most general framework is to think of the parameter described above as a "random variable" as being a possible **outcome** of an **experiment**, so to two new words have appeared. An experiment is "something one does" and an outcome is "something one measures after the experiment is done". A simple example is experiment = **roll one die**, outcome = number facing up after the die is rolled = random variable n. The set of all possible outcomes is called the **sample space**, often indicated by Ω (element = ω) or S (element = s). For the **die** experiment, we have Ω = {1,2,3,4,5,6} as the sample space. Since this is a set, various concepts regarding sets can be applied. A set we know has subsets. Each subset of the sample space has a peculiar name, each subset is called an **event**. In the die experiment, a possible event would be {2,4,6} which event is that the die rolled an even number. Another event would be {6} which event is that the die rolled a 6. For this experiment, the sample space is **discrete**.

Next, consider experiment = one spins a "**spinner**". This is a traditional (in probability texts) physical object that one might think of as a horizontal Lazy Susan having an arrow from center to some point on the rim, or a Roulette Wheel with some point marked on the rim, or just a metal arrow you spin and drop on a table top. The outcome of this experiment is the direction in which the arrow points in range (0, 2π) perhaps clockwise relative to North. In this example, the sample space is all real numbers in the interval (0,2π). For this experiment, the sample space is a **continuous** set which one could regard as the limit of a discrete set as the number of elements increases.

We shall now assume that in the die experiment the die was a "loaded" die of some sort, and that in the spinner experiment, the arrow was perhaps influenced by magnetic fields or by bad bearings on the Lazy Susan. If we do each experiment many times and each time we observe and write down the outcome, we can then "bin" these outcomes and create a distribution by dividing each bin count by the total number of experiments done (see section (f) below) . Here are possible distributions for our two experiments.

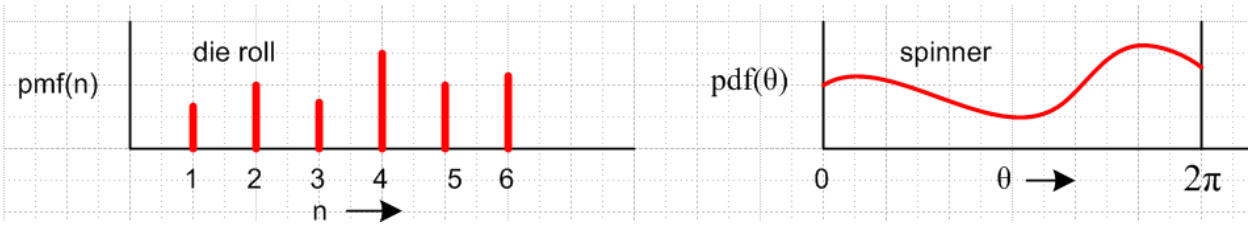


Fig G.1

For the spinner, the distribution is continuous and is called a **probability density function** or **pdf**. For the die, the distribution is discrete as is called a **probability mass function** or **pmf** . One could in fact write down a pdf for the left picture assuming n was a continuous variable in this manner,

$$\text{pdf}(n) = \Sigma_{i=1}^{6} \text{ pmf}(n)\, \delta(i-n) \ . \qquad \text{(G.1)}$$

Both a pmf and pdf must sum to one, since the probability of *some* outcome is always 1,

$$\Sigma_{k=1}^{6} \text{pmf}(k) = 1 \qquad \int_0^{2\pi} \text{pdf}(x)dx = 1.$$

Notice that the above equations are consistent in that

$$1 = \int_0^{2\pi} \text{pdf}(n)dn = \int_0^{2\pi} [\Sigma_{i=1}^{6} \text{ pmf}(n)\, \delta(i-n)]dn = \Sigma_{i=1}^{6} \text{pmf}(n) \int_0^{2\pi} \delta(i-n)]dn$$

$$= \Sigma_{i=1}^{6} \text{pmf}(n) = 1 \ .$$

The motivation for these names pdf and pmf comes from physics. One can have a continuous distribution of mass in some compressible fluid, say, where it would be called a mass **density** ρ. But in the physics of idealized point particles of **mass** m_i, the mass is congealed at specific points in space. For a set of particles of mass m_i at spatial locations $\mathbf{r}_i$ one could still write a mass density $\rho(\mathbf{r})$ as

$$\rho(\mathbf{r}) = \Sigma_i \, m_i \, \delta(\mathbf{r}-\mathbf{r}_i)$$

and one sees the analogy with (G.1) above.

Here then we have seen two "random variables" in action. One, n, is the discrete outcome of a die roll, the other θ is the continuous outcome of a spinner spin. Each of these is a real parameter and for each there exists an associated probability distribution, as plotted above. Thus, each of these parameters fulfills the requirements of our opening definition of a "random variable".

(b) What is a Random Variable ? Part II: the Capital Letter Notation

We shall continue to examine various experiments to see what new concepts appear.

Consider experiment = **coin toss** with sample space S = {heads,tails} = possible outcomes. The new feature here is that the outcomes are not real numbers and therefore don't fit our requirement that a random variable be a real parameter. This is easily remedied by assigning real numbers to each outcome, such as heads → 1 and tails → 0. If s is an element of the sample space S = {heads,tails}, then we can define a little function h = H(s) such that 1 = H(heads) and 0 = H(tails). In general this function is a mapping from the sample space S to the real numbers, so we can say H:S→R. In this situation, we can think of h as being a "random variable" as defined above. It is a real parameter and it has an associated probability distribution which in this case is a pmf since S is discrete. For a "loaded" coin we might have

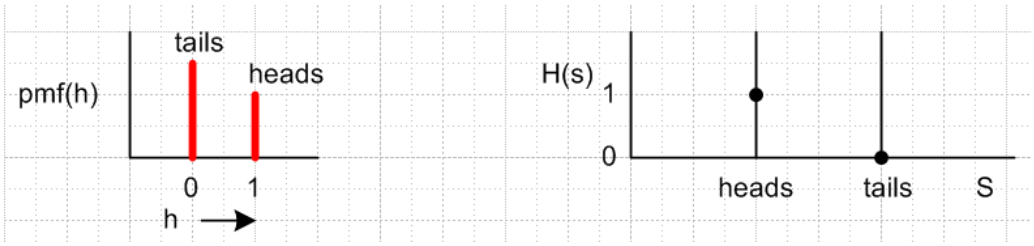


Fig G.2

We are now going to slightly alter our definition of a random variable so it is more precise. As a preliminary, forget probability and just consider some function y = f(x). In this equation, the object f is clearly a function, while the object y is a value that function can take. Thus, y and f are not the same thing. They are equal in the sense that y = f(x) for some x, but y and f belong to completely different classes of objects. Perhaps y is a real *variable* in R, whereas f is a *mapping* (function) from R to R. It is quite useful to have different symbols for y and f. If one were to write y = y(x), the reader would understand what was meant, but now y has two separate meanings (function, and value of function). To maintain precision, if is better that these symbols be different. One choice is to represent the function f(x) as Y(x), so the function name is now Y, while y is a value the function Y can take. Then y = Y(x) and things are clear.

Now we return to h = H(s). H is a function or mapping from S to R, whereas h is a real number. We would not in general say that h and H are the same object. For the coin toss experiment, it is really the function H that is the random variable, whereas h is a value that random variable can take. In our previous experiments with die and spinner, it happened that the function H was the identity function, so we didn't "see it" in the discussion. But still for the spinner, we would write θ = Θ(arrow position) and, as we set things up, if an arrow position s is in the range (0,2π), then we happen to have Θ(φ) = φ and Θ = 1. The function Θ would have been more visible had we perhaps taken arrow position in degrees and put θ in radians.

So one more time with our **improved definition**: a random variable X is the *function* which maps the sample space S to the real axis, X: S→R. The values that the random variable X takes are called x. For each element of the sample space, there is some corresponding real number x. In considering x = X(s), we make a distinction between the function X and the value x that it takes for some particular sample space element s. The other ingredient to the definition of a random variable remains unchanged: there is some probability distribution associated with the values x that the random variable X takes if an experiment is repeated many times. That is, x does not take a single fixed value. The distribution need not be a flat distribution, which one might call a random distribution in which every value of x is equally likely. Thus,

for this latter flat concept of "random" we arrive at the famous statement that a random variable is neither a variable (it is a function) nor is it random (distribution can be non-random).

What different name might one use? Perhaps a "probabilistic function" which maps the possibly non-numeric outcomes of an experiment to a real parameter, which parameter has some non-trivial probability distribution which can be observed by doing the experiment many times. The problem with this phrase is that it does not focus on that parameter which is the main item of interest, the variable of random variable.

For some general experiment, the outcomes are likely to be non-numerical in nature, and a function like H(s) will be required to map the outcomes onto the real number axis. For a discrete sample space, we might represent this situation as follows:

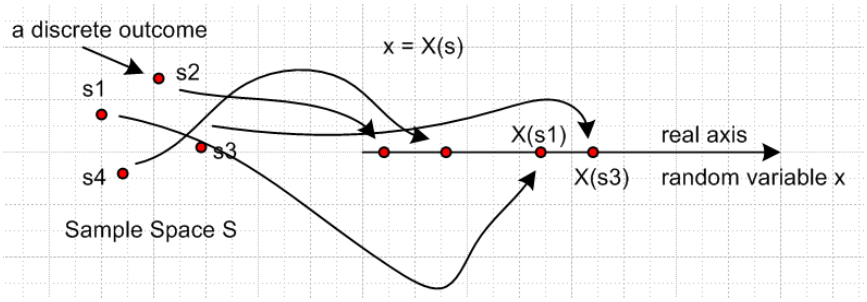


Fig G.3

In this case, an event which is any subset of the outcomes in the sample space will map into some set of points on the real axis. For a continuous sample space, the picture is a little different,

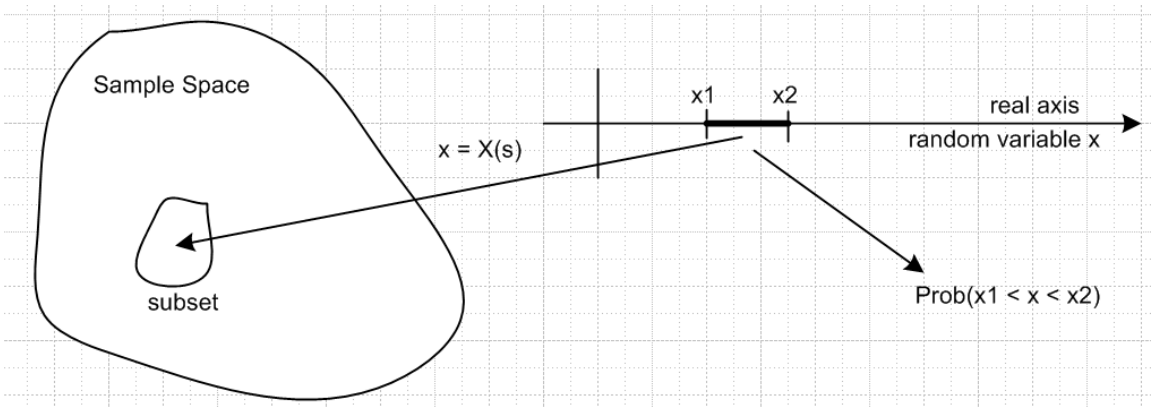


Fig G.4

The new feature here is a certain continuity requirement: every interval of the random variable x (like the interval shown) must map back into some subset of the sample space. Since subsets are events, this means that every interval of the x axis must map back to some well-defined event. Going the other way, not every event (subset) maps into an interval of the x axis. The continuity idea is that if two points are close together on the real x axis, then they must be close together in the sample space. This in turn means that the sample space has to have some kind of metric to allow a notion of distance between two points. Furthermore, as the subset is expanded, the interval it maps to cannot becomes smaller!

Having now discussed random variables, outcomes, experiments, sample spaces, and events, we are in a good position to review some general probability theory.

(c) Basic Probability Theory

In the above discussion, if we have random variables A,B,C.... , we can write a **joint probability** distribution function in this manner

pdf(A=a, B=b, C=c.....) (G.2)

where now we don't distinguish whether the various sample spaces are continuous or discrete, we just write anything as a pdf (with the understanding of (G.1) above). The meaning here is that pdf(...) is the probability that random variable A has value a, while at the same time (the same experiment) random variable B has value b, and so on. For the continuous case, pdf(...) da db dc... is the same probability but for the range da of a and db of b, etc.

If all these random variables are **statistically independent** (such as A = number of dust particles on your pillow and B = temperature at some location on Pluto), this joint probability distribution factors,

pdf(A=a,B=b,C=c.....) = pdf(A=a) pdf(B=b) pdf(C=c) .... (G.3a)

For two random variables A and B, one would have

pdf(X=x,Y=y) = pdf(X=x) pdf(Y=y) . (G.4a)

In this case, A and B are statistically independent and are said to be **uncorrelated**. If one regards pdf(X=x,Y=y) as a function $f_{\mathbf{Y}}(x)$ for various fixed values of y, (G.4a) says that the *shape* of this function is not influenced by the values of y, only the *overall scale* of f(x) is affected by y. When we formally define the notion of correlation below, we will see that the official correlation between two random variables X and Y vanishes if (G.4a) is true.

We now adopt a **shorthand notation**:

p(a,b,c....) ≡ pdf(A=a,B=b,C=c.....) = $\mathrm{pdf}_{\mathbf{ABC}\ldots}(a,b,c...)$ (G.5)

When one sees p(2, -4, 0. 4...) *one must remember* that the arguments correspond to values of specific random variables, and if things become unclear, one must revert to the fuller notation. One trick is to use a parameter name that reminds the reader of the random variable name, such as a for A: p(a) = P(A=a).

Using this shorthand notation we may express the two equations above as

p(a,b,c.....) = p(a)p(b)p(c)..... statistical independence (G3.b)

p(x,y) = p(x)p(y) uncorrelated (G.4b)

Any pdf is **normalized** to 1 since the probability of all possible outcomes (mapped from sample spaces to the random variables) is 1. Thus,

$$\int\int .... \; p(x,y....)\, dx\, dy... \; = 1 \qquad \text{or} \qquad \Sigma_{x,y....}\; p(x,y....) \; = 1 \; . \tag{G.6}$$

An example is that $\int p(x)dx = 1$ or $\Sigma_x p(x) = 1$. If x is discrete and y is continuous, $\Sigma_x \int p(x,y)dy = 1$, but we won't bother to show all such "mixed" cases below.

In general, if p(a,b,c...) is some N^{th} order joint probability, one can find lower order probabilities by summing over some of the variables:

Fact: To obtain a lower-order joint probability from a higher-order one,

$$p(x,y,z...) = \Sigma'_{a,b,c...} \; p(a,b,c.....) \tag{G.7}$$

Here {x,y,z...} is some subset of {a,b,c....} and the notation Σ' means that we sum over all variables in {a,b,c.... } *except* those in the subset { x,y,z...} . One could state (G.7) in a set notation this way

$$p(S') = \Sigma_{S-S'}\; p(S) \qquad \text{where } S' \subset S \tag{G.8}$$

<u>Proof</u>: The proof is merely the observation that if we are only interested in {x,y,z...}, we go ahead and let all the other variables {a,b,c....} – {x,y,z...} take all possible values and we add up the probability of each of these cases to get its contribution to p(x,y,z...). This is just a case of adding probabilities of outcomes to get a total probability of interest.

Examples:

$$p(x) = \Sigma_y \; p(x,y)$$

$$p(x) = \Sigma_{y,z} \; p(x,y,z)$$

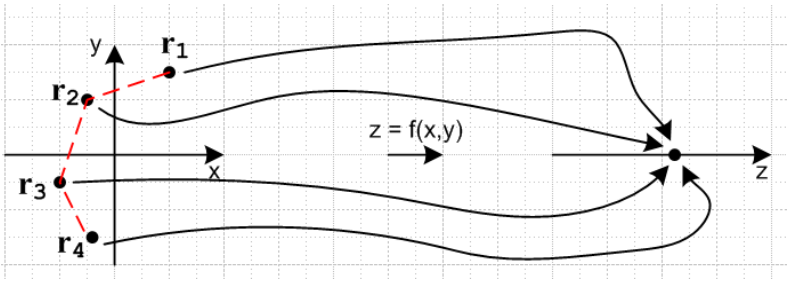
$$p(x,y) = \Sigma_z \; p(x,y,z)$$

$$p(x,y) = \Sigma_{a,b,c...\neq x,y} \; p(a,b,c.....) \tag{G.9}$$

Before we can continue our little presentation, we need to state and prove an important theorem about random variables which are functions of other random variables. To this end, we must first digress on a set of math Lemmas.

Some Math Lemmas

Consider z = f(x,y) = f(**r**) where **r** = (x,y). Here f is a "function" which means it is a single-valued function which means under the mapping f: $R^2 \rightarrow R$, every vector **r** in the domain lands in some unique location in the range. It is possible and in fact likely that f will be a many-to-one function, meaning for a given z in the range, there might be several $\mathbf{r}_i$ in the domain such that z = f($\mathbf{r}_i$). If the domain is discrete, then so is the range. We might have this situation:



domain D_f range R_f Fig G.5

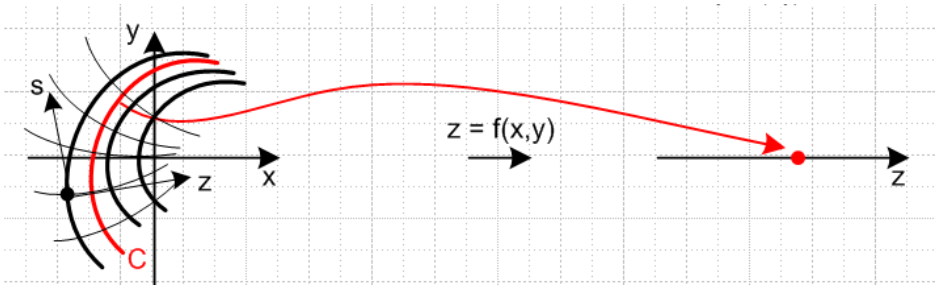
We could then talk about $\Sigma_{x,y}$ being the sum of all points $\mathbf{r} = (x,y)$ in the domain such that f(x,y) = z for some z in the range. We can exhaust all points in the range R_f by doing $z_i = f(\mathbf{r}_i)$ and letting $\mathbf{r}_i$ exhaust all (x,y) in the domain D_f. This is how we discover the extent of the range R_f.

We now make this claim, where we have in mind some unseen quantity being acted upon by both sides of the equation (symbol ∋ means "such that")

$$\sum_{z \in R_f} \left[\sum_{x,y \ni f(x,y) = z} \right] = \sum_{x,y \in D_f} \qquad \text{(G.10)}$$

Proof: If on the left we first sum over the points (x,y) corresponding to z according to z = f(x,y), and *then* we sum over all z in the range of f, our sum includes every point (x,y) in D_f exactly once. The double sum is simply a certain ordering of the total sum shown on the right. The same point (x,y) cannot show up twice in the double sum for two different values of z, because if it did, f(x,y) would map that point into those two z different values, but f is supposedly single-valued. Nor can any point (x,y) be omitted in the double sum on the left because the range R_f for z was created by exhausting all points (x,y) in D_f, so any (x,y) in D_f corresponds to some z in the range R_f.

Now for a continuous domain D_f, we might have



domain D_f range R_f Fig G.6

where now an entire continuous curve C (red) in the domain maps to some z in R_f. The equation of the red curve C is z = f(x,y) for some fixed value of z. The heavy curves including this red one are curves of constant z. We can imagine some other function s = g(x,y) whose curves of constant s form an orthogonal curvilinear coordinate system (right angles at any point) with the curves of z = f(x,y) = constant. The

parameter s would then vary along each constant-z curve, marking off points along the curve. The analogous statement to the discrete statement made above is this (only a claim, we shall not prove it)

$$\int_{z \in R_f} dz \left[\int_{C(z)} ds\ J \right] = \iint_{x,y \in D_f} dxdy \qquad \text{(G.11)}$$

where J is the Jacobian between coordinates (x,y) and (z,s). This is the same Jacobian idea that appears in taking (x,y) to polar coordinates (r,θ) where dxdy = rdrdθ with J = r. We shall not pursue this Jacobian matter further other than to claim it is possible to find a g(x,y) that works and that certain technical issues arise concerning the reasonableness of the function f(x,y).

Comment: In order for the Jacobian to exist and be well-behaved, function f(x,y) must be continuous and differentiable (C^1) in both variables, *and* the mapping between (x,y) and (z,s) must be essentially one- to- one (invertible) so that given (z,s) one can compute (x,y) and vice versa.

Example: Suppose $z = f(x,y) = 2x^3 + 3y^2$ and we are interested in the curve C(z=2). This curve C is the intersection of the surfaces $z = 2x^3 + 3y^2$ (red) and z = 2 (gray), as illustrated here:

```
p1 := plot3d(2*x^3+3*y^2,  x=-2..2,y=-2..2, color=red):
p2 := plot3d(2,  x=-2..2,y=-2..2, color = gray,style = patchnogrid):
display(p1,p2);
```

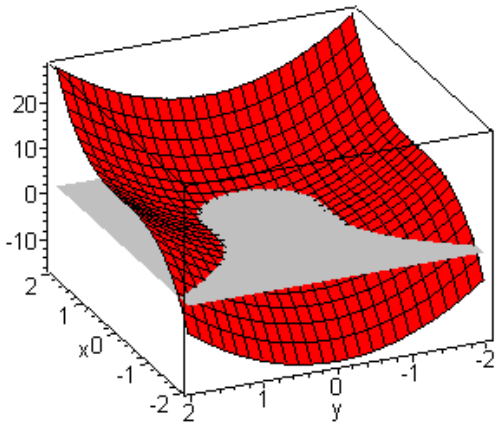


Fig G.7

If we had z = f(x,y,w) so Z is a function of *three* random variables X,Y and W, then our two results above would become

$$\sum_{z \in R_f} \left[\sum_{x,y,w\ni f(x,y,w)=z} \right] = \sum_{x,y,w \in D_f}$$

$$\int_{z \in R_f} dz \left[\int_{S(z)} dS\ J \right] = \iint_{x,y,w \in D_f} dxdydw \qquad \text{(G.12)}$$

The first line seems straightforward, but the second is more complicated. Inside the square bracket we now have dS being a differential patch of area, and S(z) is a 2D surface in the (x,y,z) space on which z is a constant according to z = f(x,y,w). For example, in spherical coordinates we write $r = x^2+y^2+z^2$ and a surface of constant r is a spherical shell. Now we have to imagine two other coordinates $s_1 = g(x,y,w)$ and $s_2 = g(x,y,w)$ so that (z,s_1,s_2) form an orthogonal coordinate system and then the new J is the Jacobian between (x,y,w) and (z,s_1,s_2).

In a more economical notation, and changing to z = f(a,b,c) we can write (G.10) and (G.11) as

$$\Sigma_z [\Sigma^z_{x,y}] = \Sigma_{x,y} \tag{G.13a}$$

$$\int dz [\int_{C(z)} ds\, J] = \int\int da\, db \tag{G.13b}$$

and then for (G.9),

$$\Sigma_z [\Sigma^z_{a,b,c}] = \Sigma_{a,b,c} \tag{G.14a}$$

$$\int dz [\int_{S(z)} dS\, J] = \int\int\int da\, db\, dc \tag{G.14b}$$

where $\Sigma^z_{a,b,c}$ indicates that the sum is restricted to those (a,b,c) values for which f(a,b,c) = z.

Hopefully it is clear how this general idea can be extended to Z = f(A,B,C,D....) where random variable Z is a function of some arbitrary number N of random variables A,B,C,D... .

Fact: One can organize a total sum/integral over the space of N random variables in this manner:

$$\Sigma_z [\Sigma^z_{a,b,c\ldots}] = \Sigma_{a,b,c\ldots} \tag{G.15a}$$

$$\int dz [\int_{S(z)} dS\, J] = \int\int \ldots \int da\, db\, dc \ldots\ldots \tag{G.15b}$$

In the first line, Σ^z means the sum is constrained to be only over those a,b,c... such that f(a,b,c...) = z. In the second line, S(z) is an N-1 dimensional surface located within the N dimensional space of (a,b,c....), which surface is defined by f(a,b,c....) = z .

With these Lemmas out of the way, we can now resume our basic probability theory review.

Fact: Any reasonable real function of random variables is a random variable. (G.16)

Proof: First consider Z = f(X,Y), where X and Y are random variables. If x is an allowed value of X, and y of Y, then the allowed values of Z will be z = f(x,y). Since x and y must be real, and since f is a real function, the allowed values of z are real, one of the requirements for Z to be a random variable. At our level of rigor, it only remains to find the probability distribution associated with Z. We claim this is given by,

$$p(z) = \Sigma^z_{x,y}\, p(x,y) \tag{G.17a}$$

$$p(z) = \int_{C(z)} ds\, J\, p(x(s,z),y(s,z)) = \int_{C(z)} J\, p(x,y) \quad . \qquad \text{(G.17b)}$$

Looking at (G.17a), for a *given value* of z, in z = f(x,y) only certain x and y values are possible. Thus, only these values of x and y appearing in p(x,y) can contribute to p(z). Other values of x and y will contribute to p(z') for some other z'. We can check normalization as follows using (G.13a)

$$\Sigma_z\, p(z) = \Sigma_z \{ [\Sigma^z{}_{x,y}]\, p(x,y)\} = \{\Sigma_z\, [\Sigma^z{}_{x,y}]\}\, p(x,y) = \Sigma_{x,y}\, p(x,y) = 1 \; .$$

What is the sample space for Z ? We can select one according to the following plan. Let R_A and R_B be the total sets of values of parameters a and b. Then $R_Z = f(R_A,R_B)$ treated as an equation of sets tells us the set R_Z. We could then just select the sample space for Z to be $S_Z = R_Z$ which would be a numerical sample space.

More generally, for Z = f(A,B,C....) with N arguments, the distribution for Z is given by

$$p(z) = \Sigma^z{}_{a,b,c\ldots}\; p(a,b,c,....) \qquad \text{(G.18a)}$$

$$p(z) = \int_{C(z)} dS\, J\, p(a,b,c....) \qquad a = a(z, s_1, s_2 ..... s_{N-1}), \text{ etc.} \qquad \text{(G.18b)}$$

The function f must be C^1 in all its arguments (as noted above) and must be such that the Jacobian J is well behaved (finite and non-zero) over the entire regions of interest in both spaces (a,b,c...) and (z,s_1,s_2...). This is why (G.16) contains the word "reasonable". **QED**

Definition : The **expected value** of a random variable X is given by

$$E(X) \equiv \int x\, p(x)\, dx \quad \text{or} \quad E(X) \equiv \Sigma_x\, x\, p(x) \quad = \textbf{the mean}, \text{ often called } \mu_X \qquad \text{(G.19)}$$

Comment 1: The expected value is sometimes called the expectation or the expectation value. In quantum mechanics the phrase expectation value predominates, but elsewhere it is the expected value. In quantum mechanics, all physical observables are random variables (position, momentum, energy, etc) except in quantum states which are eigenstates of the observable's quantum operator, in which case the observable takes a fixed value.

Comment 2: In (G.19) above for the mean (and generally below), we could write $E(X) \equiv \Sigma_i\, x_i\, p(x_i)$ (as in Appendix D) where the x_i are the possible values that random variable X can take. But in this section we use $E(X) \equiv \Sigma_x\, x\, p(x)$ where then x itself represents the values X can take. This notation puts the summation form on a little more equal footing with the integral form $\int x\, p(x)\, dx$.

Fact: If Z = f(X,Y), the expected value of Z is given by (G.20)

$E(Z) = \Sigma_{\mathbf{z}}\ z\ p(z) = \Sigma_{\mathbf{x,y}}\ f(x,y)\ p(x,y)$ discrete

or

$E(Z) = \int z\ p(z)\ dz = \int\int f(x,y)\ p(x,y)\ dx\ dy$. continuous

Proof: For the discrete case

$E(z) = \Sigma_{\mathbf{z}}\ z\ p(z)$ // by the definition of an expected value of a random variable

$= \Sigma_{\mathbf{z}}\ z\ [\ \Sigma^{\mathbf{z}}{}_{\mathbf{x,y}}\ p(x,y)]$ // by (G.17a)

$= \Sigma_{\mathbf{z}}\ [\ \Sigma^{\mathbf{z}}{}_{\mathbf{x,y}}]\ f(x,y)\ p(x,y)$ // move the x,y sum to the left, and replace z = f(x,y)

$= \Sigma_{\mathbf{x,y}}\ f(x,y)\ p(x,y)$ // by (G.13a)

For the continuous case

$E(Z) = \int z\ p(z)\ dz = \int dz\ z\ [\int_{\mathbf{C(z)}} ds\ J\ p(x,y)\ ]$ // by (G.17b)

$= \int dz\ [\int_{\mathbf{C(z)}} ds\ J\ ]\ f(x,y)\ p(x,y)$ // move $\int ds$ to the left, replace z = f(x,y)

$= \int\int dxdy\ f(x,y)\ p(x,y)$ // by (G.13b) **QED**

We now generalize this fact to obtain:

Theorem: If Q = f(X,Y,Z....) where X,Y,Z... are random variables, and if f is a reasonable real valued function, then

(1) Q is a random variable and

(2) $E(f(X,Y,Z....)) = \Sigma_{\mathbf{x,y,z...}}\ f(x,y,z....)\ p(x,y,z....)$

$E(f(X,Y,Z....)) = \int\int\int ...dxdydz....\ f(x,y,z....)\ p(x,y,z....)$ (G.21)

Proof: This is a straightforward generalization (G.20) based on (G.18) and (G.15). One just mimics the proof of (G.20).

Special cases:

$E(f(X)) = \Sigma_{\mathbf{x}}\ f(x)\ p(x)$

$E(f(X)) = \int dx\ f(x)\ p(x)$ (G.22)

$$E(f(X,Y)) = \Sigma_{x,y}\ f(x,y)\ p(x,y)$$

$$E(f(X,Y)) = \int\int dx\ dy\ f(x,y)\ p(x,y) \qquad \text{(G.23)}$$

Next we define the **covariance** of X and Y, and evaluate it using (G.23),

$$cov(X,Y) \equiv E((X-\mu_x)(Y-\mu_y)) = \Sigma_{x,y}\ (x-\mu_x)\ (y-\mu_y)\ p(x,y)$$

$$\text{or } \int\int\ (x-\mu_x)\ (y-\mu_y)\ p(x,y)\ dx\ dy \qquad \text{(G.24a)}$$

Notice that

$$cov(X,Y) \equiv E(XY) - \mu_x E(Y) - \mu_y E(x) + \mu_x\mu_y = E(XY) - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y = E(XY) - \mu_x\mu_y$$

so we have this alternate method of computing covariance

$$cov(X,Y) = E(XY) - \mu_x\mu_y \quad . \qquad \text{(G.24b)}$$

The **variance** of X is the covariance of X with itself and is the square of the **standard deviation** σ(X),

$$var(X) \equiv [\sigma(X)]^2 = cov(X,X) = E((X-\mu_x)(X-\mu_x)) = E((X-\mu_x)^2)$$

$$= \Sigma_x\ (x-\mu_x)^2\ p(x) \quad \text{or} \quad \int\ (x-\mu_x)^2\ p(x)\ dx \qquad \text{(G.25)}$$

where we use (G.22) with $f(X) = (X-\mu_x)^2$.

Finally, the **correlation** of X and Y is the covariance normalized by the two standard deviations,

$$corr(X,Y) \equiv cov(X,Y) / [\sigma(X)\ \sigma(Y)] \quad . \qquad \text{(G.26)}$$

One can take the limit of a joint distribution of random variables as two of the variables become the same. For example, for continuous and discrete (where we now use the alternate indexed notation),

$$\lim_{X\to Y} p(x,y) = p(x)\delta(x-y) \qquad \lim_{X\to Y} p(x_i,y_i) = p(x_i)\delta_{i,j} \qquad \text{(G.27)}$$

The reason is that if X and Y are the same random variable, they cannot take different values. If X,Y and Z are all the same variable, then we would have (dropping the limit notation)

$$p(x,y,z) = p(x)\delta(x-y)\delta(x-z) \qquad \text{or} \qquad p(x_i,y_j,z_k) = p(x_i)\ \delta_{i,j}\ \delta_{i,k} \qquad \text{(G.28)}$$

Note that $\delta(x-y)\delta(x-z) = \delta(x-y)\delta(y-z) = \delta(x-z)\delta(y-z)$ and $\delta_{i,j}\ \delta_{i,k} = \delta_{i,j}\ \delta_{j,k} = \delta_{i,k}\ \delta_{j,k}$.

We can apply (G.27) to obtain an alternate evaluation of the variance,

$$\text{var}(X) \equiv [\sigma(X)]^2 = \lim_{X\to Y} \text{cov}(X,Y) = \lim_{X\to Y} \left[\int\int (x-\mu_x)(y-\mu_Y)\, p(x,y)\, dx\, dy \right] \qquad \text{(G.29)}$$

$$= \int\int (x-\mu_x)(y-\mu_x)\, p(x)\delta(x-y)\, dx\, dy$$

$$= \int (x-\mu_x)^2\, p(x)\, dx \qquad \text{or} \qquad \Sigma_i (x_i-\mu_x)^2\, p(x_i) \;/\; \Sigma_x (x-\mu_x)^2\, p(x)$$

which is the same as (G.25).

Notice that the variance could in theory vanish if $p(x) = \delta(x - x_1)$ so that the entire pdf is concentrated at a single value and then $\mu_X = x_1$:

$$\text{var}(X) = \int (x-\mu_x)^2\, p(x)\, dx = \int (x-\mu_x)^2\, \delta(x - x_1)\, dx = (x_1-\mu_x)^2 = (x_1- x_1)^2 = 0 = \sigma(X) \ .$$

But then X is not a random variable since it's value is precisely determined as x_1. A pdf of this form is not very interesting and normally one has $\sigma(X) > 0$.

Fact: $\text{var}(X) = \sigma^2(X) = E(X^2) - \{E(X)\}^2$ (G.30)

Proof: $E(X^2) - \{E(X)\}^2 = \int x^2 p(x)\, dx - \{\int x\, p(x)\, dx\}^2 = \int x^2 p(x)\, dx - \mu_x^2$

$$\text{var}(X) = \int (x-\mu_x)^2\, p(x)\, dx = \int x^2 p(x)\, dx - 2\mu_x \int x\, p(x)\, dx + \mu_x^2 \int p(x)\, dx$$

$$= \int x^2 p(x)\, dx - 2\mu_x \mu_x + \mu_x^2\, 1 = \int x^2 p(x)\, dx - \mu_x^2 \ . \qquad \textbf{QED}$$

Now that we have defined correlation in (G.26), we can see why being uncorrelated corresponds to a factoring pdf as was claimed earlier in (G.4a) or (G.4b). That is one item of the following fact:

Fact: The following statements are all equivalent for random variables X and Y : (G.31)

(a) X and Y are statistically independent
(b) X and Y are uncorrelated
(c) $p(x,y) = p(x)p(y)$
(d) $E(XY) = E(X)E(Y)$
(e) $\text{cov}(X,Y) = 0$
(f) $\text{corr}(X,Y) = 0$

Proof: We defined (a) in terms of (c). If we can show (f), then we have shown (b). So we need prove only (d),(e) and (f)

(d) $E(XY) = \int\int x\, y\, p(x,y)\, dx\, dy = \int\int x\, y\, p(x)p(y)\, dx\, dy = [\int x\, p(x)\, dx]\,[\int y\, p(y)\, dy] = E(X)E(Y)$.

(e) $\text{cov}(X,Y) \equiv E((X-\mu_x)(Y-\mu_y)) = \int\int (x-\mu_x)\,(y-\mu_y)\, p(x,y)\, dx\, dy = \int\int (x-\mu_x)\,(y-\mu_y)\, p(x)p(y)\, dx\, dy$

$= [\int (x-\mu_x)\, p(x)\, dx]\,[\int (y-\mu_y)\, p(y)\, dy] = [0][0] = 0$

where for example

$[\int (x-\mu_x)\, p(x)\, dx] = \int x\, p(x)\, dx - \mu_x \int p(x)\, dx = \mu_x - \mu_x * 1 = 0.$

(f) $\text{corr}(X,Y) \equiv \text{cov}(X,Y) / [\sigma(X)\,\sigma(Y)] = 0 / [\sigma(X)\,\sigma(Y)] = 0$ **QED**

Fact: If {X,Y} are uncorrelated, and if {Y,Z} are uncorrelated, {X,Z} may be correlated. (G.32)

Proof: Knowing that p(x,y) = p(x)p(y) and p(y,z) = p(y)p(z) tells us nothing about p(x,z). It is easy to think of trivial examples of this fact. Maybe X = Z so $\text{corr}(X,Z) = \text{var}(X)/ [\sigma(X)]^2 \neq 0$.

Fact: If {X,Y} and {Y,Z} and {X,Z} are all uncorrelated pairs, X Y and Z might not be statistically independent. (G.33)

Proof: Knowing about the uncorrelated pairs says nothing about p(x,y,z). That is to say, knowing that p(x,y) = p(x)p(y) and p(y,z) = p(y)p(z) does not imply that p(x,y,z) = p(x)p(y)p(z).

This concludes our brief review of probability theory, and we now continue in our examination of experiments associated with random variables.

(d) Ensemble Experiments

We consider now a new kind of "experiment". We acquire N dice which we shall refer to as an **ensemble** of dice. We assume they are all "loaded" *differently*, perhaps with implanted weights. For each die, we perform the single-roll experiment described above for which the outcome lies in the sample space S = {1,2,3,4,5,6}. After doing these N experiments, the N dice are left lying on the green felt of a craps table, each in its final experimental state. We then survey all N dice and take note of each one's face-up number, and from that data we construct a distribution. Since these dice are all slightly different, the distribution will likely differ from that shown earlier in Fig G.1. Perhaps we get this:

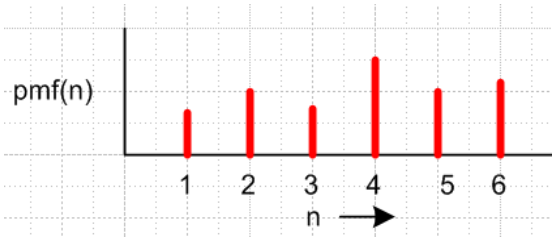


Fig G.8

As a computer algorithm, here is this new "ensemble experiment":

Acquire an ensemble of N dice (perhaps each is loaded differently)
For each die in the ensemble, carry out the single die experiment with outcome in S = {1,2,3,4,5,6}.
When the experiments are done, survey the resulting data and construct a distribution.

If the dice were identical, then this ensemble experiment would be the same as just sequentially rolling the same die N times and writing down the outcome of each roll. But the main point is that we assume the dice are *not* all the same.

Now we carry out an analogous ensemble experiment :

Acquire an ensemble of N babies (they are likely all different)
For each baby in the ensemble, carry out some experiment with outcome in some sample space S.
When the experiments are done, survey the resulting data and construct a distribution.

The experiment we have in mind is simply to let each baby grow up into an adult and then we treat some parameter of that adult as our random variable. Adults generally don't have face-up numbers, but they do have other "parameters" such as mass, height, and number of children. For each adult, the "experiment" was "growing up", analogous to rolling one die in our previous ensemble experiment. It is convenient to let nature do these experiments for us, so in practice we just assemble some group of adults into our ensemble, and then we are left with just the last item above :

When the experiments are done, survey the resulting data and construct a distribution.

Here for example we carry out three ensemble experiments with the same ensemble of people. Each ensemble experiment uses a different sample space: weight, height, number of children. and each deals with a different random variable: W, H and N. For each experiment we obtain a distribution as shown:

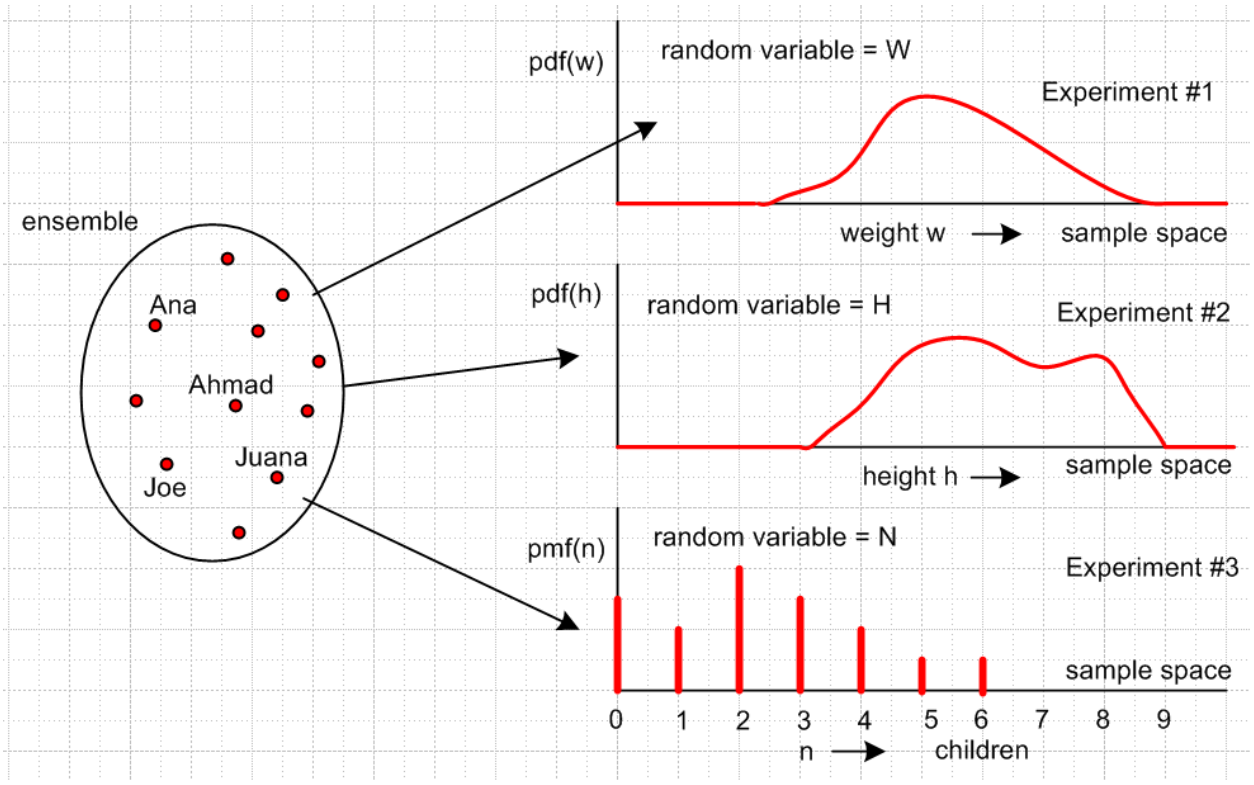


Fig G.9

In all three experiments, the outcomes are real numbers, so there is no need for any functions to translate from non-numerical outcomes to real numbers, as we required in for the {heads,tails} sample space. These experiments are all analyzing historical data, nor current "chance events" like rolling dice.

In the above scenario, sometimes the ensemble of people is called "the sample space" which is then a totally different meaning of that phrase, so we won't use it.

(e) Experiments rolling two dice at the same time

In the next set of experiments, we roll two differently weighted dice at the same time and examine particular outcomes (that is to say, we examine particular random variables, each taking real values in its numerical sample space). These two dice are not only weighted differently, but the embedded weights are magnets, which cause the two die to interact with each other during a roll. Here we get to apply various facts from the brief review of probability theory of section (c) above.

In **Experiment #1** we just look at the face up number of the X die and that number is x. The sample space as usual is S = {1,2,3,4,5,6} and there will be some $\text{pmf}_{\mathbf{X}}(x)$. Review: In this notation, the subscript indicates the random variable X, and the argument x is a value that random variable can take. Another notation is pmf(X = x), and the most compact notation is that of (G.5) which is just p(x) where one must remember what the random variable is. It is suggested by the letter used as argument: $p(x) = \text{pmf}_{\mathbf{X}}(x) = \text{pmf}(X = x)$.

In **Experiment #2** we do this for the Y die. It has the same sample space, but a different $p(y) = \text{pmf}_{\mathbf{Y}}(y)$.

In **Experiment #3** we look at z = x+y with its random variable Z = X + Y. The sum of two random variables is a random variable according to (G.16) or (G.21). The numerical sample space for this third experiment is {2,3..., 11,12}, since there are no other possible outcomes for the sum of two face-up numbers on two dice. According to (G.17a), the probability of Z taking value z in this sample space is given by $pmf_Z(z) = \Sigma^z{}_{x,y}\ p_{XY}(x,y)$, or in compact notation, $p(z) = \Sigma^z{}_{x,y}\ p(x,y)$, where the sum includes only those x,y values such that x + y = the z inside p(z) on the left. If the two die were differently weighted but contained no magnets, we would have $p(z) = \Sigma^z{}_{x,y}\ p_X(x)p_Y(y)$ since the two dice are "independent" or "uncorrelated" as in (G.31). If the two dice were identical, then $p(z) = \Sigma^z{}_{x,y}\ p_X(x)p_X(y)$ where the two pmf's are the same. If the two dice are "fair dice" (unloaded), then $pmf_X(x) = p_X(x) = 1/6$ for any x and then we find that $p(z) = \Sigma^z{}_{x,y}\ (1/6)(1/6) = (1/36)\ \Sigma^z{}_{x,y}\ 1$. We then have a classic dice problem where we can enumerate the terms in the sum $\Sigma^z{}_{x,y}\ 1$ as follows

z	$\Sigma^z{}_{x,y}\ 1$
2	1,1 = 1
3	1,2 + 2,1 = 2
4	1,3 + 3,1 + 2,2 = 3
...	
7	1,6 + 6,1 + 5,2 + 2,5 + 3,4 + 4,3 = 6
...	
12	6,6 = 1

Below, $pmf_X(x)$ is for Experiment #1, and $pmf_Z(z)$ for Experiment #3. In each experiment, we roll a pair of identical fair dice many times and obtain these distributions.

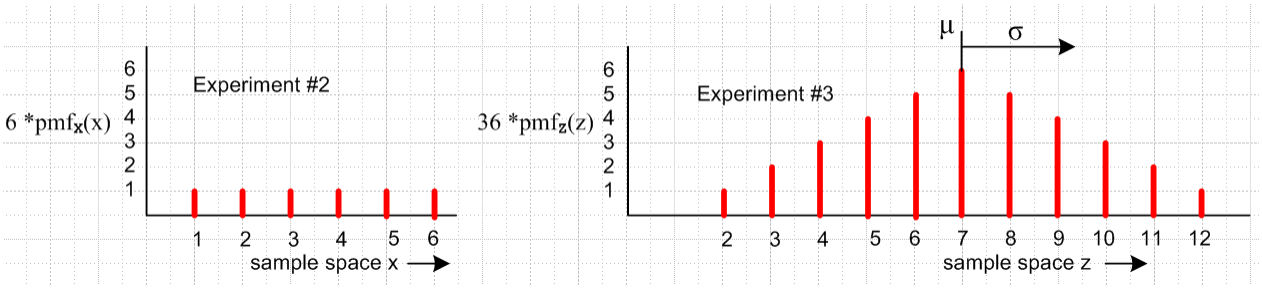


Fig G.10

We now do a few quick hand calculations to exercise some of our basic probability facts. Notice ahead of time that

$\Sigma_x\ 1 = 6$
$\Sigma_x\ x = (1 + 2 + 3 + 4 + 5 + 6) = 21$
$\Sigma_x\ x^2 = (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = 91$.

The expected value of Z for Experiment #3 is given by (G.23),

$E(Z) = E(X+Y) = \Sigma_{x,y}\ (x+y)\ p(x,y)$.

In the case of Fig G.10 where $p(x,y) = [p(x)]^2 = (1/6)^2 = (1/36)$ we have

$$E(Z) = (1/36)\,\Sigma_{x,y}(x+y) = (1/36)\,[\,(\Sigma_x x)(\Sigma_y 1) + (\Sigma_x 1)(\Sigma_y y)\,] = (1/18)\,(\Sigma_x x)(\Sigma_y 1)$$

$$= (1/18)\,(21)\,(6) = (1/3)(21) = 7 = \mu_Z$$

in agreement with the symmetric distribution on the right in Fig G.10. Next we compute,

$$E(Z^2) = \Sigma_{x,y}\,(x+y)^2\,p(x,y) = (1/36)\,\Sigma_{x,y}\,[\,x^2 + y^2 + 2xy\,]$$

$$= (1/36)\,[\,2\,(\Sigma_x x^2)(\Sigma_y 1) + 2\,(\Sigma_x x)^2\,] = (1/18)\,[91^*\,6 + 21^2] = 329/6 = 54.83.$$

The variance and standard deviation are then given by (G.30).

$$\mathrm{var}(Z) = E(Z^2) - \{E(Z)\}^2 = 329/6 - 49 = 35/6 = 5.833$$
$$\sigma(Z) = \sqrt{35/6} = 2.415 \qquad \text{(G.34)}$$

This last result seems in line with a visual inspection of Fig G.10.

Finally in **Experiment #4** we consider the outcome $z = x^*y = xy$. Things are similar to the above, but now the notation $\Sigma^z_{x,y}$ means the sum over x,y values such that the *product* of x and y is z. In the context of (G.23), we now have $f = xy$ whereas in experiment #3 we had $f = x+y$. We could of course study the situation with any reasonable function f. For $f(x,y) = xy$ we can write:

$\mathrm{pmf}_Z(z) = \Sigma^z_{x,y}\ \mathrm{pmf}_{XY}(x,y)$ dice are different and weighted with magnets

$\mathrm{pmf}_Z(z) = \Sigma^z_{x,y}\ \mathrm{pmf}_X(x) * \mathrm{pmf}_Y(y)$ dice are different but no magnets

$\mathrm{pmf}_Z(z) = \Sigma^z_{x,y}\ \mathrm{pmf}_X(x) * \mathrm{pmf}_X(y)$ dice are identical

$\mathrm{pmf}_Z(z) = \Sigma^z_{x,y}\ (1/6) * (1/6) = (1/36)\,\Sigma^z_{x,y}\,1$ dice are identical and "fair"

For this Experiment # 4 the sample space is all possible products S = {1, 2, 3.....36} with certain values missing. For "identical and fair" we make our list using our new meaning of $\Sigma^z_{x,y}$:

z	$\Sigma^z_{x,y}\,1$
1	1,1 = 1
2	1,2 + 2,1 = 2
3	1,3 + 3,1 = 2
4	1,4 + 4,1 + 2,2 = 3
5	1,5 + 5,1 = 2
6	1,6 + 6,1 + 2,3 + 3,2 = 4
7	= 0

8 2,4 + 4,2 = 2

...

36 6,6 = 1

We ask Maple to create $pmf_Z(z)$ for this Experiment #4:

```
for z from 1 to 36 do p(z) := 0 od:  # clear counters
for z from 1 to 36 do
  for x from 1 to 6 do
    for y from 1 to 6 do
      if z = x*y then p(z) := p(z) +1 fi;
    od;
  od;
od;
print(seq(p(z),z=1..36));
```

1, 2, 2, 3, 2, 4, 0, 2, 1, 2, 0, 4, 0, 0, 2, 1, 0, 2, 0, 2, 0, 0, 0, 2, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 1

```
with(stats):with(stats[statplots]):
data1 := [seq(Weight(i..i+0.3,0.3*p(i)),i = 1..36)]:
histogram(data1, color=red,tickmarks = [10,5], view = [0..37,0..5]);
```

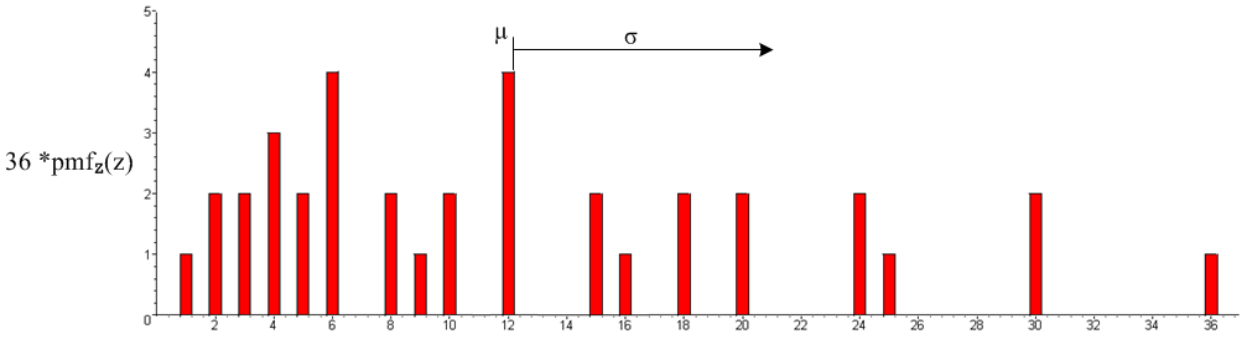


Fig G.11

We then repeat the calculations done for Experiment #3:

$$E(Z) = E(XY) = \Sigma_{x,y}\,(xy)\,p(x,y) = (1/36)\,\Sigma_{x,y}\,(xy) = (1/36)[\,(\Sigma_x x)^2 = 21^2/36 = 49/4 = 12.25 = \mu_Z$$

$$E(Z^2) = \Sigma_{x,y}\,(xy)^2\,p(x,y) = (1/36)\,\Sigma_{x,y}\,x^2y^2 = (1/36)\,[\,(\Sigma_x x^2)^2] = (1/36)(91)^2 = 8281/36 = 230.03$$

$$var(Z) = E(Z^2) - [E(Z)]^2 = 8281/36 - (49/4)^2 = 11515/144 = 79.97$$

$\sigma(Z) = 8.942$ // standard deviation (G.35)

(f) Experimental determination of discrete distribution functions

In practice, one has some list of experimental results (outcomes converted to real numbers if not already real numbers) and one puts the results into "bins" to obtain a distribution. Here we want to be more explicit about what this means.

For a **single variable** x which is the face-up value of a die, here is how we would compute p(x) after collecting data rolling the die J times :

```
restart;   J := 20: roll := rand(1..6):
x := seq( roll(),j=1..J);   # roll die J times
                    x := 4, 3, 4, 6, 5, 3, 6, 3, 2, 2, 2, 4, 4, 3, 3, 2, 1, 4, 4, 6
for n from 1 to 6 do bin[n] := 0 od:   # clear bins

for j from 1 to J do
  for n from 1 to 6 do
    if x[j] = n then bin[n] := bin[n] + 1 fi;
  od;
od;
for n from 1 to 6 do p[n] := bin[n]/J; od:
for n from 1 to 6 do printf("  p[%d] = %.3f ,",n,p[n]);od;

p[1] = .050 ,  p[2] = .200 ,  p[3] = .250 ,  p[4] = .300 ,  p[5] = .050 ,  p[6] = .150
```

which we could plot as a bar chart of the type shown in Fig G.1.

Symbolically we might write this as

$$p(n) = (1/J)\,\Sigma_{j=1}^{J}\,(\,x_j = n\,) \qquad n = 1,2,3,4,5,6\ . \tag{G.36}$$

The idea is that for each n, we count the number of times that $(\,x_j = n\,)$ is true.

For **two variables**, things are a bit more complicated. Now we have a set of J pairs $\{x_i,y_i\}$ as our collection of data, so that

$$p(n,m) = (1/J)\,\Sigma_{j=1}^{J}\,(\,[x_i,y_i] = [n,m]\,) \qquad n,m = 1,2,3,4,5,6 \tag{G.37}$$

where [...] refers to an ordered sequence. Another way to write this would be

$$p(n,m) = (1/J)\,\Sigma_{j=1}^{J}\,[(n = x_i)\ \text{and}\ (m = x_j)] \qquad n,m = 1,2,3,4,5,6 \tag{G.38}$$

Here is a sample Maple program to carry this out

```
> restart;   J := 20: roll := rand(1..6):
> x := seq( [roll(),roll()],j=1..J);   # roll two dice J times
x := [4, 3], [4, 6], [5, 3], [6, 3], [2, 2], [2, 4], [4, 3], [3, 2], [1, 4], [4, 6], [1, 1], [1, 2], [4, 2], [1, 3], [6, 3],
  [6, 4], [3, 6], [1, 6], [3, 1], [4, 6]
> for n from 1 to 6 do
    for m from 1 to 6 do
      cnt := 0;
      for j from 1 to J do
        if (x[j,1] = n) and (x[j,2] = m) then cnt := cnt + 1; fi
      od;
      p[n,m]:= cnt/J;
    od;
  od;
  for n from 1 to 6 do
     for m from 1 to 6 do printf(" p[%d,%d]=%.3f ,",n,m,p[n,m]);od;
     printf("\n");
  od;
 p[1,1]=.050 , p[1,2]=.050 , p[1,3]=.050 , p[1,4]=.050 , p[1,5]=0.000 , p[1,6]=.050 ,
 p[2,1]=0.000 , p[2,2]=.050 , p[2,3]=0.000 , p[2,4]=.050 , p[2,5]=0.000 , p[2,6]=0.000 ,
 p[3,1]=.050 , p[3,2]=.050 , p[3,3]=0.000 , p[3,4]=0.000 , p[3,5]=0.000 , p[3,6]=.050 ,
 p[4,1]=0.000 , p[4,2]=.050 , p[4,3]=.100 , p[4,4]=0.000 , p[4,5]=0.000 , p[4,6]=.150 ,
 p[5,1]=0.000 , p[5,2]=0.000 , p[5,3]=.050 , p[5,4]=0.000 , p[5,5]=0.000 , p[5,6]=0.000 ,
 p[6,1]=0.000 , p[6,2]=0.000 , p[6,3]=.100 , p[6,4]=.050 , p[6,5]=0.000 , p[6,6]=0.000 ,
```

The result can be visualized as a 3D bar chart

```
with(stats):with(stats[statplots]):
for n from 1 to 6 do
  data[n] := [seq(Weight(m..m+0.3,0.3*p[n,m]),m = 1..6)]:
od:
histogram(seq(data[n],n=1..6),color = red);
```

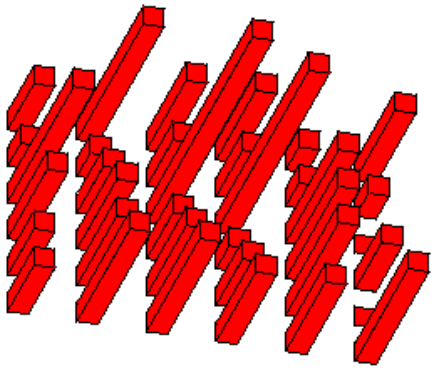


Fig G.12

A distribution function like p(a,b,c,d,e) is a 6D bar chart, not easy to display.

(g) Experiments with Sequences of Pulse Train Amplitudes

Our pulse trains have a set of amplitudes y_n which form a sequence. If the sequence has N elements, then a sequence can be written $[y_1, y_2.....y_N]$. When many pulse trains are generated in some line code, one finds that the values of y_n in position n of the sequence can vary, and that there is some probability distribution associated with the parameter y_n. Thus, the position n in the pulse train or sequence is associated with a random variable we must call Y_n which takes values y_n.

Here is our first experiment. We have an Apparatus which generates sequences of length N according to some set of rules implemented within the Apparatus (perhaps these are the rules for generating AMI linecode sequences, and the AMI encoder inputs random input streams). Every sequence it generates is "legal" according to its rules. Earlier we discussed an ensemble experiment in which M dice were rolled one at a time and were left on the craps table for study and from what we saw on the table, we were able to construct a distribution function. The ensemble was M dice. In our current **ensemble experiment**, we let the Apparatus crank away and generate an ensemble of M sequences each of length N, and these M sequences are left sitting on the same craps table for us to inspect. This is our "preliminary" ensemble of sequences.

The notation Y_n leads to some confusion. When we had a random variable H which took values h, we might have enumerated the set of h values of the sample space as h_i. To maintain clarity, if we have random variable Y_n which takes values y_n, we should enumerate those values as $(y_n)_i$. Here, y_n is the name of a parameter, just as h was the name of a parameter. Thus, to write down a specific sequence "i" we really should say

$$[(y_1)_i, (y_2)_i.....(y_N)_i] \quad = \text{sequence "i"}$$
$$\text{or} \qquad (G.39)$$
$$[y_1^{(i)}, y_2^{(i)}..... y_2^{(i)}] \quad = \text{sequence "i"}$$

where $(y_2)_i = y_2^{(i)}$ is some specific symbol value like "5". We used the latter notation when talking about an ensemble of pulse trains in Section 35 and in Appendix D.

The Useful Ensemble. To construct a "useful" ensemble of pulse trains, we first imagine generating some number M of pulse trains of length N from our Apparatus as described above. We then take this ensemble of M pulse trains and considerably increase the size of the ensemble by including in it all cyclic permutations (rotations) of the M pulse trains, so now the enlarged ensemble contains I = M*N pulse trains.

Depending on the rules of the Apparatus generating the pulse trains, these rotated pulse trains might be "illegal" at the boundary where the pulse train wraps around on itself. For example with N = 5 we obtain this sequence plus its rotations,

```
A B C D E
B C D E|A
C D E|A B
D E|A B C
E|A B C D   .
```
(G.40a)

Here the seam is marked by a bar | . If the seam were "illegal", we could probably repair the seam by replacing A with some other symbol Q to get a legal sequence of symbols,

```
A B C D E
B C D E Q
C D E Q B
D E Q B C
E Q B C D
```
(G.40b)

Perhaps several symbols near the seam would have to be "repaired" in this manner, but we shall assume just one as shown above. Notice that each column has at most one bad symbol Q, and that is 1/N of the symbols in the column. If N = 100, then at most 1% of the symbols in a column are wrong after adjustment. If a 2-symbol repair were needed, then at most 2% in a column would be wrong.

We imagine doing this for all M of our initial pulse trains to get our final set of I = M*N pulse trains. In this "useful ensemble", in each vertical column, 1/N of the symbols might be wrong. We of course have in mind that N is very large, but our examples will always have N very small, such as N = 5.

Now we associate each column of our ensemble of sequences with a random variable Y_n. For (G.40),

$$Y_1\ Y_2\ Y_3\ Y_4\ Y_5 \tag{G.41}$$

Fact: The expected value $\langle y_n \rangle \equiv E(Y_n)$ is independent of n. (G.42)

Proof: Each column has at most 1 bad symbol like Q for each initial ensemble sequence. So if there are now I = M*N pulse trains in the ensemble, then in each column there will be at most M bad symbols. When the expected value is computed by summing the elements of the column n and dividing by I,

$$E(Y_n) = (1/I) \sum_{i=1}^{I} y_n{}^{(i)} = (1/I) \sum_{i=1}^{I} (y_n)_i \qquad \text{// two notations for the same thing}$$

one finds that the sums for different n are almost identical with an error on the order of 1/N. In our example the first column sums to A+B+C+D+E whereas the other columns are Q+B+C+D+E . Of course the exact error depends on what symbol Q was used to do the repair and what the palette of symbol values is, and so on, but the general order is 1/N. As N is made large, this error approaches 0 and then all the columns have the same expected value of their Y_n. **QED**

Now recall the probability distribution for our N=5 ensemble, written in two ways as in (G.5),

$$p(a,b,c,d,e) \equiv p(Y_1=a, Y_2= b, Y_3= c, Y_4= d, Y_5= e)\ . \tag{G.43}$$

This is the height of a specific "bar" in the 6D bar chart that is $p(y_1,y_2,y_3,y_4,y_5)$, see Fig G.12

Fact: For our useful ensemble, the probability distribution p(a,b,c ...) has cyclic symmetry. (G.44)

Example: p(a,b,c,d,e) = p(e,a,b,c,d) = p(d,e,a,b,c) = p(c,d,e,a,b) = p(b,c,d,e,a)

Proof: This might seem obvious, but a formal proof is warranted. We show this for the case N = 5; the general case should then be obvious. Using the notation introduced in (G.37) we can imagine evaluating p(a,b,c,d,e) this way, where [....] indicates an ordered sequence of numbers,

$$p(a,b,c,d,e) = (1/I)\,\Sigma_{i=1}^{I}\,(\,[\,a,b,c,d,e] = [\,(y_1)_i, (y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i\,]\,)\,.$$

Since our ensemble contains all cyclic rotations of $[\,(y_1)_i, (y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i\,]$, we don't change the above sum by replacing this [...] with $[\,(y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i, (y_1)_i\,]$, we merely re-order the sum. In our example above, we can rewrite the 5 sequences on the left in this fancier notation,

		$(y_1)_i$	$(y_2)_i$	$(y_3)_i$	$(y_4)_i$	$(y_5)_i$
	$Y_1Y_2Y_3Y_4Y_5$					
α	A B C D E	$(y_1)_1$	$(y_2)_1$	$(y_3)_1$	$(y_4)_1$	$(y_5)_1$
β	B C D E A	$(y_1)_2$	$(y_2)_2$	$(y_3)_2$	$(y_4)_2$	$(y_5)_2$
δ	C D E A B	$(y_1)_3$	$(y_2)_3$	$(y_3)_3$	$(y_4)_3$	$(y_5)_3$
γ	D E A B C	$(y_1)_4$	$(y_2)_4$	$(y_3)_4$	$(y_4)_4$	$(y_5)_4$
ε	E A B C D	$(y_1)_5$	$(y_2)_5$	$(y_3)_5$	$(y_4)_5$	$(y_5)_5$

We now rotate the above columns one column to the left to get

		$(y_2)_i$	$(y_3)_i$	$(y_4)_i$	$(y_5)_i$	$(y_1)_i$
	$Y_1Y_2Y_3Y_4Y_5$					
β	B C D E A	$(y_2)_1$	$(y_3)_1$	$(y_4)_1$	$(y_5)_1$	$(y_1)_1$
δ	C D E A B	$(y_2)_2$	$(y_3)_2$	$(y_4)_2$	$(y_5)_2$	$(y_1)_2$
γ	D E A B C	$(y_2)_3$	$(y_3)_3$	$(y_4)_3$	$(y_5)_3$	$(y_1)_3$
ε	E A B C D	$(y_2)_4$	$(y_3)_4$	$(y_4)_4$	$(y_5)_4$	$(y_1)_4$
α	A B C D E	$(y_2)_5$	$(y_3)_5$	$(y_4)_5$	$(y_5)_5$	$(y_1)_5$

As it most easily seen on the left, this rotation by one column has just reordered the ensemble sequences from the original order α,β,γ,δ,ε to the new order β,γ,δ,ε,α. The same set of samples is included in the sum above regardless of how they are ordered.

Consider then our starting expression,

$$p(a,b,c,d,e) = (1/I)\,\Sigma_{i=1}^{I}\,(\,[\,a,b,c,d,e] = [\,(y_1)_i, (y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i]\,)\,.$$

Note that c is a specific number, as is $(y_5)_3$. Neither of these is a variable. Now suppose the ensemble has the property just noted above, that we can reorder the sum this way,

$$p(a,b,c,d,e) = (1/I)\,\Sigma_{i=1}^{I}\,(\,[\,a,b,c,d,e] = [(y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i, (y_1)_i]\,)\,.$$

The equality of the two sequences of numbers shown in the truth test [...] = [...] does not change if we alter each sequence in the exact same way, so rewrite again as

$$p(a,b,c,d,e) = (1/I)\,\Sigma_{i=1}^{I}\,(\,[\,a,b,c,d,e] = [(y_1)_i, (y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i,]\,)\,.$$

But the expression on the right is the definition of p(e,a,b,c,d). Thus we have shown that

$$p(a,b,c,d,e) = p(e,a,b,c,d) .$$

We can then show in the same way that p(e,a,b,c,d) = p(d,e,a,b,c) and so on, so in the end all cyclic permutations (rotations) of p(a,b,c,d,e) are equal for our useful ensemble. **QED**

Comment: In the special case that all the sequence random variables Y_n are statistically independent, we can write

$$p(y_1,y_2,y_3,y_4,y_5) = p(y_1)\, p(y_2)\, p(y_3)\, p(y_4)\, p(y_5) .$$

In this special case it is obvious that $p(y_1,y_2,y_3,y_4,y_5)$ is symmetric under cyclic permutations, as it is under *any* permutation. A point to note is that we are *not* assuming statistical independence of the Y_n and that therefore there can exist "correlation" between pairs of random variables in our sequences (pulse train amplitudes). Nevertheless, in this general case $p(y_1,y_2,y_3,...)$ our useful ensemble has cyclic symmetry, and this ensemble is a realizable ensemble from our experiments for large N with only some very small error. As $N\to\infty$, that error goes to 0.

Fact: The probability distribution $p(Y_n = x)$ is independent of n. (G.45)

Proof by Example: When we write $p(Y_n = x) = p(x)$, then p(x) is deceptively independent of n, but we must show this in the true full notation where n dependence is not concealed. So consider using our N= 5 example, where we use (G.7) to obtain $p(Y_3 = x)$ by summing over all the "other variables" of the full 5th order joint probability distribution $p(y_1,y_2,y_3,y_4,y_5)$,

$p(Y_3 = x)$

$= \Sigma_{y1,y2,y4,y5}\, p(Y_1=y_1,Y_2= y_2,Y_3 = x,Y_4= y_4,Y_5= y_5)$ // definition of $p(Y_3 = x)$

$= \Sigma_{y2,y3,y5,y1}\, p(Y_1=y_2,Y_2= y_3,Y_3= x,Y_4= y_5,Y_5= y_1)$ // rename dummy summation variables

$= \Sigma_{y1,y2,y3,y5}\, p(Y_1=y_2,Y_2= y_3,Y_3= x,Y_4= y_5,Y_5= y_1)$ // trivially reorder sums on Σ

$= \Sigma_{y1,y2,y3,y5}\, p(Y_1=y_1,Y_2= y_2,Y_3= y_3,Y_4= x,Y_5= y_5)$ // use cyclic property of p function

$= p(Y_4 = x)$ // definition of $p(Y_4 = x)$

Thus we have shown that $p(Y_3 = x) = p(Y_4 = x)$ and either by iterating or using other cyclic permutations in the middle step above we conclude that all the $p(Y_n = x)$ are equal and thus $p(Y_n = x)$ is independent of n. As usual, this is for our "useful distribution".

Corollary 1: $\langle y_n \rangle \equiv E(Y_n)$ is independent of n, since $E(Y_n) = (1/I)\Sigma_x\ x\ p(Y_n = x)$ and $p(Y_n = x)$ is independent of n. Thus we have an alternate proof of (G.42) above. (G.46)

Corollary 2: $\langle y_n^2 \rangle \equiv E(Y_n^2)$ is independent of n, since $E(Y_n^2) = (1/I)\Sigma_x\ x^2\ p(Y_n = x)$ and $p(Y_n = x)$ is independent of n. (G.47)

Corollary 3: $\langle f(y_n) \rangle \equiv E(f(Y_n))$ is independent of n, since $E(f(Y_n)) = (1/I)\Sigma_x\ f(x)\ p(Y_n = x)$, etc. See (G.22). (G.48)

Fact: The probability distribution $p(Y_n = x, Y_m = z)$ depends only on m-n mod N. (G.49)

Proof by Example: Using the same example as above, and again applying (G.7), we find that

$p(Y_2 = x, Y_4 = z)$ // m-n = 4-2 = 2

$= \Sigma_{y1,y3,y5}\ p(Y_1=y_1,Y_2= x,Y_3 = y_3,Y_4= z,Y_5= y_5)$ // definition of $p(Y_3 = x)$

$= \Sigma_{y3,y5,y1}\ p(Y_1=y_3,Y_2= x,Y_3 = y_5,Y_4= z,Y_5= y_1)$ // rename dummy sum indices

$= \Sigma_{y1,y3,y5}\ p(Y_1=y_3,Y_2= x,Y_3 = y_5,Y_4= z,Y_5= y_1)$ // trivially reorder sums on Σ

$\qquad\qquad\qquad\qquad\quad \backslash \quad \backslash \quad \backslash \quad \backslash$

$= \Sigma_{y1,y3,y5}\ p(Y_1=y_1,Y_2= y_3,Y_3 = x,Y_4= y_5,Y_5= z)$ // use cyclic property of p function

$= p(Y_3 = x, Y_5 = z)$ // m - n = 5-3 = 2

In general, one has

$$p(Y_n = x, Y_m = z) = p(Y_{n+k} = x, Y_{m+k} = z)$$

again for our useful ensemble.

Corollary: $\langle y_n y_m \rangle \equiv E(Y_n Y_m)$ for $m \neq n$ depends only on m-n mod N. (G.50)

$$E(Y_n Y_m) = (1/I)\ \Sigma_{x,z}\ (xz)\ p(Y_n = x, Y_m = z) = (1/I)\ \Sigma_{x,z}\ (xz)\ p(Y_{n+k} = x, Y_{m+k} = z)$$

$$= E(Y_{n+k} Y_{m+k})\ .$$

Conclusion: When we experimentally create a huge "statistical ensemble" of pulse trains which have a very long length N, this ensemble is statistically very similar to our "useful ensemble" described above. Every possible legal sequence of amplitudes will appear many times in the ensemble, and all the cyclic versions of each sequence will appear (ignoring the tiny seam corrections) on an equal footing. As N gets very large, the notion of cyclic permutation is more associated with a *shift* of the pulse train, when we observe some finite window near the center of the pulse train. For $N = \infty$, the shift notion is all that remains, and there are no seams to worry about. For such pulse trains, we have established these important facts:

$\langle a_n \rangle$ = independent of n // general

$\langle a_n^2 \rangle$ = independent of n // general

$\langle a_n a_{n+k} \rangle$ = independent of n, dependent in general on k // general

$\langle a_n a_m \rangle = \langle a_n \rangle \langle a_m \rangle = \langle a_n \rangle^2$ = independent of n // uncorrelated only (G.51)

_____ summary _____

Appendix G: Random Variables, Probability Theory and Pulse Train Amplitudes

Subsection (a) provides an opening definition of a "random variable" as being a parameter associated with a probability distribution. The terms experiment, outcome, sample space and event are defined and two simple experiments considered: rolling a die, and spinning a spinner. The corresponding pdf and pmf distributions are discussed.

Subsection (b) refines the definition of "random variable" as a mapping from the sample space to the real numbers, and the coin toss experiment is used as an example. Drawings show examples of discrete and continuous samples spaces and the random variable mappings.

Subsection (c) contains a brief review of basic probability theory with an emphasis on notation. The notions of statistical independence, correlation and normalization are discussed. Some Math Lemmas are presented relating the ordering of terms in a sum $\Sigma_{x,y}$ in the context of some function z = f(x,y) by first summing over a surface of constant z, then over different z values, so $\Sigma_{x,y} = \Sigma_z \Sigma_{surface(z)}$, These lemmas are needed in the following discussion of random variables defined as functions of other random variables. After this, the usual suspects of probability theory are rolled out for inspection: expected values, mean, covariance, variance, standard deviation and correlation.

Subsection (d) treats "ensemble experiments" such as rolling N loaded dice, and a transition is then made to the analysis of random variables associated with an ensemble of people.

Subsection (e) treats the rolling of two weighted, magnetically interacting dice, after which special cases are considered.

Subsection (f) gives examples of the computation of probability mass density functions from experimental data.

Subsection (g) applies the notions of random variables and probability theory to the sequences which are the amplitudes of pulse trains. A special "useful ensemble" is constructed for which the full probability density function is cyclic in its arguments, and it is shown how this ensemble matches a real world ensemble for long sequences. It is shown that $\langle y_m \rangle$ is independent of m and $\langle y_m y_{m+k} \rangle$ can only depend on k.

_____ compact overview ----------

Appendix D: How α and β are related to μ and σ (6 p)

Appendix D reviews some basic probability theory and relates the α and β parameters of the Section 35 uncorrelated spectral power density formula to statistical properties μ and σ of pulse train amplitudes. The probability theory part is significantly expanded in Appendix G.

Subsection (a) discusses random variables X and Y and their expected values like $\mu_x = E(X) = \langle x \rangle$ and E(XY). Certain statistical measures are defined and related to each other. The concept of X and Y being independent (uncorrelated) random variables is explored.

Subsection (b) translates the results of subsection (a) into the context where $X = Y_m$ and $Y = Y_n$ where Y_m is the random variable associated with position m in a pulse train, and whose values are the pulse amplitudes y_m .

Subsection (c) shows that for an uncorrelated pulse train the following facts are true for the coefficients α and β used in Section 35 on spectral power density,

$$\alpha = \mu^2 \qquad \beta = \mu^2 + \sigma^2 \qquad (\beta-\alpha) = \sigma^2 \qquad \mu \equiv \langle y_m \rangle \qquad \sigma^2 = \langle y_m^2 \rangle - \langle y_m \rangle^2$$

These relations provide a useful interpretation of the parameters α and β in terms of the statistics of the pulse train amplitudes and allow comparison of our results to those of other sources.