Random Variables. . . in a Nutshell

AT Patera, M Yano[1]

September 22, 2014

1 Preamble

It is often the case in engineering analysis that the outcome of a (random) experiment is a numerical value: a displacement or a velocity; a Young's modulus or thermal conductivity; or perhaps the yield of a manufacturing process. In such cases we denote our experiment a *random variable*. In this nutshell we develop the theory of random variables from definition and characterization to simulation and finally estimation.

In this nutshell:

We describe univariate and bivariate discrete random variables: probability mass functions (joint, marginal, conditional); independence; transformations of random variables.

We describe continuous random variables: probably density function; cumulative distribution function; quantiles.

We define expectation generally, and the mean, variance, and standard deviation in particular. We provide a frequentist interpretation of mean. We connect mean, variance, and probability through Chebyshev's Inequality and (briefly) the Central Limit Theorem.

We summarize the properties and relevance of several ubiquitous probability mass and density functions: univariate and bivariate discrete uniform; Bernoulli; binomial; univariate and bivariate continuous uniform; univariate normal.

We introduce pseudo-random variates: generation; transformation; application to hypothesis testing and Monte Carlo simulation.

We define a random sample of "i.i.d." random variables and the associated sample mean. We derive the properties of the sample mean, in particular the mean and variance, relevant to parameter estimation.

We present a procedure for estimation of the Bernoulli parameter: sample-mean estimate; confidence level and confidence interval; convergence with sample size; considerations for rare events.

[1]We thank Ms Debra Blanchard for the preparation of the figures.

Several related concepts, including estimation of parameters for a normal population, are reserved for a subsequent nutshell on regression.

Prerequisites: *In a Nutshell. . . Introduction to Probability and Statistics*; univariate and multivariate calculus; elementary combinatorics.

2 Discrete Random Variables

2.1 Probability Mass Function

2.1.1 Univariate Case

We shall denote our random variable by X. A random variable is a particular kind of random experiment for which the outcomes are *numerical values*. We can not predict for any particular experiment the outcome, however we can describe the frequency of outcomes in the limit of many experiments: outcome probabilities.

The sample space — the set of all possible outcomes of an experiment — is given by $\{x_1^o, x_2^o, \ldots, x_L^o\}$: each experiment yields x_ℓ^o for some $\ell, 1 \leq \ell \leq L$. Note that the outcomes are *numerical* values. We denote the corresponding outcome probabilities as $p_\ell = P(x_\ell^o), 1 \leq \ell \leq L$. The outcome probabilities must satisfy $0 \leq p_\ell \leq 1, 1 \leq \ell \leq L$, and

$$\sum_{\ell=1}^{L} p_\ell = 1 \; . \tag{1}$$

We recall the frequentist interpretation of the outcome probabilities: we perform an infinite number of experiments; we calculate, for $\ell = 1, \ldots, L$, the cumulative frequency function $\widetilde{\varphi}_k(x_\ell^o)$, the fraction of the first k experiments for which the experiment yields outcome x_ℓ^o; we identify, for $\ell = 1, \ldots, L$, p_ℓ as the limit of $\widetilde{\varphi}_k(x_\ell^o)$ as $k \to \infty$.

We can summarize the relationship between outcomes and outcome probabilities in a *probability mass function*, f_X:

$$f_X(x_\ell^o) = p_\ell, \quad 1 \leq \ell \leq L \; . \tag{2}$$

The input is an outcome (in our sample space) — a numerical value — and the output is the corresponding probability. Note that f_X is defined only for the L arguments $x_\ell^o, 1 \leq \ell \leq L$.

It follows from our conditions on the $p_\ell, 1 \leq \ell \leq L$, that $0 \leq f_X(x_\ell^o) \leq 1, 1 \leq \ell \leq L$, and furthermore

$$\sum_{\ell=1}^{L} f_X(x_\ell^o) = 1 \; . \tag{3}$$

We may think of a probability mass function as a distribution of a total mass of unity amongst L point particles at locations $x_\ell^o, 1 \leq \ell \leq L$.

The subscript X of the probability mass function indicates the random variable to which the probability mass function is associated: we say that X is distributed according to f_X. As always, we must distinguish between our random variable X — a *procedure* — and a realization — application of the procedure to a particular instance to yield an outcome in our sample space, $\{x_i^o, 1 \leq i \leq L\}$. The outcome of a realization, x, also often referred to as simply the realization, is denoted a *random variate*; note x is a real number. We describe a realization as $X \to x$.

As an example of a probability mass function we consider the (discrete) uniform probability mass function,

$$f_X^{\text{unif};L}(\frac{\ell}{L}) = \frac{1}{L}\ , \quad 1 \leq \ell \leq L\ : \tag{4}$$

we distribute our mass (probability) uniformly — $p_\ell = 1/L, 1 \leq \ell \leq L$ — over L uniformly placed point masses (outcomes), $x_\ell^o = \ell/L, 1 \leq \ell \leq L$. Note that indeed all outcome probabilities are non-negative and less than unity (L is a positive integer), and furthermore the sum of the outcome probabilities is $L(1/L) = 1$, as required. The superscript to f indicates the particular probability mass function of interest and the parameter value L.

There are many phenomena which might be plausibly approximated (from mechanistic considerations) by the uniform density. We provide a classical example: the rolling of a single die. The experiment is represented by a random variable X which records the number of dots on the face which lands "up"; the sample space of all possible outcomes is thus given by $\{x_\ell^o = \ell, 1 \leq \ell \leq 6\}$; we posit, from symmetry arguments, that X is distributed according to a uniform probability mass function, $X \sim f_X^{\text{unif};6}$.

As a second example of a probability mass function, we consider the Bernoulli probability mass function,

$$f_X^{\text{Bernoulli};\theta}(x) = \begin{cases} 1-\theta & \text{if } x = x_1^o \equiv 0 \quad (= p_1) \\ \theta & \text{if } x = x_2^o \equiv 1 \quad (= p_2) \end{cases}\ , \tag{5}$$

where θ is a real number, $0 \leq \theta \leq 1$: we distribute our mass (probability) — $p_1 = 1 - \theta, p_2 = \theta$ — over $L = 2$ point masses (outcomes) placed at $x_1^o \equiv 0, x_2^o \equiv 1$, respectively. Note that indeed all outcome probabilities are non-negative, thanks to our assumptions on θ, and furthermore the sum of the outcome probabilities is $1 - \theta + \theta = 1$, as required. The superscript to f indicates the particular probability mass function of interest and the parameter value θ.

The Bernoulli probability mass function may appear rather simplistic but in fact it admits an interpretation with wide applicability: we may interpret the two outcomes, $x_1^o = 0$ and $x_2^o = 1$, as "indicator" functions, in which 0 encodes False (or Off) and 1 encodes True (or On). The choice for sample space $\{0, 1\}$ — rather than any other two values — creates a built-in number function, or frequency function, which is very convenient in practice. We provide a classical example: the flipping of a coin. The experiment is represented by a random variable X which records the face which lands "up"; the sample space of all possible outcomes is given by $\{x_1^o \equiv 0 \text{ Tail}, x_2^o \equiv 1 \text{ Head}\}$; we posit that X is distributed as

$f_X^{\text{Bernoulli};1/2}$ — equal likelihood of a Tail or a Head. We choose $\theta = 1/2$ for a *fair* coin; if the coin is not fair — somehow modified to bias the outcome — then θ will differ from $1/2$.

In principle, there are infinitely many different probability mass functions. In practice, there are a few families of parametrized probability mass functions which typically suffice to "model" most random phenomena — most experiments — of interest; we have presented here two of the more common, uniform (parametrized by the number of outcomes, L), and Bernoulli (parametrized by the probability of a 1, θ). The parameters associated with a probability mass function are determined either by empirical, mechanistic, or subjective approaches. Most commonly, we combine the empirical and mechanistic technology: mechanistic serves to identify the most appropriate family; empirical serves to identify, or estimate, the parameter — informed by the connection between cumulative frequency function and probability. In some fortuitous situations, mechanistic considerations alone suffice to suggest both the appropriate family and the good parameter value.

2.1.2 Bivariate Case

We now consider a discrete random vector, (X, Y), where X and Y are each random variables.

The sample space — the set of all possible outcomes of an experiment — is now given by $\{(x_i^o, y_j^o), 1 \leq i \leq L_X, 1 \leq j \leq L_Y\}$: an $L_X \times L_Y$ grid of values in x and y. We denote the corresponding outcome probabilities as $p_{i,j}, 1 \leq i \leq L_X, 1 \leq j \leq L_Y$. We can assemble these results in a *joint* probability mass function

$$f_{X,Y}(x_i^o, y_j^o) = p_{i,j}, \quad 1 \leq i \leq L_X, 1 \leq j \leq L_Y \ . \tag{6}$$

We say that (X, Y) is (jointly) distributed according to $f_{X,Y}$. We know that $0 \leq f_{X,Y}(x_i^o, y_j^o) \leq 1, 1 \leq i \leq L_X, 1 \leq j \leq L_Y$, and furthermore

$$\sum_{i=1}^{L_X} \sum_{j=1}^{L_Y} f_{X,Y}(x_i^o, y_j^o) = 1 \ . \tag{7}$$

We may think of our joint probability mass function as a distribution of a total mass of unity amongst $L_X L_Y$ point particles at locations $(x_i^o, y_j^o), 1 \leq i \leq L_X, 1 \leq j \leq L_Y$.

We can next define marginal probability mass functions as

$$f_X(x_i^o) = \sum_{j=1}^{L_Y} f_{X,Y}(x_i^o, y_j^o), \quad 1 \leq i \leq L_X \ , \tag{8}$$

$$f_Y(y_j^o) = \sum_{i=1}^{L_X} f_{X,Y}(x_i^o, y_j^o), \quad 1 \leq j \leq L_Y \ . \tag{9}$$

Note that $f_X(x_i^o)$ is the probability of event $\{(x_i^o, y_j^o), 1 \leq j \leq L_Y\}$: $x = x_i^o$ and y may take on any value; similarly, $f_Y(y_j^o)$ is the probability of event $\{(x_i^o, y_j^o), 1 \leq i \leq L_X\}$: $y = y_j^o$ and x may take on any value.

We may also define conditional probability mass functions for the random variables $X \mid Y$ and $Y \mid X$ as

$$f_{X \mid Y}(x_i^o \mid y_j^o) = \frac{f_{X,Y}(x_i^o, y_j^o)}{f_Y(y_j^o)}, \quad 1 \le i \le L_X, 1 \le j \le L_Y \ , \text{ and} \tag{10}$$

$$f_{Y \mid X}(y_j^o \mid x_i^o) = \frac{f_{X,Y}(x_i^o, y_j^o)}{f_X(x_i^o)}, \quad 1 \le i \le L_X, 1 \le j \le L_Y \ , \tag{11}$$

respectively. Note that $f_{X \mid Y}(x_i^o \mid y_j^o)$ is the probability of the event $x = x_i^o$ given that $y = y_j^o$; $f_{Y \mid X}(y_j^o \mid x_i^o)$ admits a similar interpretation.

Finally, we introduce the notion of independence of two random variables. We say that X and Y are independent if and only if

$$f_{X,Y}(x_i^o, y_j^o) = f_X(x_i^o) f_Y(y_j^o), \quad 1 \le i \le L_X, 1 \le j \le L_Y \ . \tag{12}$$

Note that independence of random variables X and Y means that events $x = x_i^o$ and $y = y_j^o$ are independent for all i and j, $1 \le i \le L_X, 1 \le j \le L_Y$. It follows from (12) that

$$f_{X \mid Y}(x_i^o \mid y_j^o) = f_X(x_i^o) \ , \tag{13}$$

$$f_{Y \mid X}(y_j^o \mid x_i^o) = f_Y(y_j^o) \ ; \tag{14}$$

the distribution of X (respectively, Y) is not affected by the value of Y (respectively, X).

We provide here one example of joint probability mass function. We consider the draw of a single card from a shuffled deck. The draw experiment is described by a bivariate random variable (X, Y), where X represents the suit and Y represents the denomination. Our sample space is thus $\{(x_i^o \equiv i, y_j^o \equiv j), 1 \le i \le L_X, 1 \le j \le L_Y\}$ for $L_X = 4$ and $L_Y = 13$: we encode (say) clubs, diamonds, hearts, and spades as $x_1^o = 1, x_2^o = 2, x_3^o = 3$, and $x_4^o = 4$, respectively, and the denomination as $y_j^o = j, 1 \le j \le 13$. We can plausibly assume that the suit and denomination of a card drawn from a well shuffled deck are *independent*, and furthermore that any suit and any denomination are equally likely. We thus choose $f_{X,Y}$ as the (discrete) bivariate uniform probability mass function,

$$f_{X,Y}^{\text{unif};L_X,L_Y}(x_i^o, y_j^o) = \frac{1}{L_X L_Y} \ , \quad 1 \le i \le L_X, 1 \le j \le L_Y \ . \tag{15}$$

We note that $f_{X,Y}^{\text{unif};L_X,L_Y}(x, y) = f_X^{\text{unif};L_X}(x) f_Y^{\text{unif};L_Y}(y)$. (In this case we could either suppose independence to derive the bivariate probability mass function, or indeed suppose "equally likely" and then deduce independence.)

2.1.3 Random Sample

We shall consider a particular, but very important, case of an n-variate random variable, $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$, for $X_i, 1 \le i \le n$, *independent* random variables *identically distributed* according to the (discrete univariate) probability mass function f_X. It follows from

these two assumptions that

$$f_{\mathbf{X}_n}(\mathbf{x}_n \equiv (x_1, x_2, \ldots, x_n)) = \prod_{k=1}^{n} f_X(x_i) \ , \tag{16}$$

where $x_i, 1 \leq i \leq n$, may take on any value in the sample space associated with f_X. As always, our random vector $\mathbf{X}_n$ represents a procedure, and $\mathbf{x}_n$ shall represent an associated (outcome of a) realization. It is important to note that a single realization of $\mathbf{X}_n$, $\mathbf{X}_n \rightarrow \mathbf{x}_n$, requires n realizations of the the random variable X, $(X \rightarrow x_i)_{i=1,\ldots,n}$. For example, for X distributed according to the Bernoulli probability mass function, a single realization $\mathbf{x}_n$ represents n coin flips.

The random variable $\mathbf{X}_n$ is denoted a *random sample*: the "random" summarizes the requirement that the $X_i, 1 \leq i \leq n$, are independent and identically distributed, typically abbreviated as "i.i.d." in the statistical literature. Similarly, the random variate $\mathbf{x}_n$ is denoted (the outcome of) a *random sample realization*. We say, when we create a random sample of (i.i.d.) random variables distributed according to f_X, that we draw the sample from the f_X probability mass function, or equivalently, from an f_X "population." For example, if X is a Bernoulli random variable, we would say that we draw our sample from a Bernoulli population.

A random sample in some sense *defines* a random experiment: the (frequency) probability of outcomes in any given experiment X_i is not affected by the outcomes of the other experiments $X_{i'}, i' \neq i$. This is simple to state, but less simple to ensure or confirm, in particular in the case in which X represents a *physical* experiment (we discuss synthetic experiments below): are the experiments indeed independent? do the outcome probabilities (as limits of frequencies) adhere to the designated univariate probability mass function? In practice, we must do our best to verify these assumptions, and to reflect any "doubts" in subsequent inferences and decisions.

2.1.4 Functions of Random Variables

Univariate Case. We can also consider functions of random variables. We consider first the univariate case: $V = g(X)$, for X distributed according to prescribed probability mass function f_X, and g a given univariate function. Note that, since X is a random variable, so too is V.

We assume for the moment that the function g is invertible, in which case the $g(x_\ell^o), 1 \leq \ell \leq L$, are distinct. It then follows that the probability mass function for V is given by

$$f_V(v_\ell^o) = f_X(g^{-1}(v_\ell^o)), \quad 1 \leq \ell \leq L \ , \tag{17}$$

where $\{v_1^o \equiv g(x_1^o), v_2^o \equiv g(x_2^o), \ldots, v_L^o = g(x_L^o)\}$ is the sample space associated with V. The derivation is simple: the event $v = v_\ell^o \equiv g(x_\ell^o)$ happens if and only if the event $x = g^{-1}(v_\ell^o) = x_\ell^o$ happens; hence the probability of event $v = g(v_\ell^o)$ is equal to the probability of event $x = x_\ell^o$; but the probability of event $x = x_\ell^o$ is $f_X(x_\ell^o) = f_X(g^{-1}(v_\ell^o))$. In some sense our mapping is simply a "renaming" of events: $g(x_\ell^o)$ inherits the probability of x_ℓ^o.

The case in which g is not invertible is a bit more complicated. Groups of many outcomes in the sample space for X will map to corresponding single outcomes in the sample space for V; the corresponding outcome probabilities for X will sum to single outcome probabilities for V. We consider several examples below.

CYAWTP 1. We consider the experiment in which we roll two dice simultaneously. The random variable D_1 represents the number of dots on the face which lands "up" of the first die; the random variable D_2 represents the number of dots on the face which lands "up" on the second die. You may assume that D_1 and D_2 are independent and each described by the discrete uniform probability density for $L = 6$. Now introduce a new random variable V which is the sum of D_1 and D_2. Find the probability mass function for V. Next define a Bernoulli random variable W as a function of V: $W = 0$ if $V < 7$ and $W = 1$ if $V \geq 7$. Find the Bernoulli parameter θ for W.

The Sample Mean. We may also consider functions of bivariate and n-variate random variables. We consider here a special but important case.

Our point of departure is our random sample, in particular the n-variate random variable $\mathbf{X}_n \equiv (X_1, X_2, \ldots, X_n)$ for $X_i, 1 \leq i \leq n$, i.i.d. random variables drawn from a prescribed probability mass function f_X. We may then define the sum of our random variables, Z_n, as a function (from n variables to a single variable) of $\mathbf{X}_n$,

$$Z_n = \sum_{i=1}^{n} X_i \; ; \tag{18}$$

similarly, we may define the sample mean, $\bar{X}_n$, as a function (from n variables to a single variable) of $\mathbf{X}_n$,

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \; . \tag{19}$$

Note that $\bar{X}_n$ is a random variable which is the average of the elements in our n-variate random variable — our random sample — $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$: $\bar{X}_n = Z_n/n$. It is important to note that a single realization of the sample sum or sample mean, $Z_n \to z_n$ or $\bar{X}_n \to \bar{x}_n$, requires n realizations of the the random variable X, $(X \to x_i)_{i=1,\ldots,n}$.

We shall develop later some general properties of the sample mean. However, in the remainder of this section, we shall consider the particular case in which the $X_i, 1 \leq i \leq n$, are drawn from a Bernoulli population with parameter θ. To provide a more intuitive description, we shall often equate the experiment X_i (for any i) with the flip of a coin, and equate outcome 0 to a Tail and outcome 1 to a Head. Our random sample $\mathbf{X}_n$ then represents n independent coin flips. Each outcome $(\mathbf{x}_n)^o_\ell, 1 \leq \ell \leq 2^n$, may be represented as a binary vector $(x_1, x_2, \ldots, x_n), x_i = 0$ (Tail) or 1 (Head), $1 \leq i \leq n$; for example, outcome $(\mathbf{x}_n)^o_1$ (say) $= (1, 0, \ldots, 0)$ corresponds to a Head on the first flip and Tails on all the remaining flips.

Armed with this description, we may now construct the probability mass function for $\mathbf{X}_n$. In particular, $f_{\mathbf{X}_n}$ is given by

$$f_{\mathbf{X}_n}(\mathbf{x}_n) = (1-\theta)^{n-k}\theta^k \quad \text{for} \quad k \equiv \mathcal{H}(\mathbf{x}_n), \quad \mathbf{x}_n = (\mathbf{x}_n)^o_\ell, \quad 1 \leq \ell \leq 2^n \ , \tag{20}$$

where $\mathcal{H}(\mathbf{x}_n)$ is the number of 1's (Heads) in an outcome $\mathbf{x}_n$. The first factor in (20) accounts for the $n-k$ Tails (each of which occur with probability $1-\theta$) in outcome $\mathbf{x}_n$; the second factor accounts for the k Heads (each of which occur with probability θ) in outcome $\mathbf{x}_n$; the probabilities mutliply because the $X_i, 1 \leq i \leq n$, are *independent*.

We next note that, in this Bernoulli case,

$$Z_n = \sum_{i=1}^{n} X_i \tag{21}$$

is simply $\mathcal{H}(\mathbf{X}_n)$, the number of Heads in our random sample of n coin flips: the Tails contribute 0 to the sum, and each Head contributes 1 to the sum. Hence, upon division by n, $\bar{X}_n$ is the *fraction* of coin flips in our random sample which are Heads. (Recall that $\bar{X}_n$ is a random variable: the number of Heads will be different for each sample realization of n coin flips.) We could plausibly expect from our frequentist arguments that for large n, $\bar{X}_n$ will approach θ, the probability of a Head in each flip of the coin. This is indeed the case, and this simple observation will form the basis of our estimation procedures.

We can now readily derive the probability mass function for Z_n. We first note that the sample space for Z_n is $\{k, k = 0, \ldots, n\}$, since the the number of Heads in our n coin flips, k, may range from $k = 0$ — all Tails — to $k = n$ — all Heads. To obtain the probability that Z_n takes on the particular value (outcome) k, we must now sum (20) over all outcomes $\mathbf{x}_n$ — perforce mutually exclusive — which correspond to k Heads. We thus arrive at

$$f_{Z_n}(k) = f_{Z_n}^{\text{binomial};\theta,n}(k) \ , \tag{22}$$

where

$$f_{Z_n}^{\text{binomial};\theta,n}(k) = \binom{n}{k}(1-\theta)^{n-k}\theta^k \ . \tag{23}$$

We deconstruct this formula: there are "n choose k" — $n!/(n-k)!k!$ — outcomes for which k Heads occur (equivalently, "n choose k" distinguishable ways to arrange $n-k$ 0's and k 1's); the probability of each such outcome is given by (20). We note that (20) depends only on k, which is why our sum over all outcomes for which we obtain k Heads is simply a multiplication (of (20)) by the number of outcomes for which we obtain k Heads. The probability mass function (23) is known as the binomial probability mass function.

Finally, we note that the sample space for the sample mean, $\bar{X}_n$, is $\{(\bar{x}_n)^o_k = k/n, 0 \leq k \leq n\}$. The probability mass function for the sample mean, $f_{\bar{X}_n}(\bar{x}_n)$, is $f_{Z_n}^{\text{binomial};\theta,n}(n\bar{x}_n)$: we simply identify the outcome $\bar{x}_n = k/n$ for $\bar{X}_n$ with the outcome k for Z_n.

2.1.5 Pseudo-Random Variates

We now ask, given some probability mass function f_X, how might we create a random sample, $\mathbf{X}_n$, and for what purposes might this sample serve?

More classically, the random experiment X will be a "physical" experiment — the administration of a survey, the inspection of a part, the measurement of a displacement — and $\mathbf{X}_n$ will then be a collection of n independent experiments. A sample realization ($\mathbf{X}_n \to \mathbf{x}_n$) yields a collection of n random variates of X, $\{x_1, x_2, \ldots, x_n\}$. We can then exploit this sample realization — data — to estimate parameters (for example, the Bernoulli θ) or probabilities for purposes of prediction and inference. We elaborate on parameter estimation in a subsequent section.

The advent of the digital computer has created a new approach to the construction of random sample realizations: "pseudo-random variates." In particular, there are algorithms which can create "apparently random" sequences of numbers uniformly distributed between (say) 0 and 1. (We shall discuss the continuous uniform probability density function shortly.) Methods also exist which can then further transform these pseudo-random numbers to pseudo-random variates associated to any selected probability mass function f_X in order to generate pseudo-random sample realizations $(x_1, x_2, \ldots, x_n)$.

Sequences (or samples) or pseudo-random variates are in fact not random. The sequence is intiated and indeed completely determined by a seed.[2] However, absent knowledge of this seed, the pseudo-random variates appear random with respect to various metrics.[3] Note in particular that these pseudo-random variates do not simply reproduce the correct frequencies, but also replicate the necessary independence. (It follows that the first half, or second half, or "middle" half, of a pseudo-random sample realization is also a (smaller) pseudo-random sample realization and will also thus approximately reproduce the requisite outcome frequencies.) These pseudo-random variates can serve in lieu of "truly" random variates generated by some physical process. We indicate here a few applications.

A first application of pseudo-random variates: *pedagogy.* We can develop our intuition easily, rather than through many laborious physical coin flips or die rolls. For example, consider a sample realization $\mathbf{x}_n$ drawn from a Bernoulli population with prescribed parameter θ. We may then readily visualize our frequentist claim that the cumulative frequency $\widetilde{\varphi}_j(0)$ (the fraction of Tails in our sample realization) and $\widetilde{\varphi}_j(1)$ (the fraction of Heads in our sample realization) approaches $1-\theta$ and θ, respectively, as j tends to infinity. (To replicate our frequentist experiments of *Introduction to Probability and Statistics* we would choose n very large, and then consider $\widetilde{\varphi}_j(0), \widetilde{\varphi}_j(1), j = 1, \ldots, n$; not equivalently, but similarly, we can directly investigate $\varphi_n(0)$ and $\varphi_n(1)$ for increasing values of n. Note the former considers a nested sequence of subsets of a given sample, whereas the latter considers a sequence of difference and independent samples.)

[2] In practice, the reproducibility is desirable in the development of code and in particular for debugging purposes: we can perform tests for the same data. Once debugging is complete, it is possible to regularly change the seed say based on the time of day.

[3] In fact, for sufficiently long sequences of pseudo-random numbers, a pattern will indeed emerge, however typically the periodic cycle is extremely long and only of academic concern.

Numerical Experiment 2. Invoke the Bernoulli GUI for $\theta = 0.4$. Visualize the convergence of $\widetilde{\varphi}_j(0)$ to $1-\theta$ and $\widetilde{\varphi}_j(1)$ to θ for $j = 1, \ldots, n$. More quantitatively, evaluate $(\theta - \widetilde{\varphi}_n(1))/\theta$ for $n = 100, n = 400$, and $n = 1600$. Now repeat these experiment for $\theta = 0.1$.

A second application of pseudo-random variates: *methodology development*. We can readily test and optimize algorithms, say for parameter estimation, with respect to synthetic data. Pseudo-random variates are of course no substitute for the actual data — random variates — associated with a particular physical experiment: the former will need to assume precisely what the latter are intended to reveal. However, the synthetic data can serve to develop effective techniques in anticipation of real data.

A third application of pseudo-random variates: *hypothesis testing*. We can often take advantage of pseudo-random variates to generate assumed distributions — a null hypothesis — with respect to which we can then "place" our data to determine statistical significance. We provide a simple example of ths approach, which is very simply implemented and thus a natural first step in the consideration of a hypothesis. (However, it is admittedly a somewhat lazy, analysis-free approach, and inasmuch not enthusiastically endorsed by mathematical statisticians.)

Consider the distribution of birthdays through the year. We assume, our null hypothesis, that we may model birthmonth, X, by $f_X^{\text{unif};L=12}$ (ignoring small differences in the number of days in different months). If this hypothesis is true, then we expect that, for any random sample realization $\mathbf{x}_n$, the goodness-of-fit measure

$$d(\mathbf{x}_n) = (\frac{1}{12}\sum_{i=1}^{12}(\varphi_n(i) - \frac{1}{12})^2)^{1/2} \tag{24}$$

shall be rather small; here $\varphi_n(i)$ is the frequency of outcome (month) i. But how can we distinguish whether a deviation of $d(\mathbf{x}_n)$ from zero is due to a faulty hypothesis, or perhaps just a small sample — the usual "random" fluctations associated with our random experiment? (We assume that the individuals in the sample are selected in some rational way — not, for example, at a small birthday party for quintuplets.)

To proceed, we create many pseudo-random sample realizations $d(\mathbf{x}_n)$ — say m realizations — each of which corresponds to n pseudo-random variates from the (assumed) uniform distribution. We then create a plot of the frequency *versus* outcome d — a histogram — of these results; for sufficiently large m, this histogram will approach the probability mass function for the goodness-of-fit random variable, D (of which d is a realization). We may then place our actual data — the true random variate, $d_n(\mathbf{x}_n^*)$, associated with our sample of individuals — on our plot, and ask whether deviations as large as, or larger than, $d_n(\mathbf{x}_n^*)$, are likely. (Note that n for our pseudo-random variates and random variates must be the same.) If yes, then there is no reason to reject the hypothesis; if no, then perhaps we should reject the hypothesis — we can not explain the lack of fit to "chance."

We now consider results for a particular sample realization, $\bar{x}_{n=51}^*$, from the Spring 2013 2.086 class: students at lecture on particular day in February. For this particular sample realization the frequencies do not appear overly uniform: $\varphi_{n=51}(i), 1 \leq i \leq n$, is given by 3,

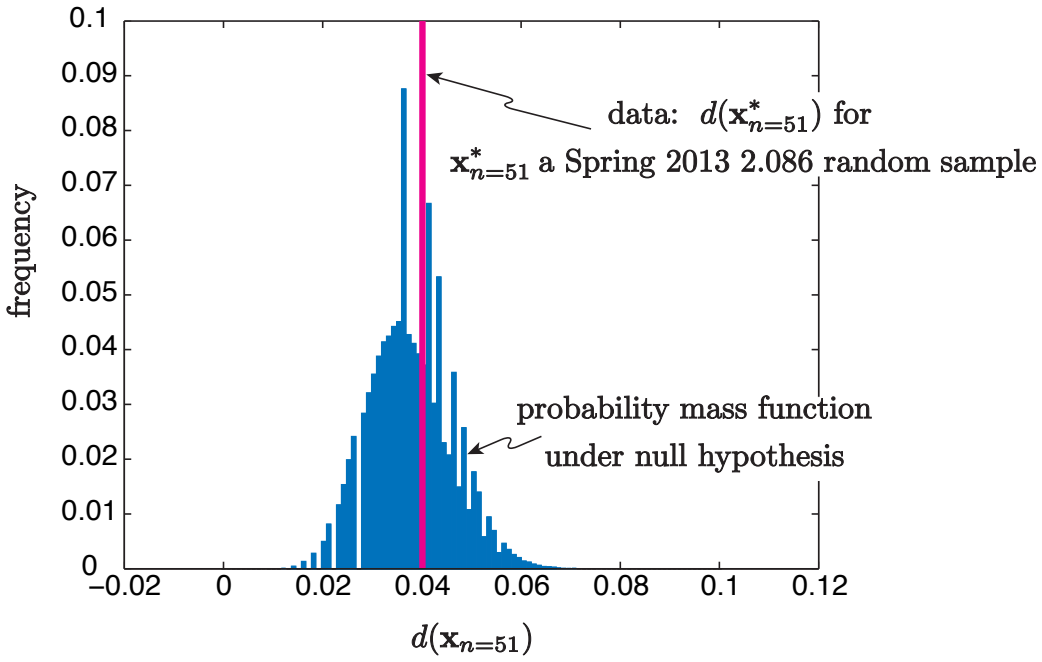

Figure 1: Test of goodness of fit for uniform distribution of birthdays over the twelve months of the calendar year. Plot adapted from William G Pritchett, 2.086 Spring 2013.

1, 7, 2, 7, 5, 3, 3, 6, 2, 6, 6. (For example, seven individuals in the sample are born in each of March and May.) We present in Figure 1 the frequency histogram — approximate probability mass function function for D — derived from $m = 500,000$ pseudo-random variates $d(\mathbf{x}_{n=51})$; we also plot, as the magenta vertical line, the actual data, $d(\mathbf{x}^*_{n=51})$. We observe that, under our hypothesis of a uniform distribution, it would not be particularly unlikely to realize a goodness-of-fit measure as large as $d(\mathbf{x}^*_{n=51})$. In fact, the goodness of fit of the data, $d(\mathbf{x}^*_{n=51})$, lies reasonably close to the "center" of histogram and within the "spread" of the distribution. (We will better understand these characteristics after the discussion of the next section.) In short, the apparent non-uniformity of the data is well within the expected fluctuations for a sample of $n = 51$ individuals and indeed a deviation of zero would be highly unlikely.

A fourth application of pseudo-random variates: *Monte Carlo simulation*. It is often important to recreate a random environment to test, design, or optimize systems in realistic circumstances. A poker player might wish to simulate deals and hence hands similar to those which will be encountered in actual games. A pilot in a flight simulator might wish to respond to gusts which are similar to the real wind patterns which will be encountered in the air. Monte Carlo Simulation also serves to calculate various quantities — failure probabilities, mean performance — which are difficult to evaluate in closed form. (In this sense, the birthmonth analysis above is an example of Monte Carlo simulation.) Finally, Monte Carlo simulation can address certain deterministic problems — transformed to correspond to the expectation of a (pseudo-) random variable — which are intractable by standard deterministic approaches; an important example is integration in many dimensions, the topic

of the next nutshell.

2.2 Expectation

2.2.1 General Definition

We first consider a univariate discrete random variable, X. Given some univariate function g, we define the expectation of $g(X)$ as

$$E(g(X)) \equiv \sum_{\ell=1}^{L} f_X(x_\ell^o) g(x_\ell^o) \; ; \tag{25}$$

we may sometimes write $E_X(g(X))$ to emphasize the random variable and hence probability mass function with respect to which we evaluate the expectation.[4] Thus the expectation of $g(X)$ is a probability-weighted sum of g over the outcomes. We emphasize that $E(g(X))$ is not a random variable, or a random quantity; rather, it is a property of our function g and our probability mass function. Finally, we note the property that if $g(X) = g_1(X) + g_2(X)$, then $E(g(X)) = E(g_1(X)) + E(g_2(X))$; more generally, the expectation of a sum of M functions of X will be equal to the sum of the individual expectations.

In the next section we shall elaborate upon the notion of mean as one particular, and particularly important, expectation. But we present already here the result in order to better motivate the concept of expectation. The mean of a univariate discrete random variable X, μ_X, is the expectation of $g(X) \equiv X$. Hence

$$\mu_X \equiv \sum_{\ell=1}^{L} f_X(x_\ell^o) x_\ell^o \; . \tag{26}$$

We observe that if we interpret f_X as a distribution of mass, rather than probability, then μ_X is precisely the center of mass. (Note that, in some sense, we have already divided through by the total mass since our probability mass function is perforce normalized to sum to unity). We can thus interpret our mean μ_X as a "center of probability." (There are other ways to define a center which we will discuss subsequently.)

Similarly, in the next section we shall elaborate upon the intepretation of variance and standard deviation, but again we already present the result here. The variance of a univariate discrete random variable X, σ_X^2, is the expectation of $g(X) \equiv (X - \mu_X)^2$. Hence

$$\sigma_X^2 \equiv \sum_{\ell=1}^{L} f_X(x_\ell^o)(x_\ell^o - \mu_X)^2 \; . \tag{27}$$

Thus, continuing our physical interpretation, we see that σ_X^2 is a kind of moment of inertia, a measure of how much probability is near (or far) from the center of mass. We also define

[4] In this context, we note that we can also evaluate $E_X(g(X))$ as $E_V(V)$ for $V = f(X)$; however typically this apparently more direct route is in fact much more complicated.

the standard deviation, σ_X, as $\sigma_X \equiv \sqrt{\sigma_X^2}$, which is now directly in the same units as X — a kind of "root-mean-square" deviation. (We note that the expectation of $(X - \mu_X)$ is zero — there are no torques about the center of mass — and hence we must arrive at a standard deviation through a variance.)

In the bivariate case, the definition of expectation is very similar. Given some bivariate function g, we define the expectation of $g(X, Y)$ as

$$E(g(X,Y)) = \sum_{i=1}^{L_X} \sum_{j=1}^{L_Y} f_{X,Y}(x_i^o, y_j^o) g(x_i^o, y_j^o) \ . \tag{28}$$

As before, we note the important property that if $g(X,Y) = g_1(X,Y) + g_2(X,Y)$, then $E(g(X,Y)) = E(g_1(X,Y)) + E(g_2(X,Y))$. But we also have two important additional properties. First, if g is solely a function of X, $g(X)$, then $E(g(X))$ is simply $E_X(g(X))$, the univariate expectation of $g(X)$ with respect to the marginal probability mass function f_X (and similarly if g is only a function of Y). Second, if $g(X,Y)$ is of the form $g_1(X)g_2(Y)$ *and* X and Y are independent, then $E(g(X,Y)) = E_X(g_1(X))E_Y(g_2(Y))$. These bivariate relations are readily extended to the n-variate case.

In the bivariate case, we define the mean (μ_X, μ_Y) as $(E(X), E(Y))$,

$$\mu_X \equiv E(X) \equiv \sum_{i=1}^{L_X} \sum_{j=1}^{L_Y} f_{X,Y}(x_i^o, y_j^o) x_i^o \ , \tag{29}$$

$$\mu_Y \equiv E(Y) \equiv \sum_{i=1}^{L_X} \sum_{j=1}^{L_Y} f_{X,Y}(x_i^o, y_j^o) y_j^o \ . \tag{30}$$

We can also define a variance, which in fact is now a 2×2 *covariance* matrix, as

$$\mathrm{Cov}_{k,\ell} = \sum_{i=1}^{L_X} \sum_{j=1}^{L_Y} f_{X,Y}(x_i^o, y_j^o) \begin{pmatrix} (x_i^o - \mu_X)^2 & (x_i^o - \mu_X)(y_j^o - \mu_Y) \\ (x_i^o - \mu_X)(y_j^o - \mu_Y) & (y_j^o - \mu_Y)^2 \end{pmatrix} \ . \tag{31}$$

The center of mass and moment of inertia interpretations remain valid. Note that if the covariance matrix is diagonal then we say that X and Y are uncorrelated — a technical, not lay, term; furthermore, if X and Y are independent, then X and Y are uncorrelated (though the converse is not necessarily true).

2.2.2 Mean, Variance, Standard Definition: Further Elaboration

We have already defined the mean, variance, and standard deviation. We provide here a few examples and interpretations.

Uniform, Bernoulli, and Binomial Distributions. For the discrete uniform probability mass function, $f_X^{\mathrm{unif};L}$, we obtain $\mu = (L+1)/2$ and $\sigma^2 = (L^2 - 1)/12$ (and, as always, $\sigma = \sqrt{\sigma^2}$).

We observe that the mean is at the "center" of the distribution, and the variance (or standard deviation) increases with L.

We turn now to the Bernoulli probability mass function, $f^{\text{Bernoulli};\theta}$. Here, we find $\mu = \theta$ and $\sigma^2 = \theta(1-\theta)$. We note that if either $\theta = 0$ or $\theta = 1$ then there is no uncertainty in the Bernoulli distribution: if $\theta = 0$, we obtain a Tail with probability one; if $\theta = 1$, we obtain a Head with probability one. In these certain cases, $\sigma^2 = 0$, as all the mass is concentrated on a single outcome.

CYAWTP 3. Derive the mean and variance of the Bernoulli probability mass function.

Finally, we consider the binomial probablility mass function, $f^{\text{binomial};\theta,n}$. In this case we find $\mu = n\theta$ and $\sigma^2 = n\theta(1-\theta)$.

Frequentist Interpretation of Mean. Let us consider a very simple game of chance. We consider the flip of a possibly unfair coin — hence Tail and Head distributed according to Bernoulli with parameter θ, $0 \leq \theta \leq 1$. If the outcome is a 0, a Tail, you neither win nor lose; if the outcome is a 1, a Head, you win \$1.00. Hence after k flips your average revenue per flip, R_k, will be $R_k = \widetilde{\#}_k(1)/k = \widetilde{\varphi}_k(1)$, where we recall that $\widetilde{\#}_k(\cdot)$ and $\widetilde{\varphi}_k(\cdot)$ are respectively the cumulative number function and cumulative frequency function associated with a particular event (here, a 1, or Head). However, we know that in the limit $k \to \infty$ the cumulative frequency function will approach $P(1) = \theta$. Hence the mean of the probability mass function, θ, is the average revenue per flip in the limit of many flips.

We can show more generally that if we consider a game of chance with L outcomes in which for outcome ℓ the pay-off (per game) and probability is x_ℓ^o and p_ℓ, respectively, then in the limit of many games the average revenue per game is the mean of the probability mass function $f_X(x_\ell^o) = p_\ell, 1 \leq \ell \leq L$.

Chebyshev's Inequality. We can understand that the variance (or standard deviation) indicates the extent to which the probability mass function is "spread out." We also understand that the variance is the expected square deviation from the mean. But we can also establish a more quantitative connection between variance and the magnitude and likelihood of deviations from the mean: Chebyshev's Inequality.

Assume that X is distributed according to probability mass function for which the mean and standard deviation are given by μ_X and σ_X, respectively. Given some positive number κ, Chebyshev's Inequality guarantees that

$$P(|x - \mu_X| > \kappa\sigma_X) \leq \frac{1}{\kappa^2} \, . \tag{32}$$

We recall that $P(\mathcal{E})$ is read "probability of event $\mathcal{E}$"; in (32), $\mathcal{E}$ is the union of all outcomes x_ℓ^o such that $|x_\ell^o - \mu| > \kappa\sigma$. This inequality is typically not very sharp, or even useful, but it is illustrative: the standard deviation σ is the relevant scale of the probability mass function such that outcomes several σ from the mean are very unlikely.

Sample Mean: Mean and Variance. As an example of the various relations proposed, we may calculate the mean and sample variance of the sample mean $\bar{X}_n$. To begin, we note that

$$E(\bar{X}_n) = \frac{1}{n} E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) \ , \tag{33}$$

since the expectation of a sum is the sum of the expectations. We next note that $E(X_i) = E_{X_i}(X_i)$: the expectation may be evaluted in terms of the marginal probability mass function of X_i. But for $i = 1, \ldots, n$, $E_{X_i}(X_i) = \mu_X$, the mean of the common probability mass function f_X from which we draw our random sample. Thus we obtain $\mu_{\bar{X}_n} = (1/n) n \mu_X = \mu_X$. We will later rephrase this result in the context of estimation: $\bar{X}_n$ is an unbiased estimator for μ_X. Note that we did not take advantage of independence in this demonstration.

We may also calculate the variance of our sample mean. To begin, we note that

$$\sigma^2_{\bar{X}_n} = E(((\frac{1}{n}\sum_{i=1}^{n} X_i) - \mu_X)^2) = E((\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X))^2) \tag{34}$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} E((X_i - \mu_X)(X_{i'} - \mu_X)) \ . \tag{35}$$

But we now invoke independence to note that, for $i \neq i'$, $E((X_i - \mu_X)(X_{i'} - \mu_X)) = E(X_i - \mu_X)E(X_{i'} - \mu_X)$; but $E(X_k - \mu_X) = 0, 1 \leq k \leq n$, by definition of the mean. Thus only the terms $i = i'$ in our double sum survive, which thus yields

$$\sigma^2_{\bar{X}_n} = \frac{1}{n^2} \sum_{i=1}^{n} E((X_i - \mu_X)^2) = \frac{1}{n^2} n \sigma^2_X = \frac{\sigma^2_X}{n} \ . \tag{36}$$

Note that in this derivation independence plays a crucial role. We observe that the variance of $\bar{X}_n$ is reduced by a factor of n relative to the variance of X. This result, derived so simply, is at the very center of all statistical estimation. We can readily understand the reduction in variance — the greater concentration about the mean: the (independent) fluctuations of the sample cancel in the sample-mean sum; alternatively, to obtain an outcome far from the mean, all the (independent) $X_i, 1 \leq i \leq n$, must be far from the mean — which is very unlikely.

3 Continuous Random Variables

A continuous random variable is a random variable which yields as outcome no longer a finite (or countably finite) number of values, but rather a continuum of outcomes. This allows us to describe a much broader range of phenomenon.

3.1 General Univariate Distribution

We denote our random variable by X. We assume that X yields an outcome in the interval $[a, b]$: X may take on any value x, $a \leq x \leq b$. We next introduce a probability density function. A probability density function is to a continuous random variable as a probability mass function is to a discrete random variable: a mapping between outcomes in the sample space and probabilities. However, in the the case of continuous variables, a few subtleties arise.

We shall denote our probability density function associated with random variable X by f_X. We shall require that $f_X(x) \geq 0, a \leq x \leq b$, and furthermore

$$\int_a^b f_X(x)\,dx = 1. \tag{37}$$

We then express the probability of event $[x, x + dx]$, which for clarity we also write as $x \leq X \leq x + dx$, as

$$P(x \leq X \leq x + dx) = f_X(x)\,dx. \tag{38}$$

(Note that $P(\cdot)$ refers to the probability of an event described in terms of outcomes in the sample space. For continuous random variables we shall explicitly include the random variable in the definition of the event so as to avoid conflation of the outcomes and realizations.) In the same way that a continuous distribution of mass is described by a density, so a continuous distribution of probability is described by a density. We may then express the probability of event $[a', b']$, for $a \leq a' \leq b' \leq b$, as

$$P(a' \leq X \leq b') = \int_{a'}^{b'} f_X(x)\,dx. \tag{39}$$

In the same way that we identify the mass associated with one segment of a (say) bar as the integral of the density over the segment, so we identify the probability associated with one segment of our interval as the integral of the probability density over the segment. Finally, we note from our definitions that

$$P(a \leq x \leq b) = \int_a^b f_X(x)\,dx = 1; \tag{40}$$

the probability that some event in our sample space occurs is, and must be, unity.

We next introduce a *cumulative distribution function*,

$$F_X(x) = \int_a^x f_X(x')\,dx' \ , \tag{41}$$

for $a \leq x \leq b$. We may then express

$$P(x' \leq x) = F_X(x) \ . \tag{42}$$

We note that $F_X(a) = 0$, $F_X(b) = 1$, $F_X(x)$ is a non-decreasing function of x (since f_X is non-negative), and finally

$$P(a' \leq x \leq b') = F_X(b') - F_X(a'). \tag{43}$$

In general F_X need not be (left) continuous, and can exhibit jumps at values of x for which there is a concentrated mass; we shall restrict attention to random variables for which F_X is continuous.

We now introduced the expection of a function $g(X)$. In general, sums over probability mass functions for discrete random variables are replaced with integrals over probability density function for continuous random variables — as we have already seen for the calculation of the probability of an event. For expectation, we obtain

$$E(g(X)) = \int_a^b g(x) f_X(x)\, dx\ . \tag{44}$$

As in the discrete case, there are several functions $g(x)$ of particular interest. The choice $g(X) = X$ yields the mean (center of mass),

$$\mu_X \equiv E(X) = \int_a^b x f_X(x)\, dx\ ; \tag{45}$$

the choice $g(X) = (X - \mu_X)^2$ yields the variance,

$$\sigma_X^2 \equiv E((X - \mu_X)^2) = \int_a^b (x - \mu_X)^2 f_X(x)\, dx. \tag{46}$$

The standard deviation is then simply $\sigma_X \equiv \sqrt{\sigma_X^2}$. The interpretation of these quantities is very similar to the discrete case. (In the continuous case it is possible that the variance may not exist, however we shall only consider probability mass functions for which the variance is finite.)

CYAWTP 4. Let X be any random variable and associated probability mass function with mean μ_X and variance σ_X^2. Introduce a new random variable $V = (X - \mu_X)/\sigma_X$. Demonstrate that $\mu_V = 0$ and $\sigma_V^2 = 1$.

The cumulative distribution function can serve to define the α-quantile of X, $\tilde{x}_\alpha$:

$$F_X(\tilde{x}_\alpha) = \alpha\ ; \tag{47}$$

in words, the probability of the event $x \leq \tilde{x}_\alpha$ is α; more informally, α of the population takes on values less than $\tilde{x}_\alpha$. We note that $\tilde{x}_{\alpha=1/2}$ is the median: half the population takes on values less than $\tilde{x}_{\alpha=1/2}$, and half the population takes on values greater than $\tilde{x}_{\alpha=1/2}$. To facilitate the calculation of $\tilde{x}_\alpha$ we may introduce the inverse of the cumulative distribution function, $F_X^{-1}(p)$: $F_X^{-1}(0) = a$, $F_X^{-1}(1) = b$, and $F_X^{-1}(\alpha) = \tilde{x}_\alpha$.

3.2 The Continuous Uniform Distribution

We begin with the univariate case. The probability density function is very simple:

$$f_X^{\text{unif};a,b}(x) \equiv \frac{1}{b-a} \ , \quad a \leq x \leq b \ . \tag{48}$$

It directly follows that the cumulative distribution function is given by $F_X^{\text{unif};a,b}(x) = (x - a)/(b-a)$ for $a \leq x \leq b$. Finally, for this particular simple density, we may explicitly evaluate (39) as

$$P(a' \leq x \leq b') = \frac{b'-a'}{b-a} \ . \tag{49}$$

This last expression tells us that the probability that X lies in a given interval is proportional to the length of the interval and *independent* of the location of the interval: hence the description "uniform" for this density.

We may readily calculate the mean of X, $\mu_X = (a+b)/2$, the variance, $\sigma_X^2 = (1/12)(b^2 - a^2)$, and hence the standard deviation, $\sigma_X \equiv \sqrt{\sigma_X^2}$. We note the similarity of these expressions to the case of a discrete uniform random variable. As expected, the center of mass is the center of the "bar," and the moment of inertia increases as the length of the bar increases. We can also readily deduce from the cumulative distribution function that the median of the distribution is given by $\tilde{x}_{\alpha=1/2} = (a+b)/2$; in this case, the median and mean coincide.

A random variable U which is distributed according to the continuous uniform density over the unit interval $[0,1]$, sometimes referred to as the *standard* uniform distribution, is of particular interest. In that case $a = 0$ and $b = 1$ and hence $f_U(u) = 1$. If we are given a random variable U distributed according to the continuous uniform density over $[0,1]$, then a random variable X distributed according to the continuous uniform distribution over $[a,b]$ may be expressed a function of U:

$$X = a + (b-a)U \ . \tag{50}$$

Note it is clear that since U takes on values over the interval $[0,1]$, then X takes on values over $[a,b]$. But the statement (50) is stronger: X defined in terms of U according to (50) will be distributed according to the continuous *uniform* density (but now) over $[a,b]$. This result is intuitive: the shift a does not affect the probability; the "dilation" $b-a$ just scales the probability — still uniform — to ensure $\int_a^b f_X(x)\,dx = 1$. Given an pseudo-random sample realization of U, $(u_1, u_2, \ldots, u_n)$, we can generate a pseudo-random sample realization of X, $(x_1, x_2, \ldots, x_n)$, for $x_i = a + (b-a)u_i, 1 \leq i \leq n$.

The continuous uniform distribution is in fact very useful in the generation of pseudo-random variates for an arbitrary (non-uniform) discrete random variable. Let us say that we wish to consider a discrete random variable with sample space $v_\ell^o, 1 \leq \ell \leq L$, and associated probabilities $p_\ell, 1 \leq \ell \leq L$, such that $f_V(v_\ell^o) = p_\ell, 1 \leq \ell \leq L$. We may readily express V as a function of a uniform random variable U over the interval $[0,1]$, $V = g(U)$. In particular,

$g(U) = v_1^o$ if $U \leq p_1$, and then

$$g(U) \equiv v_\ell^o \quad \text{if} \quad \sum_{i=1}^{\ell-1} p_i \leq U \leq \sum_{i=1}^{\ell} p_i, \quad 2 \leq \ell \leq L \; . \tag{51}$$

We can see that our function makes good sense: for example, the probability that $V = v_1^o$ is the probability that U lies in the interval $[0, p_1]$ — which is simply p_1, as desired; the probability that $V = v_2^o$ is the probability that U lies in the interval $[p_1, p_1 + p_2]$, which is simply (the length of the interval) p_2. Given a pseudo-random sample realization of U, $u_1, u_2, \ldots, u_n$, we can generate a pseudo-random sample realization of V, $v_1, v_2, \ldots, v_n$, as $v_i = g(u_i), 1 \leq i \leq n$.

CYAWTP 5. Consider the distribution for the random variable V of **CYAWTP 1**. Find a function g such that V may be expressed as $g(U)$ for U a continuous uniform random variable over the unit interval $[0, 1]$. Indicate how you can generate a pseudo-random sample realization for V from a pseudo-random sample realization of U.

We now consider the bivariate continuous uniform probability density. We denote our random variable as (X, Y). We assume that X takes on values $a_1 \leq x \leq b_1$ and Y takes on values $a_2 \leq Y \leq b_2$ such that (X, Y) takes on values in the rectangle $R \equiv \{(x, y) \,|a_1 \leq x \leq b_1, a_2 \leq y \leq b_2\}$. Our probability density is then given by

$$f_{X,Y}(x, y) = \frac{1}{(b_1 - a_2)(b_2 - a_2)} \equiv \frac{1}{A_R} \; , \tag{52}$$

where A_R is the area of our rectangle R. Then

$$P(x \leq X \leq x + dx, y \leq Y \leq y + dy) = f_{X,Y}(x, y) dx dy \; , \tag{53}$$

and, for any subdomain ("event") D of R,

$$P((X, Y) \in D) = \int_D f_{X,Y} \, dx dy = \frac{A_D}{A_R} \; , \tag{54}$$

where A_D is the area of D. We see that the probability that (X, Y) lies in a given subdomain D is proportional to the area of the subdomain and *independent* of the location or shape of the subdomain: hence the appellation "uniform." Note that $P((X, Y) \in R) = 1$, as must be the case.

We can find the marginal probability density functions,

$$f_X(x) = \int_{a_2}^{b_2} f_{X,Y}(x, y) dy = \int_{a_2}^{b_2} \frac{1}{(b_1 - a_1)(b_2 - a_2)} = \frac{1}{b_1 - a_1} \; ; \tag{55}$$

$$f_Y(y) = \int_{a_1}^{b_1} f_{X,Y}(x, y) dx = \int_{a_1}^{b_1} \frac{1}{(b_1 - a_1)(b_2 - a_2)} = \frac{1}{b_2 - a_2} \; . \tag{56}$$

We directly notice that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, and hence X and Y are independent. Again, as in the discrete case, independence follows from uniformity.

We can take advantage of independence in the generation of sample realizations. In particular, we can first generate random (or pseudo-random) sample realizations $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ for X and Y respectively by *independent* application of our translation-dilation transformation of (50). Then, armed with these random sample realizations for X and Y, $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$, respectively, we can form a random (or pseudo-random) sample realization for (X, Y) as $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$.

3.3 The Normal Distribution

The normal density, also known as the Gaussian distribution, is perhaps the most well-known of any probability density function. We consider here only the univariate case. We note that the normal density is defined over the entire real axis, so that we now take $a = -\infty$ and $b = \infty$.

The probability density function for a normal random variable X is given by

$$f^{\text{normal};\mu,\sigma} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty . \tag{57}$$

We note that the density function is symmetric about μ. The associated cumulative distribution function is denoted $F_X^{\text{normal};\mu,\sigma}(x)$.

It can be shown that $\mu_X = \mu$, and $\sigma_X^2 = \sigma$: hence, for the normal density, the two parameters which define the family are precisely the mean and variance of the density. We also conclude from symmetry of the probability density function that values above and below the mean are equally likely, and the hence the median is also given by μ. We can evaluate additional quantiles from the cumulative distribution function: $\tilde{x}_{0.841} \approx \mu+\sigma$, $\tilde{x}_{0.977} \approx \mu+2\sigma$, and $\tilde{x}_{0.9985} \approx \mu + 3\sigma$; also $\tilde{x}_{0.975} \approx \mu + 1.96\sigma$. From this last result, and symmetry, we know that only 5% of the population resides in the two "tails" $x \leq \mu - 1.96\sigma$ and $x \geq \mu + 1.96\sigma$.

CYAWTP 6. Sketch the normal density for $\mu = 1, \sigma = 1$, for $\mu = 0, \sigma = 4$, and for $\mu = 3, \sigma = .2$. In the last case, indicate roughly the two (symmetric) tails which contain roughly 5% of the population.

A random variable, say Z, which is distributed according to the normal density with mean 0 and variance 1, hence $f_Z(z) = f_Z^{\text{normal};0,1}(z)$, is know as a *standard* normal random variable. The cumulative distribution function for Z is given a special name: $\Phi(z)$. It can be shown that if Z is a standard normal variable, then

$$X = \mu + \sigma Z \tag{58}$$

is a normal variable with mean μ and variance σ^2. Note that, conversely, if X is a normal random variable with mean μ and variance σ^2, then $Z = (X - \mu)/\sigma$ is a standard normal variable: this is a special case of **CYAWTP 4**. We can also deduce the quantiles of X from the quantiles of Z: $\tilde{x}_\alpha = \mu + \sigma\Phi^{-1}(p)$; for example, $\Phi^{-1}(0.975) \approx 1.96$, and hence

$\tilde{x}_{0.975} = \mu + 1.96\sigma$. The transformation (58) is invoked often: given a pseudo-random sample realization of Z, $(z_1, z_2, \ldots, z_n)$, we can generate a pseudo-random sample realization of X, $(x_1, x_2, \ldots, x_n)$, for $x_i = \mu + \sigma z_i, 1 \leq i \leq n$.

The normal density is ubiquitous for many reasons. The domain is infinite, so there is no need to artifically truncate. The density is defined by a single "location" parameter, μ, also the mean, and a single "scale" parameter, σ (the standard deviation); it is often easy to choose these parameters in a plausible fashion even absent extensive experimental data. Gaussians also play well together: the sum of M independent Gaussian random variables is, in fact, a Gaussian random variable, with mean (respectively, variance) the sum of the means (respectively, the sum of the variances).

But the Gaussian is perhaps most noteworthy for its universality, as summarized in the "Central Limit Theorem," of which we provide here a particular, relatively restrictive, statement. We are given a random variable X distributed according to a prescribed probability mass or density function, f_X. (This mass or density must satisfy certain criteria; we do not provide technical details.) We next construct a random sample and form the associated sample mean, $\bar{X}_n$. It can then be shown that, as $n \to \infty$, and we consider larger and larger samples,

$$P\left(\frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \leq z\right) \to \Phi(z) \; ; \tag{59}$$

in other words, the cumulative distribution function of the sample mean (shifted and scaled to zero mean and unit variance) approaches the cumulative distribution function of a standard normal random variable. This approximation (59) is often valid even for quite small n: for example, for a Bernoulli random variable, X, the normal approximation is accurate to a few percent if $n\theta \geq 10$ and $n(1-\theta) \geq 10$. Note that, like Chebyshev's Inequality, (59) illustrates the connection between mean, variance, and large deviations; however (when applicable), (59) is asymptotically an equality.

The Central Limit Theorem also lends some credence to the choice of a Gaussian to model "unknown" effects. If we think of randomness as arising from many independent sources, all of which add (and cancel) to yield the final result, then the Central Limit Theorem suggests that these many additive sources might well be approximated by a Gaussian. And indeed, it is often the case that phenomena can be well modeled by a normal probability density. However, it is also often the case that there particular constraints present in any paticular system — related to positivity, or (un)symmetry, or correlation — which create significant non-normality in the distribution.

4 Estimation of the Mean

4.1 Parameter Estimation: Motivation

We must often estimate a parameter associated with a probability mass function or probability density function. These estimates can serve two important but quite different purposes:

to calibrate a probability mass or density function for subsequent service say in Monte Carlo studies; to make inferences about the underlying population and ultimately to make decisions informed by these inferences.

As an example of calibration, we refer to the distribution of wind gust velocity. We may wish to take advantage of the probability density function in a flight simulator, but we must first determine an appropriate underlying probability density function, and subsequently estimate the associated parameters. For example, we may say that the wind gust velocity follows a normal distribution, which in turn requires calibration of the mean — zero by definition of a gust — and the variance — *a priori* unknown. As an example of inference, we consider a Bernoulli population. The parameter θ may represent a probability of failure, or the fraction of parts from an assembly line which do not satisfy quality standards, or indeed the fraction of a (real, people) population who will vote for a particular candidate. In these cases, the parameter is directly of interest as regards subsequent decisions.[5]

In this section we develop estimators for the mean of a probability mass or density function. In the particular case of a Bernoulli population, the mean is simply the Bernoulli parameter, θ.

4.2 Sample Mean Estimator: General Case

In our discussion of discrete random variables, we demonstrated that the sample mean has very nice properties. In fact, these properties extend to the case of probability density functions as well.

We introduce a univariate random variable, discrete or continuous, X, distributed according to a probability mass function or probability density function f_X. We then introduce a random sample $\mathbf{X}_n \equiv (X_1, X_2, \ldots, X_n)$. We then define the sample mean $\bar{X}_n$ as $\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$: $\bar{X}_n$ is simply the usual arithmetic average of the $X_i, 1 \leq i \leq n$. We can then demonstrate — as already derived for the discrete case — the following properties: $\bar{X}_n \to \mu_X$ as $n \to \infty; E(\bar{X}_n) = \mu_X; \sigma_{\bar{X}_n} = \sigma_X/\sqrt{n}$.

It is thus natural to define an estimator for μ_X, $\hat{\mu}_X$, as $\hat{\mu}_X \equiv \bar{X}_n$. (Because the expectation of $\bar{X}_n = \mu_X$, the quantity we wish to estimate, $\bar{X}_n$ is denoted an *unbiased* estimator.) As we take larger and larger samples, $\hat{\mu}_X$ will approach μ_X: the variance of our estimator decreases, and thus the *expected* deviation of our estimator, $\hat{\mu}_X$, from $\mu_{\bar{X}_n} = \mu_X$ — which we wish to estimate — will be smaller and smaller; we can further state from Chebyshev's Inequality that the *probability* that $\hat{\mu}_X$ will differ from μ_X by more than, say $10\sigma_X/\sqrt{n}$, will be less than .01. We thus have a kind of probabilistic convergence of our estimator. It is important to note that $\hat{\mu}_X$ is a random variable: each realization $\bar{X}_n \to (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ will be different.

The results summarized above suggest that the sample mean is indeed a good estimator for the mean of a population, and that our estimator will be increasingly accurate as we increase the sample size. It is possible to develop more quantitative, and sharp, indicators of the error in our estimator: confidence intervals, which in turn are based on some estimate of

[5] Another example of a Bernoulli parameter is $P(R \mid T)$ of the nutshell *Introduction to Probability and Statistics.*

the variance of the population. We shall consider the latter in the particular, and particularly simple, case of a Bernoulli population.

4.3 Sample Mean Estimator: Bernoulli Population

For a Bernoulli population, as already indicated, the mean is simply our parameter θ. We shall thus denote our sample mean estimator as $\hat{\Theta}_n$; we denote our sample estimate as $\hat{\theta}_n$. Note $\hat{\Theta}_n \rightarrow \hat{\theta}_n$: an application of our sample mean estimator, $\hat{\Theta}_n$ — a random variable — yields a sample mean estimate, $\hat{\theta}_n$ — a real number.

Our results for the general sample-mean estimator directly apply to the Bernoullli population, and hence we know that $\hat{\Theta}_n$ will converge to θ as n increases, that $E(\hat{\Theta}_n) = \theta$, and that $E((\bar{X}_n - \theta)^2) = \sqrt{\theta(1-\theta)/n}$. But we now also provide results for confidence intervals: we consider two-sided confidence intervals, though it is also possible to develop one-sided confidence intervals; we consider normal-approximation confidence intervals, though it is also possible to develop (less transparent) exact confidence intervals.

We first introduce our confidence interval

$$[CI]_n(\hat{\Theta}_n, \gamma) \equiv [\frac{\hat{\Theta}_n + \frac{z_\gamma^2}{2n} - z_\gamma \sqrt{\frac{\hat{\Theta}_n(1-\hat{\Theta}_n)}{n} + \frac{z_\gamma^2}{4n^2}}}{1 + \frac{z_\gamma^2}{n}}, \frac{\hat{\Theta}_n + \frac{z_\gamma^2}{2n} + z_\gamma \sqrt{\frac{\hat{\Theta}_n(1-\hat{\Theta}_n)}{n} + \frac{z_\gamma^2}{4n^2}}}{1 + \frac{z_\gamma^2}{n}}], \tag{60}$$

where γ is our desired confidence level, $0 < \gamma < 1$, say $\gamma = 0.95$, and $z_\gamma = \Phi^{-1}((1+\gamma)/2)$. Note that $\gamma = 0$ yields $z_\gamma = 0$, and as $\gamma \rightarrow 1$ — much confidence — $z_\gamma \rightarrow \infty$; for $\gamma = 0.95$, $z_\gamma = 1.96$. We note that $\hat{\Theta}_n$ is a random variable, and hence $[CI]_n(\hat{\Theta}, \gamma)$ is a random interval. We can then state that

$$P(\theta \in [CI]_n(\hat{\Theta}_n, \gamma)) \approx \gamma \; . \tag{61}$$

The $\approx$ in (61) is due to our (large-sample) normal approximation, as elaborated in the Appendix. We shall consider the approximation valid if $n\theta \geq 10$ and $n(1-\theta) \geq 10$; under these requirements, the errors induced by the large sample approximation (say in the length of the confidence interval) are on the order of 1% and we may interpret $\approx$ as $=$.

We can now state the practical algorithm. We first perform the realization: we draw the necessary sample from the Bernoulli population, $\mathbf{X}_n \rightarrow \mathbf{x}_n \equiv (x_1, x_2, \ldots, x_n)$, and we subsequently evaluate the associated sample-mean estimator, $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n x_i$. We next compute the confidence interval associated with our sample-mean estimate,

$$[ci]_n(\hat{\theta}_n, \gamma) \equiv [\frac{\hat{\theta}_n + \frac{z_\gamma^2}{2n} - z_\gamma \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n} + \frac{z_\gamma^2}{4n^2}}}{1 + \frac{z_\gamma^2}{n}}, \frac{\hat{\theta}_n + \frac{z_\gamma^2}{2n} + z_\gamma \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n} + \frac{z_\gamma^2}{4n^2}}}{1 + \frac{z_\gamma^2}{n}}] \; . \tag{62}$$

(In cases in which we consider confidence intervals for several different quantities, we will denote the confidence interval for θ more explicitly as $[ci_\theta]_n$.) We can then state that, with confidence level γ, the Bernoulli parameter θ will reside within the confidence interval $[ci]_n(\hat{\theta}_n, \gamma)$. Recall that, say for $\gamma = 0.95$, $z_\gamma = z_{\gamma=0.95} \approx 1.96$. We should only provide

(or in any event, quantitatively trust) the confidence intervals if the criterion $n\hat{\theta}_n \geq 10$, $n(1-\hat{\theta}_n) \geq 10$, is satisfied. (In principle, the criterion should be stated in terms of θ; however, since we are not privy to θ, we replace θ with $\hat{\theta}_n \approx \theta$.)

CYAWTP 7. Invoke the Bernoulli GUI for $\theta = 0.4$. Consider first $n = 400$: is θ inside the confidence interval (62) — recall $\hat{\theta}_{n=400} \equiv \varphi_{n=400}(1)$ — for confidence level $\gamma = 0.95$? for confidence level $\gamma = 0.5$? for confidence level $\gamma = 0.1$? Now consider $n = 4000$: is θ inside the confidence interval for $\gamma = 0.95$?

We next characterize this confidence interval result. First, there is the *confidence level*, γ; γ is, roughly, the probability that our statement is correct — that θ really does reside in $[ci]_n(\hat{\theta}_n, \gamma)$. We can be a bit more precise, and provide a frequentist interpretation. If we were to construct many sample-mean estimates and associated confidence intervals — in other words, repeat (or repeatedly realize) our entire estimation procedure m times — in a fraction γ of these $m(\to \infty)$ realizations the parameter θ would indeed reside in $[ci]_n(\hat{\theta}_n, \gamma)$. (Note that for *each* realization of the estimation procedure we perform n Bernoulli realizations $(X \to x_i)_{i=1\,...,n}$ to form $\hat{\theta}_n = (1/n)\sum_{i=1}^{n} x_n$.) Conversely, in a fraction of $1-\gamma$ of our estimation procedures, the Bernoulli parameter θ will not reside in the confidence interval. In actual practice, we will only conduct one realization — not $m > 1$ realizations — of our estimation procedure. How do we know that our particular estimation procedure is not in the unlucky $1-\gamma$ fraction? We do not. But note that in most real-life experiments there are many uncertainties, mostly unquantified; at least for our confidence interval we can assess and control the uncertainty. We also remark that even if θ does not reside in the confidence interval, it may not be far outside.

Second, there is the *length of the confidence interval*, which is related to the accuracy of our prediction, as we now quanity. In particular, we now note that since (with confidence level γ) θ can reside anywhere in the confidence interval, the extremes of the confidence interval constitute an error bound. In what follows, in order to arrive at a more transparent result, we shall neglect in the confidence interval the terms z_γ^2/n relative to $\hat{\theta}_n$. We may obtain an absolute error bound,

$$|\theta - \hat{\theta}_n| \leq \mathrm{AbsErr}_n(\hat{\theta}_n, \gamma) \;, \tag{63}$$

for

$$\mathrm{AbsErr}_n(\hat{\theta}_n, \gamma) \sim z_\gamma \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \;. \tag{64}$$

(Given the terms neglected, this result is, in principle, valid only asymptotically as $n \to \infty$; in practice, rather modest n suffices.) We may also develop a relative error bound,

$$\frac{|\theta - \hat{\theta}_n|}{\hat{\theta}_n} \leq \mathrm{RelErr}_n(\hat{\theta}_n, \gamma) \;, \tag{65}$$

for

$$\mathrm{RelErr}_n(\hat{\theta}_n, \gamma) \sim z_\gamma \sqrt{\frac{(1-\hat{\theta})}{\hat{\theta} n}} \; . \tag{66}$$

(Again, the result is, in principle, valid only asymptotically as $n \to \infty$; in practice, rather modest n suffices.)

We note that both the absolute and relative error bounds scale with z_γ: as we demand more confidence, $\gamma \to 1$, z_γ will tend to infinity. In short, we pay for inceased confidence, or certainty, with decreased accuracy. Alternatively, we might say that as we become more certain of our statement we become less certain of the actual value of θ. Typically γ is chosen to as $\gamma = 0.95$, in which case $z_\gamma = 1.96$; however, there may be circumstances in which more confidence is desired.

In general, convergence of $\hat{\theta}_n \to \theta$ is quite slow: the error decreases only as $1/\sqrt{n}$; to double our accuracy, we must increase the size of our sample fourfold. If we wish to incur an absolute error no larger than ϵ_{Abs}, then we should anticipate (for $\gamma = 0.95$) a sample size of roughly $1/\epsilon^2$. The situation is more problematic for small θ, for which we must consider the relative error. In particular, if we wish to incur a relative error no larger than ϵ_{Rel}, then we should anticipate (for $\gamma = 0.95$) a sample size of roughly $4/(\theta \epsilon^2_{\mathrm{Rel}})$. From the latter we conclude, correctly, that it is very difficult to estimate accurately the probability of rare events; an important example of a rare event is failure of engineering systems.

5 Perspectives

Our treatment of random variables — from definition and characterization to simulation and estimation — is perforce highly selective. For a comprehensive introduction to probability, random variables, and statistical estimation, in a single volume, we recommend *Introduction to the Theory of Statistics*, AM Mood, FA Graybill, and DC Boes, McGraw-Hill, 1974.

6 Appendix: Derivation of Confidence Interval

To start, we note from (59) that, for sufficiently large n,

$$P\left(\frac{(\hat{\Theta}_n - \theta)}{\sqrt{\frac{\theta(1-\theta)}{n}}} \leq z \right) \approx \Phi(z) \; . \tag{67}$$

We may take advantage of the symmetry of the standard normal density to rewrite (67) as

$$P\left(-z \leq \frac{\hat{\Theta}_n - \Theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \leq z \right) \approx \Phi(z) - \Phi(-z) = \Phi(z) - (1 - \Phi(z)) = 2\Phi(z) - 1 \; . \tag{68}$$

We now choose $2\Phi(z_\gamma) - 1 = \gamma$, where γ shall denote our *confidence level*; thus $z_\gamma = \Phi^{-1}((1+\gamma)/2)$. Note that $\gamma = 0$ yields $z_\gamma = 0$, and as $\gamma \to 1$ — much confidence — $z_\gamma \to \infty$; for $\gamma = 0.95$, $z_\gamma = 1.96$.

We next note that the event in (68) can be "pivotted" about θ to yield an equivalent statement

$$P\left(\hat{\Theta}_n - z_\gamma\sqrt{\frac{\theta(1-\theta)}{n}} \leq \theta \leq \hat{\Theta}_n + z_\gamma\sqrt{\frac{\theta(1-\theta)}{n}}\right) \approx \gamma \; ; \tag{69}$$

note that in (69) we also substitute γ for $2\Phi(z_\gamma) - 1$. In the present form, (69) is not useful since θ appears both "in the middle" — as desired — but also in the limits of the interval. To eliminate the latter we note that the event in (69) is in fact a quadratic inequality for θ,

$$\theta^2(1 + \frac{z_\gamma^2}{n}) + \theta(-2\hat{\Theta}_n - \frac{z_\gamma^2}{n}) + \hat{\Theta}_n^2 \leq 0 \; , \tag{70}$$

which has solution

$$\frac{\hat{\Theta}_n + \frac{z_\gamma^2}{2n} - z_\gamma\sqrt{\frac{\hat{\Theta}_n(1-\hat{\Theta}_n)}{n} + \frac{z_\gamma^2}{4n^2}}}{1 + \frac{z_\gamma^2}{n}} \leq \theta \leq \frac{\hat{\Theta}_n + \frac{z_\gamma^2}{2n} + z_\gamma\sqrt{\frac{\hat{\Theta}_n(1-\hat{\Theta}_n)}{n} + \frac{z_\gamma^2}{4n^2}}}{1 + \frac{z_\gamma^2}{n}} \; . \tag{71}$$

We may thus define our confidence interval as

$$[CI]_n(\hat{\Theta}_n, \gamma) \equiv [\frac{\hat{\Theta}_n + \frac{z_\gamma^2}{2n} - z_\gamma\sqrt{\frac{\hat{\Theta}_n(1-\hat{\Theta}_n)}{n} + \frac{z_\gamma^2}{4n^2}}}{1 + \frac{z_\gamma^2}{n}}, \frac{\hat{\Theta}_n + \frac{z_\gamma^2}{2n} + z_\gamma\sqrt{\frac{\hat{\Theta}_n(1-\hat{\Theta}_n)}{n} + \frac{z_\gamma^2}{4n^2}}}{1 + \frac{z_\gamma^2}{n}}] \; . \tag{72}$$

We then conclude from (69), (71), and (72) that

$$P(\theta \in [CI]_n(\hat{\Theta}_n, \gamma)) \approx \gamma \; . \tag{73}$$

Note that our confidence interval $[CI]_n(\hat{\Theta}_n, \gamma)$ is a *random* interval.

The $\approx$ in (73) is due to the normal approximation in (67). In practice, if $n\theta \geq 10$ and $n(1-\theta) \geq 10$, then the error induced in the confidence interval by the normal approximation will be on the order of several percent at most.

2.086 Numerical Computation for Mechanical Engineers
Fall 2014