University of Illinois Department of Economics
Spring 2014 Econ 574 Roger Koenker

Lecture 1
"13 ways of looking at a random variable"

0. Some Obligatory Formalism

It is customary to begin with some priestly incantations about Kolmogorov's axiomatic treatment of probability. We will keep this brief, since it won't play a large role in subsequent developments. It is essential at some stage however since it is what really makes probability a proper branch of mathematics.

In the beginning, before the chaos, there was the *probability space* $(\Omega, \mathcal{A}, P)$ where

- Ω is the sample space with elements ω
- $\mathcal{A}$ is a σ-field of subsets of Ω with elements A called "events" and
- P is a probability measure which assigns probabilities to elements of $\mathcal{A}$ according to the following rules:
  - (i) $0 \leq P(A) \leq 1$
  - (ii) $P(\Omega) = 1$
  - (iii) If the "events" $A_1, \ldots, A_n$ are disjoint (pairwise mutually exclusive, i.e., $A_i \cap A_j = \phi$) Then $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n)$

From these humble beginnings a great sand castle may be built. Other sand castles may be built from alternative axioms – see for example Savage (1954, *Foundations of Statistics*) who develops probability as a degree of personal belief rather than based on frequency in some hypothetical reproducible experiment.

Remark It is sometimes helpful to extend the axioms above from "finitely additive" to "countably additive". Note that there is an appendix (L1a) to this lecture that attempts to review some basics of the underlying measure theory.

The standard example is the identification of an experiment in which a potentially infinite sequence of coin flips occurs and the probability space is taken as follows

$$\begin{aligned} \Omega &= [0,1) \\ \mathcal{A} &= \text{Borel sets on } \Omega \\ P &= \text{Lebesgue measure} \end{aligned}$$

How does this work? We have on the i^{th} toss $\delta_i = 1$ if the toss lands heads and 0 if tails, and write for n tosses

$$X_i = \sum_{i=1}^{n} \delta_i / 2^i$$

Thus the sequence HHTTH would be expressed as

$$x = \frac{1}{2} + \frac{1}{4} + \frac{0}{8} + \frac{0}{16} + \frac{1}{32}$$

or in a diadic (binary) expansion

$$x = .11001$$

We thereby have a way to translate the physical experiment of coin flipping, as in the first scene of Rosencrantz and Guildenstern Are Dead, into formal mathematical language: an infinite sequence of coin flips becomes just a number somewhere in the interval [0,1].

This seemingly innocent formalism yields some remarkably deep insights into number theory as well as probability. See for example, Kac (1959, *Statistical Independence in Probability and Number Theory*) or Billingsley (1979). E.g., the set of rational numbers (which have terminating expansions consisting entirely of zeros) is known to have Lebesgue measure 0. We are now ready to begin our 13 ways of looking at X.

1. A function $X : \Omega \to \mathbb{R}$ such that images $X^{-1}(B)$ of any Borel set are elements of $\mathcal{A}$ is called a *random variable.* A p-tuple of r.v.s. is called a *random vector.*

2 Associated with a random vector X on $(\Omega, \mathcal{A}, P)$ is a *distribution function*, df,

$$F(x) = F_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = P(\omega : X_1(\omega) \leq x_1, \ldots, X_p(\omega) \leq x_p)$$

Note that F is *right* continuous. This is a convention, but a *useful* one. For students with a little knowledge of French there is a helpful *aide memoire:* cadlag, which is an acronym for *continué à droit, limites à gauche*, continuous from the right, limits from the left.

3 For any scalar r.v. X with df F, the quantity

$$Q(u) = F^{-1}(u) = \inf\{x : \ F(x) \geq u\}$$

is called the u^{th} *quantile* of X, or F. In particular, $Q(1/2)$ is the *median.* Note Q is left continuous caglad. For this it is important that we have the weak inequality in the definition. Drawing a picture is valuable to clarify this. If you draw a piecewise constant F and then flip the page over to see the Q, you can amaze your friends with your deep knowledge of "how to invert functions which don't have inverses." Note that we have, relying on our continuity conventions: $F(Q(u)) = u$ for any $u \in (0, 1)$.

We note in passing, in the hope that it may be deemed relevant at some later point, that the random variable X^* whose quantile function is,

$$F^{-1}_{X^*}(u) = (1-u)^{-1} \int_u^1 F_X^{-1}(v)dv$$

is called the Hardy-Littlewood transform of X.

4 If the df F is absolutely continuous with respect to the measure μ, then F has a *density*, f, with respect to μ. We will only be concerned with the case in which μ is

Lebesgue measure in which case we may write

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

and thus we may regard the density $f(t)$ as the derivative $F'(t)$.

5. The *expectation* of a random variable X is

$$EX = \int_{\Omega} X(\omega)dP(\omega) = \int_{-\infty}^{\infty} xdF(x) = \int_{-\infty}^{\infty} xf(x)dx$$

expectations of functions X may be computed similarly

$$Eg(X) = \int_{\Omega} g(X(\omega))dP(\omega) = \int g(x)dF(x).$$

An important special case is $g(x) = I_B(x)$ where $I_B(x)$ is the "indicator function" of the set B and takes the value 1 if $x \in B$ and 0 otherwise. In this case we have, setting $A = X^{-1}(B)$,

$$EI_B(x) = \int_B dF(x) = \int_A dP(\omega) = P(A)$$

Example: (The most famous homework problem of economics.) Daniel Bernoulli, while visiting Euler in St Petersburg in the 1730's, made famous the problem of valuing a gambling prospect, X, that paid 2^i with probability 2^{-i} for $i = 1, 2, \ldots$. This prospect can be made a bit more explicit by imagining someone with a large bankroll flipping coins against a equally wealthy casino; each flip that our prospective gambler loses, he doubles the stakes of the bet in the next round until finally he guesses correctly. One has to make a "willing suspension of disbelief" to imagine the wealth and patience of the players in this scenario. The paradoxical aspect of the problem is that $EX = +\infty$, but it seems quite clear that few would want to pay this much. Bernoulli's proposed solution was to suggest that individuals evaluating the gamble X might instead compute expected *utility* which would be a concave function of the monetary payoff. The (natural) logarithm provided a convenient example, and one can easily compute $E\log(X) \approx 1.40$ utils, and $\exp(1.40) \approx 4.00$ ducats in the original monetary units of the problem. Of course, we should realize that this "certainty equivalent" value is also suspect, and should be modified, presumably reduced, to account for the variability of the payoff. And we should also account for the individuals initial wealth in the utility calculation, but these complications takes us too far away from the original purpose.

6. Moments. Expectations of higher powers of $X - \mu$ are often used to describe the basic characteristics of the distributions of r.v.s.

$$\mu_k = E(X - EX)^k$$

in particular,

measures

$\text{Var}(x) = \mu_2 = \sigma^2$	dispersion-spread
$\text{Skewness}(X) = \mu_3/\sigma^3$	asymmetry
$\text{Kurtosis}(X) = \mu_4/\sigma^4$	peakedness & tail length.

7. Moment generating function To compute moments it is often convenient to the the moment generating function, mgf,

$$m_X(t) = Ee^{tX} = \int e^{tX} dF(x)$$

when $m_X(t)$ and its derivatives exist in some neighborhood of 0 we have

$$\nu_k = m_X^{(k)}(0) = EX^k \qquad k = 0, 1, 2, \ldots$$

i.e., the moments about the origin are simply the derivatives of the mgf. To get the moments about the mean $EX = \mu$ we may use

$$m_{X-\mu}(t) = Ee^{t(X-\mu)} = e^{-\mu t} m_X(t)$$

Note that the following properties hold

(i) For constants μ, σ
$$m_{\mu+\sigma X}(t) = e^{\mu t} m(\sigma t)$$
(ii) For independent X, Y
$$m_{X+Y}(t) = m_X(t) m_Y(t)$$

Since the moment generating function always seems a bit mysterious it is worth taking a few moments to try to demystify it. I will try to do this by starting with a more elementary setting: discrete random variables on the nonnegative integers.

Suppose we have a discrete *r.v.* on $\{0, 1, 2, \ldots\}$ with

$$P(X = j) = a_j$$

We will define the *generating function* of X as,

$$g(z) = \sum_{j=0}^{\infty} a_j z^j$$

Since $\sum a_j = 1$ it is clear that

$$|g(z)| \leq \sum_j |a_j| \ |z|^j \leq \sum a_j = 1 \qquad \text{for } |z| \leq 1.$$

Now consider derivatives:

$$g'(z) = a_1 + 2a_2 z + 3a_3 z^2 + \ldots = \sum_{n=1}^{\infty} n a_n z^{n-1}$$

$$g''(z) = 2a_2 + 6a_3 z + \ldots = \sum_{n=1}^{\infty} n(n-1) a_n z^{n-2}$$

$$\vdots$$

$$g^{(j)}(z) = \sum_{n=j}^{\infty} n(n-1)\ldots(n-j+1) a_n z^{n-j} = \sum_{n=j}^{\infty} \left(\begin{array}{c} n \\ j \end{array} \right) (j!) a_n z^{n-j}$$

Thus,

$$g^{(j)}(0) = j! a_j \qquad \text{or} \qquad a_j = (j!)^{-1} g^{(j)}(0)$$

so all the information about the a_j's are contained within the function g and is made accessible by simply differentiating and evaluating at 0. Already this justifies the comment by K.L. Chung that the generating function is a "true gimmick".

Note also that we can get moments, provided they exist, by evaluating derivatives at $z = 1$,

$$\begin{aligned} g'(1) &= \sum_{n=0}^{\infty} n a_n = EX \\ g''(1) &= \sum n^2 a_n - \sum n a_n = EX^2 - EX \end{aligned}$$

so $EX = g'(1)$ and $EX^2 = g''(1) + g'(1)$, etc.

Thm: The distribution of a nonnegative integer valued *r.v.* is uniquely determined by its generating function.

Pf: Follows from the fact that $a_j = (j!)^{-1} g^{(j)}(0)$.

Now we begin to see how to use the gimmick. Suppose we multiply two generating functions

$$\begin{aligned} g(z)h(z) &= \sum_i a_i z^i \sum_j b_j z^j = \sum_i \sum_j a_i b_j z^{i+j} \\ &= \sum c_k z^k \end{aligned}$$

where $c_k = \sum_{i+j=k} a_i b_j = \sum_{i=0}^{k} a_i b_{k-i}$. This series $\{c_k\}$ is called the convolution of $\{a_i\}$ and $\{b_i\}$. The reason it is interesting is that it arises from adding independent *r.v.*'s together. First, note that

$$\begin{aligned} c_k &= \sum_{i=0}^{k} P(X = i) P(Y = k - i) \\ &= \sum_{i=0}^{k} P(X = i, Y = k - i) \qquad \text{by } \perp\!\!\!\perp \\ &= P(X + Y = k) \end{aligned}$$

Note that the last step uses the fact that X, Y don't take negative values.

Thm: If $X_1, \ldots, X_n$ are $\perp\!\!\!\perp$ with generating functions $g_1, \ldots, g_n$, then $X_1 + \ldots + X_n$ has generating function $\prod g_i$.

Example: For a single die the a_i's are all $1/6$ so

$$g(z) = \frac{1}{6}\sum_{i=1}^{6} z^i = \frac{z(1-z^6)}{6(1-z)}$$

The generating function for the sum of 3 dice is:

$$\begin{aligned} g^3(z) &= \frac{z^3(1-z^6)^3}{6^3(1-z)^3} = \frac{z^3}{6^3}(1-3z^6+3z^{12}-z^{18})(1-z)^{-3} \\ &= \frac{z^3}{6^3}(1-3z^6+3z^{12}-z^{18})\sum_{k=0}^{\infty}\binom{k+2}{2} z^k. \end{aligned}$$

Now to get, e.g., the probability the 3 dice yield the sum 9 we must determine the coefficient on z^9, this means either $k = 6$ or $k = 0$ for the last term multiplied into the first and second terms respectively so we have

$$c_9 = \frac{1}{6^3}\left(1 \cdot \binom{6+2}{2} - 3 \cdot \binom{0+2}{2}\right) = \frac{28-3}{6^3} = \frac{25}{6^3}$$

Note that this is fully automatic, meaning that it is easily implementable in *Mathematica* among other things. Three dice problems like this are tractable by enumeration, but just barely at the margin of human patience. If you like this sort of mathematics you should look at Graham, Knuth, and Patashnik (1989) which contains a wealth of it.

Now note that we can write, in a slightly fancier notation,

$$g(x) = Ez^X = \sum_{i=0}^{\infty} a_i z^i$$

thus we can now revisit the last theorem and write

$$\begin{aligned} Ez^{X_1+\ldots+X_n} &= Ez^{X_1} \cdot z^{X_2} \cdot \ldots \cdot z^{X_n} \\ &= Ez^{X_1} \cdot Ez^{X_2} \cdot \ldots \cdot Ez^{X_n} \qquad \text{by } \perp\!\!\!\perp \end{aligned}$$

We will consider a further extension. Suppose X takes arbitrary real values, and consider $0 < z \leq 1$, any such z can be written as $e^{-\lambda}$ for $0 \leq \lambda < \infty$, thus instead of writing

$$Ez^X = Ee^{-\lambda X} \qquad 0 \leq \lambda < \infty$$

so in the previous case,

$$Ee^{-\lambda X} = \sum_{j=0}^{\infty} a_j e^{-j\lambda}$$

but more generally if X takes values x_j with probability p_j, we have,

$$Ee^{-\lambda X} = \sum_j p_j e^{-\lambda x_j}$$

thus if X has density f

$$Ee^{-\lambda X} = \int e^{-\lambda u} f(u) du$$

so we can also deal with continuous case. This formulation is the Laplace Transform of X, of f, or F. If we now replace λ by $i\theta$, we have, $Ee^{-i\theta X}$ and obtain the Fourier Transform or characteristic function of the *r.v.* These notes follow Chung (1978 §6.5), this book is sometimes called "baby Chung" since it is much more elementary that Chung's standard text in probability, but nevertheless contains much useful wisdom.

Chung, K.L. (1978). Elementary Probability Theory with Stochastic Processes, Springer.

Graham, R.L., D.E. Knuth, and O. Patashnik (1989). Concrete Mathematics, Addison-Wesley.

8. *Cumulants.* Rather than moments it is sometimes useful to consider cumulants. The cumulant generating function is simply the logarithm of the mgf,

$$k_Y(t) = \log m_Y(t)$$

$$k_{Y-\mu}(t) = \log m_{Y-\mu}(t) = \log(e^{-\mu t} m_Y(t)) = -\mu t + k_Y(t)$$

The r^{th} cumulant of Y is the coefficient of $t^r/r!$ in the Taylor expansion of $k_Y(t)$, i.e.,

$$\begin{aligned} \sum k_r t^r/r! &= \log(1 + \sum \mu'_r t^r/r!) \\ &= \mu t + \log(1 + \sum \mu_r t^r/r!) \end{aligned}$$

Example: For standard normal $k_Z(t) = \frac{1}{2}t^2$ so $k = \{0, 1, 0, 0, \dots\}$.

Remark: For sums of independent rv's cumulants *add.*

9. Characteristic Functions. The Fourier transform of the df, F is called a characteristic function:

$$\phi(t) = Ee^{itX} = \int e^{itx} dF(x)$$

where, as conventional, $i^2 = -1$ and $e^{itx} = \cos(tx) + i\sin(tx)$. This behaves very much like the mgf, but is defined more generally. Existence follows from $|a+ib|^2 \equiv (a+ib)(a-ib) = a^2 + b^2$, and

$$|Ee^{itX}| \le E|e^{itX}| = E|\cos tX + i \sin tX| = E(\cos^2 tX + \sin^2 tX) = E1 = 1$$

One of the more elegant expressions in mathematics is the identity $e^{2\pi i} = 1$ which is usually attributed to Euler. Verify. An important property of the cf is that it (essentially) determines its corresponding df through the inversion formulae for the density

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt$$

or, more generally,

$$F(x) - F(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx} - e^{-ity}}{-it}\phi(t)dt$$

if x and y are continuity points of F. There is a nice discussion of this in Williams.

Example:

For	$Z \sim \mathcal{N}(0,1)$	$\phi_Z(t) = \exp(-t^2/2)$.
	$Z \sim \mathcal{N}(\mu, \sigma^2)$	$\phi_Z(t) = \exp(it\mu - \sigma^2 t^2/2)$
	$Z \sim \mathcal{N}(\mu, V)$	$\phi_Z(t) = \exp(it'\mu - t'Vt/2)$
	$Z = a$ w.p.1	$\phi(t) = e^{iat}$
	$Z \sim (1-p)\delta_0 + p\delta_1$	$\phi_Z(t) = 1 + p(e^{it} - 1)$
	$Z \sim t_1$	$\phi(t) = e^{-\lvert t\rvert}$.

A crucial property of the characteristic function and the reason that it proves to be such an important tool is that for independent random variables X and Y the characteristic function of $Z = X+Y$ is simply the product of the characteristic functions of X and Y, i.e. $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$. To illustrate how this may prove relevant in econometric modeling, consider a model in which productivity Z is determined by an ability component, X, and a luck component, Y. If we assume that we know that the luck component has a known distribution, say Gaussian, then the question: When can we consistently estimate the ability distribution? can be answered by noting that as long as the characteristic function of the ability distribution isn't zero on an open interval the characteristic function of ability can be obtained simply by dividing the observed productivity chf by the chf of the luck distribution. Then inversion yields the distribution as above. This process of unraveling one distribution from that of sum of two, or more, components is called deconvolution.

Moment Expansions

We begin by recalling some useful facts about *Taylor's expansion.* When we write

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \cdots + \frac{(x-x_0)^n}{n!}f^{(n)}(x_0) + R_n$$

we may write R_n in two possible ways:

(1) $$R_n = \frac{(x-x_0)^{n+1}}{(n+1)!}f^{n+1}(\xi) \qquad \xi \in (x_o, x)$$

or

(2) $$R_n = \int_{x_0}^{x} f^{(n+1)}(t)\frac{(x-t)^n}{n!}dt.$$

The latter form is usually called Taylor expansion with integral remainder. It has the virtue of being *exact* provided of course that the derivatives exist. The following result proves convenient in working with expansions of cfs. It shows that under certain conditions the moments fully characterize the distribution they arise from – since the characterize the cf, but one needs to be careful about interpreting this too generously as some of the problems on the first problem set show.

Theorem: If $E|X|^m < \infty$ for an integer $m > 0$, then setting $\nu_j = EX^j$,

$$\phi(t) = \sum_{j=0}^{m} \frac{(it)^j}{j!}\nu_j + o(t^m).$$

Remark One can also, using the same tricks as in §8 for the cumulant generating function, write the moment expansion in terms of the log characteristic function as,

$$\log \phi(t) = \sum \kappa_j (it)^j/j!$$

or equivalently,

$$\phi(t) = \exp\{\sum \kappa_j (it)^j/j!\}.$$

It is in this form that we will use the expansion for the CLT in L4.

Proof: Whittle claims that by Taylor's theorem,

$$e^{itx} = \sum_{j=0}^{m} \frac{(itx)^j}{j!} + \frac{(it)^m}{(m-1)!}\int_0^1 (e^{itxs}x^m - x^m)(1-s)^{m-1}ds$$

Now he replaces x by X and takes expectations, evaluating the remainder as,

$$\frac{(it)^m}{(m-1)!}\int_0^1 (\phi_m(ts) - \phi_m(0))(1-s)^{m-1}ds$$

Under the conditions $\phi_m(t) = EX^m e^{itX}$ is uniformly continuous so the integral, see e.g., Whittle (pp. 124-5), for small t, is $o(1)$ and thus the remainder is $o(t^m)$.

A more detailed argument from Chung goes like this: begin by showing that $\phi^{(k)}(t) = \int_{-\infty}^{\infty}(ix)^k e^{itx}dF(x)$ To see this consider the case $k = 1$. The result follows from

$$\frac{f(t+h) - f(t)}{h} = \int_{-\infty}^{\infty} \frac{e^{i(t+h)x} - e^{itx}}{h} dF(x),$$

by taking limits as $h \to 0$ and using L'Hopital on the right-hand side. The validity of passing the limit inside the integral is argued as follows. The integrand is bounded by $|x|$, so if $E|x| < \infty$ we can take limit inside the $\int$ to get the result. Uniform continuity in t is proved as follows;

$$\begin{aligned} f(t+h) - f(t) &= \int (e^{i(t+h)x} - e^{itx})dF(x) \\ |f(t+h) - f(t)| &\leq \int |e^{itx}|\ |e^{ihx} - 1|dF(x) \\ &= \int |e^{ihx} - 1|dF(x). \end{aligned}$$

since

$$|e^{itx}| = \cos^2(tx) + \sin^2 t(x) = 1$$

The result for general k then follows by induction.

Next we argue that if $E|X|^k \equiv \mu_k < \infty \quad k \geq 1$, then for $|t| < 1$,

$$\phi(t) = \sum_{j=0}^{k} \frac{(it)^j}{j!}\nu_j + o(|t|^k)$$

From G.H. Hardy's *Course of Pure Mathematics*, p.290, see below, if ϕ has finite k^{th} derivative at $t = 0$,

$$(*) \qquad \phi(t) = \sum \frac{\phi^{(j)}(0)}{j!}t^j + o(|t|^k)$$

Since $\mu_j < \infty \quad 1 \leq j \leq k, \quad \phi^{(j)}(0) = i^j\mu_j$ – this follows from T 6.4.1. So substituting in (*) we have the result.

So the mystery in so far as there is one is really (*). Thus is argued as follows in Hardy.

Suppose $f(x)$ has n derivatives $f'(0), \ldots, f^{(n)}(0)$ at $x = 0$. Existence of $f^{(k)}(x)$ at any point x_0 entails existence of $f^{(k-1)}(x)$ in an interval containing x_0. Let $h > 0$, and consider

$$R_n(h) = f(h) - f(0) - hf'(0) - \ldots - \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(0)$$

then $R_n(h)$ and its first $n-1$ derivatives vanish at $h = 0$, and $R_n(0) = f^{(n)}(a)$. Now, write

$$G(h) = R_n(h) - \frac{h^n}{n!}(f^{(n)}(0) - \delta)$$

where δ is positive. We have

$$G(0) = 0, G'(0) = 0, \ldots, G^{(n-1)}(0) = 0$$

and

$$G^{(n)}(0) = \delta > 0.$$

Thus $G^{(n-1)}(h)$ is increasing at $h = 0$, and positive for sufficiently small $h > 0$. Next, $G^{(n-2)}(0) = 0$, and $G^{(n-1)}(h) > 0$ for small $h > 0$ so $G^{(n-2)}(h) > 0$ for small $h > 0$. Repeating we find $G^{(n-3)}(h), \ldots, G(h)$ are positive, so

$$R_n(h) > \frac{h^n}{n!}(f^{(n)}(0) - \delta).$$

Similarly,

$$R_n(h) < \frac{h^n}{n!}(f^n(0) + \delta)$$

for small $h > 0$. Treating negative values of h similarly we have

$$f(h) = f(0) + hf'(0) + \ldots + \frac{h^{(n-1)}}{(n-1)!}f^{(n-1)}(0) + \frac{h^n}{n!}(f^{(n)}(0) + \eta)$$

where $\eta \to 0$ as $h \to 0$, or

$$f(h) = f(0) + hf'(0) + \ldots + \frac{h^{(n-1)}}{(n-1)!} f^{(n-1)}(0) + \sigma(h^n)$$

This would follow easily from usual form of Taylor theorem if we just strengthened the assumptions slightly to ensure that $f^{(n)}(h)$ was continuous in a neighborhood of 0. The above result only needs the existence of the derivitive at 0.

10. *Conditional Probability*

The conditional probability of an event B given that an event A has occurred is

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

where by convention $P(B|A) = 0$ when $P(A) = 0$.

But this definition breaks down for uncountable Ω and we need something new. The conventional approach is the Radon-Nikodym theorem which uses the fact that for measures ν, μ if $\nu << \mu$, i.e., ν is absolutely continuous with respect to μ on $\mathcal{A}$[1] there exists a nonnegative function φ such that

$$\nu(A) = \int_A \varphi d\mu \qquad \text{for any} A \in \mathcal{A}.$$

or

$$\frac{d\nu}{d\mu} = \varphi$$

Thus, if, for example, we have (X, Y) with joint density $f(x, y)$ and X with marginal density $g(x)$, then the conditional density

$$\varphi(y|x) = f(x, y)/g(x).$$

This works when limits can be taken *very carefully* but is generally quite problematic. For an alternative approach see Pollard's recent text, *A User's Guide to Measure Theoretic Probability*.

An alternative approach using expectations has been advanced by Whittle who suggests that one may view the conditional expectation of Y given X, which we write as $\varphi(X) = E(Y|X)$ and can be interpreted as a random variable which takes the value $E(Y|X = x)$ with probability $P(X = x)$, can be defined as the scalar function of X which satisfies

$$E[(Y - \varphi(X))H(X)] = 0 \qquad (\perp \text{ condition})$$

for *any* scalar function $H(X)$.

Remark 1: This coincides with the prior definition when Y is discrete. Take $H(X) = I(X = x)$ then the $\perp$ condition yields

$$EYI(X = x) = E\varphi(X)I(X = x)$$

[1] That is ν assigns probability zero to any set that μ assigns probability zero.

which at $X = x$ yields

$$\varphi(X) = \frac{EYI(X = x)}{EI(X = x)} = E(Y|X = x)$$

Remark 2: If $EY^2 < \infty$, then $E(Y|X)$ can be interpreted as the best $\mathcal{L}_2$ approximant of Y in terms of X, i.e., the function $\varphi(X)$ which minimizes

$$(*) \qquad E(Y - \varphi(X))^2$$

The necessity of the $\perp$ condition follows from the fact that if $EY|X$ is a solution, then stationarity of $(*)$ under the perturbation $E(Y|X) + \varepsilon H(X)$ requires it, i.e.,

$$\frac{d}{d\varepsilon} E(Y - E(Y|X) - \varepsilon H(X))^2|_{\varepsilon=0} = 0$$

$$\Rightarrow -2E(Y - EY|X)H(X) = 0$$

which should hold for any perturbation function $H(X)$. This is analogous to the usual arguments in the calculus of variations.

11. *Independence* A crucial concept is the following.

Definition: The r.v.s. $X_1, \ldots, X_n$ on $(\Omega, \mathcal{F}, P)$ are *independent* if for any sets $B_1, \ldots, B_n \in \mathcal{B}$,

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i).$$

As a notational matter it is convenient to use $X \perp\!\!\!\perp Y$ to denote independence, a concept which is stronger than the uncorrelatedness, which we denoted, $X \perp Y$, in the previous section. One can also define independence of σ-fields, i.e., $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ are independent if for any sets $A_1 \in \mathcal{F}_1, \ldots$

$$P(A_1 \cap \cdots \cap A_n) = \prod P(A_i)$$

and $X_1, \ldots, X_n$ are independent iff $\mathcal{F}(X_1), \ldots, \mathcal{F}(X_n)$ are independent. An important implication of independence is that conditional probabilities and marginal probabilities are equal, i.e., generally

$$P(A \cap B) = P(B|A)P(A)$$

but under independence

$$P(A \cap B) = P(A)P(B)$$

so if the events A and B are independent $P(B|A) = P(B)$ which conveys the idea that A isn't informative about B, i.e. knowing A occurred doesn't help in evaluating the probability of B.

12. *Correlation and least squares projection.* Relationship between rv's is often expressed in terms of correlation. Suppose we wish to approximate the r.v. Y by the rv's $X_1, \ldots, X_p$. In particular, we wish to restrict ourselves to approximations of the form

$$\hat{Y} = \sum_{i=1}^{p} X_i \beta_i = X'\beta$$

which minimize

$$\begin{aligned} D(b) &= E(Y - \hat{Y})^2 \\ &= EY^2 - 2EY\hat{Y} + E\hat{Y}^2 \\ &= S_Y - 2b'S_{XY} + b'S_{XX}b \end{aligned}$$

where $S_Y = EY^2$, $S_{XY} = (EYX_i)$, $S_{XX} = E(X_iX_j)$. Any b minimizing $D(\cdot)$ must satisfy

$$S_{XX}b = S_{XY}$$

If the matrix S_{XX} is nonsingular, we may write this as $b = S_{XX}^{-1}S_{XY}$. This is a linear approximation to the conditional expectation $EY|X$, i.e., $X'\beta \approx EY|X\beta$. If one of the X_i's is degenerate, taking the value 1 with probability 1 then the p by p matrix S_{XX} may be interpreted as the covariance matrix of X and S_{XY} as the vector of covariances of Y with X. At the minimum

$$D\hat{\beta} = S_Y - \hat{\beta}'S_{XY} = S_Y - S'_{XY}S_{XX}^{-1}S_{XY}$$

so if X "contains a intercept" S_Y is variance of Y, and $S'_{XY}S_{XX}^{-1}S_{XY}$ is the amount of $V(Y)$ explained by X In the scalar X case

$$\rho^2 = \frac{S_{XY}^2}{S_{XX}S_Y}$$

is proportion of variance of Y explained by X.

13. *Tail Behavior.* For scalar r.v.s X with df F we say F, or X, has an *exponential tail* if

$$\lim_{a\to\infty} \frac{-\log(1 - F(a))}{ca^r} = 1 \qquad \text{for some } c > 0; r > 0$$

and an *algebraic tail* if

$$\lim_{a\to\infty} \frac{-\log(1 - F(a))}{m \log a} = 1 \qquad \text{for some } m > 0$$

Examples:

Exponential df:	$F(a) = 1 - e^{-\lambda a}$	so $c = \lambda, r = 1$.
Gaussian df:	$c = 2, r = 2$	exponential
Student t:	$m = \nu$	algebraic
Pareto:	$F(x) = 1 - (a/x)^\theta$	algebraic with $m = \theta$

The following result clarifies the connection between tail behavior and the existence of moments.

Theorem: For any r.v. X and real number $r > 0$

$$E|X|^r = r \int_0^\infty x^{r-1} P(|X| \geq x) dx$$

Proof: By induction and the following lemma.

Lemma: Let X be a nonnegative random variable with distribution function F, then

$$EX = \int_0^\infty (1 - F(x)) dx$$

Proof: Integrating by parts, for any fixed $c > 0$,

$$\begin{aligned} (*) \int_0^c x f(x) dx &= -\int_0^c F(x) dx + xF(x)|_0^c = \int_0^c (1 - F(x)) dx - c(1 - F(c)) \\ &\leq \int_0^c (1 - F(x)) dx. \end{aligned}$$

And, it is clear that $c(1 - F(c)) \leq \int_c^\infty xF(x)$, hence if $EX < \infty$, $(1 - F(c)) = o(c^{-1})$, and thus letting $c \to \infty$ yields the result. If $EX = \infty$, the result follows from (*).

Another interesting variation on this theme is the following result from Lindsey and Basak (2000).

Theorem Let two distribution functions, F and G, have the same first $2p$ moments:

$$m_i(F) = m_i(G) = m_i \qquad\qquad i = 0, 1, \ldots, 2p$$

where $m_i(F) = \int x^i dF$, and $m_0 = 1$. Then for all x,

$$|F(x) - G(x)| \leq (V_p(x)' M_p^{-1} V_p(x))^{-1}$$

where $V_p(x) = (1, x, x^2, \ldots, x^p)'$ and

$$M_p = \begin{bmatrix} 1 & m_1 & \ldots & & m_p \\ m_1 & m_2 & & & \\ m_2 & \ldots & & & \\ \vdots & & & & \\ m_p & m_{p+1} & \ldots & \ldots & m_{2p} \end{bmatrix}$$

Remark I won't provide a full proof, see Lindsey and Basak, but I will sketch some salient features of the argument.

A general construction – sometimes called the method of moment spaces leads to the following characterization of the bounds. Given any point, x_0, and target df F we can construct a new, discrete $p + 1$ point distribution, $F_p(x)$ with the properties:

(*i*) $F_p(x)$ has mass $\omega_p(x_0)$ at x_0

where

$$\omega_p(x) = (\sum_{K=1}^{P} |P_k(x)|^2)^{-1}$$

and

$$\begin{aligned} P_0(x) &\equiv 1 \\ P_1(x) &= (x - m_1)/\sqrt{D_1} \\ &\vdots \\ P_k(x) &= E_k(x)\sqrt{D_{k-1}D_k} \qquad\qquad k = 2, \ldots, 3 \end{aligned}$$

with

$$D_p = \text{ det } M_p$$

and

$$E_p(x) = \text{ det } \tilde{M}_p(x)$$

where $\tilde{M}_p$ denotes the matrix M_p with the last column replaced by the vector $V_p(x)$.

(ii) $F_p(x)$ has the same first $2p$ moments as F.

(iii) for every df G with given moments $F_p(x^-) \leq G(x) \leq F_p(x)$ at every mass point of F_p.

These discrete $F_p(x)$ distributions constitute the worst case behavior. In particular, for any continuous df G.

$$\max\{G(x_0^-) - F_p(x_0^-), F_p(x_0) - G(_0)\} \geq \frac{\omega_p(x_0)}{2}$$

Note by construction $F_p(x_0) - F_p(x_0^-) = \omega_p(x_0)$ so either $G(x_0)$ falls half way between and the bound is achieved or it falls on one side or the other of halfway and the max{ , } is then strictly greater than the bound.

Since (iii) provides the bound

$$F_p(x_0^-) \leq G(x_0) \leq F_p(x_0)$$

we have that for all x_0

$$|F(x) - G(x)| \leq \omega_p(x).$$

To relate this to the bound given in the *Theorem*, one can show that

$$\begin{aligned} \sum |P_k(x)|^2 &= -D_p^{-1} \text{ det } \begin{bmatrix} 0 & V_p(x)' \\ V_p(x) & M_p \end{bmatrix} \\ &= -D_p^{-1}|M_p| \ |0 - V_p' M_p V_p| \\ &= V_p' M_p V_p \end{aligned}$$

The last steps are trivial – see e.g. Rao (1973, p. 32) the first step isn't, or at least doesn't appear so to me.

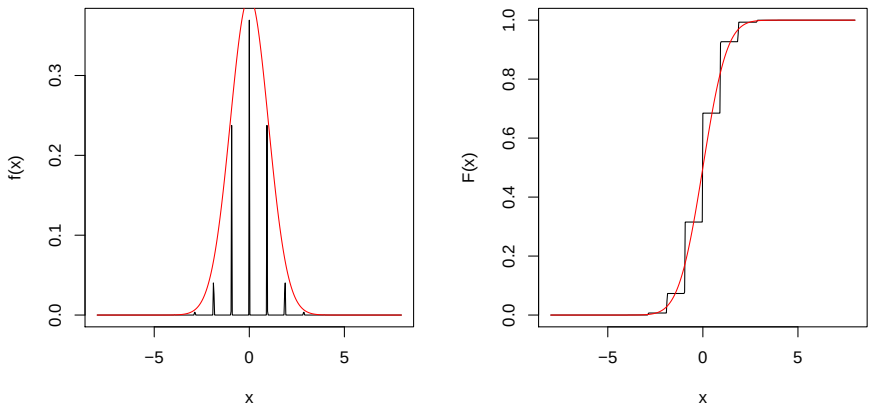

Figure 1: Approximate discrete distributions matching the first 20 moments of the standard Gaussian distribution: The left panel illustrates the location and magnitude of the mass points and the right panel plots the two distribution functions. The distributions differ by about 0.18 at zero, quite a large difference given that they have the same first 20 moments.

It is an interesting numerical problem to find such $F_p(\cdot)$'s. Rough approximations to these least favorable densities can be found by solving for the discrete distribution supported on the grid points: $x_1, \ldots, x_q, 0, x_{q+1}, \ldots x_n$. This produces the following linear programming problem:

$$\min_p \{c^\top p \mid Mp = m_p, p \in \mathcal{S}\}$$

where p is a n vector of probability masses associated with the discrete points, M is the moment matrix with typical element x_i^k with $k = 1, \ldots, K$, m_p is the K vector of moments of a standard Gaussian random variable, $c = (1_q, 0_{q+1})$, and $\mathcal{S}$ is the $n-1$ dimensional simplex. Matching $K = 20$ moments of the normal distribution yields a discrete distribution with the mass points in the left panel of Figure 1 and distribution function as shown in the right panel. Evaluating at zero, the discrete distribution differs from the normal df by about 0.18, quite a large difference given that we have matched so many moments. A convenient strategy for computing these solutions can be formulated in Mathematica, where it is possible to do "exact" linear programming. Recall that once we have an optimal basis for the LP, the solution simply requires a solution to a linear system, and for problems defined on the rationals the solutions are also in the rationals. Further details appear in a forthcoming paper by Steve Portnoy in *Am. Statistician.*

One can conclude from this that the moments do characterize well the tails of the distribution – at least when the moments are finite, but they do a poor job of characterizing the middle of the distribution as our prior examples illustrate.

Lindsey, B. and P. Basak (2000). Moments determine the tail of a Distribution (but not

much else). *Am. Statistician*, 54, 248-52.