Fundamental Probability: A Computational Approach

Marc S. Paolella

University of Zürich

Course Outline

Course Outline

The Factorial

- The number of ways that n distinguishable objects can be ordered is given by
$$n\,(n-1)\,(n-2)\ldots 2\cdot 1 = n!,$$
pronounced "n factorial".
- The number of ways that k objects can be chosen from n, $k \leq n$, when order is relevant is
$$n\,(n-1)\ldots(n-k+1) =: n_{[k]} = \frac{n!}{(n-k)!},$$
which is referred to as the **falling, or descending factorial**.
- Similarly, we denote the **rising, or ascending factorial**, by
$$n^{[k]} = n\,(n+1)\ldots(n+k-1)\,.$$

The Binomial Coefficient

- If the order of the k objects is irrelevant, then $n_{[k]}$ is adjusted by dividing by $k!$, the number of ways of arranging the k chosen objects. Thus, the total number of ways is

$$\frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!\,k!} =: \binom{n}{k},$$

which is pronounced "n choose k" and referred to as a **binomial coefficient** for reasons which will become clear below.

- Notice that, both algebraically and intuitively, $\binom{n}{k} = \binom{n}{n-k}$.

Example From a group of two boys and three girls, how many ways can two be chosen such that at least one boy is picked? There are $\binom{2}{1}\binom{3}{1} = 6$ ways with exactly one boy, and $\binom{2}{2}\binom{3}{0} = 1$ ways with exactly two boys, or 7 altogether. Alternatively, from the $\binom{5}{2} = 10$ total possible combinations, we can subtract $\binom{3}{2} = 3$ ways such that no boy is picked.

A Useful Identity

- A very useful identity is

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}, \quad k < n. \qquad ✠$$

- It follows because

$$\begin{aligned} \binom{n}{k} &= \frac{n!}{(n-k)!\,k!} \cdot 1 = \frac{n!}{(n-k)!\,k!} \cdot \left(\frac{n-k}{n} + \frac{k}{n} \right) \\ &= \frac{(n-1)!}{(n-k-1)!\,k!} + \frac{(n-1)!}{(n-k)!\,(k-1)!} = \binom{n-1}{k} + \binom{n-1}{k-1}. \end{aligned}$$

- Some intuition: For any single particular object, either it is among the chosen k or not. First include it in the k choices so that there are $\binom{n-1}{k-1}$ ways of picking the remaining $k-1$. Alternatively, if the object is not one of the k choices, then there are $\binom{n-1}{k}$ ways of picking k objects. As these two situations exclude one another yet cover all possible situations, their sum equals the total number of possible combinations.

A Useful Identity: Pascal's Triangle

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1
1 10 45 120 210 252 210 120 45 10 1
1 11 55 165 330 462 462 330 165 55 11 1
1 12 66 220 495 792 924 792 495 220 66 12 1
1 13 78 286 715 1287 1716 1716 1287 715 286 78 13 1
1 14 91 364 1001 2002 3003 3432 3003 2002 1001 364 91 14 1
1 15 105 455 1365 3003 5005 6435 6435 5005 3003 1365 455 105 15 1
1 16 120 560 1820 4368 8008 11440 12870 11440 8008 4368 1820 560 120 16 1

A Useful Identity (2)

- By applying the identity $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$ recursively,

$$\begin{aligned}
\binom{n}{k} &= \binom{n-1}{k} + \binom{n-1}{k-1} \\
&= \binom{n-1}{k} + \binom{n-2}{k-1} + \binom{n-2}{k-2} \\
&= \binom{n-1}{k} + \binom{n-2}{k-1} + \binom{n-3}{k-2} + \binom{n-3}{k-3} \\
&\vdots \\
&= \sum_{i=0}^{k} \binom{n-i-1}{k-i}, \quad k < n.
\end{aligned}$$

The Multinomial Coefficient

- If a set of n distinct objects is to be divided into k distinct groups, whereby the size of each group is n_i, $i = 1, \ldots, k$ and $\sum_{i=1}^{k} n_i = n$, then the number of possible divisions is given by

$$\binom{n}{n_1, n_2, \ldots, n_k} := \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_k}{n_k} = \frac{n!}{\prod_i n_i!}.$$

- This reduces to the familiar combinatoric when $k = 2$:

$$\binom{n}{n_1, n_2} = \binom{n}{n_1}\binom{n-n_1}{n_2} = \binom{n}{n_1}\binom{n_2}{n_2} = \binom{n}{n_1} = \frac{n!}{n_1!\,(n-n_1)!}$$

Example A small factory employs 15 people and produces three goods on three separate assembly lines, A,B and C. Lines A and B each require six people, line C needs three. How many ways can the workers be arranged?

$$\binom{15}{6}\binom{15-6}{6}\binom{15-6-6}{3} = \frac{15!}{6!9!}\,\frac{9!}{6!3!} = \frac{15!}{6!6!3!} = 420420.$$

Further Examples: Flags

- Imagine you have a set of flags, which are the same size, shape, etc., and differ only with respect to color.
- If you have 2 red and 3 green flags, and you arrange all 5 in a row, how many "signals" can you make?
- Of the 5 positions, you must choose 2 to be red, so $\binom{5}{2} = 10$.
- Notice that $\binom{5}{2} = 5!/(2!3!)$, so we can interpret the answer as being the $5! = 120$ possible orderings of the 5 flags, and then adjusted for the redundancy by dividing by 2! to account for the fact that we cannot differentiate between the two red flags, and similarly with green.
- From a set of 12 flags, 4 blue, 4 red, 2 yellow, and 2 green, all hung out in a row, the number of different "signals" of length 12 you could produce is given by $12! \,/\, (4!\,4!\,2!\,2!) = 207900$.

Further Examples: Marbles

- From a set of 2 red and 3 green marbles, how many different nonempty combinations can be placed into an urn? (In other words, the ordering does not count.)
- Letting R denote red and G denote green, all possibilities are: of size 1, R, G; of size 2, RR, RG, GG; of size 3, RRG, RGG, GGG; of size 4, RRGG, RGGG; and of size 5, RRGGG; for a total of 11.
- This total can be obtained without enumeration by observing that, in each possible combination, either there are 0,1, or 2 reds, and 0,1,2, or 3 greens, or, multiplying, $(2+1)\cdot(3+1) = 12$, but minus one, because that would be the zero-zero combination.
- In general, if there are $n = \sum_{i=1}^{k} n_i$ marbles, where the size of each distinguishable group is n_i, $i = 1, \ldots, k$, then there are $\prod_{i=1}^{k}(n_i + 1) - 1$ combinations ignoring the ordering.

The Binomial Theorem

- The relation

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

is referred to as (Newton's) binomial theorem. It is a simple, yet fundamental result which arises in numerous applications.

Examples

$$(x+(-y))^2 = x^2 - 2xy + y^2, \qquad (x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3,$$

$$0 = (1-1)^n = \sum_{i=0}^{n} \binom{n}{i} (-1)^{n-i}, \qquad 2^n = (1+1)^n = \sum_{i=0}^{n} \binom{n}{i}.$$

Example from Calculus

- Let $f(x) = x^n$ for $n \in \mathbb{N}$. The binomial theorem implies

$$f(x+h) = (x+h)^n = \sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i = x^n + nhx^{n-1} + \cdots + h^n,$$

so that

$$\begin{aligned} f'(x) &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h \to 0} \left(nx^{n-1} + \cdots + h^{n-1} \right) \\ &= nx^{n-1}. \end{aligned}$$

Induction

- Induction is a method of proof for identities which are a function of an integer variable, say n, and first entails verifying the conjecture for $n = 1$, and then demonstrating that it holds for $n + 1$, assuming it holds for n.

Example

- Let $S_n = \sum_{k=0}^{n} \binom{n}{k}$. We wish to prove $S_n = 2^n$. From (✠),

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \left[\binom{n-1}{k} + \binom{n-1}{k-1} \right].$$

Using the fact that $\binom{m}{i} = 0$ for $i > m$,

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n-1} \binom{n-1}{k} + \sum_{k=1}^{n} \binom{n-1}{k-1} = 2^{n-1} + \sum_{k=1}^{n} \binom{n-1}{k-1},$$

and, with $j = k - 1$, the latter term is

$$\sum_{k=1}^{n} \binom{n-1}{k-1} = \sum_{j=0}^{n-1} \binom{n-1}{j} = 2^{n-1},$$

so that $\sum_{k=0}^{n} \binom{n}{k} = 2^{n-1} + 2^{n-1} = 2\left(2^{n-1}\right) = 2^n$.

Proof of Binomial Theorem

The theorem clearly holds for $n = 1$. Assuming it holds for $n - 1$,

$$\begin{aligned}(x+y)^n &= (x+y)(x+y)^{n-1} = (x+y)\sum_{i=0}^{(n-1)}\binom{n-1}{i}x^i y^{(n-1)-i} \\ &= \sum_{i=0}^{n-1}\binom{n-1}{i}x^{i+1}y^{n-(i+1)} + \sum_{i=0}^{n-1}\binom{n-1}{i}x^i y^{n-1-i+1}.\end{aligned}$$

Then, with $j = i + 1$,

$$\begin{aligned}(x+y)^n &= \sum_{j=1}^{n}\binom{n-1}{j-1}x^j y^{n-j} + \sum_{i=0}^{n-1}\binom{n-1}{i}x^i y^{n-i} \\ &= x^n + \sum_{j=1}^{n-1}\binom{n-1}{j-1}x^j y^{n-j} + \sum_{i=1}^{n-1}\binom{n-1}{i}x^i y^{n-i} + y^n \\ &= x^n + \sum_{i=1}^{n-1}\left\{\binom{n-1}{i-1} + \binom{n-1}{i}\right\}x^i y^{n-i} + y^n \\ &= x^n + \sum_{i=1}^{n-1}\binom{n}{i}x^i y^{n-i} + y^n = \sum_{i=0}^{n}\binom{n}{i}x^i y^{n-i}. \qquad \blacksquare\end{aligned}$$

Example of Induction with Co-Skewness Matrices

There is an enormous literature on the effect and use of skewness in asset allocation / portfolio selection. The concept of *co-skewness* is central to the discussion.

Recall that, if $\mathbf{R} = (R_1, \ldots, R_n)'$ denotes the vector of n individual asset returns, then $\mathbb{E}[\mathbf{R}]$ is the n-length **vector** of means with ith element $\mu_i := \mathbb{E}[R_i]$, and $\mathrm{Var}(\mathbf{R})$ is the $n \times n$ **matrix** of covariances with (ij)th element

$$\sigma_{ij} = \mathrm{Cov}\left(R_i, R_j\right) = \mathbb{E}\left[\left(R_i - \mu_i\right)\left(R_j - \mu_j\right)\right]$$

and $\sigma_i^2 = \sigma_{ii}$.

For $n = 3$,

$$\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{R}) = \mathbb{E}\left[(\mathbf{R} - \boldsymbol{\mu})(\mathbf{R} - \boldsymbol{\mu})'\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}.$$

Let $\mathbf{a}$ be a vector of portfolio weights. The variance of portfolio $P = \mathbf{a}'\mathbf{R}$ is $\mathrm{Var}(P) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$.

Example of Induction with Co-Skewness Matrices

The extension of the variance to third-order cross products, i.e., the co-skewness, would involve an $n \times n \times n$ **cube**, with (ijk)th element $s_{ijk} = \mathbb{E}\left[(R_i - \mu_i)\left(R_j - \mu_j\right)(R_k - \mu_k)\right]$. It is more far more useful to work with a matrix version of this expression obtained by placing the n matrices of size $n \times n$ side by side in a particular, agreed-upon order.

This is denoted by M_3 (M is for moment, and 3 indicates the order; M_4 is used for kurtosis, etc.) and given by

$$M_3 = \mathbb{E}\left[(\mathbf{R} - \boldsymbol{\mu})(\mathbf{R} - \boldsymbol{\mu})' \otimes (\mathbf{R} - \boldsymbol{\mu})'\right].$$

Recall: Let $\mathbf{A} = \left(a_{ij}\right)$ be an $m \times n$ matrix and $\mathbf{B}$ a $p \times q$ matrix. Then the (right) Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is the $mp \times nq$ block matrix given by

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}.$$

This is known as tensor notation and is used throughout the literature.[1]

[1] See, e.g., Gustavo M. de Athaydea and Renato G. Flôres, Jr., *Finding a Maximum Skewness Portfolio—A General Solution to Three-Moments Portfolio Choice*, Journal of Economic Dynamics and Control, 2004, Vol 28(7), 1335–1352; and Eric Jondeau, Ser-Huang Poon and Michael Rockinger, *Financial Modeling Under Non-Gaussian Distributions*, Springer, 2007, Section 2.5.

Example of Induction with Co-Skewness Matrices

Note that $(\mathbf{R}-\mu)(\mathbf{R}-\mu)'$ is $n \times n$ (and whose expectation is $\mathrm{Var}(\mathbf{R})$), so that $(\mathbf{R}-\mu)(\mathbf{R}-\mu)' \otimes (\mathbf{R}-\mu)'$ is $n \times n^2$.

To illustrate, for $n=4$, with $s_{ijk} = \mathbb{E}\left[(R_i-\mu_i)\left(R_j-\mu_j\right)(R_k-\mu_k)\right]$, M_3 is given by

$$\begin{aligned} M_3 &= \mathbb{E}\left[(\mathbf{R}-\boldsymbol{\mu})(\mathbf{R}-\boldsymbol{\mu})' \otimes (\mathbf{R}-\boldsymbol{\mu})'\right] \\ &= \left[\begin{array}{c|c|c|c} S_{1jk} & S_{2jk} & S_{3jk} & S_{4jk} \end{array}\right], \quad S_{ijk} = \begin{bmatrix} s_{i11} & s_{i12} & s_{i13} & s_{i14} \\ s_{i21} & s_{i22} & s_{i23} & s_{i24} \\ s_{i31} & s_{i32} & s_{i33} & s_{i34} \\ s_{i41} & s_{i42} & s_{i43} & s_{i44} \end{bmatrix}. \end{aligned}$$

Writing the entire M_3 matrix out but just writing the indices (leave off the s) and marking the i index with boldface, we have

$$\left[\begin{array}{cccc|cccc|cccc|cccc} \mathbf{1}11 & \mathbf{1}12 & \mathbf{1}13 & \mathbf{1}14 & \mathbf{2}11 & \mathbf{2}12 & \mathbf{2}13 & \mathbf{2}14 & \mathbf{3}11 & \mathbf{3}12 & \mathbf{3}13 & \mathbf{3}14 & \mathbf{4}11 & \mathbf{4}12 & \mathbf{4}13 & \mathbf{4}14 \\ \mathbf{1}21 & \mathbf{1}22 & \mathbf{1}23 & \mathbf{1}24 & \mathbf{2}21 & \mathbf{2}22 & \mathbf{2}23 & \mathbf{2}24 & \mathbf{3}21 & \mathbf{3}22 & \mathbf{3}23 & \mathbf{3}24 & \mathbf{4}21 & \mathbf{4}22 & \mathbf{4}23 & \mathbf{4}24 \\ \mathbf{1}31 & \mathbf{1}32 & \mathbf{1}33 & \mathbf{1}34 & \mathbf{2}31 & \mathbf{2}32 & \mathbf{2}33 & \mathbf{2}34 & \mathbf{3}31 & \mathbf{3}32 & \mathbf{3}33 & \mathbf{3}34 & \mathbf{4}31 & \mathbf{4}32 & \mathbf{4}33 & \mathbf{4}34 \\ \mathbf{1}41 & \mathbf{1}42 & \mathbf{1}43 & \mathbf{1}44 & \mathbf{2}41 & \mathbf{2}42 & \mathbf{2}43 & \mathbf{2}44 & \mathbf{3}41 & \mathbf{3}42 & \mathbf{3}43 & \mathbf{3}44 & \mathbf{4}41 & \mathbf{4}42 & \mathbf{4}43 & \mathbf{4}44 \end{array}\right].$$

Example of Induction with Co-Skewness Matrices

The skewness of the portfolio $\mathbf{a}'\mathbf{R}$ is $s_P^3 = \mathbf{a}' M_3 (\mathbf{a} \otimes \mathbf{a})$.
To confirm this for $n = 2$, we have: $P = \mathbf{a}'\mathbf{R} = a_1 R_1 + a_2 R_2$, and, assuming $E[\mathbf{R}] = \mathbf{0}$ for simplicity, we have

$$\begin{aligned} \mathrm{Skew}(P) &= \mathrm{Skew}(a_1 R_1 + a_2 R_2) = \mathbb{E}\left[(a_1 R_1 + a_2 R_2)^3\right] \\ &= \mathbb{E}\left[R_1^3 a_1^3 + R_2^3 a_2^3 + 3 R_1 R_2^2 a_1 a_2^2 + 3 R_1^2 R_2 a_1^2 a_2\right] \\ &= a_1^3 s_{111} + a_2^3 s_{222} + 3 a_1 a_2^2 s_{122} + 3 a_1^2 a_2 s_{112}. \end{aligned}$$

We see this is indeed equal to

$$\begin{aligned} s_P^3 &= \mathbf{a}' M_3 (\mathbf{a} \otimes \mathbf{a}) \\ &= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} s_{111} & s_{112} & s_{211} & s_{212} \\ s_{121} & s_{122} & s_{221} & s_{222} \end{pmatrix} \begin{pmatrix} a_1 a_1 \\ a_1 a_2 \\ a_2 a_1 \\ a_2 a_2 \end{pmatrix} \\ &= a_1^3 s_{111} + a_2^3 s_{222} + a_1^2 a_2 s_{112} + a_1^2 a_2 s_{121} + a_1^2 a_2 s_{211} + a_1 a_2^2 s_{122} + a_1 a_2^2 s_{212} + a_1 a_2^2 s_{221} \\ &= a_1^3 s_{111} + a_2^3 s_{222} + 3 a_1^2 a_2 s_{112} + 3 a_1 a_2^2 s_{122}. \end{aligned}$$

Example of Induction with Co-Skewness Matrices

Recall that $\mathrm{Var}(\mathbf{R})$ is symmetric, and so there are not N^2 unique elements. Without the diagonal, there are $n^2 - n$ elements, of which half are unique. Adding back on the diagonal gives

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

unique elements.

Another way to see this is to observe that the unique elements are on the diagonal and upper diagonal part of the matrix, of which there are $n + (n-1) + (n-2) + \cdots + 1 = (n+1)\,n/2$ elements.

Prove that the number of unique elements in M_3 is $\binom{n+2}{3}$.

Make a computer program (in matlab, call it `function M3 = coskewness(n)`) which calculates the M_3 matrix taking the redundancies into account, and for now, just fill the ijkth entry with $100i + 10j + k$. It should also confirm that the number of nonredundant terms to compute is $\binom{n+2}{3}$.

Example of Induction with Co-Skewness Matrices

Working with the $n = 4$ case for illustration, observe that S_{1jk} has, like the covariance matrix, $\binom{n+1}{2} = \binom{5}{2} = 10$ unique elements. Now consider S_{2jk}. We would expect that its first row (for which $j = 1$) and first column (for which $k = 1$) will have redundancies with elements in S_{1jk} (for which $i = 1$). (Of course, S_{ijk} is a symmetric matrix, so its jth row is just the transpose of the jth column.) Inspection shows that the first row of S_{2jk} is indeed equivalent to the second row of S_{1jk}. We see this clearly in the following expression for $\left[\ S_{1jk}\ \middle|\ S_{2jk}\ \right]$

$$\left[\begin{array}{cccc|cccc}
111 & \boxed{112} & 113 & 114 & \boxed{211} & \boxed{212} & \boxed{213} & \boxed{214} \\
\boxed{121} & \boxed{122} & \boxed{123} & \boxed{124} & \boxed{221} & 222 & 223 & 224 \\
131 & \boxed{132} & 133 & 134 & \boxed{231} & 232 & 233 & 234 \\
141 & \boxed{142} & 143 & 144 & \boxed{241} & 242 & 243 & 244
\end{array}\right]$$

Thus, from S_{2jk}, we only need to consider the principle submatrix obtained by "striking out" the first row and column. There are 6 unique elements here, or $\binom{n}{2}$.

Example of Induction with Co-Skewness Matrices

Continuing with S_{3jk}, we would expect that its first row (and column) (for which $j = 1$ or $k = 1$), will have redundancies with elements in S_{1jk} (for which $i = 1$). In particular, the first row (and column) of S_{3jk} should be equivalent to the third row (and column) of S_{1jk}. Likewise, we expect the second row (and column) of S_{3jk} to be equivalent to the third row (and column) of S_{2jk}. We mark these in the following expression for $\left[\begin{array}{c|c|c} S_{1jk} & S_{2jk} & S_{3jk} \end{array} \right]$

$$\left[\begin{array}{cccc|cccc|cccc}
111 & 112 & \boxed{113} & 114 & 211 & 212 & \boxed{213} & 214 & \boxed{311} & \boxed{312} & \boxed{313} & \boxed{314} \\
121 & 122 & \boxed{123} & 124 & 221 & 222 & \boxed{223} & 224 & \boxed{321} & \boxed{322} & \boxed{323} & \boxed{324} \\
\boxed{131} & \boxed{132} & \boxed{133} & \boxed{134} & \boxed{231} & \boxed{232} & \boxed{233} & \boxed{234} & \boxed{331} & \boxed{332} & 333 & 334 \\
141 & 142 & \boxed{143} & 144 & 241 & 242 & \boxed{243} & 244 & \boxed{341} & \boxed{342} & 343 & 344
\end{array} \right].$$

Thus, from S_{3jk}, we only need to consider the principle submatrix obtained by striking out the first two rows and columns, leaving 3 unique elements, or $\binom{n-1}{2}$. Continuing this argument, we see that M_3 for $n = 4$ contains

$$\begin{aligned}
\binom{n+1}{2} + \binom{n}{2} + \binom{n-1}{2} + \binom{n-2}{2} &= \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} \\
&= 10 + 6 + 3 + 1 = 20 = \binom{n+2}{3}
\end{aligned}$$

unique elements, as claimed.

Example of Induction with Co-Skewness Matrices

Applying the same argument to the general n case, M_3 has

$$\sum_{i=1}^{n} \binom{n+2-i}{2}$$

unique elements. We now wish to prove that

$$\binom{n+2}{3} = \sum_{i=1}^{n} \binom{n+2-i}{2}. \tag{1}$$

This is best done by induction.

It is easy to see that it holds for $n = 1$ and $n = 2$. Assume it holds for n. Then, recalling the facts that

$$\binom{a}{b} = \binom{a-1}{b} + \binom{a-1}{b-1} \quad \text{and} \quad \sum_{i=1}^{n} i = \frac{(n+1)\,n}{2},$$

we have, for the $n+1$ case, ...

Example of Induction with Co-Skewness Matrices

$$
\begin{aligned}
\sum_{i=1}^{(n+1)} \binom{(n+1)+2-i}{2} &= \sum_{i=1}^{n+1} \binom{n+3-i}{2} \\
&= \sum_{i=1}^{n+1} \binom{n+2-i}{2} + \sum_{i=1}^{n+1} \binom{n+2-i}{1} \\
&= \sum_{i=1}^{n} \binom{n+2-i}{2} + \sum_{i=1}^{n+1} (n+2-i) \\
&= \binom{n+2}{3} + (n+2)(n+1) - \frac{(n+2)(n+1)}{2} \\
&= \binom{n+2}{3} + \frac{(n+2)(n+1)}{2} \\
&= \binom{n+2}{3} + \binom{n+2}{2} \\
& \binom{n+3}{3},
\end{aligned}
$$

as was to be shown.

Example of Induction with Co-Skewness Matrices

Another way to prove that the number of unique elements in M_3 is $\binom{n+2}{3}$ is as follows.

Before solving the problem for M_3, let's do it again for the variance-covariance matrix $\mathrm{Var}\left(\mathbf{R}\right)$. Above, we saw that a way to see that the total number of unique elements in $\mathrm{Var}\left(\mathbf{R}\right)$ is is to note that the unique elements are on the diagonal and upper diagonal part (or lower diagonal part) of the matrix, of which there are $n + (n-1) + (n-2) + \cdots + 1 = (n+1)\,n/2$ elements.

What we have essentially done is to say that, for each of $i = 1, 2, \ldots, n$ rows, we want to count those elements in the jth column for which $j \leq i$. Thus, we want

$$\sum_{i=1}^{n} \sum_{j=1}^{i} 1 = \sum_{i=1}^{n} i = \frac{n\,(n+1)}{2}.$$

Example of Induction with Co-Skewness Matrices

Similarly, for M_3, we need to "run through" all $i = 1, \ldots, n$ but then only through the $j \leq i$ and $k \leq j$.
Note that, if you take any $i, j, k \in \{1, \ldots, n\}$, then $S_{i,j,k} = S_{i^*,j^*,k^*}$, where $i^* = \max(i, j, k)$, j^* is the second largest, and k^* is the last one remaining. Thus, $1 \leq i^* \leq n$, $j^* \leq i^*$, and $k^* \leq j^*$. So, we want

$$\sum_{i=1}^{n} \sum_{j=1}^{i} \sum_{k=1}^{j} 1 = \sum_{i=1}^{n} \sum_{j=1}^{i} j = \sum_{i=1}^{n} \frac{1}{2} i(i+1) = \frac{1}{2} \sum_{i=1}^{n} \left(i^2 + i\right)$$
$$= \frac{1}{2} \frac{n(n+1)(2n+1)}{6} + \frac{1}{2} \frac{n(n+1)}{2}.$$

Simplifying this yields

$$\frac{n(n+1)(n+2)}{6} = \binom{n+2}{3}.$$

Example of Induction with Co-Skewness Matrices

The co-kurtosis matrix, M_4, extends M_3 in a natural way, and is defined as

$$M_4 = \mathbb{E}\left[(R-\mu)(R-\mu)' \otimes (R-\mu)' \otimes (R-\mu)'\right],$$

with elements

$$\kappa_{ijk\ell} = \mathbb{E}\left[(R_i-\mu_i)\left(R_j-\mu_j\right)(R_k-\mu_k)(R_\ell-\mu_\ell)\right].$$

The number of unique elements is

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{i}\sum_{k=1}^{j}\sum_{\ell=1}^{k} 1 &= \sum_{i=1}^{n}\sum_{j=1}^{i}\sum_{k=1}^{j} k = \sum_{i=1}^{n}\sum_{j=1}^{i}\frac{j\,(j+1)}{2} \\ &= \sum_{i=1}^{n}\frac{i\,(i+1)\,(i+2)}{6} \\ &= \frac{n\,(n+1)\,(n+2)\,(n+3)}{24} = \binom{n+3}{4},\end{aligned}$$

where we used the fact that (see pages 21-22 of the text for derivation)

$$\sum_{k=1}^{n} k^3 = \frac{n^2\,(n+1)^2}{4}.$$

Vandermonde's Theorem

This states that, for $0 < k \leq \min(N, M)$,

$$\begin{aligned}\binom{N+M}{k} &= \sum_{i=0}^{k}\binom{N}{i}\binom{M}{k-i} \\ &= \binom{N}{0}\binom{M}{k} + \binom{N}{1}\binom{M}{k-1} + \cdots + \binom{N}{k}\binom{M}{0}.\end{aligned}$$

Proof I (intuitive): Assume $N \geq k$ and $M \geq k$. Then either

- all k objects are chosen from the group of size M and none from the group of size N, or
- $k-1$ objects are chosen from the group of size M and 1 from the group of size N, or

$$\vdots \quad \vdots \quad \vdots$$

- no objects are chosen from the group of size M and all k from the group of size N.

Summing these disjoint events yields the desired formula.

Vandermonde's Theorem: Proof II

- This algebraic proof follows from a more general result:
- A non–obvious extension of the binomial theorem is

$$(x+y)^{[n]} = \sum_{i=0}^{n} \binom{n}{i} x^{[i]} y^{[n-i]}, \quad x^{[n]} = \prod_{j=0}^{n-1} (x + ja), \qquad (2)$$

  for $n = 0, 1, 2, \ldots$, and x, y, a are real numbers.
- It holds trivially for $n = 0$, and is easy to see for $n = 1$ and $n = 2$, but otherwise appears difficult to verify, and induction gets messy and doesn't (seem to) lead anywhere.
- It can be proven using properties of the beta function, introduced below.

Vandermonde's Theorem: Proof II

With $a = -1$,

$$k^{[n]} = (k)(k-1)(k-2)\cdots(k-(n-1)) = \frac{k!}{(k-n)!} = \binom{k}{n} n!$$

so that (2) reads

$$\binom{x+y}{n} n! = \sum_{i=0}^{n} \binom{n}{i} \binom{x}{i} i! \binom{y}{n-i} (n-i)! = n! \sum_{i=0}^{n} \binom{x}{i} \binom{y}{n-i}$$

or

$$\binom{x+y}{n} = \sum_{i=0}^{n} \binom{x}{i} \binom{y}{n-i},$$

which is Vandermonde's theorem.

Vandermonde's Theorem: Proof III

- Equate coefficients of x^r from each side of the identity

$$(1+x)^{N+M} = (1+x)^N (1+x)^M$$

and use the binomial theorem to get

$$(1+x)^{N+M} = \sum_{i=0}^{N+M} \binom{N+M}{i} x^i, \tag{3}$$

with the coefficient of x^r, $0 \leq r \leq N+M$ being $\binom{N+M}{r}$; and

$$\begin{aligned}(1+x)^N (1+x)^M &= \left(\sum_{j=0}^{N} \binom{N}{j} x^j\right) \left(\sum_{k=0}^{M} \binom{M}{k} x^k\right) \\ &=: \left(a_0 + a_1 x + \cdots a_N x^N\right) \left(b_0 + b_1 x + \cdots b_M x^M\right),\end{aligned}$$

where $a_j = \binom{N}{j}$ and $b_k = \binom{M}{k}$ are defined for convenience.

Vandermonde's Theorem: Proof III

- Again:

$$(1+x)^N (1+x)^M = \left(\sum_{j=0}^{N} \binom{N}{j} x^j \right) \left(\sum_{k=0}^{M} \binom{M}{k} x^k \right)$$
$$=: \left(a_0 + a_1 x + \cdots a_N x^N\right) \left(b_0 + b_1 x + \cdots b_M x^M\right), \tag{4}$$

  where $a_j = \binom{N}{j}$ and $b_k = \binom{M}{k}$.
- Observe that the coefficient of x^0 in (4) is $a_0 b_0$; the coefficient of x^1 is $a_0 b_1 + a_1 b_0$; and, in general, the coefficient of x^r is $a_0 b_r + a_1 b_{r-1} + \cdots + a_r b_0$, which is valid for $r \in \{0, 1, \ldots, N+M\}$ if we define $a_j := 0$ for $j > N$ and $b_k := 0$ for $k > M$.
- Thus, the coefficient of x^r is

$$\sum_{i=0}^{r} a_i b_{r-i} = \sum_{i=0}^{r} \binom{N}{i} \binom{M}{r-i}.$$

Vandermonde's Theorem: Proof III

- Again: The coefficient of x^r is

$$\sum_{i=0}^{r} a_i b_{r-i} = \sum_{i=0}^{r} \binom{N}{i} \binom{M}{r-i}.$$

- From (3) above, recall that the coefficient of x^r is $\binom{N+M}{r}$.
- Thus, it follows that

$$\binom{N+M}{r} = \sum_{i=0}^{r} \binom{N}{i} \binom{M}{r-i}, \quad 0 \leq r \leq N+M,$$

as was to be shown.

A fourth proof is given in the text.

Course Outline

The Gamma Function

- The gamma function is a smooth function (it is continuous as are all its derivatives) of one parameter, say a, on $\mathbb{R}_{>0}$ which coincides with the factorial function when $a \in \mathbb{N}$.
- We take as the definition of the gamma function the improper integral expression of one parameter

$$\Gamma(a) := \int_0^\infty x^{a-1} e^{-x} \, \mathrm{d}x, \quad a \in \mathbb{R}_{>0}.$$

- There exists no closed form expression for $\Gamma(a)$ in general, so that it must be computed using numerical methods. However,

$$\Gamma(a) = (a-1)\,\Gamma(a-1), \quad a \in \mathbb{R}_{>1},$$

and, in particular,

$$\Gamma(n) = (n-1)!, \quad n \in \mathbb{N}.$$

The Gamma Function (2)

- To see the latter, apply integration by parts with $u = x^{a-1}$ and $\mathrm{d}v = e^{-x}\,\mathrm{d}x$. This gives $\mathrm{d}u = (a-1)\,x^{a-2}\,\mathrm{d}x$, $v = -e^{-x}$ and

$$\begin{aligned}\Gamma(a) &= \int_0^\infty x^{a-1}e^{-x}\,\mathrm{d}x = uv\big|_{x=0}^{\infty} - \int_0^\infty v\,\mathrm{d}u = -e^{-x}x^{a-1}\big|_{x=0}^{\infty} \\ &\quad + \int_0^\infty e^{-x}(a-1)\,x^{a-2}\,\mathrm{d}x \\ &= 0 + (a-1)\,\Gamma(a-1).\end{aligned}$$

- For example, $\Gamma(1) = 0! = 1$, $\Gamma(2) = 1! = 1$ and $\Gamma(3) = 2! = 2$.
- It can be shown that $\Gamma(1/2) = \sqrt{\pi}$.

The Gamma Function (3)

Example To calculate $I = \int_0^\infty x^n e^{-mx}\,\mathrm{d}x$, $m > 0$, define $u = mx$. Then

$$I = m^{-1} \int_0^\infty (u/m)^n \, e^{-u} \, \mathrm{d}u = m^{-(n+1)} \Gamma(n+1),$$

which is a simple result, but required for the scaled gamma distribution which we will see later.

- The gamma function can be approximated by **Stirling's approximation**, given by

$$\Gamma(n) \approx \sqrt{2\pi}\, n^{n-1/2} \exp(-n),$$

which, clearly, provides an approximation to $n!$ for integer n. This is a famous result which is important, if not critical, in a variety of contexts in probability and statistics.

The Incomplete Gamma Function

- The **incomplete gamma function** is defined as

$$\Gamma_x(a) = \int_0^x t^{a-1} e^{-t} \, \mathrm{d}t, \quad a, x \in \mathbb{R}_{>0}$$

  and also denoted by $\gamma(x, a)$.
- The **incomplete gamma ratio** is the standardized version, given by

$$\bar{\Gamma}_x(a) = \Gamma_x(a) / \Gamma(a).$$

- In general, both functions $\Gamma(a)$ and $\Gamma_x(a)$ need to be evaluated using numerical methods.

A Recursive Expression

- For integer a, $\Gamma_x(a)$ can be easily computed: Using integration by parts with $u = t^{a-1}$ and $\mathrm{d}v = e^{-t}\,\mathrm{d}t$ gives

$$\begin{aligned} \int_0^x t^{a-1} e^{-t}\,\mathrm{d}t &= -t^{a-1}e^{-t}\Big|_0^x + (a-1)\int_0^x e^{-t} t^{a-2}\,\mathrm{d}t \\ &= -x^{a-1}e^{-x} + (a-1)\,\Gamma_x(a-1), \end{aligned}$$

i.e.,

$$\Gamma_x(a) = (a-1)\,\Gamma_x(a-1) - x^{a-1}e^{-x},$$

so that $\Gamma_x(a)$ can be evaluated recursively, noting that

$$\Gamma_x(1) = \int_0^x e^{-t}\,\mathrm{d}t = 1 - e^{-x}.$$

The Beta Function

- The **beta function** is an integral expression of two parameters, denoted $B(\cdot, \cdot)$ and defined to be

$$B(a, b) := \int_0^1 x^{a-1} (1-x)^{b-1} \, \mathrm{d}x, \quad a, b \in \mathbb{R}_{>0}.$$

- Closed–form expressions do not exist for general a and b; however, the identity

$$B(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}$$

can be used for its evaluation in terms of the gamma function.

Example and Exercises

- Example: To express $\int_0^1 \sqrt{1-x^4}\,\mathrm{d}x$ in terms of the beta function, let $u = x^4$ and $\mathrm{d}x = (1/4)u^{1/4-1}\,\mathrm{d}u$, so that

$$\int_0^1 \sqrt{1-x^4}\,\mathrm{d}x = \frac{1}{4}\int_0^1 u^{-3/4}\,(1-u)^{1/2}\,\mathrm{d}u = \frac{1}{4}B\left(\frac{1}{4},\frac{3}{2}\right).$$

- Exercise I: Express

$$I = \int_0^s x^a\,(s-x)^b\,\mathrm{d}x, \qquad s \in (0,1), \quad a, b > 0$$

in terms of the beta function.

- Exercise II: Express

$$I = \int_{-1}^1 \left(1-x^2\right)^a (1-x)^b\,\mathrm{d}x$$

in terms of the beta function.

Solution

1. To compute $I = \int_0^s x^a (s-x)^b \, \mathrm{d}x$, $s \in (0,1)$, $a, b > 0$, use $u = 1 - x/s$ (so that $x = (1-u)\,s$ and $\mathrm{d}x = -s\,\mathrm{d}u$) to get

$$\begin{aligned} I &= \int_0^s x^a (s-x)^b \, \mathrm{d}x = -s \int_1^0 ((1-u)\,s)^a \, (s - (1-u)\,s)^b \, \mathrm{d}u \\ &= s^{a+b+1} \int_0^1 (1-u)^a \, u^b \, \mathrm{d}u = s^{a+b+1} B\,(b+1, a+1)\,. \end{aligned}$$

2. To compute $I = \int_{-1}^1 (1-x^2)^a (1-x)^b \, \mathrm{d}x$, express $1 - x^2$ as $(1-x)(1+x)$ and use $u = (1+x)/2$, (so that $x = 2u - 1$ and $\mathrm{d}x = 2\,\mathrm{d}u$) to get

$$I = 2^{2a+b+1} \int_0^1 u^a (1-u)^{a+b} \, \mathrm{d}u = 2^{2a+b+1} B\,(a+1, a+b+1)\,.$$

The Incomplete Beta Function

- Similar to the incomplete gamma function, the **incomplete beta function** is

$$B_x(p,q) = \mathbb{I}_{[0,1]}(x) \int_0^x t^{p-1} (1-t)^{q-1} \, \mathrm{d}t.$$

- The normalized function $B_x(p,q)/B(p,q)$ is the **incomplete beta ratio**, which we denote by $\bar{B}_x(p,q)$.

Course Outline

Basic Counting

- R unique balls in an urn: we randomly draw n of them.
- We wish to know how many ways there are of doing this, but we need to specify if the balls are drawn with or without replacement, and also if the ordering of the balls is relevant or not.
- Consider first drawing the balls "*Ordered without replacement*": The first draw is one of R possibilities; the second is one of $(R-1)$, ..., the n^{th} is one of $R-n+1$. In total: $R_{[n]} = R! \,/\, (R-n)!$.
- "*Ordered with replacement*": the first draw is one of R possibilities; the second is one of R, etc., so R^n possibilities.
- "*Unordered without replacement*": similar to ordered without replacement, but we need to divide $R_{[n]}$ by $n!$ to account for the irrelevance of order, giving "R choose n", $\frac{R!}{(R-n)!n!} = \binom{R}{n}$.

Unordered Drawing with Replacement

- The last case is "*Unordered with replacement*".
- Let the balls be labeled **1**, **2**, ..., **R**.
- Because order does not matter, the result after n draws can be condensed to just reporting the number of times ball **1** was chosen, the number of times ball **2** was chosen, etc.
- Denote the number of times ball i was chosen as x_i, $i = 1, \dots, R$.
- We are actually seeking the number of nonnegative integer solutions to $x_1 + x_2 + \cdots + x_R = n$.
- First let $R = 2$, i.e., we want the number of solutions to $x_1 + x_2 = n$.
- Imagine placing the n balls along a straight line, like $\bullet\bullet\bullet\bullet\bullet$ for $n = 5$. Now, by placing a partition among them, we essentially "split" the balls into two groups.

Unordered Drawing with Replacement (2)

- We can place the partition in any of the $n+1$ "openings", such as $\bullet\bullet \mid \bullet\bullet\bullet$, which indicates $x_1 = 2$ and $x_2 = 3$, or as $\mid \bullet\bullet\bullet\bullet\bullet$, which indicates $x_1 = 0$ and $x_2 = 5$, etc.
- If $R = 3$, we would need two partitions, e.g.,

$$\bullet\bullet \mid \bullet\bullet \mid \bullet \quad \text{or} \quad \bullet\bullet \mid \bullet\bullet\bullet \mid \quad \text{or} \quad \mid \mid \bullet\bullet\bullet\bullet\bullet .$$

- In general, we need $R-1$ partitions interspersed among the n balls. This can be pictured as randomly arranging n balls and $R-1$ partitions along a line, for which there are $(n+R-1)!$ ways. But, as the balls are indistinguishable, and so are the partitions, there are

$$\frac{(n+R-1)!}{n!\,(R-1)!} = \binom{n+R-1}{n}$$

ways to arrange them, or $\binom{n+R-1}{n}$ nonnegative integer solutions to the equation $\sum_{i=1}^{R} x_i = n$.

Probability Spaces

- A formal description of a probability space is "mathematical" and somewhat abstract; it is based however on a "common sense" frequentistic description. Let us introduce it by way of two examples:

Example Draw a card from a shuffled deck of 52 cards. Each of the cards is equally likely to be drawn. Probability of getting the Ace of Spades, is just $1/52$, denoted $\Pr(\clubsuit\mathrm{A}) = 1/52$.

- This assignment of probability appears quite reasonable; it can be thought of as the "long run" fraction of times the Ace of Spades is drawn when the experiment is repeated indefinitely under the same conditions
- Each card represents one of the 52 possible outcomes and said to be one of the **sample points** making up the **sample space**.
- Assign labels to each of the sample points: $\omega_1 = \clubsuit\mathrm{A}$, $\omega_2 = \clubsuit 2, \ldots, \omega_{13} = \clubsuit\mathrm{K}$, $\omega_{14} = \spadesuit\mathrm{A}, \ldots, \omega_{52} = \heartsuit\mathrm{K}$ and denote the sample space as $\Omega = \{\omega_i,\ i = 1, \ldots, 52\}$.

Probability Spaces (2)

- To compute the probability that the card drawn is an Ace, sum the number of relevant sample points $(\omega_1, \omega_{14}, \omega_{27}, \omega_{40})$ and divide by 52, i.e., $4/52 = 1/13$.
- Any such case of interest is termed an **event**. The event that a picture card is drawn has probability $12/52 = 3/13$, etc.
- There are $2^{52} \approx 4.5 \times 10^{15}$ possible events for this sample space and experiment; this totality of possible events is simply called the **collection of events**.

Example Draw two cards **without** replacement and order **relevant**.

- Sample points consist of all $52 \cdot 51 = 2652$ possibilities, all of which are equally likely.
- A possible event is that both cards have the same suit.

Probability Spaces (3)

- Two ways of computing its probability are as follows.
  - There are 52 ways of drawing the first card; its suit determines the 12 cards which can be drawn next so that both have the same suit. This gives a probability of
  $$\Pr(\text{both cards have same suit}) = \frac{52}{52}\frac{12}{51} = \frac{12}{51}.$$
  - There are $\binom{4}{1}$ possible suits to "choose from" times the number of ways of getting two cards from this suit, or $\binom{13}{2}$. This gets doubled because order matters. Thus
  $$\Pr(\text{both cards have same suit}) = \frac{\binom{4}{1}\cdot\binom{13}{2}}{\binom{52}{2}} = \frac{12}{51}.$$
- The number of events for this experiment is now astronomical: $2^{2652} \approx 2.14 \times 10^{798}$, but is still finite.
- This intuitive method of assigning probabilities to events is valid for any finite sample space in which each sample point is equally likely to occur and also works for non–equally likely sample points.

Finite Probability Spaces

- If we denote the sample space as the set Ω with elements, or possible outcomes, ω_i, $i = 1, \ldots, N$, then Ω and a function Pr with domain Ω and range $[0, 1]$ such that $\sum_{i=1}^{N} \Pr(\omega_i) = 1$ is referred to as a **finite probability space**.
- If $\Pr(\omega_i)$ is the same for all i (in which case $\Pr(\omega_i) = 1/N$), then the finite probability space $(\Omega, \Pr)$ is also called a **symmetric probability space**.

Exercise (Problem 2.12)

A lottery consists of 100 tickets, labeled $1, 2, \ldots, 100$, three of which are "winning numbers". You buy 4 tickets. Calculate the probability, p, that you have at least one winning ticket.

Solution

Think of the complement: You draw no winning tickets.

First way: From the 3 winning tickets, you choose none, from the 97 losing tickets, you draw 4. Thus,

$$1-p = \frac{\binom{3}{0}\binom{97}{4}}{\binom{100}{4}} = \frac{97 \cdot 96 \cdot 95 \cdot 94}{4!} \frac{4!}{100 \cdot 99 \cdot 98 \cdot 97} = \frac{96 \cdot 95 \cdot 94}{100 \cdot 99 \cdot 98} = 0.8836.$$

Second way: The problem can be thought of as follows. Instead of the 100 tickets being produced with 3 winners, 97 losers, imagine they are produced as 4 belonging to you, and 96 not. That is, their "characteristic" is changed from being winner/loser to yours/not yours. Then, 3 tickets are chosen to be labeled as winner, and we want the probability that, from your 4 tickets, none are chosen. This is

$$1-p = \frac{\binom{4}{0}\binom{96}{3}}{\binom{100}{3}} = \frac{96 \cdot 95 \cdot 94}{3!} \frac{3!}{100 \cdot 99 \cdot 98} = \frac{96 \cdot 95 \cdot 94}{100 \cdot 99 \cdot 98}.$$

Countable Probability Spaces

- For some experiments of interest, it might be the case that the sample space Ω is **denumerable** or **countably infinite**, i.e., each $\omega_i \in \Omega$ can be put in a 1–1 correspondence with the elements of $\mathbb{N}$.
- For example, consider tossing a fair coin until a tail appears.
- It is theoretically possible that the first 10,000 trials will result in heads, or that a tail may never occur.
- Letting ω_i be the total number of required tosses, $i = 1, 2, \ldots$, we see that Ω is countable.
- If associated with Ω is a function Pr with domain Ω and range $[0, 1]$ such that $\sum_{i=1}^{\infty} \Pr(\omega_i) = 1$, $(\Omega, \Pr)$ is referred to as a **probability space**. For example, taking $\Pr(\omega_i) = (1/2)^i$ is valid.

Uncountable Probability Spaces

- Matters increase in complexity when the sample space is **uncountable**.
- Examples of such sample spaces include
  - "random" numbers between zero and one,
  - times until an electrical component fail,
  - measurement of a patients blood pressure using an analog device.
- In all these cases, one could argue that a finite number of significant digits has to ultimately be used to measure or record the result, so that the sample space, however large, is actually finite.
- This is valid but becomes cumbersome: By allowing for a continuum of sample points, the powerful techniques of calculus may be employed, which significantly eases many computations of interest.
- It is no longer conceptually clear what probability is assigned to a "random" number between 0 and 1.

Uncountable Probability Spaces (2)

Example Let R such a random number with its decimal representation truncated at the third place. There are then 1000 possible outcomes, 0.000, 0.001, ..., 0.999, each equally likely.

- The probability that $R \leq 0.400$ is then $401/1000$. As the number of digits is increased, the probability of any particular outcome goes to zero. The calculation of $\Pr(R \leq 0.4)$, however, approaches 0.4.
- Thus, with uncountable sample spaces, we assume that a probability has been **pre–assigned** to each possible subset of Ω: for the random number between 0 and 1, each point in $[0, 1]$ is assigned probability zero while each interval of $[0, 1]$ is assigned probability corresponding to its length.
- Not all subsets of Ω can be assigned probabilities. The class of subsets for which assignment is possible is termed a σ–**field** and containing essentially all subsets relevant for practical problems.

Axiomatic Description of Probability

- A **realization** is the result of some well-defined **trial** or **experiment** performed under a given **set of conditions**, whose **outcome** is not known in advance, but belongs to a **set of possibilities** or **set of outcomes** which are known in advance.
- The set of possible outcomes could be **countable** (either **finite** or **denumerable**, i.e., **countably infinite**) or **uncountable**.
- Denote the **sample space** as Ω, the set of all possible outcomes, with individual outcomes or **sample points** $\omega_1, \omega_2, \ldots$. A **subset** of the sample space is known as an **event**, usually denoted by a capital letter, possibly with subscripts, i.e., A, B_1, etc., the totality of which under Ω will be denoted $\mathcal{A}$, and forms the **collection** of events—also called the **collection of measurable events**.

Axiomatic Description of Probability (2)

- An outcome $\omega \in \Omega$ may belong to many events, always belongs to the **certain event** Ω, and never to $\emptyset$, the **empty set** or **impossible event**.
- The usual operations in set theory can be applied to two events, i.e., **complement**, **intersection**, **union**, **difference**, **symmetric difference**, **inclusion**, etc.
- As in general set theory, two events are **mutually exclusive** if $A \cap B = \emptyset$.
- If a particular set of events A_i, $i \in J$, are such that $\bigcup_{i \in J} A_i \supseteq \Omega$, they **cover** (or **exhaust**) the same space Ω.
- If events A_i, $i \in J$, are such that $\bigcup_{i \in J} A_i = \Omega$ are mutually exclusive and exhaust Ω, they **partition** Ω, i.e., one and only one of the A_i will occur on a given trial.

Probability Function

- A **probability measure** is a set function which assigns a real number $\Pr(A)$ to each event $A \in \mathcal{A}$ such that
  - $\Pr(A) \geq 0$,
  - $\Pr(\Omega) = 1$, and
  - for a countable infinite sequence of mutually exclusive events A_i,

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i).$$

- The latter requirement is known as **(countable) additivity**.
- If $A_i \cap A_j = \emptyset$, $i \neq j$ and $A_{n+1} = A_{n+2} = \cdots = \emptyset$, then $\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr(A_i)$, which is **finite additivity**.
- The triplet $\{\Omega, \mathcal{A}, \Pr(\cdot)\}$ refers to the **probability space** or **probability system** with sample space Ω, collection of measurable events $\mathcal{A}$ and probability measure $\Pr(\cdot)$.

Example

A fair die is tossed once and the number of dots is observed.

- The set of outcomes is $\Omega = \{1, 2, 3, 4, 5, 6\}$ with natural ordering $\omega_i = i$, $i = 1, \ldots, 6$. Each outcome is equally likely: $\Pr(\omega_i) = 1/6$.
- Possible events include $E = \{2, 4, 6\}$, $O = \{1, 3, 5\}$ and $A_i = \{1, 2, \ldots, i\}$. E and O partition the sample space, i.e., $E^c = O$ and $\Omega = E \cup O$, while $A_i \subseteq A_j$, $1 \leq i \leq j \leq 6$.
- Events A_i exhaust Ω since $\bigcup_{i=1}^{6} A_i = \Omega$, but do not partition it.
- Defining events $B_1 = A_1 = \omega_1$ and $B_i = A_i \setminus A_{i-1} = \omega_i$, it follows that B_i, $i = 1, \ldots, 6$, partition Ω since precisely one and only one B_i can occur on a given trial and $\bigcup_{i=1}^{6} B_i = \Omega$.

Example (cont.)

- The collection of all events, $\mathcal{A}$, has 64 elements: Any event in $\mathcal{A}$ specifies which of the six ω_i are included.
- By associating a binary variable with each of ω_i, say 1 if present, 0 otherwise, we see that the number of possible elements in $\mathcal{A}$ is the same as the number of possible codes from a 6-length string of binary numbers, or 2^6.
- For example, 000000 would denote the empty set $\emptyset$, 111111 denotes the certain event Ω, 010101 denotes event E, etc.

Example: Uncountably Infinite Sample Space

Imagine an automobile factory in which all new cars coming off a particular assembly line are subjected to a battery of tests.

- The proportion of cars that pass all the tests is a random with countable sample space $\Omega^0 = \{m/n; m \in \{0 \cup \mathbb{N}\}, n \in \mathbb{N}, m \leq n\}$, i.e., all rational numbers between and including 0 and 1.
- If m and n are typically quite large, it is usually mathematically more convenient to use the uncountable sample space

$$\Omega = \{x : x \in [0,1]\} .$$

Example (cont.)

- Then $\mathcal{A}$ can be described as all interval subsets of $[0,1]$ along with their complements and their finite unions and intersections.
- If A is any such event, $\Pr(A) = \int_A f$ for a suitable function f.
- For example, with $f(x) = 20x^3(1-x)$ and

$$A = \{\text{proportion greater than } 0.8\},$$

$\Pr(A) = \int_{0.8}^{1} f(x)\,\mathrm{d}x \approx 0.263$. Observe also that $\Pr(A) \geq 0\ \forall$ $A \in \mathcal{A}$ and $\Pr(\Omega) = \int_0^1 f(x)\,\mathrm{d}x = 1$.

Example: Countably Infinite Sample Space

Your newly married colleague has decided to have as many children as it takes until she has a son.

Exclude any factors which would prevent her from having an unlimited number of children.

- Denote a boy as B and a girl as G. The set of possible outcomes can be listed as $\omega_1 = B$, $\omega_2 = GB$, etc., i.e., the collection of all sequences which end in a boy but have no previous occurrence of such.
- The sample space is Ω is clearly countable.

Example on Sons, cont.

- Events of interest might include
  $A_i = \{\text{at most } i \text{ children}\} = \{\omega_1, \omega_2 \ldots, \omega_i\}$ and
  $O = \{\text{odd number of children}\} = \{\omega_1, \omega_3, \ldots\}$.
- Events O and O^c partition the sample space, but events A_i do not (they do exhaust it though).
- If we define $B_1 = A_1$ and $B_i = A_i \setminus A_{i-1}$, $i \geq 2$, we see that $B_i = \omega_i$ and B_i, $i = 1, 2, \ldots$, partition the sample space.
- Let $\Pr(\omega_i) = 2^{-i}$; then $\Pr(A_i) = \sum_{j=1}^{i} \Pr(\omega_j) = 1 - 2^{-i}$ and $\Pr(\Omega) = \sum_{j=1}^{\infty} \Pr(\omega_j) = 1$.

Example: 3 daughters in a row

Not to be outdone by your colleague, you and your spouse decide to have as many children as it takes until you have three daughters in row.

The set of possible outcomes is the collection of all sequences which end in three girls but have no previous occurrence of such.

- The total number of children "required" is of interest.
- Let f_n be the probability that $n \geq 3$ children will born under this family planning strategy.

Example: 3 daughters (cont.)

- Assume that p is the probability of getting a girl and $q := 1 - p$ is that of a boy.
- Clearly, $f_1 = f_2 = 0$, while $f_3 = p^3$, i.e., all three children are girls or, in obvious notation, GGG is obtained.
- For $n = 4$, the first child must be a boy and the next three are girls, i.e., BGGG, with $f_4 = qp^3 = qf_3$.
- For $n = 5$, the situation must end as BGGG but the first child can be either B or G, i.e., either BBGGG or GBGGG occurs, so that

$$f_5 = q^2 f_3 + qpf_3 = qf_4 + qpf_3.$$

Example: 3 daughters (cont.)

- For $n = 6$, either BBBGGG or BGBGGG or GBBGGG or GGBGGG can occur. The first two of these start with a boy and are accounted for in the expression qf_5.
- Series GBBGGG corresponds to qpf_4 and series GGBGGG to $qp^2 f_3$, i.e.,

$$f_6 = qf_5 + qpf_4 + qp^2 f_3.$$

- This may seem to be getting complicated, but we are practically done. Notice so far, that, for $4 \leq n \leq 6$,

$$f_n = qf_{n-1} + qpf_{n-2} + qp^2 f_{n-3}.$$

- But this holds for all $n \geq 4$ because either a boy on the first trial occurs and is followed by one of the possibilities associated with f_{n-1}; or the first two trials are GB followed by one of the possibilities associated with f_{n-2}; or GGB followed by f_{n-3}.

Example: 3 daughters (cont.)

This solution, along with the *initial conditions* $f_1 = f_2 = 0$, $f_3 = p^3$, is a *difference equation*.

The recursive scheme is easy to implement in a computer program. To compute and plot it in Matlab, the following code can be used:

```
p=0.5; q=1-p; f=zeros(60,1);
f(3)=p^3;
for i=4:60
  f(i)=q*f(i-1)+q*p*f(i-2)+q*p^2*f(i-3);
end
bar(f)
```

It produces a graphic like the following:

Example: 3 daughters (cont.)

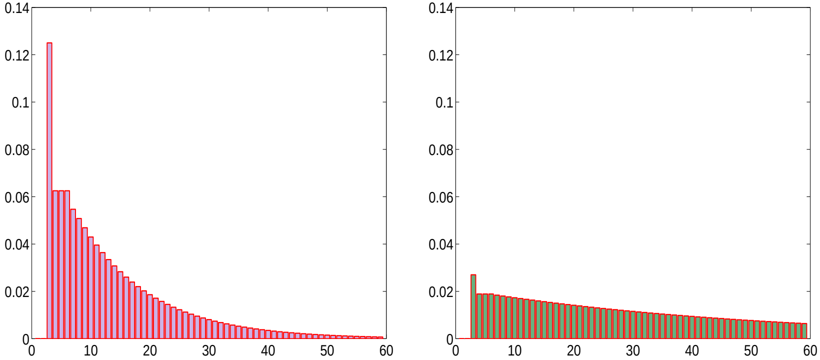


Figure: Recursion for $p = 0.5$ (left) and $p = 0.3$ (right).

Example: 3 daughters (cont.)

- Another quantity of interest is the "average" number of children a family can expect if they implement such a strategy.
- Assuming $p = 1/2$, i.e., that girls and boys are equally likely, an example in Chapter 8 shows the solution to this latter question to be 14.
- If instead of three girls in a row, m are desired, $m \geq 1$, then

$$f_n(m) = \begin{cases} 0, & \text{if } n < m, \\ p^m, & \text{if } n = m, \\ q \sum_{i=0}^{m-1} p^i f_{n-i-1}, & \text{if } n > m. \end{cases}$$

Using this, $f_n(m)$ can be computed for any $n, m \in \mathbb{N}$.

Example: 3 daughters (cont.)

- It is not difficult to extend the previous example in such a way that the solution is either quite complicated, or even intractable.
- In such cases, *simulation* could be used either to corroborate a tentative answer, or just to provide an approximate, numeric solution if an analytic solution is not available.
- In our case with $m = 3$, simulation involves getting the computer to imitate the family planning strategy numerous times (independently of one another), and then tabulate the results.
- A function in Matlab is given next to accomplish this.

Example: 3 daughters (cont.)

```
function len = consecutive (p,m,sim)
len=zeros(sim,1);
for i=1:sim
  if mod(i,100)==0, i, end
  cnt=0; mcount=0;
  while mcount<m
    r=rand; cnt=cnt+1;
    if r>p, mcount=0; else mcount=mcount+1; end
  end
  len(i)=cnt;
end
hist(len,max(len)-m+1)
```

Example: 7 children

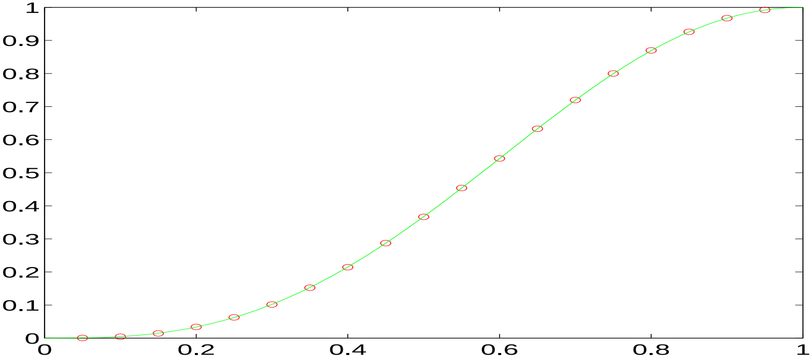
Example Your neighbors are somewhat more family planning conscious and decide to have exactly seven children:

- You (as the watchful neighbor) are interested in the sequence of sexes of the children, i.e., Ω consists of the 2^7 possible binary sequences of "b" and "g".
- Your interest centers on the probability that they have (at least) three girls in a row.

Example: 7 children (cont.)

- A later example will show this to be about 0.367, assuming that boys and girls are equally likely.
- More generally, we will see that, if the probability of getting a girl is p, $0 < p < 1$, then at least three girls occur with probability $P = 5p^3 - 4p^4 - p^6 + p^7$.
- This is plotted as the solid line in the figure below.
- The dots in the figure were obtained from simulation from the program shown below.

Example: 7 children (cont.)



```
function f=threegirls(p,sim,kids,inarow)
if nargin<3  % nargin is a pre-defined variable in Matlab
  kids=7;    %    which indicates the Number of ARGuments INput
end          %    to the function.
if nargin<4  % It is used to help assign default values
  inarow=3;  %   to variables if they are not passed to the function
end
sum = 0;
for i=1:sim
  k = binornd(1,p,kids,1); % iid Bin(1,p) vector of length kids
  bool=0;   % boolean variable, i.e., either true or false
  for j=1:kids-inarow+1;
    bool = ( bool | all(k(j:j+inarow-1) == ones(inarow,1)) );
    if bool
      sum=sum+1;
      break % BREAK terminates the inner FOR loop, which
    end     % is useful because once 3 girls in a row are
  end       % found, there is no further need to search.
end
f = sum/sim;
```

Properties: Basic

- From the definition of a probability measure, many properties of $(\Omega, \mathcal{A}, \Pr)$ are intuitive and easy to see, notably so with the help of a **Venn diagram**, such as

(i) $\Pr(\emptyset) = 0$,

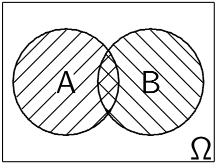
(ii) If $A \subset B$, then $\Pr(A) \leq \Pr(B)$.

(iii) $\Pr(A) \leq 1$,

(iv) $\Pr(A^c) = 1 - \Pr(A)$,

(v) $\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \Pr(A_i)$,

(vi) $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2)$.



Properties: Basic (2)

- From (i) and (ii), it follows that $\Pr(A) \geq 0$. Result (v) is equivalent to saying that $\Pr(\cdot)$ is **subadditive** and is also referred to as Boole's inequality; if the A_i are disjoint, then equality holds.
- By taking complements of both sides (i.e., $1-$), Booles's inequality can also be written as $\Pr\left(\bigcap_{i=1}^n A_i^c\right) \geq 1 - \sum_{i=1}^n \Pr(A_i)$.
- Another useful result which is particularly clear from the Venn diagram is that event A can be partitioned into two disjoints sets AB and AB^c, i.e., $\Pr(A) = \Pr(AB) + \Pr(AB^c)$ or $\Pr(A \setminus B) = \Pr(AB^c) = \Pr(A) - \Pr(AB)$.
- If $B \subset A$, then $\Pr(AB) = \Pr(B)$, so that

$$\Pr(A \setminus B) = \Pr(A) - \Pr(B), \quad B \subset A.$$

Properties: Advanced

- Less obvious results which build on (i) through (vi) above include the following powerful results.
  1. Bonferroni's inequality,
  2. Poincaré's theorem,
  3. Sieve theorem or de Moivre–Jordan theorem, and
  4. The Problem of Coincidences.

1. Bonferroni's Inequality

- For $n = 2$, this is given by

$$\Pr(A_1 \cap A_2) \geq \Pr(A_1) + \Pr(A_2) - 1,$$

which follows from properties (iii) and (vi) above:

$$\begin{aligned} \Pr(A_1 A_2) &= \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cup A_2) \\ &\geq \Pr(A_1) + \Pr(A_2) - 1. \end{aligned}$$

- In general,

$$\Pr\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} \Pr(A_i) - (n-1) = 1 - \sum_{i=1}^{n} \Pr\left(\bar{A}_i\right),$$

which can be shown by induction.

- Bonferroni's inequality is of great importance in the context of **joint confidence intervals** and **multiple testing**.

2. Poincaré's Theorem or the Inclusion–Exclusion Principle

- Generalizing property (vi) above,

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} (-1)^{i+1} S_i,$$

where

$$S_j = \sum_{i_1 < \cdots < i_j} \Pr\left(A_{i_1} \cdots A_{i_j}\right),$$

i.e.,

$$\begin{aligned}\Pr\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} \Pr(A_i) - \sum_{i<j} \Pr(A_i \cap A_j) \\ &\quad + \sum_{i<j<k} \Pr(A_i A_j A_k) - \cdots \\ &\quad + \cdots + (-1)^{n+1} \Pr(A_1 A_2 \cdots A_n).\end{aligned}$$

Example

- Assume that each of n indistinguishable balls is randomly placed in one of r distinguishable cells, with $n \geq r$. Let U_i be the event that cell i is empty, $i = 1, \ldots, r$.
- Then $p_1 := \Pr(U_i) = \left(\frac{r-1}{r}\right)^n$, $i = 1, \ldots, r$, $p_2 := \Pr(U_i U_j) = \left(\frac{r-2}{r}\right)^n$ for $1 \leq i < j \leq r$, etc., and Poincaré's theorem then gives the probability of at least one empty cell as

$$P_{1,r} = r^{-n} \sum_{i=1}^{r-1} (-1)^{i+1} \binom{r}{i} (r-i)^n ,$$

noting that $\Pr(U_1 U_2 \cdots U_r) = 0$.

3. Sieve or de Moivre–Jordan Theorem

- For events $B_1, \ldots, B_n$, the probability that exactly m of the B_i occur, $m = 0, 1, \ldots, n$, is given by

$$p_{m,n}\left(\{B_i\}\right) = \sum_{i=m}^{n} (-1)^{i-m} \binom{i}{i-m} S_i = \sum_{i=m}^{n} (-1)^{i-m} \binom{i}{m} S_i,$$

where $S_j = \sum_{i_1 < \cdots < i_j} \Pr\left(B_{i_1} \cdots B_{i_j}\right)$ with $S_0 = 1$.

- The probability that **at least** m of the B_i occur is given by

$$P_{m,n} = \sum_{i=m}^{n} (-1)^{i-m} \binom{i-1}{i-m} S_i.$$

- Observe that $P_{m,n} = \sum_{i=m}^{n} p_{i,n}$ and $p_{m,n} = P_{m,n} - P_{m+1,n}$.

4. The Problem of Coincidences

- If n objects, labelled 1 through n are randomly arranged in a row, the probability that the position of exactly m of them coincide with their label, $0 \leq m \leq n$, is

$$c_{m,n} = \frac{1}{m!}\left(\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots + \frac{(-1)^{n-m}}{(n-m)!}\right).$$

- Let A_i denote the event that the i^{th} object is arranged correctly and observe that $\Pr(A_1) = (n-1)!/n! = \Pr(A_2) = \cdots = \Pr(A_n)$; $\Pr(A_1 A_2) = (n-2)!\,/\,n!$, etc.
- In general,

$$\Pr\left(A_{i_1} A_{i_2} \cdots A_{i_m}\right) = \frac{(n-m)!}{n!}.$$

4. The Problem of Coincidences (2)

- For $m = 0$, the probability of no coincidences is, from Poincaré's theorem,

$$\begin{aligned} c_{0,n} &= 1 - \Pr(\text{at least one}) = 1 - \Pr\left(\cup_{i=1}^{n} A_i\right) \\ &= 1 - \left[\binom{n}{1}\frac{(n-1)!}{n!} - \binom{n}{2}\frac{(n-2)!}{n!} + \cdots + (-1)^{n+1}\frac{0!}{n!}\right] \\ &= 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!}. \end{aligned}$$

- For the general case, use the Sieve theorem.
- For large $n - m$, note that $c_{m,n}$ is approximately $e^{-1} / m!$.
- Imagine a party with four married couples. The four wives each grab a different man to dance with. The probability can be calculated that m of the women, $m = 0, \ldots, 4$ happen to pair off with their own husbands.

Continuity of Probability Measure

- Recall that a probability measure is a set function which assigns a real number $\Pr(A)$ to each event $A \in \mathcal{A}$ such that (i) $\Pr(A) \geq 0$, (ii) $\Pr(\Omega) = 1$, and (iii) for *mutually exclusive* A_i,

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i),$$

  where the latter requirement is known as countable additivity.

- The property of countable additivity is a crucial assumption for showing the following important result:
  If $A_1, A_2, \ldots$ is a monotone sequence of (measurable) events, then

$$\lim_{i\to\infty} \Pr(A_i) = \Pr\left(\lim_{i\to\infty} A_i\right).$$

Continuity of Probability: Increasing Events

Let $A_1, A_2, \ldots$ be a seq. of increasing events, i.e., $A_1 \subset A_2 \subset \cdots$.

Let $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1} = A_n A_{n-1}^c$, $n = 2, 3, \ldots,$. Thus, B_2 is the part of A_2 which is "new", i.e., not already in A_1. Thus,

$$\begin{aligned} A_n &= A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \cdots \cup (A_n \setminus A_{n-1}) \\ &= B_1 \cup B_2 \cup \cdots \cup B_n = \bigcup_{i=1}^{n} B_i, \end{aligned}$$

and, by construction, the B_i are disjoint, so that $\Pr(A_n) = \sum_{i=1}^{n} \Pr(B_i)$. Then

$$\Pr\left(\lim_{n\to\infty} A_n\right) = \Pr\left(\lim_{n\to\infty} \bigcup_{i=1}^{n} B_i\right) = \Pr\left(\bigcup_{i=1}^{\infty} B_i\right)$$

and, from countable additivity,

$$\Pr\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \Pr(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \Pr(B_i) = \lim_{n\to\infty} \Pr(A_n).$$

Continuity of Probability: Decreasing Events

Now consider the case for monotone decreasing A_i.

Recall (see Appendix A.1 on sets), if $A_1 \subset A_2 \subset \cdots$, so that the A_n are monotone increasing, then

$$\lim_{i \to \infty} A_i = \bigcup_{i=1}^{\infty} A_i.$$

Similarly, if $A_1 \supset A_2 \supset \cdots$, (monotone decreasing), then

$$\lim_{i \to \infty} A_i = \bigcap_{i=1}^{\infty} A_i.$$

Recall also De Morgan's laws,

$$\left(\bigcup_{n=1}^{\infty} A_n \right)^c = \bigcap_{n=1}^{\infty} A_n^c \quad \text{and} \quad \left(\bigcap_{n=1}^{\infty} A_n \right)^c = \bigcup_{n=1}^{\infty} A_n^c.$$

Continuity of Probability: Decreasing Events

For monotone decreasing A_i, $A_1 \supset A_2 \supset \cdots$, and $A_1^c \subset A_2^c \subset \cdots$, so that

$$\lim_{i\to\infty} \Pr\left(A_i^c\right) = \Pr\left(\lim_{i\to\infty} A_i^c\right)$$

from the previous result. Then

$$\lim_{i\to\infty} \Pr\left(A_i^c\right) = \lim_{i\to\infty} \left(1 - \Pr\left(A_i\right)\right) = 1 - \lim_{i\to\infty} \Pr\left(A_i\right)$$

and, from the above results,

$$\Pr\left(\lim_{i\to\infty} A_i^c\right) = \Pr\left(\bigcup_{i=1}^{\infty} A_i^c\right) = 1 - \Pr\left(\bigcap_{i=1}^{\infty} A_i\right) = 1 - \Pr\left(\lim_{i\to\infty} A_i\right),$$

so that

$$1 - \lim_{i\to\infty} \Pr\left(A_i\right) = 1 - \Pr\left(\lim_{i\to\infty} A_i\right).$$

Exercise with Continuity I

Let $A_n = \left[0, 1 + n^{-1}\right]$, $n = 1, 2, \ldots$.

Show that $\{A_n\}$ is monotone and compute $L := \lim_{n\to\infty} A_n$.

Let $B_n := A_n \setminus A_{n+1}$, $n = 1, 2, \ldots$.

Express B_n as an interval and express A_n in terms of the B_i and L.

Exercise with Continuity I: **Solution**

As $1/(n+1) < 1/n$, $A_n = [0, 1 + 1/n] \supset [0, 1 + 1/(n+1)] = A_{n+1}$ and so $\{A_n\}$ is monotone decreasing for $n = 1, 2, \ldots$.

So, $L = \lim_{n\to\infty} A_n = \cap_{i=1}^{\infty} A_i$ which, as $\lim_{n\to\infty} n^{-1} = 0$, is $[0, 1]$.

We have $B_1 = [0, 2] \setminus [0, 1.5] = (1.5, 2]$; $B_2 = (1 + 1/3, 1 + 1/2]$; and

$$B_n = [0, 1 + 1/n] \setminus [0, 1 + 1/(n+1)] = (1 + 1/(n+1), 1 + 1/n].$$

Also,

$$\begin{aligned} A_n &= \left[0, 1 + \frac{1}{n}\right] \\ &= [0,1] \cup \left(1 + \frac{1}{n+1}, 1 + \frac{1}{n}\right] \cup \left(1 + \frac{1}{n+2}, 1 + \frac{1}{n+1}\right] \cup \cdots \\ &= L \cup \bigcup_{i=n}^{\infty} B_i. \end{aligned}$$

It should be clear that L and the B_i are mutually exclusive.

Exercise with Continuity II

Let $\{A_n\}$ be a monotone decreasing sequence of events. Show that event A_n can be expressed in terms of the $B_n := A_n \setminus A_{n+1}$ and $L := \lim_{n\to\infty} A_n$.

Exercise with Continuity II: **Solution**

We claim that $A_n = L \cup \bigcup_{i=n}^{\infty} B_i$. We first show $A_n \subset L \cup \bigcup_{i=n}^{\infty} B_i$.

Fix an $n \in \mathbb{N}$ and assume $\omega \in A_n$. In general, the two events $A_n A_{n+1}$ and $A_n A_{n+1}^c$ are mutually exclusive and partition A_n, so either $\omega \in A_n A_{n+1}$ or $\omega \in A_n A_{n+1}^c$. This says, in our case with $A_n \supset A_{n+1}$ for all n, that either $\omega \in A_n A_{n+1} = A_{n+1}$ or $\omega \in A_n A_{n+1}^c = A_n \setminus A_{n+1} = B_n$.

If the latter is true, i.e., $\omega \in B_n$, then obviously $\omega \in L \cup \bigcup_{i=n}^{\infty} B_i$. If the former is true, i.e., $\omega \in A_{n+1}$, then we repeat the argument; either $\omega \in A_{n+1} A_{n+2} = A_{n+2}$ or $\omega \in A_{n+1} A_{n+2}^c = A_{n+1} \setminus A_{n+2} = B_{n+1}$.

Continuing, we see that it must be the case that either $\omega \in B_n$ or $\omega \in B_{n+1}$ or ..., or $\nexists m \geq n$ such that $\omega \in B_m = A_m \setminus A_{m+1}$, i.e., $\nexists m \geq n$ such that $\omega \notin A_{m+1}$, in which case it is in $A_n, A_{n+1}, \ldots$, and in $A_1, \ldots, A_{n-1}$ because the A_n are monotone decreasing, i.e., $\omega \in \cap_{i=1}^{\infty} A_i = \lim_{n\to\infty} A_n = L$. Thus, $\omega \in L \cup \bigcup_{i=n}^{\infty} B_i$.

To prove $A_n \supset L \cup \bigcup_{i=n}^{\infty} B_i$, simply note that, by definition, $L \subset A_n$, and as $A_n \supset A_{n+1}$, $B_i \subset A_n$ for $i = n, n+1, \ldots$.

Exercise with Continuity III

Let $\{A_n\}$ be a monotone decreasing sequence of events. Show that $\lim_{n\to\infty} \Pr(A_n) = \Pr(\lim_{n\to\infty} A_n)$ using the property of countable additivity.

Exercise with Continuity III: **Solution**

With $B_n := A_n \setminus A_{n+1}$, from countable additivity and the result in the previous problem,

$$\Pr(A_n) = \Pr\left(L \cup \cup_{i=n}^{\infty} B_i\right) = \Pr(L) + \sum_{i=n}^{\infty} \Pr(B_i) = \Pr(L) + \lim_{k \to \infty} \sum_{i=n}^{k} \Pr(B_i).$$

Taking limits of both sides,

$$\lim_{n \to \infty} \Pr(A_n) = \Pr(L) + \lim_{n \to \infty} \lim_{k \to \infty} \sum_{i=n}^{k} \Pr(B_i) = \Pr(L) = \Pr\left(\lim_{n \to \infty} A_n\right),$$

which follows because $\sum \Pr(B_i)$ is convergent (and the Cauchy criterion applies; see, e.g., Appendix A.2, page 384).

Exercise with Continuity IV

We wish to prove that the continuity of probability and the property of countable additivity are equivalent:

Prove: For any sequence of increasing measurable events $A_1 \subset A_2 \subset \cdots$, if $\lim_{i\to\infty} \Pr(A_i) = \Pr(\lim_{i\to\infty} A_i)$, then countable additivity holds, i.e., for an arbitrary sequence of mutually exclusive measurable events B_i, $\Pr\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \Pr(B_i)$.

Exercise with Continuity IV: **Solution**

Let $\{B_i\}$ denote an arbitrary sequence of disjoint measurable events, and define $A_n = \cup_{i=1}^{n} B_i$, $n = 1, 2, \ldots$, so that $\{A_n\}$ is an increasing sequence of measurable events.

Then

$$\bigcup_{i=1}^{\infty} B_i = \lim_{n\to\infty} \bigcup_{i=1}^{n} B_i = \lim_{n\to\infty} A_n$$

and, as $\lim_{i\to\infty} \Pr(A_i) = \Pr(\lim_{i\to\infty} A_i)$,

$$\begin{aligned} \Pr\left(\bigcup_{i=1}^{\infty} B_i\right) = \Pr\left(\lim_{n\to\infty} A_n\right) &= \lim_{n\to\infty} \Pr\left(\cup_{i=1}^{n} B_i\right) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} \Pr(B_i) = \sum_{i=1}^{\infty} \Pr(B_i), \end{aligned}$$

which is countable additivity.

Course Outline

Conditional Probability

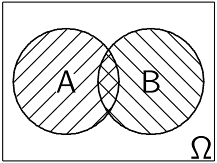
- In most applications, there will exist information which, when taken into account, alters the assignment of probability to events of interest. As a simple example, the number of customer transactions requested per hour from an on-line bank might be associated with an *unconditional* probability which was ascertained by taking the average of a large collection of hourly data.
- However, the *conditional* probability of receiving a certain number of transactions might well depend on the time of day, the arrival of relevant economic or business news, etc. If these events are taken into account, then more accurate probability statements can be made. Other examples include the number of years a manufacturing product will continue to work, conditional on various factors associated with its operation, and the batting average of a baseball player conditional on the opposing pitcher, etc.

Conditional Probability (2)

- If $\Pr(B) > 0$, then the *conditional probability of event A given the occurrence of event B*, or just the *probability of A given B*, is

$$\Pr(A \mid B) = \frac{\Pr(AB)}{\Pr(B)}.$$

- This definition is motivated by observing that the occurrence of event B essentially reduces the relevant sample space, as indicated in the Venn diagram.
- The probability of A given B is the intersection of A and B, scaled by $\Pr(B)$. If $B = \Omega$, then the scaling factor is just $\Pr(\Omega) = 1$, which coincides with the unconditional case.



Independent Events

- If the occurrence or "non-occurrence" of event B does not influence that of A, and visa-versa, then the two events are said to be *independent*, i.e., $\Pr(A \mid B) = \Pr(A)$ and $\Pr(B \mid A) = \Pr(B)$.
- From the definition of conditional probability, if events A and B are independent, then $\Pr(AB) = \Pr(A)\Pr(B)$. This is also referred to as *pairwise* independence.
- In general, events A_i, $i = 1, \ldots, n$ are *mutually* or *completely* independent if, and only if, for every collection $A_{i_1}, A_{i_2}, \ldots, A_{i_m}$, $1 \leq m \leq n$,

$$\Pr\left(A_{i_1} A_{i_2} \cdots A_{i_m}\right) = \prod_{j=1}^{m} \Pr\left(A_{i_j}\right).$$

Independent Events (2)

- For $n = 3$, this means that

$$\Pr(A_1 A_2) = \Pr(A_1)\Pr(A_2),$$

$$\Pr(A_1 A_3) = \Pr(A_1)\Pr(A_3),$$

$$\Pr(A_2 A_3) = \Pr(A_2)\Pr(A_3),$$

  and $\Pr(A_1 A_2 A_3) = \Pr(A_1)\Pr(A_2)\Pr(A_3)$.

- That pairwise independence does not imply mutual independence is referred to as *Bernstein's Paradox*.
- Letting B_i be either A_i or A_i^c, $i = 1, \ldots, n$, the events B_i are mutually independent.

Example

- Independent events $\{A, B, C\}$ occur with $\Pr(A) = a$, $\Pr(B) = b$, $\Pr(C) = c$. Let E be the event that at least one of $\{A, B, C\}$ occur. Using the complement, we can write

$$\begin{aligned} \Pr(E) &= 1 - \Pr\{\text{neither A nor B nor C occur}\} \\ &= 1 - \Pr(A^c B^c C^c) = 1 - (1-a)(1-b)(1-c) \end{aligned}$$

  from the independence of A,B and C.

- Alternatively, from Poincaré's theorem,

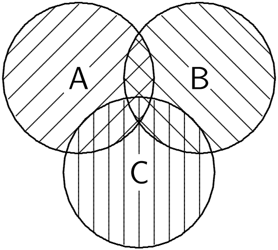
$$\begin{aligned} \Pr(E) &= \Pr(A \cup B \cup C) \\ &= a + b + c - ab - ac - bc + abc. \end{aligned}$$

Example (2)

- A third method results by observing that $A \cup B \cup C$ is the same as $A \cup A^c B \cup A^c B^c C$, as best seen using the Venn diagram. As the three events A, $A^c B$ and $A^c B^c C$ are nonoverlapping,

$$\begin{aligned}\Pr(E) &= \Pr(A) + \Pr(A^c B) + \Pr(A^c B^c C) \\ &= a + (1-a)\,b + (1-a)(1-b)\,c.\end{aligned}$$

- Straightforward algebra shows the equivalence of all the solutions.



Total Probability

- From a Venn diagram with (overlapping) events A and B, event A may be partitioned into mutually exclusive events AB and AB^c, so that

$$\begin{aligned}\Pr(A) &= \Pr(AB) + \Pr(AB^c) \\ &= \Pr(A \mid B)\Pr(B) + \Pr(A \mid B^c)(1 - \Pr(B)).\end{aligned}$$

- This is best understood as expressing $\Pr(A)$ as a *weighted sum of conditional probabilities in which the weights reflect the occurrence probability of the conditional events*.
- In general, if events B_i, $i = 1, \ldots, n$ are exclusive and exhaustive, then the **law of total probability** states that

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i).$$

Example

- Interest centers on the probability of getting at least three girls in a row among seven children. Assume that each child's sex is independent of the all the others and let $p = \Pr(\text{girl on any trial})$.
- Denote the event that three girls in a row occur as R and the total number of girls as T. Then, from the law of total probability,

$$\Pr(R) = \sum_{t=0}^{7} \Pr(R \mid T = t) \Pr(T = t).$$

  Clearly, $\Pr(R \mid T = t) = 0$ for $t = 0, 1, 2$ and $\Pr(R \mid T = 6) = \Pr(R \mid T = 7) = 1$.
- For $T = 3$, there are only 5 possible configurations, i.e.,

$$gggbbbb, \; bgggbbb, \ldots, bbbbggg,$$

  so that

$$\Pr(R \mid T = 3) = \frac{5p^3 (1-p)^4}{\binom{7}{3} p^3 (1-p)^4} = \frac{5}{35}.$$

Example (2)

- Some work shows that $\Pr(R \mid T=4) = 16/\binom{7}{4} = 16/35$ and

$$\Pr(R \mid T=5) = 18/\binom{7}{5} = 18/21,$$

so that

$$\begin{aligned}\Pr(R) &= 0+0+0+\frac{5}{35}\binom{7}{3}p^3(1-p)^4 \\ &\quad + \frac{16}{35}\binom{7}{4}p^4(1-p)^3 + \frac{18}{21}\binom{7}{5}p^5(1-p)^2 \\ &\quad + \binom{7}{6}p^6(1-p) + p^7 = 5p^3 - 4p^4 - p^6 + p^7.\end{aligned}$$

- For $p = 1/2$, $\Pr(R) \approx 0.367$.

Bayes' Rule

- From the law of total probability, **Bayes' rule** is given by

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A \mid B)\Pr(B) + \Pr(A \mid B^c)\Pr(B^c)}.$$

- For mutually exclusive and exhaustive events B_i, $i = 1, \ldots, n$, the *general Bayes' rule* is given by

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i)}.$$

Example of Bayes' Rule

- A very important example of Bayes' rule is the following.
- A test for a disease possesses the following accuracy. If a person has the disease (event D), the test detects it 95% of the time; if a person does *not* have the disease, the test will falsely detect it 2% of the time.
- Let d_0 denote the *prior probability* of having the disease before the test is conducted. (This could be taken as an estimate of the proportion of the relevant population believed to have the disease).
- Assume that, using this test, a person is detected as having the disease (event $+$).
- To find the probability that the person actually has the disease, given the positive test result, we use Bayes' rule,

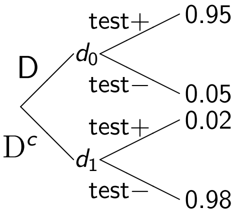
$$\Pr(D \mid +) = \frac{0.95 d_0}{0.95 d_0 + 0.02 (1 - d_0)}.$$

Example (cont.)

The answer may also be graphically depicted via a tree diagram.

The probability of an end-node is the product of the "branch" probabilities starting from the left, where $d_1 = 1 - d_0$.

Then $\Pr(D \mid +)$ is obtained as the ratio of the end-node $\{D \cap +\}$ and all branches (including $\{D \cap +\}$) leading to $+$.



Example (cont.)

- For a rare disease such that $d_0 = 0.001$, $\Pr(D \mid +)$ is only 0.045!
- There is evidence to suggest that many medical doctors are not capable of this calculation and vastly overestimate the probability of having a disease given a positive test result; see Gerd Gigerenzer's "Reckoning with Risk" (2002) for numerous examples and some of the social and economic implications of this.
- To vastly aid understanding of Bayes' rule in this context, Gigerenzer recommends expressing things not in terms of probabilities, but rather in "natural frequencies".
- For example...

Example (cont.)

- (Gigerenzer, 2002, p. 41) Consider posing the following question to a physician (let alone a layperson):
- Referring to asymptomatic (when the patient does not experience any noticeable symptoms) women aged 40 to 50 undergoing a routine mammography screening:

  The probability that one of these women has breast cancer is 0.8 percent. If a woman has breast cancer, the probability is 90 percent that she will have a positive mammogram. If a woman does not have breast cancer, the probability is 7 percent that she will still have a positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?

Example (cont.)

- Bayes' rule of course gives the answer: With $d_0 = 0.008$,

$$\Pr(D \mid +) = \frac{0.90 d_0}{0.90 d_0 + 0.07 (1 - d_0)} = 0.094,$$

  i.e., less than a 1 in 10 chance!
- When the question is posed as above, it is not obvious for most people to apply Bayes' rule.

Example (cont.)

- In terms of "natural frequencies", the same problem can be expressed as:

  Eight out of every 1,000 women have breast cancer. Of these 8 women with breast cancer, 7 will have a positive mammogram. Of the remaining 992 women who don't have breast cancer, some 70 will still have a positive mammogram.
  Imagine a sample of women who have positive mammograms in screening. How many of these women actually have breast cancer?

- In studies conducted by Gigerenzer, doctors are much more successful in getting the correct answer with this formulation:
- Roughly, 7 out of 70+7 women actually have breast cancer, or $1/11 = 0.091$.

Gambler's Ruin Problem

- Let two people, say A and B, repeatedly play a game, where each game is independent and $p = \Pr(\text{A wins}) = 1 - \Pr(\text{B wins})$.
- If, on any given round, A wins, she collects one dollar from B; if B wins, he gets one dollar from A.
- There is a total of T dollars at stake, and person A starts with i dollars and person B starts with $T - i$ dollars.
- Play continues until someone losses all his or her money.
- The probability that A winds up with all the money (and B goes bankrupt, or is *ruined*) is, with $r = (1-p)/p$,

$$\frac{1-r^i}{1-r^T}, \quad p \neq 1/2, \qquad \text{and} \qquad \frac{i}{T}, \quad p = 1/2.$$

Gambler's Ruin Problem

Similarly, the probability that A goes bankrupt is

$$\frac{r^i - r^T}{1 - r^T}, \quad p \neq 1/2, \qquad \text{and} \qquad \frac{T-i}{T}, \quad p = 1/2. \tag{5}$$

(Problem 3.19). Derive (5) first for $p \neq 1/2$. Hint: Let A be the event that person A is ruined, let W be the event that A wins the first round played, let $q = 1 - p$, and define

$$s_i := \Pr_i(A) := \Pr(A \mid \text{A starts with } i \text{ dollars and B starts with } T - i \text{ dollars}).$$

Use the law of total probability to derive the difference equation

$$s_i = ps_{i+1} + qs_{i-1}, \qquad 1 \leq i \leq T.$$

With $r = q/p$ and $d_i = s_{i+1} - s_i$, show that $d_i = r^i d_0$. Then determine the boundary conditions s_0 and s_T and use that $s_0 - s_T = \sum_{i=0}^{T-1} d_i$ to derive an expression for d_0. Finally, write a similar telescoping expression for $s_j - s_T$, from which the answer follows.

Gambler's Ruin Problem: Solution to Derivation

Note that, if A wins the first round played, then (because of independence of trials) the game can be viewed as "starting over" but such that now A has $i+1$ dollars and B has $T-i-1$ dollars. Thus, $\Pr_i(A \mid W) = \Pr_{i+1}(A) = s_{i+1}$. Using the law of total probability,

$$\begin{aligned} s_i = \Pr_i(A) &= \Pr_i(A \mid W)\Pr(W) + \Pr_i\left(A \mid \bar{W}\right)\Pr\left(\bar{W}\right) \\ &= s_{i+1}\; p + s_{i-1}\, q, \end{aligned}$$

i.e., $s_i = ps_{i+1} + qs_{i-1}$, $1 \leq i \leq T$, or, as $s_i = ps_i + qs_i$,

$$qs_i - qs_{i-1} = ps_{i+1} - ps_i.$$

With $r = q/p$ and $d_i = s_{i+1} - s_i$, this yields $d_i = rd_{i-1}$ or $d_i = r^i d_0$.

Gambler's Ruin Problem: Solution to Derivation

Conditioning on $i = 0$, we see that $s_0 = 1$. Similarly, $s_T = 0$. Then

$$1 = s_0 - s_T = -\sum_{i=0}^{T-1} d_i = -d_0 \sum_{i=0}^{T-1} r^i = -d_0 \frac{1 - r^T}{1 - r}$$

so that

$$d_0 = -\frac{1 - r}{1 - r^T}.$$

Similarly,

$$s_j = s_j - 0 = s_j - s_T = -d_0 \sum_{i=j}^{T-1} r^i = -d_0 \frac{r^j - r^T}{1 - r},$$

so that

$$s_j = (-1) \frac{1 - r}{1 - r^T} (-1) \frac{r^j - r^T}{1 - r} = \frac{r^j - r^T}{1 - r^T}, \qquad 0 \leq j \leq T.$$

Gambler's Ruin Problem for $p = 1/2$

The probability that A winds up with all the money (and B goes bankrupt) is, with $r = (1-p)/p$,

$$\frac{1-r^i}{1-r^T}, \quad p \neq 1/2, \qquad \text{and} \qquad \frac{i}{T}, \quad p = 1/2.$$

Apply l'Hôpital's rule to derive the latter result.

Gambler's Ruin Problem: Solution to Derivation for $p = 1/2$

We have

$$\lim_{p \to \frac{1}{2}} \frac{1 - \left(\frac{1-p}{p}\right)^i}{1 - \left(\frac{1-p}{p}\right)^T}$$

is indeterminant, l'Hôpital's rule implies

$$\lim_{p \to \frac{1}{2}} \frac{\frac{\mathrm{d}}{\mathrm{d}p}\left(1 - \left(\frac{1-p}{p}\right)^i\right)}{\frac{\mathrm{d}}{\mathrm{d}p}\left(1 - \left(\frac{1-p}{p}\right)^T\right)} = \lim_{p \to \frac{1}{2}} \frac{\frac{i}{p(1-p)}\left(\frac{1}{p}(1-p)\right)^i}{\frac{T}{p(1-p)}\left(\frac{1}{p}(1-p)\right)^T} = \frac{4i}{4T} = \frac{i}{T}.$$

Course Outline

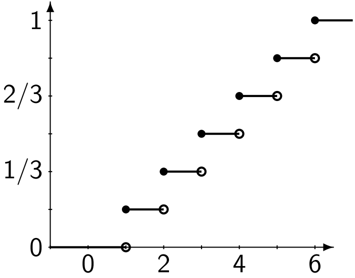
Random Variables

- A random variable (r.v.) X is in general a **function** from the sample space Ω to the real line $\mathbb{R}$. A **univariate random variable** X is such that for every $x \in \mathbb{R}$, the set $\{\omega \in \Omega : X(\omega) \leq x\}$ belongs in $\mathcal{A}$, i.e., is measurable.
- The most important function associated with a r.v. X is the **cumulative distribution function**, or **cdf**, denoted $F_X(\cdot)$ or $F(\cdot)$. It is defined to be $\Pr(X \leq x)$ for some point $x \in \mathbb{R}$.
- A cdf F has the following properties:
  - (i) $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$,
  - (ii) F is nondecreasing, i.e., if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$,
  - (iii) F is **right continuous**, i.e., $\lim_{x \to x_0^+} F(x) = F(x_0)$ for all $x_0 \in \mathbb{R}$,
  - (iv) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

Example

- A fair, 6-sided die is thrown once with the value of the top face being of interest, so that $\Omega = \{1, 2, \ldots, 6\}$ with each element being equally likely to occur. If the random variable X is this value, then $X(\omega) = \omega$ and F_X is given by the right continuous step function depicted below.

$$F_X(x) = \begin{cases} 0, & \text{if} \quad x < 1, \\ 1/6, & \text{if} \quad 1 \leq x < 2, \\ 2/6, & \text{if} \quad 2 \leq x < 3, \\ \vdots & \quad \vdots \\ 5/6, & \text{if} \quad 5 \leq x < 6, \\ 1, & \text{if} \quad x \geq 6. \end{cases}$$



Example (2)

- Note that lower case x refers to a particular point on the real line, while X is the random variable of interest; this is the usual notation used.
- Let x_i be a bounded decreasing sequence with $\lim_{i\to\infty} x_i = x_0 = 2$. Then $\lim_{i\to\infty} F(x_i) = 2/6 = F(2)$. It is easy to see that $\lim_{x\downarrow x_0} F(x) = F(x_0)$ for all $x_0 \in \mathbb{R}$, i.e., that F_X is **right continuous**.
- It is not, however, **left continuous**: If x_i is an increasing sequence with $\lim_{i\to\infty} x_i = 2$, then $\lim_{i\to\infty} F(x_i) = 1/6 \neq 2/6 = F(2)$.

Discrete Random Variables

- A random variable is said to be **discrete**, and has a **discrete distribution** if it takes on either a finite or countably infinite number of values. Thus, r.v. X in the previous example is discrete.
- If a discrete distribution (i.e., the pmf of a discrete r.v.) has support only on a set of equidistant points, then it is more precise to refer to it as a **lattice distribution**. Most of the discrete r.v.s one encounters (and all of them herein) will have lattice distributions.
- The **probability mass function**, or **pmf**, of discrete r.v. X is given by $f_X(x) = f(x) = \Pr(X = x)$.
- The **support** $\mathcal{S}$ of r.v. X is most simply defined as the subset $x \in \mathbb{R}$ such that $f_X(x) > 0$.
- Note that $f_X(x) = 0$ for any $x \notin \mathcal{S}$ and that the pmf sums to unity: $\sum_{x \in \mathcal{S}} f_X(x) = 1$.
- If there exists a number $B > 0$ such that $|x| < B$ for all $x \in \mathcal{S}_X$, then X is said to have **bounded support**, otherwise, it has unbounded support.

Definitions

- If the cdf of r.v. X is absolutely continuous, i.e., there exists function f such that, for all $x \in \mathbb{R}$,

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \quad \text{and} \quad f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x) = \frac{\mathrm{d}}{\mathrm{d}x} \int_{-\infty}^{x} f(t)\,\mathrm{d}t,$$

  then X is a **continuous random variable** and function f is denoted a **probability density function**, or **pdf**, of X.
- The **median** of a random variable X is any value $m \in \mathbb{R}$ such that the two conditions

$$\Pr(X \le m) \ge \frac{1}{2} \quad \text{and} \quad \Pr(X \ge m) \ge \frac{1}{2}$$

  are satisfied.
- If there exists a pdf f_X such that, for all a, $f_X(m + a) = f_X(m - a)$, then density $f_X(x)$ and random variable X are said to be **symmetric** about m. This is equivalent to the condition that $\Pr(X \le m - a) = \Pr(X \ge m + a)$ for all a.

Example of a Continuous r.v.

- If X follows a **standard normal** distribution, then its pdf is

$$f_X(x; 0, 1) = (2\pi)^{-1/2} \exp\left\{-\frac{x^2}{2}\right\},$$

while the more general form, sometimes referred to as the **Gaussian** distribution, is given by

$$f_N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}.$$

It is easy to see that $f_X\left(x; \mu, \sigma^2\right)$ is symmetric about μ, which is also its median.

Example of a Continuous r.v. (2)

- Now let X follow an **exponential distribution**. Its pdf is

$$f_X(x; \lambda) = \lambda \exp\{-\lambda x\} \mathbb{I}_{[0,\infty)}(x)$$

with $\lambda \in \mathbb{R}_{>0}$. The cdf is

$$F_X(x; \lambda) = \int_0^x \lambda \exp\{-\lambda t\} \, \mathrm{d}t = 1 - \exp\{-\lambda x\},$$

so that the median is given by the solution to

$$1/2 = 1 - \exp\{-\lambda x\},$$

or $x = \lambda^{-1} \ln 2$.

Discrete Sampling Schemes

- Assume that objects are consecutively and randomly drawn from a known population, either **with** or **without replacement**.
- We assume that each element is equally likely to be drawn.
- If a sampling scheme were desired in which $\Omega = \{\text{red}, \text{white}, \text{blue}\}$ and such that red is 1.4 times as likely as either white or blue, then an urn with 14 red, 10 white and 10 blue marbles could be used.
- For both sampling schemes, either a fixed number of draws, n, is set in advance, ...
- or trials continue until given numbers of objects from each class are obtained. This gives rise to four possible sampling schemes.
- For the common univariate r.v.s, the population consists of only two different distinguishable elements, labeled **"success"** and **"failure"**.

Bernoulli

- A **Bernoulli** r.v. X has support $\{0,1\}$ and takes on the value one ("success") with probability p or zero ("failure") with probability $1-p$. The mass function is

$$\Pr(X=x)=\begin{cases} p, & \text{if } x=1, \\ 1-p, & \text{if } x=0, \end{cases}$$

  and zero otherwise.

- By using the indicator function, the pmf can be written as

$$\Pr(X=x)=f_X(x)=p^x(1-p)^{1-x}\,\mathbb{I}_{\{0,1\}}(x).$$

- Only a single draw, therefore: no need to specify whether trials are conducted with or without replacement.
- For notational convenience, the abbreviation $X\sim\mathrm{Ber}(p)$ means "the random variable X is Bernoulli distributed with parameter p".

Binomial

- If an urn contains N white and M black marbles, and n of them are randomly withdrawn **with** replacement, then $X =$ **the number of white marbles drawn** is a random variable with a **binomial** distribution.
- Letting $p = N/(N+M)$ be the probability of drawing a white one (on any and all trials), we write $X \sim \text{Bin}(n, p)$ with pmf

$$f_X(x) = f_{\text{Bin}}(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \mathbb{I}_{\{0,1,\ldots,n\}}(x).$$

- This follows from the independence of trials and noting that there are $\binom{n}{x}$ possible orderings with x white and $n-x$ black marbles.

Example

- Imagine a tournament between A and B in which rounds, or games, are repeatedly played against one another.
- Assume the probability that, in a particular round, A wins is 0.3, that B wins is 0.2, and that a tie results is 0.5.
- If 10 rounds are played, the probability that there are exactly 5 ties can be computed by noting that the number of ties is binomial with $p = 0.5$ and $n = 10$, yielding

$$\binom{10}{5} 0.5^5 0.5^5 = \frac{252}{1024} \approx 0.246.$$

Useful result

- Consider a fixed number of trials with replacement, but with possibly more than two possible outcomes.
- Let A_i be the event that outcome type i, $i = 1, 2$, occurs on any particular trial. Then, in a sequence of trials, the probability that A_1 occurs **before** A_2 occurs is given by

$$\frac{\Pr(A_1)}{\Pr(A_1) + \Pr(A_2)}.$$

- If there are only two possible outcomes for each trial, then this reduces to just $\Pr(A_1)$.

Example Persons A and B conduct their tournament in a "sudden death" manner. The probability that A wins a particular round is 0.3, that B wins is 0.2, and that a tie results is 0.5. The probability that A wins the tournament is then $0.3/(0.3 + 0.2) = 0.6$ from the previous formula.

Example: Continued

What if, instead of sudden death, they play 20 rounds. What is the probability that A won 7 rounds, given that 10 of 20 ended in a tie?

- Similarly, the answer follows from the binomial quantity $\binom{10}{7}(0.6)^7(0.4)^3 \approx 0.215$.
- Alternatively, the conditional probability is given by

$$\frac{\Pr(\text{A wins 7} \cap \text{ B wins 3} \cap \text{ 10 ties})}{\Pr(\text{10 ties in 20 rounds})}.$$

- Note that the numerator is similar to binomial probabilities, but there are three possible outcomes on each trial instead of just two. Generalizing in an obvious way,

$$\frac{\Pr(\text{A wins 7} \cap \text{ B wins 3} \cap \text{ 10 ties})}{\Pr(\text{10 ties in 20 rounds})} = \frac{\binom{20}{7,3,10}(0.3)^7(0.2)^3(0.5)^{10}}{\binom{20}{10}(0.5)^{20}}$$

  which reduces to $\binom{10}{7}(0.6)^7(0.4)^3 \approx 0.215$.
- This is actually a special case of the **multinomial** distribution.

Hypergeometric

- An urn contains N white and M black balls.
- n balls are randomly withdrawn **without** replacement.
- $X =$ **the number of white balls drawn** is a random variable with a **hypergeometric** distribution.
- We write $X \sim \text{HGeo}\,(N, M, n)$ with

$$f_{\text{HGeo}}\,(x; N, M, n) = \frac{\binom{N}{x}\binom{M}{n-x}}{\binom{N+M}{n}} \mathbb{I}_{\{\max(0,n-M),1,\ldots,\min(n,N)\}}\,(x)\,.$$

- The range of x follows from the constraints in the two numerator combinatorics, i.e., $0 \leq x \leq N$ together with $0 \leq n - x \leq M \Leftrightarrow n - M \leq x \leq n$.
- These are very intuitive: If $N = M = 5$ and $n = 6$, then x has to be at least one and can be at most five.
- That $\sum_x f_X\,(x) = 1$ follows directly from Appendix A.
- Notice that, on any trial i, $i = 1, \ldots, n$, the probability of drawing a white ball depends on the outcomes of trials $1, 2, \ldots, i - 1$.

Exercise

In a certain village, there are n young men and n young women who are available for marriage. A plague strikes, randomly killing d of the $2n$ young people, $0 < d < 2n$. Let X be the number of possible marriages which can occur after the plaque.

- First start with $n = 3$ and $d = 2$, and show that $\Pr(X = 1) = 2/5$, $\Pr(X = 2) = 3/5$, and $\Pr(X = x) = 0$ for $x \notin \{1, 2\}$.
- Derive the p.m.f. of X for general n and d, and be sure to work out the support of X.
- Finally, write a Matlab program which simulates "the plague" to help confirm your algebraic results.

Solution

- The event $X = x$ is the same as "there are exactly x remaining men, and at least x remaining women, or visa versa".
- Working with the former option, there are $\binom{n}{n-x}$ ways of choosing $n-x$ men for death (so that x remain), and thus $d-(n-x)$ women have to die, which there are $\binom{n}{d-(n-x)}$ ways. This would imply that

$$\Pr(X = x) \stackrel{?}{=} 2\frac{\binom{n}{n-x}\binom{n}{d-(n-x)}}{\binom{2n}{d}}, \tag{6}$$

  where the factor of two enters because of the "visa versa" role of men and women.
- However, in the simple case with $n = 3$ and $d = 2$ worked out above, (6) yields $\Pr(X = 1) = 2/5$, which is correct, but $\Pr(X = 2) = 6/5$, which is off by a factor of two. We return to this after we work out the support of X.

Solution (2)

- From the two combinatorics in the numerator of (6), e.g., for the latter, we require that $d - (n - x) \geq 0$ and $n \geq d - (n - x)$, we see that
$$0 \leq x \leq n \quad \text{and} \quad n - d \leq x \leq 2n - d.$$
- However, there is another constraint: Recall that event $X = x$ implies that there are least x remaining women, i.e., $n - [d - (n - x)] \geq x$, or $x \leq (2n - d)/2$.
- Combining these, we see that the support of X is
$$\mathcal{S}_X = \left\{ x \in \mathbb{N} : \max(0, n - d) \leq x \leq \frac{2n - d}{2} \right\}. \qquad (7)$$

Solution (3)

- If d is even, then x can be exactly $(2n-d)/2$, whereas when d is odd, x can be at most $(2n-d-1)/2$.
  For d even and x at its maximum of $x=(2n-d)/2$, this means that an equal number, $d/2$, of men and women are killed, so we do not have to take into account the "visa versa" in the description of event $X=x$ above.
- Putting all this together, we get

$$f_X(x) = \mathbb{I}_{\mathcal{S}_X}(x) \frac{\binom{n}{n-x}\binom{n}{d-(n-x)}}{\binom{2n}{d}} \times \begin{cases} 1, & \text{if } d \text{ is even and } x = \frac{2n-d}{2}, \\ 2, & \text{otherwise,} \end{cases}$$

  where $\mathcal{S}_X$ is given in (7).

Matlab Solution (1)

This function simulates one plague.

```
function pairs=bernoullijar(n,d)
% n indistinguishable black and n indistinguishable white
%    balls in an urn, remove d of them.
% This simulates how many black/white pairs are left.
jar=[ones(n,1); 2*ones(n,1)];
  % Our jar, or urn, contains n ones, and n twos
rem=2*n;  % how many balls remaining in the jar
for i=1:d % take d of them out
  w=unidrnd(rem,1,1); % random number (index) between 1 and rem
  jar=[jar(1:w-1) ; jar(w+1:end)];  % remove that ball
  rem=rem-1; % and so there is one less ball in the jar
end
pairs = min( sum(jar==1), sum(jar==2) );
 % possible number of black/white pairs
 % is dictated by minimum of the two.
```

Matlab Solution (2)

The following code can be used to compare the empirical and true probabilities.

```
n=12; d=14; sim=1e5; pairs=zeros(sim,1);
for i=1:sim
  pairs(i)=bernoullijar(n,d);
end, tabulate(pairs)
lo=max(0,n-d); hi=floor((2*n-d)/2);
p=[]; for x=lo:hi; p=[p C(n,n-x)*C(n,d-n+x) ]; end
p=2*p/C(2*n,d);
if d==2*(floor(d/2)) % in which case, it is even
  p(end)=p(end)/2;
end
true_prob_times_100 = [lo:hi ; 100*p]
```

Sample Output

```
true_prob_times_100 = [lo:hi ; 100*p]
  Value     Count    Percent
      0         4      0.00%
      1       265      0.27%
      2      3417      3.42%
      3     17840     17.84%
      4     46377     46.38%
      5     32097     32.10%

true_prob_times_100 =
     0    1.0000    2.0000    3.0000    4.0000    5.0000
0.0067    0.2692    3.3315   17.7682   46.6415   31.9828
```

Geometric

- The random variable X follows a **geometric** distribution with parameter $p \in (0,1)$ and pmf

$$f_{\text{Geo}}(x;p) = p\,(1-p)^{x}\,\mathbb{I}_{\{0,1,\ldots\}}(x)$$

  if it represents the number of "failed" Bernoulli trials required until (**and not including**) a "success" is observed.
- This is denoted $X \sim \text{Geo}\,(p)$.
- Alternatively, X can be defined as the number of trials which are observed until (**and including**) the first "success" occurs, i.e.,

$$f_X\,(x;p) = p\,(1-p)^{x-1}\,\mathbb{I}_{\{1,2,\ldots\}}(x)\,.$$

- Recall the example in which your colleague will keep having children until the first son. The total number of children can be modeled with this mass function. With $p = 0.5$, there is a 1 in 16 chance that she will have more than four children.

Negative Binomial

- A generalization of the geometric is the **negative binomial**, which represents the number of "failures" observed until r "successes" are observed.
- For $X \sim \text{NBin}(r, p)$, the pmf of X is given by

$$f_{\text{NBin}}(x; r, p) = \Pr(X = x) = \binom{r+x-1}{x} p^r (1-p)^x \mathbb{I}_{\{0,1,\ldots\}}(x)$$

- Function $f_{\text{NBin}}(x; r, p)$ takes its form from the independence of trials and noting that the $(r+x)^{\text{th}}$ trial must be a success and that $r-1$ successes must occur within the first $r+x-1$ trials.

Inverse Hypergeometric

- The 4th sampling scheme is for when trials continue until k "successes" are obtained, but sampling is **without** replacement.
- This is referred to as the **inverse hypergeometric** distribution and arises in many useful applications.
- If an urn contains w white and b black balls, the probability that a total of x balls need to be drawn to get k white balls, $1 \leq k \leq w$, is given by

$$\Pr(X = x) = \binom{x-1}{k-1} \frac{\binom{w+b-x}{w-k}}{\binom{w+b}{w}} \mathbb{I}_{\{k,k+1,\ldots,b+k\}}(x), \quad 1 \leq k \leq w,$$

denoted $X \sim \mathrm{IHyp}(k, w, b)$.

Inverse Hypergeometric (2)

- To see this for general k and x, where $1 \leq k \leq w$ and $k \leq x \leq b + k$, the x^{th} draw must be the k^{th} white ball, while the previous $x - 1$ trials must have produced $k - 1$ white balls and $x - k$ black balls (in any order).
- This latter requirement is hypergeometric; thus,

$$\Pr(X = x) = \frac{w - (k-1)}{w + b - (x-1)} \frac{\binom{w}{k-1}\binom{b}{x-k}}{\binom{w+b}{x-1}}$$

which, upon rewriting, is precisely the pmf above.

Poisson

For $\lambda > 0$, density of $X \sim \text{Poi}(\lambda)$ is

$$f_{\text{Poi}}(x;\lambda) = \Pr(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}\mathbb{I}_{\{0,1,\dots\}}(x).$$

- There is no obvious sampling scheme which gives rise to this distribution. Instead, it turns out that all four previous schemes (and many others) asymptotically behave like a Poisson.
- It is often an accurate approximation involving far less computation.
- Consider the binomial. If $X \sim \text{Bin}(n, p)$ and $np = \lambda$, then

$$\begin{aligned}\Pr(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{(n-x)!x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \underbrace{\frac{n(n-1)\cdots(n-x+1)}{n^x}}_{\to 1} \frac{\lambda^x}{x!} \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-x}}_{\to 1},\end{aligned}$$

as $n \to \infty$ and $p \to 0$, so that, for large n and small p, $\Pr(X = x) \approx e^{-\lambda}\lambda^x/x!$.

Example

- A new border patrol is set up to inspect 1000 cars a day. If it is known that the probability of finding a violator of some sort is 0.001, we can approximate the binomial r.v. X, the number of violators caught, as a Poisson with $\lambda = np = 1$.
- The probability of finding at least three violators is then approximately

$$1 - \left(\frac{e^{-1} 1^0}{0!} + \frac{e^{-1} 1^1}{1!} + \frac{e^{-1} 1^2}{2!} \right) = 1 - \frac{5}{2} e^{-1} \approx 0.0803014.$$

The exact answer is 0.0802093.

Moments: Expected Value

For convenience, define the integral

$$\int_{-\infty}^{\infty} g(x)\, \mathrm{d}F_X(x) = \begin{cases} \int_{\mathcal{S}} g(x)\, f_X(x)\, \mathrm{d}x, & \text{if } X \text{ is continuous,} \\ \sum_{i \in \mathcal{S}} g(x_i)\, f_X(x_i), & \text{if } X \text{ is discrete,} \end{cases}$$

(assuming the right hand side exists), where $\mathcal{S}$ is the support of X and $g(x)$ is a real-valued function.

- The **expected value** of random variable X is

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x\, \mathrm{d}F_X(x).$$

Moments: Examples

- For the geometric distribution

$$f_{\mathrm{Geo}}(x;p) = p\,(1-p)^x\,\mathbb{I}_{\{0,1,\ldots\}}(x)$$

with $q = 1 - p$,

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x f_X(x) = p \sum_{x=0}^{\infty} x q^x =: pS_1,$$

where

$$\begin{aligned} S_1 &= q + 2q^2 + 3q^3 + \cdots \\ qS_1 &= \quad\;\; q^2 + 2q^3 + 3q^4 + \cdots \\ S_1 - qS_1 &= q + q^2 + q^3 + \cdots = \frac{q}{1-q} \end{aligned}$$

so that

$$S_1 = \frac{q}{(1-q)^2}, \qquad \mathbb{E}[X] = pS_1 = \frac{1-p}{p}.$$

Moments: Examples (2)

- For binomial, negative binomial, hypergeometric and inverse hypergeometric, there exist far more expedient ways of evaluating the mean.
- The mean for $X \sim \text{Bin}(n, p)$ will be shown later to be $\mathbb{E}[X] = np$.
- If $X \sim \text{Poi}(\lambda)$, then

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} \frac{xe^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda.$$

This value should not be surprising given the relationship between the binomial and Poisson.

Moments

- Let $g(X)$ be a real–valued function of r.v. X, so that $Y = g(X)$ is itself a random variable.
- To compute $\mathbb{E}[g(X)]$, one could first obtain the density of $g(X)$ and then use the definition of expected value above.
- However, it can be shown that

$$\mathbb{E}[g(X)] = \int_{\mathcal{S}_X} g(x) \, \mathrm{d}F_X(x),$$

i.e., a direct computation is possible without having to first compute the density f_Y.

Moments

- By taking $Y = g(X) = X^r$, the r^{th} **raw moment** of r.v. X is defined as
$$\mu_r' = \mathbb{E}[X^r] = \int_{\mathcal{S}} x^r \, \mathrm{d}F_X(x),$$
while the r^{th} **central moment** of X is defined by
$$\mu_r = \mathbb{E}\left[(X-\mu)^r\right] = \int_{\mathcal{S}} (X-\mu)^r \, \mathrm{d}F_X(x),$$
recalling that $\mu = \mathbb{E}[X]$.
- The second central moment, μ_r, plays an important role in many statistical models and is referred to as the **variance** of X:
$$\mu_2 = \mathrm{Var}(X) = \int_{\mathcal{S}} (x-\mu)^2 \, \mathrm{d}F_X(x) = \mu_2' - \mu^2$$
and often denoted by σ^2.
- The **standard deviation** of a r.v. is defined to be $\sigma := \sqrt{\mu_2}$.

Moments: Example

- Using geometric pmf

$$f_X(x;p) = p(1-p)^{x-1}\,\mathbb{I}_{\{1,2,\dots\}}(x),$$

the 2nd raw moment $\mathbb{E}\left[X^2\right]$ is

$$\mathbb{E}\left[X^2\right] = \sum_{x=1}^{\infty} x^2 f_X(x) = p\sum_{x=1}^{\infty} x^2 q^{x-1} = p\sum_{x=1}^{\infty} x^2 q^{x-1} =: pS_2,$$

where

$$\begin{aligned}
S_2 &= 1 + 4q + 9q^2 + 16q^3 + \cdots + x^2 q^{x-1} + \cdots \\
qS_2 &= q + 4q^2 + 9q^3 + \cdots + (x-1)^2 q^x + \cdots \\
S_2 - qS_2 &= 1 + 3q + 5q^2 + 7q^3 + \cdots + (2x-1)\,q^{x-1} + \cdots \\
&= \sum_{i=0}^{\infty} (2i+1)\,q^i = 2S_1 + \sum_{i=0}^{\infty} q^i = 2\frac{q}{(1-q)^2} + \frac{1}{1-q}.
\end{aligned}$$

Moments: Example (2)

Therefore

$$S_2 = 2\frac{q}{(1-q)^3} + \frac{1}{(1-q)^2} = \frac{1+q}{(1-q)^3},$$

so that

$$\mathbb{E}\left[X^2\right] = \frac{2-p}{p^2}$$

and

$$\mathrm{Var}\,(X) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2},$$

which holds for both geometric pmf forms because $\mathrm{Var}\,(X-1) = \mathrm{Var}\,(X)$.

Moments

- A common measure of the extent to which a pdf deviates from symmetry is $\mu_3/\mu_2^{3/2}$, called **skewness**
- The **kurtosis** measures the **heaviness of the tails of the distribution**, and is given by μ_4/μ_2^2.
- Raw and central moments of X (if they exist) are related by

$$\begin{aligned} \mu_n &= \mathbb{E}\left[(X-\mu)^n\right] = \mathbb{E}\left[\sum_{i=0}^{n}\binom{n}{i}X^{n-i}(-\mu)^i\right] \\ &= \sum_{i=0}^{n}\binom{n}{i}\mathbb{E}\left[X^{n-i}\right](-\mu)^i = \sum_{i=0}^{n}(-1)^i\binom{n}{i}\mu'_{n-i}\mu^i \end{aligned}$$

and μ'_n given similarly by

$$\mathbb{E}\left[(X-\mu+\mu)^n\right] = \mathbb{E}\left[\sum_{i=0}^{n}\binom{n}{i}(X-\mu)^{n-i}\mu^i\right] = \sum_{i=0}^{n}\binom{n}{i}\mu_{n-i}\mu^i.$$

Moments

- For low order moments, these simplify to

$$\begin{aligned} \mu_2' &= \mu_2 + \mu^2, & \mu_2 &= \mu_2' - \mu^2, \\ \mu_3' &= \mu_3 + 3\mu_2\mu + \mu^3, & \mu_3 &= \mu_3' - 3\mu_2'\mu + 2\mu^3, \\ \mu_4' &= \mu_4 + 4\mu_3\mu + 6\mu_2\mu^2 + \mu^4, & \mu_4 &= \mu_4' - 4\mu_3'\mu + 6\mu_2'\mu^2 - 3\mu^4, \end{aligned}$$

  using $\mu' = \mu$ and $\mu_1 = 0$.
- Not all random variables possess finite moments of all order; examples include the Pareto and Student's t distributions.

Jensen's Inequality

- A function f is (strictly) concave on $[a, b]$ if

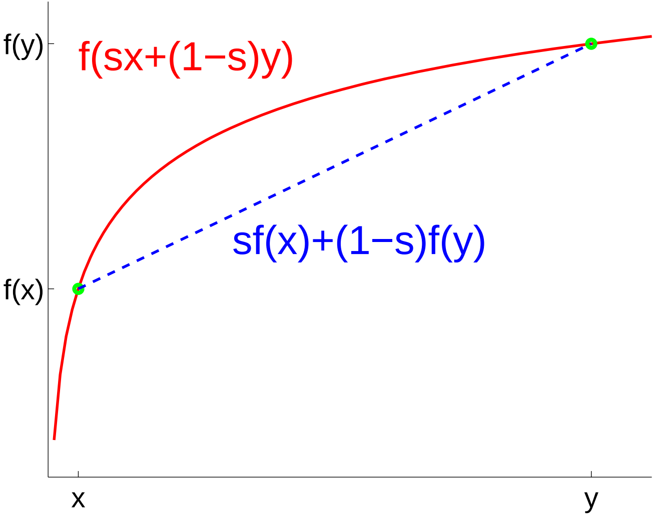
$$\forall x, y \in [a, b] \text{ and } \forall s \in [0, 1] \text{ with } t = 1 - s,$$

$$f(sx + ty) > sf(x) + tf(y).$$

- Function f is convex on $[a, b]$ iff $-f$ is concave on $[a, b]$.
- In particular, a (possibly piecewise) differentiable function f is concave on an interval if its derivative f' is non-increasing on that interval; a twice-differential function f is concave on an interval if $f'' \leq 0$ on that interval.
- *Jensen's inequality* states that, for any r.v. X with finite mean,

$$\begin{aligned} \mathbb{E}[g(X)] &\geq g(\mathbb{E}[X]), \qquad g(\cdot) \text{ convex}, \\ \mathbb{E}[g(X)] &\leq g(\mathbb{E}[X]), \qquad g(\cdot) \text{ concave}. \end{aligned}$$

Jensen's Inequality: Graphic Illustration

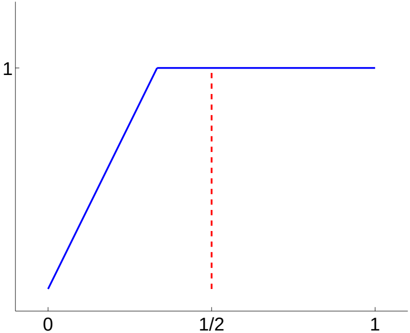


Jensen's Inequality: Intuition from the Graphic

Let $X \sim \mathrm{Unif}(0,1)$ and consider the concave function

$$g(x) = \begin{cases} 3x, & \text{if} \quad x \leq 1/3, \\ 1, & \text{if} \quad x > 1/3. \end{cases}$$

Then $g(\mathbb{E}[X]) = 1$ and $\mathbb{E}[g(X)] = \int_0^{1/3} 3x \, dx + \int_{1/3}^{1} dx = 5/6$, i.e., $\mathbb{E}[g(X)] < g(\mathbb{E}[X])$.



Jensen's Inequality (cont.)

- Let's verify Jensen's inequality for g convex and assuming $g''(x)$ exists with $g''(x) \geq 0$ for all x.
- We wish to show $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$.
- Let X be a r.v. with finite mean μ. Then, for g a twice differentiable, convex function, there exists a value ξ such that

$$g(x) = g(\mu) + g'(\mu)(x-\mu) + \frac{1}{2} g''(\xi)(x-\xi)^2 .$$

- That is, $g(x) \geq g(\mu) + g'(\mu)(x-\mu)$ for all x.
- Thus, $g(X) \geq g(\mu) + g'(\mu)(X-\mu)$.
- Take expectations of both sides.

Jensen's Inequality (cont.)

- **Example**: Assume X is a r.v. with finite mean μ. Let $g(x) = x^2$ with $g''(x) = 2$, so that g is convex. Then $\mathbb{E}\left[X^2\right] \geq \mu^2$. Note: if $\mathbb{E}\left[X^2\right] < \infty$, then $\mathbb{E}\left[X^2\right] = \mathbb{V}(X) + \mu^2$, which shows the result immediately.
- **Example**: Let X be a nonnegative r.v., and take $g(x) = \sqrt{x}$. As $g''(x)$ is negative for $x > 0$, g is concave and $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mu}$. Note: As

  $$0 \leq \mathbb{V}(\sqrt{X}) = \mathbb{E}[|X|] - (\mathbb{E}[\sqrt{X}])^2 = \mathbb{E}[X] - (\mathbb{E}[\sqrt{X}])^2,$$

  the result follows immediately.
- **Example**: Let $g(x) = \ln(x)$ for $x > 0$. Then g is concave because $g''(x) = -x^{-2}$ and $\mathbb{E}[\ln X] \leq \ln \mu$.

Jensen's Inequality: Exercise

Let X be a positive random variable. What can you say about the relative magnitudes of $\mathbb{E}[1/X]$ and $1/\mathbb{E}[X]$?

By computing $\mathrm{Cov}\left(X, X^{-1}\right)$, derive an exact expression for their difference.

Jensen's Inequality: Solution

As the function $f(x) = 1/x$ is convex, Jensen's inequality implies $\mathbb{E}[1/X] - 1/\mathbb{E}[X] > 0$.

With $\mu = \mathbb{E}[X]$, we have

$$\begin{aligned} \mathrm{Cov}\left(X, X^{-1}\right) &= \mathbb{E}\left[(X-\mu)\left(X^{-1} - \mathbb{E}\left[X^{-1}\right]\right)\right] \\ &= \mathbb{E}\left[1 - X\mathbb{E}\left[X^{-1}\right] - \mu X^{-1} + \mu\mathbb{E}\left[X^{-1}\right]\right] \\ &= 1 - \mu\mathbb{E}\left[X^{-1}\right], \end{aligned}$$

so that

$$\frac{\mathrm{Cov}(X, 1/X)}{\mathbb{E}[X]} = \frac{1}{\mathbb{E}[X]} - \mathbb{E}\left[X^{-1}\right];$$

and it is intuitively clear that $\mathrm{Cov}(X, 1/X) < 0$.

Utility Functions, Jensen's Inequality and FSD

(Example A.12). In microeconomics, a *utility function*, $U(\cdot)$, is a preference ordering for different goods of choice ("bundles" of goods and services, amount of money, etc.) For example, if bundle A is preferable to bundle B, then $U(A) > U(B)$.

Let $U : A \to \mathbb{R}$, $A \subset \mathbb{R}_{>0}$, be a continuous and twice differentiable utility function giving a preference ordering for overall wealth, W. Not surprisingly, one assumes that $U'(W) > 0$, i.e., people prefer more wealth to less, but also that $U''(W) < 0$, i.e., the more wealth you have, the less additional utility you reap upon obtaining a fixed increase in wealth.

In this case, U is a concave function and the person is said to be **risk-averse**.

Utility Functions, Jensen's Inequality and FSD

A popular choice of U is $U(W;\gamma) = W^{1-\gamma}/(1-\gamma)$ for a fixed parameter $\gamma \in \mathbb{R}_{>0} \setminus 1$ and $W > 0$.

Indeed, an easy calculation verifies that $U'(W) > 0$ and $U''(W) < 0$.

Interest centers on the limit of U as $\gamma \to 1$. In this case, $\lim_{\gamma\to 1} W^{1-\gamma} = 1$ and $\lim_{\gamma\to 1}(1-\gamma) = 0$ so that l'Hôpital's rule is not applicable.

However, as utility is a relative measure, we can let $U(W;\gamma) = \left(W^{1-\gamma} - 1\right)/(1-\gamma)$ instead. Then, from Example A.11, $(\mathrm{d}/\,\mathrm{d}\gamma)\, W^{1-\gamma} = -W^{1-\gamma}\ln W$, so that

$$\begin{aligned}
\lim_{\gamma\to 1} U(W;\gamma) &= \lim_{\gamma\to 1} \frac{W^{1-\gamma} - 1}{1-\gamma} \\
&= \lim_{\gamma\to 1} \frac{(\mathrm{d}/\,\mathrm{d}\gamma)\left(W^{1-\gamma} - 1\right)}{(\mathrm{d}/\,\mathrm{d}\gamma)(1-\gamma)} = \lim_{\gamma\to 1} W^{1-\gamma}\ln W = \ln W.
\end{aligned}$$

Utility Functions, Jensen's Inequality and FSD

(Example 4.30). Let $U(W)$ be a twice–differentiable, concave function of wealth W, i.e., $U'(W) > 0$ and $U''(W) < 0$.

Letting A be a random variable associated with the payoff of a financial investment, Jensen's inequality implies that $\mathbb{E}[U(A)] \leq U(\mathbb{E}[A])$.

The intuition behind this result is that a risk-averse person (one for whom $U''(W) < 0$) prefers a sure gain of zero (the utility of the expected value of zero) to taking a fair bet (win or lose x dollars with equal probability).

Utility Functions, Jensen's Inequality and FSD

(Example 4.21). Let A be a continuous random variable with support $S \subset \mathbb{R}$ which describes the value (cash payoff) of a certain financial investment, so that $F_A(x)$ is the probability of making less than or equal to x (say) dollars.

Similarly, let B be a r.v. referring to the payoff of a different investment. If, for every $x \in S$, $F_A(x) \leq F_B(x)$, then investment A is said to *first order stochastically dominate* investment B, or A FSD B, and A would be preferred by all (rational) investors.

This is because, for any $x \in S$, $\Pr(A > x) \geq \Pr(B > x)$, i.e., the probability of making more than x dollars is higher with investment A, for all possible x.

Utility Functions, Jensen's Inequality and FSD

As a trivial example, if investments A and B are such that $A = B + k$, $k > 0$, then clearly, A is to be preferred, and indeed, A FSD B, because

$$F_B(x) = \Pr(B \leq x) = \Pr(A \leq x + k) = F_A(x + k) \geq F_A(x),$$

using the fact that F is a nondecreasing function.

Similarly, if the support of A and B is positive and if $A = Bk$, $k > 1$, then $F_B(x) = \Pr(B \leq x) = \Pr(A \leq xk) = F_A(xk) \geq F_A(x)$.

Utility Functions, Jensen's Inequality and FSD

Use of the exponential distribution with different parameters is a special case of the latter result, recalling that the parameter in the exponential distribution is an (inverse) scale parameter.

In particular, if $0 < \lambda_1 < \lambda_2 < \infty$, then it is easy to see that, $\forall\ x > 0$, $e^{-\lambda_1 x} > e^{-\lambda_2 x}$ or $1 - e^{-\lambda_1 x} < 1 - e^{-\lambda_2 x}$, so that, if $X_i \sim \mathrm{Exp}\,(\lambda_i)$, $i = 1, 2$, then, $\forall\ x > 0$, $F_{X_1}(x) < F_{X_2}(x)$, and X_1 FSD X_2.

A distribution with finite support might make more sense in this context: Let $X_i \sim \mathrm{Beta}\,(p, q_i)$, $i = 1, 2$, with $0 < q_1 < q_2$. A graphical analysis suggests that $F_{X_1}(x) < F_{X_2}(x)\ \forall\ x \in (0, 1)$ and any $p > 0$. The reader is invited to prove this, or search for its proof in the literature.

Utility Functions, Jensen's Inequality and FSD

Let $U(W)$ be a wealth utility function as in Example A.12 and let $\mathbb{E}[U(A)]$ be the expected utility of the return on investment A.

If A FSD B, then one might expect that $\mathbb{E}[U(w+A)] \geq \mathbb{E}[U(w+B)]$ for *any* increasing utility function and any fixed, initial level of wealth w.

This is easily proven when U is differentiable, with $U'(W) > 0$, and A and B are continuous r.v.s.

Utility Functions, Jensen's Inequality and FSD

First note that, if A FSD B, then $(w + A)$ FSD $(w + B)$, so we can take $w = 0$ without loss of generality. Let interval (a, b) be the union of the support of A and B.

Integrating by parts shows that

$$\begin{aligned} \int_a^b U(x) f_A(x)\, \mathrm{d}x &= U(b) F_A(b) - U(a) F_A(a) - \int_a^b F_A(x) U'(x)\, \mathrm{d}x \\ &= U(b) - \int_a^b F_A(x) U'(x)\, \mathrm{d}x, \end{aligned}$$

as $F_A(b) = 1$ and $F_A(a) = 0$.

Utility Functions, Jensen's Inequality and FSD

Similarly, $\int_a^b U(x) f_B(x)\, dx = U(b) - \int_a^b F_B(x) U'(x)\, \mathrm{d}x$, so that

$$\begin{aligned}\mathbb{E}[U(A)] - \mathbb{E}[U(B)] &= \int_a^b U(x) f_A(x)\, \mathrm{d}x - \int_a^b U(x) f_B(x)\, \mathrm{d}x \\ &= \int_a^b [F_B(x) - F_A(x)]\, U'(x)\, \mathrm{d}x \geq 0,\end{aligned}$$

which follows because $U'(x) > 0$ and $F_B(x) > F_A(x)$ by assumption.

Course Outline

Multivariate Random Variables

- Most real statistical applications involve several random variables, because
  1. more than one random quantity is associated or observed in conjunction with the process of interest or
  2. the variable(s) of interest can be expressed as functions of two or more (possibly unobserved) random variables.
- Similar to the univariate case, the n-variate vector function

$$\mathbf{X} = (X_1, X_2, \ldots, X_n) = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) = \mathbf{X}(\omega)$$

is defined to be a (**multivariate** or **vector**) **random variable**.

Joint CDF

- The joint cdf of $\mathbf{X}$ is denoted $F_{\mathbf{X}}(\cdot)$ and defined to be the function with domain $\mathbb{R}^n$ and range $[0,1]$ given by

$$F_{\mathbf{X}}(\mathbf{x}) = \Pr(\mathbf{X} \le \mathbf{x}) := \Pr(-\infty < X_i \le x_i,\ i = 1, \ldots, n)$$

  for any $\mathbf{x} \in \mathbb{R}^n$, where vector inequalities are defined to operate **elementwise** on the components.

- The multivariate cdf has properties similar to the univariate case, but with one additional constraint. In the bivariate case, this is

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \ge 0,$$

  and in the trivariate case,

$$\begin{aligned} & F(b_1, b_2, b_3) - F(a_1, b_2, b_3) - F(b_1, a_2, b_3) - F(b_1, b_2, a_3) \\ & + F(a_1, a_2, b_3) + F(a_1, b_2, a_3) + F(b_1, a_2, a_3) - F(a_1, a_2, a_3) > 0, \end{aligned}$$

  with expressions available for the general case.

Multivariate PMF and PDF

- Similar to the univariate case, $\mathbf{X}$ is discrete if it has a finite or countably infinite support, and is continuous otherwise.
- The multivariate pmf of discrete $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \Pr(\{\omega\}) = \Pr(\omega : \omega \in \Omega \mid X_i(\omega) = x_i \quad i = 1, \ldots, n),$$

with $f_{\mathbf{X}}(\mathbf{x}) = 0$ for any $X(\omega) \notin \mathcal{S}$ and

$$1 = \sum_{\omega \in \mathcal{S}} f_{\mathbf{X}}(\{\omega\}) = \sum_{j_1=-\infty}^{\infty} \sum_{j_2=-\infty}^{\infty} \cdots \sum_{j_n=-\infty}^{\infty} \Pr(X_1 = j_1, \ldots, X_n = j_n).$$

- For continuous $\mathbf{X}$, $f_{\mathbf{X}}(\mathbf{x})$ is the multivariate pdf if, for all $A \in \mathcal{B}^n$,

$$\Pr(\mathbf{X} \in A) = \underset{A}{\int \cdots \int} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

and $\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = 1$.

Multivariate PMF and PDF

- In the continuous bivariate case,

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y},$$

  with the extension to n random continuous variables $X_1, \ldots, X_n$ analogously given.

- A typical event is when A is a rectangle in $\mathbb{R}^n$, i.e.,

$$A = \{\mathbf{x} : a_i < x_i \leq b_i, i = 1, \ldots, n\},$$

  so that

$$\Pr(\mathbf{X} \in A) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f_{\mathbf{X}}(\mathbf{x}) \, \mathrm{d}x_n \cdots \mathrm{d}x_1,$$

  which should be viewed as an **iterated integral**, i.e., it is to be evaluated from the innermost univariate integral outwards, holding $x_1, \ldots, x_{n-i}$ constant when evaluating the i^{th} one.

Example

- Let (X, Y, Z) be jointly distributed with density

$$f_{X,Y,Z}(x, y, z) = k\, xy \exp\left(-\frac{x+y+z}{3}\right) \mathbb{I}_{(0,\infty)}(x)\, \mathbb{I}_{(0,\infty)}(y)\, \mathbb{I}_{(0,\infty)}(z)\,.$$

- Note that, for constants a and b, $b > 0$

$$\int_0^\infty \exp(-a - bz)\, \mathrm{d}z = e^{-a} \int_0^\infty \exp(-bz)\, \mathrm{d}z = e^{-a} \frac{1}{b}.$$

Example (2)

Thus, the constant k is determined by

$$
\begin{aligned}
1 &= k \int_0^\infty \int_0^\infty \int_0^\infty xy \exp\left(-\frac{x+y+z}{3}\right) \mathrm{d}z\, \mathrm{d}y\, \mathrm{d}x \\
&= k \int_0^\infty \int_0^\infty xy \int_0^\infty \exp\left(-\frac{x+y+z}{3}\right) \mathrm{d}z\, \mathrm{d}y\, \mathrm{d}x \\
&= k \int_0^\infty \int_0^\infty xy \left(e^{-\frac{1}{3}(x+y)} \times 3\right) \mathrm{d}y\, \mathrm{d}x \\
&= 3k \int_0^\infty x \left(\int_0^\infty y e^{-\frac{1}{3}(x+y)} \mathrm{d}y\right) \mathrm{d}x \\
&= 3k \int_0^\infty x e^{-x/3} \mathrm{d}x \int_0^\infty y e^{-y/3} \mathrm{d}y = 3k \times 9^2 = 243k.
\end{aligned}
$$

For $\Pr(X < Y < Z)$, there are 3! integral expressions, three of which are

1. $\int_0^\infty \int_x^\infty \int_y^\infty f_{X,Y,Z}(x,y,z)\, \mathrm{d}z\, \mathrm{d}y\, \mathrm{d}x = \frac{7}{108}$,
2. $\int_0^\infty \int_0^z \int_0^y f_{X,Y,Z}(x,y,z)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z = \frac{7}{108}$,
3. $\int_0^\infty \int_x^\infty \int_x^z f_{X,Y,Z}(x,y,z)\, \mathrm{d}y\, \mathrm{d}z\, \mathrm{d}x = \frac{7}{108}$.

More General Events

Let $A = \{\mathbf{x} : a_i(\mathbf{x}_{i-1}) < x_i \leq b_i(\mathbf{x}_{i-1}),\ i = 1, \ldots, n\}$, where $\mathbf{x}_j := (x_1, \ldots, x_j)$, i.e., bounds a_i and b_i are functions of $x_1, x_2, \ldots, x_{i-1}$.

Example Let $f_{X,Y}(x, y) = e^{-y}\mathbb{I}_{(0,\infty)}(x)\,\mathbb{I}_{(x,\infty)}(y)$. That is, $f_{X,Y}(x, y) > 0$ if and only if $0 < x < y < \infty$.

- Then $\Pr(aX < Y) = 1$ for $a \leq 1$ and, for $a > 1$,

$$\begin{aligned}\Pr(aX < Y) &= \iint\limits_{ax<y} f_{X,Y}(x, y)\,\mathrm{d}y\,\mathrm{d}x = \int_0^\infty \int_{ax}^\infty e^{-y}\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_0^\infty e^{-ax}\,\mathrm{d}x = \frac{1}{a}, \quad a > 1.\end{aligned}$$

Exercise Repeat with $\mathrm{d}x\,\mathrm{d}y$ instead of $\mathrm{d}y\,\mathrm{d}x$.

More General Events

Example Let

$$\begin{aligned} f_{X,Y}(x,y) &= abe^{-ax-by}\mathbb{I}_{(0,\infty)}(x)\,\mathbb{I}_{(0,\infty)}(y) \\ &= ae^{-ax}\mathbb{I}_{(0,\infty)}(x) \times be^{-by}\mathbb{I}_{(0,\infty)}(y)\,. \end{aligned}$$

- Then

$$\begin{aligned} \Pr(X<Y) &= \int_0^\infty \int_x^\infty f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_0^\infty ae^{-ax}\left(\int_x^\infty be^{-by}\,\mathrm{d}y\right)\mathrm{d}x \\ &= \int_0^\infty ae^{-ax}\left(1-F_Y(x)\right)\mathrm{d}x, \quad F_Y(t) = 1-e^{-bt} \\ &= \int_0^\infty ae^{-ax}e^{-bx}\,\mathrm{d}x = \int_0^\infty ae^{-(a+b)x}\,\mathrm{d}x = \frac{a}{a+b}. \end{aligned}$$

- For $a=b$, $\Pr(X<Y) = 1/2$, as we would expect.

Marginal Distributions

- For a given n-variate cdf $F_{\mathbf{X}}$, interest often centers on only a subset of $\{X_i, i = 1, \ldots, n\}$.
- The pdf (cdf) of this subset is referred to as the *marginal* density (distribution) for the chosen subset.
- There are a total of $2^n - 2$ marginal distributions.
- In the bivariate ($n = 2$) continuous case, the marginal pdfs are

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y, \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x.$$

- The marginal cdfs are

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}x \\ F_Y(y) &= \int_{-\infty}^{y}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{y} f_Y(y)\,\mathrm{d}y. \end{aligned}$$

- The general case is treated similarly.

Example

- Let $f_{X,Y}(x,y) = e^{-y}\mathbb{I}_{(0,\infty)}(x)\,\mathbb{I}_{(x,\infty)}(y)$.
- Then
$$f_X(x) = \mathbb{I}_{(0,\infty)}(x)\int_x^\infty e^{-y}\,\mathrm{d}y = e^{-x}\mathbb{I}_{(0,\infty)}(x),$$
so that $X \sim \mathrm{Exp}(1)$.
- With $0 < x < y < \infty$,
$$f_Y(y) = e^{-y}\int_0^y \mathrm{d}x = ye^{-y}\mathbb{I}_{(0,\infty)}(y).$$
- To check that this is a valid density, let $u = y$ (so that $\mathrm{d}u = \mathrm{d}y$) and $\mathrm{d}v = e^{-y}\,\mathrm{d}y$ (so that $v = -e^{-y}$),
$$\begin{aligned}\int_0^\infty ye^{-y}\,\mathrm{d}y &= \int u\,\mathrm{d}v = uv - \int v\,\mathrm{d}u \\ &= -ye^{-y}\Big|_0^\infty - \int\left(-e^{-y}\right)\,\mathrm{d}y \\ &= 0 + 1.\end{aligned}$$

Independence

- Informally, a set of r.v.s are independent if they have "nothing whatsoever to do with one another".
- More formally, r.v.s $X_1, \ldots, X_n$ are said to be **mutually independent** (or just independent) if, for all rectangles $I_{\mathbf{a},\mathbf{b}} = I_{(a_1,a_2,\ldots,a_n),(b_1,b_2,\ldots,b_n)}$ for which $a_i \leq b_i$, $i = 1, \ldots, n$,

$$\Pr(\mathbf{X} \in I) = \Pr(X_1 \in I_{a_1,b_1}, \ldots, X_n \in I_{a_n,b_n}) = \prod_{i=1}^{n} \Pr(X_i \in I_{a_i,b_i}).$$

- As a special case, this implies that the joint c.d.f. can be expressed as $F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} F_{X_i}(x_i)$.
- This definition is equivalent with one in terms of the pdf of $\mathbf{X}$: The r.v.s $X_1, \ldots, X_n$ are independent iff their joint density can be factored as

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

IID

- If n random variables $X_1, \ldots, X_n$ are not only independent but each follow the same distribution, i.e., $F_{X_i} = F_X$, $i = 1, \ldots, n$, for some distribution F_X, then $X_1, \ldots, X_n$ are said to be **independently and identically distributed**, which is often abbreviated as **iid**.
- This is expressed as $X_i \overset{\text{iid}}{\sim} f_X$.
- If the X_i are independent and described or **indexed** by the same family of distributions but with different parameters, say θ_i, we write $X_i \overset{\text{ind}}{\sim} f_{X_i}(x_i; \theta_i)$.
- To emphasize that the functional form of f is the same, use the distributional name, e.g., $X_i \overset{\text{ind}}{\sim} \mathrm{Ber}(p_i)$.
- Question: If $X_i \overset{\text{iid}}{\sim} \mathrm{Ber}(p)$, then how is $\sum_{i=1}^n X_i$ distributed?
- Question: If $X_i \overset{\text{iid}}{\sim} \mathrm{Geo}(p)$, then how is $\sum_{i=1}^n X_i$ distributed?

Moments

- The expected value of a function $g(\mathbf{X})$ for $g : \mathbb{R}^n \to \mathbb{R}$, with respect to the n-length vector random variable $\mathbf{X}$ with pmf or pdf $f_{\mathbf{X}}$ is defined by

$$\mathbb{E}[g(\mathbf{X})] = \int_{\mathbf{x} \in \mathbb{R}^n} g(\mathbf{X}) \, \mathrm{d}F_{\mathbf{X}}(\mathbf{x}).$$

- For example, if

$$f_{X,Y}(x, y) = abe^{-ax} e^{-by} \mathbb{I}_{(0,\infty)}(x) \mathbb{I}_{(0,\infty)}(y)$$

for $a, b \in \mathbb{R}_{>0}$, then

$$\begin{aligned} \mathbb{E}[XY] &= ab \int_0^\infty \int_0^\infty xye^{-ax} e^{-by} \, \mathrm{d}x \, \mathrm{d}y \\ &= ab \int_0^\infty xe^{-ax} \, \mathrm{d}x \int_0^\infty ye^{-by} \, \mathrm{d}y = \frac{1}{ab}, \end{aligned}$$

an easy calculation because X and Y are independent.

Moments (2)

- Often only a subset of the $\mathbf{X}$ are used in g so that the desired expected value can be expressed in terms of the marginal density of the subset. For example,

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \int_0^\infty \int_0^\infty x^2 f_{X,Y}(x,y)\, \mathrm{d}y\, \mathrm{d}x \\
&= \int_0^\infty x^2 \int_0^\infty f_{X,Y}(x,y)\, \mathrm{d}y\, \mathrm{d}x \\
&= \int_0^\infty x^2 f_X(x)\, \mathrm{d}x = \frac{2}{a^2}.
\end{aligned}$$

Covariance

- Interest often centers on the mean $\mu_i = \mathbb{E}[X_i]$ and variance $\sigma_i^2 = \mathbb{E}\left[(X_i - \mu_i)^2\right]$ of the individual components of **X**.
- A generalization of the variance is the **covariance**: For any two X_i, the covariance is given by

$$\sigma_{ij} := \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathbb{E}[X_i X_j] - \mu_i \mu_j,$$

where $\mu_i = \mathbb{E}[X_i]$, and is a measure of the **linear association** between the two variables.

Covariance (2)

- If σ_{ij} is positive, then, generally speaking, relatively large (small) values of X_1 tend to occur with relatively large (small) values of X_2, while, if $\sigma_{ij} < 0$, then relatively small (large) values of X_1 tend to occur with relatively large (small) values of X_2.
- From symmetry, $\mathrm{Cov}\,(X_i, X_j) = \mathrm{Cov}\,(X_j, X_i)$.
- Thus, for $i \neq j$, there are $\left(n^2 - n\right)/2$ unique covariance terms among n random variables.
- If X_i and X_j are independent, then, for $i \neq j$,

$$\mathrm{Cov}\,(X_i, X_j) = \mathbb{E}\,[X_i - \mu_i]\,\mathbb{E}\,[X_j - \mu_j] = 0.$$

- So, the occurrence of relatively large or small values of X_1 gives no indication as to what to expect from X_2.

Correlation

- If $i = j$, $\sigma_{ii} = \mathrm{Cov}\left(X_i, X_i\right) = \mathrm{Var}\left(X_i\right) = \sigma_i^2$. The **correlation** of two r.v.s is defined to be

$$\mathrm{Corr}\left(X_i, X_j\right) = \frac{\mathrm{Cov}\left(X_i, X_j\right)}{\sqrt{\mathrm{Var}\left(X_i\right)\mathrm{Var}\left(X_j\right)}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

- The covariance can, in general, be any value in $\mathbb{R}$, because it depends on the scaling and range of the r.v.s of interest.
- The correlation is bound between -1 and 1 , with high positive (negative) correlation associated with values near 1 (-1).
- Clearly, $\mathrm{Corr}\left(X_i, X_j\right) = 0$ if X_i and X_j are independent.
- The converse does not hold in general: A correlation (or covariance) of zero does not necessarily imply that the two r.v.s are independent.

Course Outline

Mean and Variance of Sums of r.v.s

- Let $Y = \sum_{i=1}^{n} X_i$, where the X_i are random variables. A very important result is that

$$\mathbb{E}[Y] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

- Similarly,

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j}\sum \mathrm{Cov}(X_i, X_j)$$

with important special case

$$\mathrm{Var}(X_i + X_j) = \mathrm{Var}(X_i) + \mathrm{Var}(X_j) + 2\,\mathrm{Cov}(X_i, X_j).$$

Mean and Variance of Sums of r.v.s (2)

To show the result for the expectation when $n = 2$, let $g(\mathbf{X}) = X_1 + X_2$. Then

$$
\begin{aligned}
&\mathbb{E}[X_1 + X_2] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 + x_2)\, f_{X_1,X_2}(x_1, x_2)\, \mathrm{d}x_2\, \mathrm{d}x_1 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f_{X_1,X_2}(x_1, x_2)\, \mathrm{d}x_2\, \mathrm{d}x_1 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2 f_{X_1,X_2}(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty} x_1 \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\, \mathrm{d}x_2\, \mathrm{d}x_1 + \int_{-\infty}^{\infty} x_2 \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1)\, \mathrm{d}x_1 + \int_{-\infty}^{\infty} x_2 f_{X_2}(x_2)\, \mathrm{d}x_2 \\
&= \mathbb{E}[X_1] + \mathbb{E}[X_2].
\end{aligned}
$$

The result for $n > 2$ can be similarly derived; it also follows directly from the $n = 2$ case by induction.

Mean and Variance of Sums of r.v.s (3)

- More generally,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} \mathbb{E}[a_i X_i] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i] = \sum_{i=1}^{n} a_i \mu_i$$

and

$$\mathrm{Var}(X) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(X_i) + \sum_{i \neq j}\sum a_i a_j \,\mathrm{Cov}(X_i, X_j).$$

- If X_1 and X_2 are uncorrelated and $a_1 = -a_2 = 1$,

$$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2).$$

- The covariance between two r.v.s $X = \sum_{i=1}^{n} a_i X_i$ and $Y = \sum_{i=1}^{m} b_i Y_i$ is

$$\mathrm{Cov}(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \,\mathrm{Cov}(X_i, Y_j).$$

Random Variable Decomposition

- A binomial random variable with parameters n and p can be represented as a sum of n independent, Bernoulli distributed random variables, each with parameter p.
- That is, if $X \sim \mathrm{Bin}\,(n, p)$, then $X = \sum_{i=1}^{n} X_i$, where $X_i \sim \mathrm{Ber}\,(p)$, $i = 1, \ldots, n$, and $X_i \overset{\text{iid}}{\sim} \mathrm{Ber}\,(p)$.
- As $\mathbb{E}\,[X_i] = \Pr\,(X_i = 1) = p$ for all i, it follows that $\mathbb{E}\,[X] = np$.
- Similarly, $\mathrm{Var}\,(X) = np\,(1 - p)$, because $\mathrm{Var}\,(X_i) = p\,(1 - p)$.

Random Variable Decomposition (2)

- Likewise, $X \sim \mathrm{NBin}(r, p)$ can be decomposed into r iid geometric random variables, $X_1, \ldots, X_r$, each with density

$$f_X(x) = f_{\mathrm{Geo}}(x; p) = p(1-p)^x \, \mathbb{I}_{\{0,1,\ldots\}}(x).$$

- We know that $\mathbb{E}[X_i] = (1-p)/p$ and $\mathrm{Var}(X_i) = (1-p)/p^2$, so that $\mathbb{E}[X] = r(1-p)/p$ and $\mathrm{Var}(X) = r(1-p)/p^2$.
- It follows that, if $X_i \overset{\mathrm{ind}}{\sim} \mathrm{Bin}(n_i, p)$, then $X = \sum_i X_i \sim \mathrm{Bin}(n, p)$, where $n = \sum_i n_i$.
- Also, if $X_i \overset{\mathrm{ind}}{\sim} \mathrm{NBin}(r_i, p)$, then $X = \sum_i X_i \sim \mathrm{NBin}(r, p)$, where $r = \sum_i r_i$.

Exercise

- Recall our village which gets hit by a plague. Now, consider the adults: Before the plague, there are n married couples. The plague randomly kills d of the $2n$ adults, $0 < d < 2n$.
  Let X be the number of married couples remaining.
  According to Johnson and Kotz (1977, p. 23), Daniel Bernoulli (1700–1782) proposed and answered this question around 1766.
  Calculate the support of X, its pmf, $\mathbb{E}[X]$ and $\mathbb{V}(X)$. Also make a program to simulate X, and thus compare the empirical and true pmf, expected value, and variance.
  Hint: To assist deriving the mean and variance, let

$$B_i = \mathbb{I}\left(i\text{th couple still intact}\right).$$

Solution

- First consider the support of X. If the d deaths are spread out such that the largest number of couples are affected, then $n - x$ (the number of couples affected) can be as large as d (but not larger than n), so that $n - x \leq \min(n, d)$, or $x \geq n - \min(n, d)$. For the upper bound on X, observe that, if d is even, then we want that $d/2$ couples are affected (and both the man and wife die), leaving $n - d/2$ couples intact. If d is odd and X is to be as large as possible, then there is one couple such that either the man or woman, but not both, is dead, so it suffices to imagine that d is $d+1$, i.e., if d is odd, then the upper bound is $n - (d+1)/2$. Thus,

$$\mathcal{S}_X = \left\{ x \in \mathbb{N} : n - \min(n, d) \leq x \leq n - \left\lceil \frac{d}{2} \right\rceil \right\},$$

where $\lceil a \rceil$ is the *ceiling* of a, meaning that a is rounded off towards positive infinity, e.g., $\lceil 3.2 \rceil = \lceil 4 \rceil = 4$.

Solution (2)

- In order for there to be x couples still intact, it must be the case that $n-x$ couples have lost either the man, or the woman, or both, such that the total number of deaths is d. There are $\binom{n}{n-x} = \binom{n}{x}$ ways of choosing which couples get affected (the "danger group"). Concentrating now on this group, $n-x$ of the d deaths must go to eliminating either the man or woman, so there are 2^{n-x} ways for this to occur. Of the $2(n-x)$ people in the danger group, $n-x$ have been eliminated, leaving $n-x$ people to accommodate $d-(n-x)$ remaining deaths. There are $\binom{n-x}{d-(n-x)}$ ways of doing this, but we need to divide this quantity by $2^{d-(n-x)}$ because they were implicitly "ordered" by having considered the 2^{n-x} factor.

Solution (3)

- Thus,

$$\Pr(X = x) = \frac{2^{n-x}}{2^{d-(n-x)}} \frac{\binom{n}{x}\binom{n-x}{d-(n-x)}}{\binom{2n}{d}} \mathbb{I}_{\mathcal{S}_X}(x).$$

- To compute the expected value from its definition appears difficult, but from the hint, observe that $X = \sum_{i=1}^{n} B_i$, so that

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[B_i] = n\mathbb{E}[B_1],$$

and

$$\mathbb{E}[B_1] = \Pr(B_1 = 1) = \frac{\binom{2n-2}{d}}{\binom{2n}{d}} = \frac{(2n-d)(2n-d-1)}{(2n)(2n-1)},$$

yielding

$$\mathbb{E}[X] = \frac{(2n-d)(2n-d-1)}{2(2n-1)} =: t_1.$$

Solution (4)

- For the variance, recall that $\mathbb{V}(X) = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2$, and

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} B_i\right)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[B_i^2\right] + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathbb{E}\left[B_i B_j\right] \\ &= n\mathbb{E}\left[B_1\right] + n(n-1)\,\mathbb{E}\left[B_1 B_2\right],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[B_1 B_2\right] &= \Pr\left(B_1 = 1 \text{ and } B_2 = 1\right) = \frac{\binom{2n-4}{d}}{\binom{2n}{d}} \\ &= \frac{(2n-d)(2n-d-1)(2n-d-2)(2n-d-3)}{(2n)(2n-1)(2n-2)(2n-3)} =: t_2.\end{aligned}$$

Putting all this together gives

$$\mathbb{V}(X) = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2 = t_1 + n(n-1)\,t_2 - t_1^2.$$

Matlab Solution (1)

```
function pairs=bernoullijar2(n,d)
%% n distinguishable black and n distinguishable white
%%    balls in an urn. Each set is labeled 1..n.
%% Randomly remove d of them.
%% This simulates how many black/white pairs are left.
%
urn=[1:n , 1:n]; rem=2*n;
for i=1:d
  w=unidrnd(rem,1,1); urn=[urn(1:w-1), urn(w+1:end)];
  rem=rem-1;
end
s=0;
for i=1:n, s=s+(sum(urn==i)==2); end, pairs=s;
```

- This can be used in conjunction with the following code:

Matlab Solution (2)

```
n=5; d=4; sim=1e5; pairs=zeros(sim,1);
for i=1:sim,   pairs(i)=bernoullijar2(n,d);  end
tabulate(pairs)
lo=n-min(n,d); hi=n-ceil(d/2); p=[];
for x=lo:hi
  p=[p 2^(n-x) / 2^(d-n+x) * C(n,x) * C(n-x,d-n+x)];
end
p=p/C(2*n,d);
true_prob_times_100 = [lo:hi ; 100*p]
empirical_mean=mean(pairs)
true_mean = (2*n-d)*(2*n-d-1)/2/(2*n-1)
t1=true_mean;  a=(2*n-d);  b=2*n;
t2=a*(a-1)*(a-2)*(a-3)/b/(b-1)/(b-2)/(b-3);
empirical_var=var(pairs)
true_var = t1+n*(n-1)*t2-t1^2
```

Sample Output

```
 Value     Count    Percent
      1     38297      38.30%
      2     56960      56.96%
      3      4743       4.74%

true_prob_times_100 =
    1.0000    2.0000    3.0000
   38.0952   57.1429    4.7619


empirical_mean = 1.6645
true_mean      =  1.6667

empirical_var = 0.3178
true_var      =  0.3175
```

Density and CDF of Sums of Random Variables

- For the previous binomial and negative binomial cases, if the X_i are still independent, but the p_i differ, the results no longer hold.
- In this case, the mass function of $\sum_i X_i$ is more complicated. For example, if $X_i \sim \mathrm{Bin}(n_i, p_i)$, $i = 1, 2$, then the mass function of $X = X_1 + X_2$, or the **convolution** of X_1 and X_2, can be written as

$$\begin{aligned} \Pr(X = x) &= \sum_{i=0}^{n} \Pr(X_1 = i) \Pr(X_2 = x - i) \\ &= \sum_{i=0}^{n} \Pr(X_1 = x - i) \Pr(X_2 = i), \end{aligned}$$

  where $n = n_1 + n_2$.
- This makes sense because X_1 and X_2 are independent and, in order for X_1 and X_2 to sum to x, it must be the case that one of the events $\{X_1 = 0,\, X_2 = x\}$, $\{X_1 = 1,\, X_2 = x - 1\}$, …, $\{X_1 = x,\, X_2 = 0\}$ must have occurred. These events partition the event $\{X = x\}$.

Density and CDF of Sums of Random Variables (2)

- Similarly, the cdf of $X = X_1 + X_2$ is given by

$$\begin{aligned} \Pr(X \leq x) &= \sum_{i=0}^{n} \Pr(X_1 = i) \Pr(X_2 \leq x - i) \\ &= \sum_{i=0}^{n} \Pr(X_1 \leq x - i) \Pr(X_2 = i) . \end{aligned}$$

- This result extends to any two discrete independent random variables, although if both do not have bounded support, the sums will be infinite.
- They can also be generalized to the sum of three, four, or more (discrete and independent) random variables, but will then involve double, triple, etc., sums and become computationally inefficient.

Density and CDF of Sums of Random Variables (3)

- Let $X_i \overset{\text{ind}}{\sim} \text{Poi}(\lambda_i)$ and let $Y = X_1 + X_2$.
- With $\lambda = \lambda_1 + \lambda_2$,

$$\begin{aligned}
\Pr(Y = y) &= \sum_{i=-\infty}^{\infty} \Pr(X_1 = i) \Pr(X_2 = y - i) \\
&= \sum_{i=0}^{y} \frac{e^{-\lambda_1}\lambda_1^i}{i!} \frac{e^{-\lambda_2}\lambda_2^{y-i}}{(y-i)!} = e^{-(\lambda_1+\lambda_2)} \sum_{i=0}^{y} \frac{\lambda_1^i}{i!} \frac{\lambda_2^{y-i}}{(y-i)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{y!} \sum_{i=0}^{y} \frac{y!}{i!\,(y-i)!} \lambda_1^i \lambda_2^{y-i} = \frac{e^{-(\lambda_1+\lambda_2)}}{y!} (\lambda_1 + \lambda_2)^y \\
&= \frac{e^{-\lambda}\lambda^y}{y!},
\end{aligned}$$

where the second to last equality follows from the binomial theorem.

- It follows that $\sum_{i=1}^{n} X_i \sim \text{Poi}(\lambda)$, where $\lambda = \sum_{i=1}^{n} \lambda_i$.

Course Outline

Introducing Continuous Univariate Random Variables

All phenomena are ultimately discrete; continuous distributions are "natural approximations" which are used extensively.

When "meeting" a cont. dist. for the first time, consider:

1. Its theoretical importance.
2. Its use in applications.
3. How its functional form came about. For example, it might
   - be a "base" distribution arising from mathematical simplicity, e.g., the uniform or exponential;
   - arise or be strongly associated with a particular application, e.g., the Cauchy in a geometrical context, the Pareto for income distribution, the F in the analysis of variance;
   - be a "natural" or obvious generalization of a base distribution, such as the beta or Weibull;
   - be a function of other, simpler r.v.s, such as gamma as a sum of exponentials, or Student's t as a ratio of normal and weighted χ^2;
   - be a **limiting** distribution, such as the Poisson, normal and Gumbel.

Introducing Continuous Univariate Random Variables

4. The extent to which certain characteristics can be easily ascertained and/or computed, such as: the expected value and higher moments, quantiles and the c.d.f..
5. Other properties of interest: unimodal? closed under addition? member of the exponential family?
6. Recognizing what role the associated parameters play. Continuous distributions often have a **location** parameter which shifts the density and a **scale** parameter which stretches or shrinks the density.
7. Further parameters are referred to generically as **shape** parameters but, for any particular distribution, often have standard names, e.g., the degrees of freedom for Student's t. In some densities, there is a parameter which is responsible for the skewness.
8. The behavior of the c.d.f. far into the **tails** of the distribution.

Introducing Continuous Univariate Random Variables

- Regarding the location and scale parameters just mentioned, if X is a continuous random variable with p.d.f. $f_X(x)$, then the **linearly transformed** random variable $Y = \sigma X + \mu$, $\sigma > 0$, has density

$$f_Y(y) = \frac{1}{\sigma} f_X\left(\frac{y-\mu}{\sigma}\right),$$

  to be derived later.
  The distributions of X and Y are said to be members of the same **location-scale family**, with **location parameter** μ and **scale parameter** σ.
- The **kernel** of a p.d.f. is that part of it which involves only the variables associated with the r.v.s of interest, e.g., x.

The Uniform and the Exponential Distribution

Some common univariate continuous distributions are:

1. uniform, $\text{Unif}\,(a, b)$;

$$f_{\text{Unif}}\,(x; a, b) = \frac{1}{b-a}\mathbb{I}_{(a,b)}\,(x)$$

and

$$F_{\text{Unif}}\,(x; a, b) = \frac{x-a}{b-a}\mathbb{I}_{[a,b]}\,(x) + \mathbb{I}_{[b,\infty]}\,(x)\,.$$

2. exponential, $\text{Exp}(\lambda)$, $\lambda \in \mathbb{R}_{>0}$ with density and distribution function

$$f_{\text{Exp}}\,(x; \lambda) = \lambda \exp\left\{-\lambda x\right\} \mathbb{I}_{[0,\infty)}\,(x)$$

and

$$F_{\text{Exp}}\,(x; \lambda) = 1 - \exp\left\{-\lambda x\right\},$$

where λ is a **scale parameter**.

The Gamma Distribution

3. gamma, $\text{Gam}(\alpha, \beta)$, $\alpha, \beta \in \mathbb{R}_{>0}$; the density is given by

$$f_{\text{Gam}}(x; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} \mathbb{I}_{[0,\infty)}(x),$$

β is a **scale parameter** and, with $\alpha = 1$, reduces to the exponential distribution. The gamma c.d.f. is given by

$$F_{\text{Gam}}(x; \alpha, 1) = \bar{\Gamma}_x(\alpha) := \frac{\Gamma_x(\alpha)}{\Gamma(\alpha)},$$

where $\Gamma_x(a) = \int_0^x t^{a-1} e^{-t} \, dt$ is the incomplete gamma function. The moments are straightforwardly shown to be

$$\mathbb{E}\left[X^k\right] = \frac{\Gamma(k+\alpha)}{\beta^k \Gamma(\alpha)}, \quad k > -\alpha.$$

In particular, for $k = 0$, we see that the density integrates to one; for $k = 1$, it follows that $\mathbb{E}[X] = \alpha/\beta$; with $k = 2$,
$\text{Var}(X) = \alpha(1+\alpha)/\beta^2 - (\alpha/\beta)^2 = \alpha/\beta^2$.

The Beta Distribution

4. beta, $\text{Beta}(p, q)$, $p, q \in \mathbb{R}_{>0}$;

$$f_{\text{Beta}}(x; p, q) = \frac{1}{B(p, q)} x^{p-1} (1-x)^{q-1} \mathbb{I}_{[0,1]}(x)$$

and

$$F_{\text{Beta}}(x; p, q) = \frac{B_x(p, q)}{B(p, q)} \mathbb{I}_{[0,1]}(x) + \mathbb{I}_{(1,\infty)}(x),$$

where

$$B_x(p, q) = \mathbb{I}_{[0,1]}(x) \int_0^x t^{p-1} (1-t)^{q-1} \, \mathrm{d}t$$

is the incomplete beta function.
If $p = q = 1$, the beta distribution reduces to $\text{Unif}(0, 1)$.
The k^{th} moment of $X \sim \mathrm{Beta}(p, q)$ is easily derived:

$$\mathbb{E}\left[X^k\right] = \frac{\Gamma(p+q)}{\Gamma(p)} \frac{\Gamma(p+k)}{\Gamma(p+k+q)},$$

for $k > -p$.

The Weibull Distribution

5. Weibull, $\mathrm{W}(\beta, x_0, \sigma)$, $\beta, \sigma \in \mathbb{R}_{>0}$ and $x_0 \in \mathbb{R}$; the density of $\mathrm{W}(\beta, x_0, \sigma)$ is given by

$$f_{\mathrm{W}}(x; \beta, x_0, \sigma) = \frac{\beta}{\sigma}\left(\frac{x - x_0}{\sigma}\right)^{\beta - 1} \exp\left\{-\left(\frac{x - x_0}{\sigma}\right)^{\beta}\right\} \mathbb{I}_{(x_0, \infty)}(x),$$

where x_0 and σ are location and scale parameters.

- With $\beta = 1$ and $x_0 = 0$, $\mathrm{W}(\beta, x_0, \sigma)$ reduces to the exponential distribution.
- The c.d.f. is closed form, and is given by

$$F_{\mathrm{W}}(x; \beta, x_0, \sigma) = 1 - \exp\left\{-\left(\frac{x - x_0}{\sigma}\right)^{\beta}\right\} \mathbb{I}_{(x_0, \infty)}(x).$$

- For $W(b, 0, s)$, substituting $u = (x/s)^b$ and simplifying gives

$$\mathbb{E}[X^p] = s^p \int_0^{\infty} u^{p/b} \exp\{-u\}\, \mathrm{d}u = s^p \Gamma\left(1 + \frac{p}{b}\right)$$

which exists for $p > -b$.

The Laplace and the Cauchy Distribution

6. Laplace or double exponential, $\mathrm{Lap}(\mu, \sigma)$ or $\mathrm{DExp}(\mu, \sigma)$, $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{>0}$; $f_{\mathrm{Lap}}(x; 0, 1) = \exp\{-|x|\}/2$ and

$$F_{\mathrm{Lap}}(x; 0, 1) = \frac{1}{2}\begin{cases} e^{x}, & \text{if } x \leq 0, \\ 2 - e^{-x}, & \text{if } x > 0. \end{cases}$$

7. Cauchy, $\mathrm{Cau}(\mu, \sigma)$, $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{>0}$; from

$$f_{\mathrm{Cau}}(x; 0, 1) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2}$$

and $F_{\mathrm{Cau}}(x; 0, 1) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$. A simple calculation (see example A.27) shows that the mean of a Cauchy r.v. does not exist.

Pareto Distributions

8. Type I Pareto (or Pareto distribution of the first kind) $\mathrm{Par\,I}\,(\alpha, x_0)$, $\alpha, x_0 \in \mathbb{R}_{>0}$;

$$f_{\mathrm{Par\,I}}\,(x;\alpha,x_0) = \alpha x_0^{\alpha} x^{-(\alpha+1)} \mathbb{I}_{[x_0,\infty)}\,(x)$$

and

$$F_{\mathrm{Par\,I}}\,(x;\alpha,x_0) = 1 - \left(\frac{x_0}{x}\right)^{\alpha} \mathbb{I}_{[x_0,\infty)}\,(x)\,.$$

The moments of X are given by

$$\mathbb{E}\,[X^m] = \int_{-\infty}^{\infty} x^m f_X(x)\,\mathrm{d}x = \alpha x_0^{\alpha} \int_{x_0}^{\infty} x^{-\alpha-1+m}\,\mathrm{d}x$$
$$= \frac{\alpha x_0^{\alpha}}{m-\alpha}\, x^{m-\alpha}\Big|_{x_0}^{\infty} = \frac{\alpha}{\alpha-m} x_0^m,$$

for $m < \alpha$, and do not exist for $m \geq \alpha$.

9. Type II Pareto, $\mathrm{Par\,II}\,(b)$, $b \in \mathbb{R}_{>0}$; the c.d.f. is

$$F_{\mathrm{Par\,II}}(x;b) = \left[1 - \left(\frac{1}{1+x}\right)^{b}\right] \mathbb{I}_{(0,\infty)}(x).$$

Pareto and Power Tails

- In various applications, the **survivor function** of r.v. X, defined by

$$\bar{F}_X(x) = 1 - F_X(x) = \Pr(X > x),$$

  is of particular interest.
- For $X \sim \mathrm{Par\,I}(\alpha, x_0)$, $\bar{F}(x) = Cx^{-\alpha}$, where $C = x_0^{\alpha}$.
- Can be shown: The survivor function for a number of important distributions is asymptotically of the form $Cx^{-\alpha}$ as x increases, where C denotes some constant.
- If this is the case, we say that the right **tail** of the density is **Pareto–like** or that the distribution has **power tails**.
- Somewhat informally, if $\bar{F}_X(x) \approx Cx^{-\alpha}$, then, as $\mathrm{d}\left(1 - Cx^{-\alpha}\right) / \mathrm{d}x \propto x^{-(\alpha+1)}$, the maximally existing moment of X is bounded above by α.

Exponential Tails (2)

- For many distributions, all positive moments exist, such as for the gamma and normal. These are said to have **exponential tails**.
- For instance, if $X \sim \mathrm{Exp}(\lambda)$, then $\bar{F}_X(x) = e^{-\lambda x}$.
- While the normal c.d.f. (or that of any distribution with exponential right tail) dies off rapidly, the c.d.f. with power tails tapers off slowly.
- For power tails, they are so "thick" that the probability of extreme events **never** becomes negligible.
- This is why the expected value of X raised to a sufficiently large power will fail to exist.

The Normal Distribution

10. normal or Gaussian, $\mathrm{N}\left(\mu, \sigma^2\right)$, $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{>0}$ with density

$$f_{\mathrm{N}}\left(x; \mu, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}.$$

- While μ and σ are respectively the location and scale parameters, it is standard convention to write $\mathrm{N}\left(\mu, \sigma^2\right)$ instead of $\mathrm{N}\left(\mu, \sigma\right)$.
- The normal distribution, sometimes referred to as the **"bell curve"**, enjoys certain properties which render it the most reasonable description for modeling a large variety of stochastic phenomena. As such, it plays a central role in much of statistical analysis.
- The first four moments are $\mu_1 = \mathbb{E}\left[X\right] = \mu$, $\mu_2 = \mathrm{Var}\left(X\right) = \sigma^2$, $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$. An easy way of calculating these is given later.

The Normal Distribution (2)

- While the normal c.d.f. is not expressible in closed form, the survivor function for $Z \sim \mathrm{N}(0,1)$ does have an upper bound in the right tail. For $t > 0$ and with $u = z^2/2$,

$$\begin{aligned}
\Pr(Z \geq t) &= \frac{1}{\sqrt{2\pi}} \int_t^\infty \exp\left\{-\frac{1}{2}z^2\right\} \mathrm{d}z \\
&\leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} \int_t^\infty z \exp\left\{-\frac{1}{2}z^2\right\} \mathrm{d}z \qquad \left(\text{because } \frac{z}{t} > 1\right) \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{t} \int_{t^2/2}^\infty \exp\{-u\} \, \mathrm{d}u = \frac{1}{\sqrt{2\pi}} \frac{1}{t} \exp\left\{-t^2/2\right\}.
\end{aligned}$$

The term $\exp\{-t^2/2\}$ clearly goes to zero far faster than t^{-1}, so that Z has exponential tails.

Normal–Based Distributions

The distributions χ^2, t and F are of utmost importance for statistical inference involving the wide class of **normal linear models** (which includes the familiar **two–sample t test**, **regression analysis**, **ANOVA**, **random–effects models**, as well as many **econometrics models** such as for modeling **panel** and **time–series data**.

11. chi-square with ν **degrees of freedom**, $\chi^2(\nu)$ or χ^2_ν, $\nu \in \mathbb{R}_{>0}$; the density is given by

$$f_{\chi^2_\nu}(x;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \mathbb{I}_{(0,\infty)}(x)$$

and is a special case of the gamma distribution with $\alpha = \nu/2$ and $\beta = 1/2$. In most statistical applications, $\nu \in \mathbb{N}$ and often the notation χ^2_n is used.

Normal–Based Distributions: Student's t

12. Student's t with ν **degrees of freedom**, abbreviated $t(\nu)$ or t_ν, $\nu \in \mathbb{R}_{>0}$;

$$f_t(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)\nu^{\frac{\nu}{2}}}{\sqrt{\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(\nu + x^2\right)^{-\frac{\nu+1}{2}} = \frac{v^{-\frac{1}{2}}}{B\left(\frac{\nu}{2},\frac{1}{2}\right)}\left(1 + x^2/\nu\right)^{-\frac{\nu+1}{2}}.$$

If $\nu = 1$, then the Student's t distribution reduces to the Cauchy distribution while, as $\nu \to \infty$, it **converges in distribution** to the normal. In most statistical applications $\nu \in \mathbb{N}$.

Let $T \sim t_n$. The mean of T is zero for $n > 1$, but does not otherwise exist. For the variance,

$$\mathbb{V}(T) = \frac{n}{n-2}, \qquad \text{for } n > 2.$$

It is easy to see from the p.d.f. that $f_t(x;n) \propto |x|^{-(n+1)}$, which is similar to the type I Pareto, showing that the maximally existing moment of the Student's t is bounded above by n.

Normal–Based Distributions: Student's t

Recall that the *incomplete beta function* is

$$B_x(p,q) = \mathbb{I}_{[0,1]}(x) \int_0^x t^{p-1} (1-t)^{q-1} \, \mathrm{d}t$$

and the normalized function $B_x(p,q) / B(p,q)$ is the *incomplete beta ratio*, denoted by $\bar{B}_x(p,q)$.

For $t < 0$, the c.d.f. of the Student's t is given by

$$F_T(t) = \frac{1}{2} \bar{B}_L \left(\frac{n}{2}, \frac{1}{2} \right), \quad L = \frac{n}{n+t^2}, \quad t < 0.$$

For $t > 0$, the symmetry of the t density about zero implies that $F_T(t) = 1 - F_T(-t)$.

Normal–Based Distributions: F

13. F (or variance ratio or Fisher's F) with n_1 numerator and n_2 denominator **degrees of freedom**, $\mathrm{F}(n_1, n_2)$, $n_1, n_2 \in \mathbb{R} > 0$;

$$f_{\mathrm{F}}(x; n_1, n_2) = \frac{n_1/n_2}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \frac{\left(\frac{n_1}{n_2}x\right)^{n_1/2-1}}{\left(1+\frac{n_1}{n_2}x\right)^{(n_1+n_2)/2}} \quad \text{and}$$

$$F_{\mathrm{F}}(x; n_1, n_2) = \bar{B}_y\left(\frac{n_1}{2}, \frac{n_2}{2}\right), \quad y = n_1 x / (n_2 + n_1 x).$$

In most statistical applications, $n_1, n_2 \in \mathbb{N}$.

Univariate Transformations

- If X is a continuous r.v. with p.d.f. f_X and g is a continuous differentiable function with domain contained in the range of X and $\mathrm{d}g/\,\mathrm{d}x \neq 0 \ \forall x \in \mathcal{S}_X$, then f_Y, the p.d.f. of $Y = g(X)$, can be calculated by

$$f_Y\left(y\right) = f_X\left(x\right)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|,$$

  where $x = g^{-1}\left(y\right)$ is the inverse function of Y.
- This can be intuitively understood by observing that

$$f_X\left(x\right)\triangle x \approx \Pr\left(X \in \left(x, x+\triangle x\right)\right) \approx \Pr\left(Y \in \left(y, y+\triangle y\right)\right) \approx f_Y\left(y\right)\triangle y$$

  for small $\triangle x$ and $\triangle y$, where $\triangle y = g\left(x+\triangle x\right) - g\left(x\right)$ depends on g, x and $\triangle x$.

Examples of Univariate Transformations

- The location-scale family discussed earlier is derived as follows. Let $Y = \sigma X + \mu$ and $\sigma > 0$. Then, with $x = (y - \mu)/\sigma$,

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X\left(\frac{y-\mu}{\sigma}\right)\sigma^{-1}.$$

- Let $U \sim \mathrm{Unif}(0,1)$ and define $Y = -\ln U$. Then, with $u = \exp(-y)$,

$$f_Y(y) = f_U(u)\left|\frac{\mathrm{d}u}{\mathrm{d}y}\right| = \mathbb{I}_{(0,1)}(\exp(-y))\left(e^{-y}\right) = e^{-y}\mathbb{I}_{(0,\infty)}(y),$$

so that $Y \sim \mathrm{Exp}(1)$.

Example: The lognormal Distribution

This distribution has numerous applications in statistics, but is ubiquitous in finance because a standard assumption is that the log of the (financial asset) returns are normally distributed, so that the corresponding prices are lognormal.

If Z is standard normal, then the density of $X = \exp(Z\sigma + \zeta) + \theta$ is

$$f_Z(z)\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{1}{\sqrt{2\pi}}\frac{1}{(x-\theta)\,\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(\ln(x-\theta)-\zeta\right)^2\right)\mathbb{I}_{(\theta,\infty)}(x).$$

The r^{th} moment of $X-\theta$ is given by

$$\begin{aligned}\mathbb{E}\left[(X-\theta)^r\right] &= \mathbb{E}\left[(\exp(Z\sigma+\zeta))^r\right] = \mathbb{E}\left[\exp(rZ\sigma + r\zeta)\right] \\ &= \exp(r\zeta)\int_{-\infty}^{\infty}\exp(r\sigma Z)\,f_Z(z)\,dz \\ &= \exp(r\zeta)\exp\left(\frac{1}{2}r^2\sigma^2\right) = \exp\left(r\zeta + \frac{1}{2}r^2\sigma^2\right),\end{aligned}$$

where $\mathbb{E}\left[e^{tZ}\right]$ is shown below.

Example: The lognormal Distribution

Repeating from the previous slide, we have

$$\mathbb{E}\left[(X-\theta)^r\right] = \exp\left(r\zeta + \frac{1}{2}r^2\sigma^2\right).$$

The mean and variance can be more easily expressed in terms of $w = \exp\left(\sigma^2\right)$ as follows:

$$\mathbb{E}[X] - \theta = \mathbb{E}[X-\theta] = \exp\left(\zeta + \frac{1}{2}\sigma^2\right) = e^{\zeta} w^{1/2}$$

and

$$\mathbb{V}(X) = \mathbb{V}(X-\theta) = e^{2\zeta} w^2 - e^{2\zeta} w = e^{2\zeta} w (w-1).$$

The m.g.f. of the Normal

Let $Z \sim \mathrm{N}(0,1)$ and $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$.
To compute $\mathbb{E}\left[\mathrm{e}^{tZ}\right]$, which is the so-called *moment generating function* of Z at t, denoted $\mathbb{M}_Z(t)$,

$$\mathbb{E}\left[\mathrm{e}^{tZ}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}z^2 + tz\right\} \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(z^2 - 2tz\right)\right\} \mathrm{d}z$$

and, by completing the square as $z^2 - 2tz + t^2 - t^2 = (t-z)^2 - t^2$,

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{tZ}\right] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left((t-z)^2 - t^2\right)\right\} \mathrm{d}z \\ &= \exp\left\{\frac{t^2}{2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(z-t)^2\right\} \mathrm{d}z = \exp\left\{\frac{t^2}{2}\right\}.\end{aligned}$$

As $X = \mu + \sigma Z$ is a (location-scale) transformation of X,

$$\mathbb{E}\left[\mathrm{e}^{tX}\right] = \mathbb{E}\left[\mathrm{e}^{t(\mu+\sigma Z)}\right] = \exp\{t\mu\}\,\mathbb{M}_Z(t\sigma) = \exp\left\{t\mu + \frac{t^2\sigma^2}{2}\right\}.$$

Partial Expectation of Log Normal

Example: The χ^2 Distribution

Let $Y = X^2$ and $X \sim \mathrm{N}(0,1)$.

With $y = x^2$, split up the x region as $x = \pm\sqrt{y}$ so that

$$
\begin{aligned}
f_Y(y) &= f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\
&= \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\sqrt{y}\right)^2\right\}\frac{y^{-1/2}}{2}\mathbb{I}_{(0,\infty)}\left(\sqrt{y}\right) \\
&\quad +\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(-\sqrt{y}\right)^2\right\}\frac{y^{-1/2}}{2}\mathbb{I}_{(-\infty,0)}\left(-\sqrt{y}\right) \\
&= \frac{1}{2\sqrt{2\pi}}y^{-1/2}e^{-y/2}\mathbb{I}_{(0,\infty)}(y) + \frac{1}{2\sqrt{2\pi}}y^{-1/2}e^{-y/2}\mathbb{I}_{(0,\infty)}(y) \\
&= \frac{1}{\sqrt{2\pi}}y^{-1/2}e^{-y/2}\mathbb{I}_{(0,\infty)}(y).
\end{aligned}
$$

This shows that $Y \sim \chi^2_1$.

Exercise: Folded Cauchy

Let $X \sim \mathrm{Cau}\,(0,1)$. Recall that

$$f_X\,(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \qquad F_X\,(x) = \frac{1}{\pi}\int_{-\infty}^{x}\frac{1}{1+t^2}\mathrm{d}t = \frac{1}{2}+\frac{1}{\pi}\arctan\,(x)\,.$$

From the symmetry of f_X about 0, for $x>0$, $F_X\,(x) = 1 - F_X\,(-x)$.

Let $Z = |X|$ (a folded Cauchy).
From the symmetry, we would expect $f_Z\,(z) = 2f_X\,(z)\,\mathbb{I}_{(0,\infty)}\,(z)$.
For the c.d.f., for $z>0$, $F_Z\,(z)$ is $\Pr\left(|X| \le z\right)$, which is $\int_{-z}^{z} f_X\,(x)\mathrm{d}x$ or one minus twice the tail area $\int_{-\infty}^{-z} f_X\,(x)\mathrm{d}x$ or $F_Z\,(z) = 1 - 2F_X\,(-z)$.

Verify these formally, in two ways:

1. Express F_Z directly in terms of F_X and differentiate to get f_Z.
2. Use the relation $Z = \sqrt{X^2}$; compute $f_Y\,(y)$ where $Y = X^2$.

Solution, FIRST

First way, for $z \geq 0$,

$$\begin{aligned} F_Z(z) &= \Pr(Z \leq z) = \Pr(|X| \leq z) = \Pr(-z \leq X \leq z) \\ &= F_X(z) - F_X(-z) = [1 - F_X(-z)] - F_X(-z) \\ &= 1 - 2F_X(-z) = -\frac{2}{\pi}\arctan(-z) = \frac{2}{\pi}\arctan(z). \end{aligned}$$

We need to differentiate this to get $f_Z(z)$. Note that the derivation of $F_Z(z) = 1 - 2F_X(-z)$ is valid for *any* r.v. with continuous p.d.f. symmetric about zero, i.e., $f_X(-x) = f_X(x)$, so that

$$f_Z(z) = \frac{\mathrm{d}F_Z(z)}{\mathrm{d}z} = -(-1)2f_X(-z) = 2f_X(z).$$

To confirm this in the Cauchy case, recalling that $\mathrm{d}\arctan(x)/\mathrm{d}x = 1/\left(1+x^2\right)$, which is obvious from differentiating the top integral expression for $F_X(x)$ and using the FTC, we get

$$f_Z(z) = \frac{\mathrm{d}F_Z(z)}{\mathrm{d}z} = \frac{2}{\pi}\frac{\mathrm{d}\arctan(z)}{\mathrm{d}z} = \frac{2}{\pi}\left(\frac{1}{1+z^2}\right) = 2f_X(z)\,\mathbb{I}_{(0,\infty)}(z).$$

Solution, SECOND

For the second way, as in Example 7.14, the c.d.f. of $Y = X^2$ can be computed as

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr\left(X^2 \leq y\right) \\ &= \Pr(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) \end{aligned}$$

while the density of Y is given by $\partial F_Y(y)/\partial y$, or

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right] = \pi^{-1} y^{-\frac{1}{2}} (1+y)^{-1}\, \mathbb{I}_{(0,\infty)}(y).$$

Solution, SECOND, cont.

Now let $Z = \sqrt{Y}$ so that $y = z^2$, and

$$f_Z(z) = f_Y(y)\frac{\mathrm{d}y}{\mathrm{d}z} = \pi^{-1}\left(z^2\right)^{-\frac{1}{2}}\left(1+z^2\right)^{-1}\ 2z\mathbb{I}_{(0,\infty)}(z) = \frac{2}{\pi}\frac{1}{1+z^2}\mathbb{I}_{(0,\infty)}(z),$$

which is of course the same as $2f_X(z)\,\mathbb{I}_{(0,\infty)}(z)$. Integrating (and using the integral expression for $F_X(x)$) immediately gives the c.d.f., or note that, from the above expression for $F_Y(y)$, for $z \geq 0$,

$$\begin{aligned}
F_Z(z) &= \Pr\left(\sqrt{Y} \leq z\right) \\
&= \Pr\left(Y \leq z^2\right) = F_Y\left(z^2\right) = F_X\left(\sqrt{z^2}\right) - F_X\left(-\sqrt{z^2}\right) \\
&= \frac{1}{2} + \frac{1}{\pi}\arctan(z) - \left(\frac{1}{2} + \frac{1}{\pi}\arctan(-z)\right) \\
&= \frac{2}{\pi}\arctan(z).
\end{aligned}$$

The Probability Integral Transform (PIT)

- The probability integral transform of r.v. X is defined by $Y = F_X(X)$, where $F_X(t) = \Pr(X \leq t)$.
- Here, $F_X(X)$ is not to be interpreted as $\Pr(X \leq X) = 1$, but rather as a random variable Y defined as the transformation of X, the transformation being the function F_X.
- Assume that F_X is strictly increasing. Then $F_X(x)$ is a one to one function for $x \in (0, 1)$ so that

$$\begin{aligned} F_Y(y) = \Pr(Y \leq y) = \Pr(F_X(X) \leq y) &= \Pr\left(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)\right) \\ &= \Pr\left(X \leq F_X^{-1}(y)\right) = F_X\left(F_X^{-1}(y)\right) = y, \end{aligned}$$

showing that $Y \sim \mathrm{Unif}(0, 1)$.

The PIT and Simulation

- The probability integral transform is of particular value for simulating random variables.
- By applying F^{-1} to both sides of $Y = F_X(X) \sim \mathrm{Unif}(0,1)$, we see that, if $Y \sim \mathrm{Unif}(0,1)$, then $F^{-1}(Y)$ is a realization of a random variable with c.d.f. F.
- For example, from the exponential c.d.f. $y = F(x) = F_{\mathrm{Exp}}(x;1) = 1 - e^{-x}$, we have $F^{-1}(x) = -\ln(1-y)$. Thus, taking $Y \sim \mathrm{Unif}(0,1)$, the PIT implies that $-\ln(1-Y) \sim \mathrm{Exp}(1)$.
- As $1 - Y$ is also uniformly distributed, we can use $-\ln(Y)$ instead.

Exercise: Problems 7.12 and 7.13(a)

Let X be a positive continuous r.v. with p.d.f. f_X and c.d.f. F_X.

1. Show that a necessary condition for $\mathbb{E}[X]$ to exist is $\lim_{x\to\infty} x(1 - F_X(x)) = 0$. Use this to show that the expected value of a Cauchy random variable does not exist.
2. Prove via integration by parts that $\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))\, \mathrm{d}x$ if $\mathbb{E}[X] < \infty$.
3. Prove that

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))\, \mathrm{d}x \tag{8}$$

by expressing $1 - F_X$ as an integral, and reversing the integrals.

Solution to 1, first part

Necessity means that, if $\mathbb{E}[X]$ exists, then

$$\lim_{x\to\infty} x\left(1 - F_X(x)\right) = 0.$$

If $\mathbb{E}[X] = \int_0^\infty t\ f_X(t)\ \mathrm{d}t$ exists, then

$$\lim_{x\to\infty} \int_x^\infty t\ f_X(t)\ \mathrm{d}t = 0.$$

So, as $\lim_{x\to\infty} x\left(1 - F_X(x)\right) \geq 0$ and

$$\lim_{x\to\infty} x\left(1 - F_X(x)\right) = \lim_{x\to\infty} x \int_x^\infty f_X(t)\ \mathrm{d}t \leq \lim_{x\to\infty} \int_x^\infty t\ f_X(t)\ \mathrm{d}t = 0,$$

it follows that $\lim_{x\to\infty} x\left(1 - F_X(x)\right) = 0$.

Solution to 1, for the Cauchy

The theorem is valid for nonnegative r.v.s, so: can we apply it to $\mathbb{E}\big[|X|\big]$? Observe that, if X is a random variable with density f_X symmetric about zero, then $f_X(-x) = f_X(x)$ and, substituting $u = -x$,

$$\mathbb{E}[X] = \int_{-\infty}^{0} x f_X(x)\,\mathrm{d}x + \int_{0}^{\infty} x f_X(x)\,\mathrm{d}x = -\int_{0}^{\infty} u f_X(u)\,\mathrm{d}u + \int_{0}^{\infty} x f_X(x)\,\mathrm{d}x = 0,$$

if $\int_0^\infty x f_X(x)\mathrm{d}x$ exists. If it does, then, as $\mathbb{E}[g(X)] = \int g(x) f(x)\mathrm{d}x$,

$$\begin{aligned}
\mathbb{E}[|X|] &= \int_{-\infty}^{0} |x|\, f_X(x)\,\mathrm{d}x + \int_{0}^{\infty} |x|\, f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{0} (-x)\, f_X(x)\,\mathrm{d}x + \int_{0}^{\infty} x f_X(x)\,\mathrm{d}x \\
&= -\int_{-\infty}^{0} x f_X(x)\,\mathrm{d}x + \int_{0}^{\infty} x f_X(x)\,\mathrm{d}x = 2\int_{0}^{\infty} x f_X(x)\,\mathrm{d}x,
\end{aligned}$$

where $\int_{-\infty}^{0} x f_X(x)\mathrm{d}x = -\int_0^\infty u f_X(u)\mathrm{d}u$ from above.

Solution to 1, for the Cauchy

That $\mathbb{E}[|X|] = 2\int_0^\infty x f_X(x)\,\mathrm{d}x$ also follows because, as we saw above in the context of the Cauchy, for $Z = |X|$ and f_X symmetric about zero, $f_Z(z) = 2f_X(z)$ for $z \geq 0$.

Thus, if $f_X(-x) = f_X(x)$ and $\mathbb{E}[X]$ exists, then $\mathbb{E}[|X|]$ exists.

The contrapositive then implies that, if $\mathbb{E}[|X|]$ does not exist, then $\mathbb{E}[X]$ does not exist.

Solution to 1, for the Cauchy

Above we saw that the p.d.f. and c.d.f. of a folded standard Cauchy random variable, call it X, are

$$f_X(x) = \frac{2}{\pi}\frac{1}{1+x^2} \qquad \text{and} \qquad F_X(x) = \frac{2}{\pi}\arctan(x).$$

As $\lim_{x\to\infty}(1 - F_X(x)) = 0$, l'Hôpital's rule gives

$$\lim_{x\to\infty} x(1 - F_X(x)) = \lim_{x\to\infty} \frac{\left(1 - \frac{2}{\pi}\arctan(x)\right)}{1/x} = \lim_{x\to\infty} \frac{-\frac{2}{\pi}\frac{1}{1+x^2}}{-x^{-2}} = \frac{2}{\pi},$$

which is nonzero.
The necessity of the condition implies that $\mathbb{E}[|X|]$ and, hence, $\mathbb{E}[X]$ does not exist.

Solution to 2 (first way)

Let $u = 1 - F_X(x)$ and $\mathrm{d}v = \mathrm{d}x$, so that

$$\begin{aligned} I &= \int_0^\infty (1 - F_X(x)) \cdot 1 \, \mathrm{d}x \\ &= uv|_0^\infty - \int_0^\infty v \, \mathrm{d}u = x(1 - F_X(x))|_0^\infty - \int_0^\infty (-1)\, x F_X'(x) \, \mathrm{d}x \\ &= \lim_{x\to\infty} x(1 - F_X(x)) + \int_0^\infty x f_X(x) \, \mathrm{d}x \\ &= \mathbb{E}[X], \end{aligned}$$

provided $\lim_{x\to\infty} x(1 - F_X(x)) = 0$, which is a necessary condition for the existence of the first moment, as shown in the previous question.

Solution to 2 (second way)

What if we started with the definition of $\mathbb{E}[X]$ and didn't know the other expression?

From

$$\mathbb{E}[X] = \int_0^\infty x f_X(x)\, \mathrm{d}x,$$

let $u = x$ and $\mathrm{d}v = f_X(x)\mathrm{d}x$, so that $\mathrm{d}u = \mathrm{d}x$ and $v = F_X(x)$, and

$$\mathbb{E}[X] = \int_0^\infty x f_X(x)\, \mathrm{d}x = x F_X(x)\big|_0^\infty - \int_0^\infty F_X(x)\, \mathrm{d}x.$$

Now observe that

$$x F_X(x)\big|_0^\infty = \lim_{x\to\infty} x F_X(x) - \lim_{x\to 0} x F_X(x) = \lim_{x\to\infty} x \lim_{x\to\infty} F_X(x) = \lim_{x\to\infty} x.$$

Solution to 2 (second way)

Finally, assuming $\mathbb{E}[X] < \infty$,

$$\begin{aligned}
\mathbb{E}[X] &= \lim_{x\to\infty} x + \left[-\int_0^\infty F_X(x)\,dx + \lim_{x\to\infty} x - \lim_{x\to\infty} x \right] \\
&= \lim_{x\to\infty} x + \left[-\int_0^\infty F_X(x)\,dx + \int_0^\infty dx - \lim_{x\to\infty} x \right] \\
&= \int_0^\infty dx - \int_0^\infty F_X(x)\,dx = \int_0^\infty [1 - F_X(x)]\,dx,
\end{aligned}$$

which is the result.

This is the "informal way" of writing things, which makes them clear. Somewhat more correctly...

Solution to 2 (second way)

...assuming $\mathbb{E}[X] < \infty$,

$$\begin{aligned}
\mathbb{E}[X] &= \lim_{x\to\infty} x + \left[-\int_0^\infty F_X(x)\,\mathrm{d}x + \lim_{x\to\infty} x - \lim_{x\to\infty} x\right] \\
&= \lim_{x\to\infty} x + \left[-\lim_{b\to\infty}\int_0^b F_X(x)\,\mathrm{d}x + \lim_{b\to\infty}\int_0^b \mathrm{d}x - \lim_{x\to\infty} x\right] \\
&= \lim_{b\to\infty}\int_0^b \mathrm{d}x - \lim_{b\to\infty}\int_0^b F_X(x)\,\mathrm{d}x \\
&= \lim_{b\to\infty}\left(\int_0^b \mathrm{d}x - \int_0^b F_X(x)\,\mathrm{d}x\right) = \lim_{b\to\infty}\left(\int_0^b (1 - F_X(x))\,\mathrm{d}x\right) \\
&= \int_0^\infty [1 - F_X(x)]\,\mathrm{d}x.
\end{aligned}$$

Solution to 3

Assuming the order of integration can be interchanged, a graph of the range of integration shows that

$$\begin{aligned}\int_0^\infty (1 - F_X(x))\, \mathrm{d}x &= \int_0^\infty \int_x^\infty f_X(t)\, \mathrm{d}t\, \mathrm{d}x \\ &= \int_0^\infty \left(\int_0^t \mathrm{d}x \right) f_X(t)\, \mathrm{d}t = \int_0^\infty t f_X(t)\, \mathrm{d}t,\end{aligned}$$

and the last expression is just $\mathbb{E}[X]$.

Stochastically More Variable

- Recall that function h is convex if, for all $0 < \lambda < 1$ and all x_1, x_2,

$$h(\lambda x_1 + (1-\lambda)x_2) \leq \lambda h(x_1) + (1-\lambda)h(x_2).$$

- For example, functions $h(x) = x^2$ and $h(x) = \mathrm{e}^x$ are convex for all x, and $h(x) = 1/x$ is convex for $x > 0$. Function g is concave if $-g$ is convex; $g(x) = \ln(x)$ and $g(x) = \sqrt{x}$ are concave.
- Random variable X is said to be *more variable* than Y, denoted $X \geq_v Y$, (or also $F_X \geq_v F_Y$, where F_X is the c.d.f. of X), if $\mathbb{E}[h(X)] \geq \mathbb{E}[h(Y)]$ for all **increasing** and **convex** functions h.
- Let $\bar{F}_X(x) = \Pr(X > x)$. We wish to prove: If X and Y are nonnegative r.v.s, then $X \geq_v Y$ iff, for all $a \geq 0$,

$$\int_a^\infty \bar{F}_X(x)\,\mathrm{d}x \geq \int_a^\infty \bar{F}_Y(x)\,\mathrm{d}x, \qquad (9)$$

Stochastically More Variable

To prove "only if", as in Ross, Stochastic Processes (1996, Section 9.5), assume X and Y are nonnegative r.v.s and $X \geq_v Y$. Define the increasing, convex function

$$h_a(x) = (x-a)^+ = (x-a)\,\mathbb{I}_{(a,\infty)}(x) = \max(0, x-a)$$

so that $\mathbb{E}[h_a(X)] \geq \mathbb{E}[h_a(Y)]$. Define events

$$E_x^+ = \left\{(X-a)^+ > x\right\} \quad \text{and} \quad E_x = \{X-a > x\}\,.$$

As long as $x > 0$, observe that $E_x^+ = E_x$ for both cases $X \leq a$ and $X > a$.

Stochastically More Variable

Thus, recalling (8), i.e., that $\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))\,\mathrm{d}x$,

$$\begin{aligned}
\mathbb{E}[h_a(X)] &= \int_0^\infty \Pr(h_a(X) > x)\,\mathrm{d}x \\
&= \int_0^\infty \Pr\left(E_x^+\right)\mathrm{d}x = \int_0^\infty \Pr(E_x)\,\mathrm{d}x = \int_0^\infty \Pr(X > a + x)\,\mathrm{d}x
\end{aligned}$$

and substituting $z = a + x$ in the last integral gives

$$\mathbb{E}[h_a(X)] = \int_a^\infty \Pr(X > z)\,\mathrm{d}z = \int_a^\infty \bar{F}_X(z)\,\mathrm{d}z.$$

Similarly, $\mathbb{E}[h_a(Y)] = \int_a^\infty \bar{F}_Y(z)\mathrm{d}z$, and this shows (9).

Stochastically More Variable: Continuation and Exercise

To prove "if", assume (9) for all $a \geq 0$ and let h be an increasing, convex function.

Assume further that h is twice differentiable, in which case, convexity of h implies that $h''(x) \geq 0$, so that, from (9),

$$\int_0^\infty h''(a) \int_a^\infty \bar{F}_X(x) \, \mathrm{d}x \, \mathrm{d}a \geq \int_0^\infty h''(a) \int_a^\infty \bar{F}_Y(x) \, \mathrm{d}x \, \mathrm{d}a.$$

Exercise: Show that the left hand side (lhs) is

$$\int_0^\infty h''(a) \int_a^\infty \bar{F}_X(x) \, \mathrm{d}x \, \mathrm{d}a = \mathbb{E}[h(X)] - h(0) - h'(0) \, \mathbb{E}[X].$$

(Solution after the next slide...)

Stochastically More Variable

Similarly, the rhs is $\mathbb{E}[h(Y)] - h(0) - h'(0)\,\mathbb{E}[Y]$, so that

$$\mathbb{E}[h(X)] - h'(0)\,\mathbb{E}[X] \geq \mathbb{E}[h(Y)] - h'(0)\,\mathbb{E}[Y]$$

or

$$\mathbb{E}[h(X)] - \mathbb{E}[h(Y)] \geq h'(0)\,(\mathbb{E}[X] - \mathbb{E}[Y])\,.$$

But, as h is increasing, $h'(0) \geq 0$, and setting $a = 0$ in (9) shows that $\mathbb{E}[X] \geq \mathbb{E}[Y]$, so that $\mathbb{E}[h(X)] - \mathbb{E}[h(Y)] \geq 0$, or $\mathbb{E}[h(X)] \geq \mathbb{E}[h(Y)]$, which is the same as saying that $X \geq_v Y$, as was to be shown.

Solution to the Exercise

We change the order of integration in the same manner as was done to show that, for nonnegative X, $\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))\mathrm{d}x$. That is,

$$\begin{aligned}
A &= \int_0^\infty h''(a) \int_a^\infty \bar{F}_X(x)\,\mathrm{d}x\ \mathrm{d}a = \int_0^\infty \int_0^x h''(a)\,\mathrm{d}a\ \bar{F}_X(x)\ \mathrm{d}x \\
&= \int_0^\infty [h'(x) - h'(0)]\ \bar{F}_X(x)\ \mathrm{d}x \\
&= \int_0^\infty h'(x)\,\bar{F}_X(x)\ \mathrm{d}x - \int_0^\infty h'(0)\,\bar{F}_X(x)\ \mathrm{d}x \\
&= \int_0^\infty h'(x)\,\bar{F}_X(x)\ \mathrm{d}x - h'(0)\,\mathbb{E}[X].
\end{aligned}$$

Solution, cont.

Similarly, the latter integral is

$$\begin{aligned}
B &= \int_0^\infty h'(x)\,\bar{F}_X(x)\,\mathrm{d}x = \int_0^\infty h'(x) \int_x^\infty f_X(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_0^\infty \int_0^y h'(x)\,\mathrm{d}x\, f_X(y)\,\mathrm{d}y \\
&= \int_0^\infty [h(y) - h(0)]\, f_X(y)\,\mathrm{d}y \\
&= \int_0^\infty h(y)\, f_X(y)\,\mathrm{d}y - \int_0^\infty h(0)\, f_X(y)\,\mathrm{d}y \\
&= \mathbb{E}[h(X)] - h(0).
\end{aligned}$$

Thus, $A = \mathbb{E}[h(X)] - h(0) - h'(0)\,\mathbb{E}[X]$, as was to be shown.

Stochastically More Variable: Application and Homework

- The result (or, it's "dual") is useful in the study of stochastic dominance in financial economics.
- In particular, Second-Order Stochastic Dominance, SSD, is defined as follows. For r.v.s X and Y with with support on a closed interval $[0, 1]$, X second-order stochastically dominates Y, written F_X SSD F_Y, if, for all $a \in [0, 1]$,

$$\int_0^a [F_Y(x) - F_X(x)] \, \mathrm{d}x \geq 0. \tag{10}$$

- See, e.g., Danthine and Donaldson, Intermediate Financial Theory (2005, Section 4.6) and Laffont, The Economics of Uncertainty and Information, (1989, Section 2.5).
- Let u be an increasing, concave, twice-differentiable (utility) function, so that $u'(x) \geq 0$ and $u''(x) \leq 0$. Prove that

$$\mathbb{E}[u(Y)] \leq \mathbb{E}[u(X)].$$

Stochastically More Variable: Solution to Homework

From (10),

$$\int_0^1 u''(a) \int_0^a F_Y(x)\, \mathrm{d}x\, \mathrm{d}a \leq \int_0^1 u''(a) \int_0^a F_X(x)\, \mathrm{d}x\, \mathrm{d}a \tag{11}$$

and the lhs is

$$\begin{aligned} A &= \int_0^1 u''(a) \int_0^a F_Y(x)\, \mathrm{d}x\, \mathrm{d}a = \int_0^1 \left[\int_x^1 u''(a)\, \mathrm{d}a \right] F_Y(x)\, \mathrm{d}x \\ &= \int_0^1 u'(1) F_Y(x)\, \mathrm{d}x - \int_0^1 u'(x) F_Y(x)\, \mathrm{d}x \end{aligned}$$

and the latter integral is

$$\begin{aligned} B &= \int_0^1 u'(x) F_Y(x)\, \mathrm{d}x = \int_0^1 u'(x) \int_0^x f_Y(t)\, \mathrm{d}t\, \mathrm{d}x \\ &= \int_0^1 f_Y(t) \int_t^1 u'(x)\, \mathrm{d}x\, \mathrm{d}t \\ &= \int_0^1 f_Y(t) \left[u(1) - u(t) \right] \mathrm{d}t = u(1) - \mathbb{E}\left[u(Y) \right]. \end{aligned}$$

Stochastically More Variable: Solution to Homework

The rhs is obviously similar, so that (11) implies

$$\int_0^1 u'(1)\, F_Y(x)\ \mathrm{d}x + E[u(Y)] \leq \int_0^1 u'(1)\, F_X(x)\ \mathrm{d}x + \mathbb{E}[u(X)]$$

or

$$\mathbb{E}[u(Y)] - \mathbb{E}[u(X)] \leq u'(1) \int_0^1 [F_X(x) - F_Y(x)]\ \mathrm{d}x.$$

The rhs is negative, because $u'(1)$ is positive and the integral is just the reverse of (10) for $a = 1$, and thus negative.

Exercise: Problem 7.20

The F density is given by:

$$f_{\mathrm{F}}(x; n_1, n_2) = \frac{n}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \frac{(nx)^{n_1/2-1}}{(1+nx)^{(n_1+n_2)/2}}, \quad n = \frac{n_1}{n_2},$$

and the beta is given by

$$f_{\mathrm{Beta}}(x; p, q) = \frac{1}{B(p,q)} x^{p-1} (1-x)^{q-1} \mathbb{I}_{[0,1]}(x).$$

Let $X \sim \mathrm{F}(n_1, n_2)$ and define

$$B = \frac{\frac{n_1}{n_2} X}{1 + \frac{n_1}{n_2} X}.$$

Show that $B \sim \mathrm{Beta}(n_1/2, n_2/2)$ and $(1-B) \sim \mathrm{Beta}(n_2/2, n_1/2)$.

Solution

The density of X is given by

$$f_X\left(x;\ n_1,n_2\right)=\frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{n_1/2}\frac{x^{(n_1-2)/2}}{\left(1+\frac{n_1}{n_2}x\right)^{(n_1+n_2)/2}}\mathbb{I}_{(0,\infty)}\left(x\right).$$

Defining $a=n_1/n_2$, $\Pr\left(B\leq b\right)$ is given by

$$\begin{aligned}\Pr\left(\frac{aX}{1+aX}\leq b\right) &= \Pr\left(X\leq\frac{b}{a\left(1-b\right)}\right)\\ &= \frac{1}{B\left(n_1/2,n_2/2\right)}a^{n_1/2}\int_0^{\frac{b}{a(1-b)}}\frac{x^{(n_1-2)/2}}{\left(1+ax\right)^{(n_1+n_2)/2}}\,\mathrm{d}x.\end{aligned}$$

Solution

Using the change of variable $y = y(x) = ax(1+ax)^{-1}$, $x = ya^{-1}(1-y)^{-1}$ and $\mathrm{d}x/\mathrm{d}y = a^{-1}(1+y)^{-2}$ with the limits of integration $y(0) = 0$ and $y\left(\frac{b}{a(1-b)}\right) = b$ to get, noting that $1 + y(1-y)^{-1} = (1-y)^{-1}$,

$$\begin{aligned}
\Pr(B \le b) &= \frac{1}{B(n_1/2, n_2/2)} a^{n_1/2} \int_0^b \frac{\left(\frac{y}{a(1-y)}\right)^{(n_1-2)/2}}{\left(1 + \frac{ay}{a(1-y)}\right)^{(n_1+n_2)/2}} \frac{1}{a(1-y)^2} \,\mathrm{d}y \\
&= \frac{1}{B(n_1/2, n_2/2)} a^{\left(\frac{n_1}{2} - \frac{n_1-2}{2} - 1\right)} \int_0^b y^{\frac{n_1}{2}-1} (1-y)^{-\frac{n_1-2}{2} + \frac{n_1+n_2}{2} - 2} \,\mathrm{d}y \\
&= \frac{1}{B(n_1/2, n_2/2)} \int_0^b y^{\frac{n_1}{2}-1} (1-y)^{\frac{n_2}{2}-1} \,\mathrm{d}y = \bar{B}_b(n_1/2, n_2/2),
\end{aligned}$$

so that $B \sim \mathrm{Beta}(n_1/2, n_2/2)$.

Solution

In general, if $X \sim \mathrm{Beta}(a, b)$, then

$$f_X(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \mathbb{I}_{(0,1)}(x)$$

and, with $Y = 1 - X$,

$$f_Y(x; a, b) = f_X(1-y; a, b)\,|-1| = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} (1-y)^{a-1} y^{b-1} \mathbb{I}_{(0,1)}(1-y)$$

so that $Y \sim \mathrm{Beta}(b, a)$.
Thus,

$$1 - B = 1 - \frac{(n_1/n_2)X}{1 + (n_1/n_2)X} = \frac{1}{1 + (n_1/n_2)X} \sim \mathrm{Beta}(n_2, n_1).$$

Course Outline

Conditional Distributions

- Let $F_{X,Y}$ denote a bivariate c.d.f. with corresponding p.m.f. or p.d.f. $f_{X,Y}$ and define A and B to be the events $\{(x, y) : x \in A_0\}$ and $\{(x, y) : y \in B_0\}$, respectively.
  The conditional probability $\Pr(x \in A_0 \mid y \in B_0)$ is given by

$$\frac{\Pr(x \in A_0,\, y \in B_0)}{\Pr(y \in B_0)} = \frac{\Pr(A \cap B)}{\Pr(B)},$$

  assuming $\Pr(B) > 0$, where

$$\Pr(B) = \Pr(x \in \mathbb{R}, y \in B_0) = \int_{y \in B_0} \mathrm{d}F(y)$$

  is evaluated from the marginal p.m.f. or p.d.f. of Y.

Conditional Distributions (2)

- Discrete case: $\Pr\left(x \in A_0 \mid y \in B_0\right)$ is given by

$$\frac{\sum_{x \in A_0} \sum_{y \in B_0} f_{X,Y}(x,y)}{\sum_{x \in \mathbb{R}} \sum_{y \in B_0} f_{X,Y}(x,y)} = \sum_{x \in A_0} \frac{\sum_{y \in B_0} f_{X,Y}(x,y)}{\sum_{y \in B_0} f_Y(y)} =: \sum_{x \in A_0} f_{X \mid Y \in B_0}\left(x \mid B_0\right),$$

  where $f_{X \mid Y \in B_0}$ is defined to be the **conditional p.m.f. given** $y \in B_0$.

- Now let event $B = \{(x,y) : y = y_0\}$. If event $A = \{(x,y) : x \leq x_0\}$, then the **conditional c.d.f.** of X **given** $Y = y_0$ is

$$\Pr(A \mid B) = \frac{\Pr\left(X \leq x, Y = y_0\right)}{\Pr\left(Y = y_0\right)} = \sum_{i=-\infty}^{x} \frac{f_{X,Y}\left(i, y_0\right)}{f_Y\left(y_0\right)} =: F_{X \mid Y = y_0}\left(x \mid y_0\right).$$

- Likewise, if A is the event $\{(x,y) : x = x_0\}$, then the **conditional p.m.f.** of X **given** $Y = y_0$ is

$$\Pr(A \mid B) = \frac{\Pr\left(X = x, Y = y_0\right)}{\Pr\left(Y = y_0\right)} = \frac{f_{X,Y}\left(x, y_0\right)}{f_Y\left(y_0\right)} =: f_{X \mid Y}\left(x \mid y_0\right).$$

Conditioning on the Sum

- Let X_1 and X_2 be independent r.v.s. We want the **conditional** distribution of X_1 **given** that their sum is some particular value, say s.
- From the previous conditional p.m.f. formula and the independence of X_1 and X_2,

$$\begin{aligned}\Pr(X_1 = x \mid X_1 + X_2 = s) &= \frac{\Pr(X_1 = x,\ X_2 = s - x)}{\Pr(X_1 + X_2 = s)} \\ &= \frac{\Pr(X_1 = x)\Pr(X_2 = s - x)}{\Pr(X_1 + X_2 = s)}.\end{aligned}$$

- Important special cases:
  1. If $X_i \overset{\text{iid}}{\sim} \mathrm{Bin}(n, p)$, then $X_1 + X_2 \sim \mathrm{Bin}(2n, p)$ and $X_1 \mid X_1 + X_2$ is hypergeometric.
  2. If $X_i \overset{\text{iid}}{\sim} \mathrm{Geo}(p)$, then $X_1 + X_2 \sim \mathrm{NBin}(r = 2, p)$ and $X_1 \mid X_1 + X_2$ is discrete uniform.
  3. If $X_i \overset{\text{ind}}{\sim} \mathrm{Poi}(\lambda_i)$, then $X_1 + X_2 \sim \mathrm{Poi}(\lambda_1 + \lambda_2)$ and $X_1 \mid X_1 + X_2$ is binomial with $p = \lambda_1 / (\lambda_1 + \lambda_2)$ and $x = 0, 1, \ldots, s$.

Continuous Conditional Distributions

- In the continuous case, $\Pr\left(x \in A_0 \mid y \in B_0\right)$ is given by

$$\frac{\int_{x \in A_0} \int_{y \in B_0} f_{X,Y}(x,y)\, \mathrm{d}y\, \mathrm{d}x}{\int_{x \in \mathbb{R}} \int_{y \in B_0} f_{X,Y}(x,y)\, \mathrm{d}y\, \mathrm{d}x} = \int_{x \in A_0} \frac{\int_{y \in B_0} f_{X,Y}(x,y)\, \mathrm{d}y}{\int_{y \in B_0} f_Y(y)\, \mathrm{d}y}\, \mathrm{d}x$$

$$=: \int_{x \in A_0} f_{X \mid Y \in B_0}\left(x \mid B_0\right) \mathrm{d}x$$

  and $f_{X \mid Y \in B_0}$ is referred to as the **conditional p.d.f. given** $y \in B_0$.
- If B_0 is a point in $\mathcal{S}_Y$, this is problematic:

$$\Pr\left(X \leq x \mid Y = y_0\right) = \frac{\Pr\left(X \leq x, Y = y_0\right)}{\Pr\left(Y = y_0\right)} = \frac{0}{0},$$

  so that alternative definitions are needed.

Continuous Conditional Distributions (2)

- These are

$$F_{X\mid Y=y_0}(x \mid y_0) := \int_{-\infty}^{x} \frac{f(t, y_0)}{f(y_0)} \, \mathrm{d}t$$

$$\text{and} \quad f_{X\mid Y=y_0}(x \mid y_0) := \frac{\partial}{\partial x} F_{X\mid Y}(x \mid y_0) = \frac{f(x, y_0)}{f(y_0)},$$

which should appear quite natural in light of the results for the discrete case.

- The equation for $F_{X\mid Y=y_0}$ can be justified by expressing

$$\begin{aligned}
\Pr(x \in A_0 \mid Y = y_0) &= \lim_{h \to 0^+} \int_{x \in A_0} \frac{\int_{y_0}^{y_0+h} f_{X,Y}(t, y) \, \mathrm{d}y}{\int_{y_0}^{y_0+h} f_Y(y) \, \mathrm{d}y} \, \mathrm{d}t \\
&= \int_{x \in A_0} \frac{h f_{X,Y}(x, y_0)}{h f_Y(y_0)} \, \mathrm{d}x = \int_{x \in A_0} \frac{f_{X,Y}(x, y_0)}{f_Y(y_0)} \, \mathrm{d}x,
\end{aligned}$$

using the mean value theorem for integrals.

General Definition

- Usually $F_{X|Y=y_0}$ and $f_{X|Y=y_0}$ are not given for a specific value y_0 but rather as a general function of y, denoted respectively as

$$F_{X|Y}(x \mid y) := \frac{\int_{-\infty}^{x} f_{X,Y}(t, y)\, \mathrm{d}t}{f_Y(y)}, \qquad f_{X|Y}(x \mid y) := \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

  which can then be evaluated for any particular $y = y_0$.
- In this case, we speak simply of the **conditional c.d.f.** and **conditional p.d.f.**, given as a function of y.
- By multiplying both sides of $f_{X|Y}(x \mid y) = f_{X,Y}(x, y) / f_Y(y)$ with $f_Y(y)$ and integrating with respect to y, we obtain an expression for the marginal of X as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y)\, \mathrm{d}F_Y(y).$$

General Definition (2)

- f_X is **weighted average of the conditional density of X given Y**, **weighted by density Y**.
- The analogy to the law of total probability is particularly clear in the discrete case, i.e., $\Pr(X = x) = \sum \Pr(X = x \mid Y = y) \Pr(Y = y)$ for all $x \in \mathbb{S}_X$.
- Furthermore, Bayes' rule can be generalized as

$$f_{X \mid Y=y}(x \mid y) = \frac{f_X(x) f_{Y \mid X}(y \mid x)}{\int_{-\infty}^{\infty} f_{Y \mid X}(y \mid x) \, \mathrm{d}F_X(x)},$$

which provides an expression for conditional $X \mid Y$ in terms of that for $Y \mid X$ and the marginal of X.

Exchange Paradox (Wallet Game)

The so–called *Exchange Paradox*, or *Wallet Game* provides a nice example of the value and "naturalness" of Bayesian arguments, and goes as follows.

There are two sealed envelopes, the first with m dollars inside, and the second with $2m$ dollars inside, and they are otherwise identical (appearance, thickness, weight, etc.).

You and your opponent are told this, but the envelopes are mixed up so neither of you know which contains more money. You randomly choose an envelope, and your opponent receives the other.

You open yours, and find x dollars inside. Your opponent opens hers, and finds out the contents; call it Y dollars.

Exchange Paradox (Wallet Game)

The two players are then given the opportunity to trade envelopes. You quickly reason as follows: With equal probability, her envelope contains either $Y = x/2$ or $2x$ dollars. If you trade, you thus expect to get

$$\frac{1}{2}\left(\frac{x}{2} + 2x\right) = \frac{5x}{4}, \tag{12}$$

which is greater than x, so you express your interest in trading.

Your opponent, of course, made the same calculation, and is just as eager to trade!
The paradox is that, while the rule which led to the decision to trade seems simple and correct, the result that one should *always* trade seems intuitively unreasonable.

Exchange Paradox (Wallet Game)

The Bayesian resolution of the problem involves treating the amount of money in the first envelope, m, as a random variable, M, and with a prior probability distribution.

To believe that no prior information exists on M violates all common sense. For example, if the game is being played at a party, the $3m$ dollars probably derived from contributions from a few people, and you have an idea of their disposable income and their willingness to part with their money for such a game.

Let g_M be the p.d.f. of M, and let X be the amount of money in the envelope you chose. Then

$$\Pr(X = m \mid M = m) = \Pr(X = 2m \mid M = m) = \frac{1}{2}.$$

Exchange Paradox (Wallet Game)

As X is either M or $2M$, it follows that, on observing $X = x$, M is either x or $x/2$.
From Bayes' rule, $\Pr(M = x \mid X = x)$ is

$$\frac{\Pr(X = x \mid M = x)\, g_M(x)}{\Pr(X = x \mid M = x)\, g_M(x) + \Pr(X = x \mid M = x/2)\, g_M(x/2)}$$

$$= \quad \frac{\frac{1}{2} g_M(x)}{\frac{1}{2} g_M(x) + \frac{1}{2} g_M(x/2)} = \frac{g_M(x)}{g_M(x) + g_M(x/2)},$$

and likewise,

$$\Pr(M = x/2 \mid X = x) = \frac{g_M(x/2)}{g_M(x) + g_M(x/2)}.$$

Exchange Paradox (Wallet Game)

Thus, the expected amount after trading is

$$E[Y \mid X = x] = \frac{g_M(x/2)}{g_M(x) + g_M(x/2)} \cdot \frac{x}{2} + \frac{g_M(x)}{g_M(x) + g_M(x/2)} \cdot 2x, \quad (13)$$

and when $g(x/2) = 2g(x)$, $E[Y \mid X = x] = x$.
Thus, the decision rule is:

$$\text{trade if } g(x/2) < 2g(x),$$

and keep it otherwise.

Exchange Paradox (Wallet Game)

For example, if $g_M \sim \mathrm{Exp}(\lambda)$, i.e., $g_M(m;\lambda) = \lambda e^{-\lambda m}\mathbb{I}_{(0,\infty)}(m)$, then

$$g(x/2) < 2g(x) \Rightarrow \lambda e^{-\lambda x/2} < 2\lambda e^{-\lambda x} \Rightarrow x < \frac{2\ln 2}{\lambda},$$

i.e., you should trade if $x < (2\ln 2)/\lambda$.

Note that (12) and (13) coincide when $g_M(x/2) = g_M(x)$ for all x, i.e., $g_M(x)$ is a constant, which implies that the prior distribution on M is a "noninformative", improper uniform density on $(0,\infty)$.

As mentioned above, it would defy logic to believe that, *in a realistic situation which gives cues about the amount of money involved*, someone would place equal probability on M being 10 dollars and M being 1000 dollars.

Thus, the paradox is resolved by realizing that prior information cannot be ignored if one wishes to make a rational decision.

Memoryless

- Let $X \sim \mathrm{Exp}(\lambda)$. Then, for $x, s \in \mathbb{R}_{>0}$,

$$\begin{aligned}
\Pr(X > s + x \mid X > s) &= \frac{\Pr(X > s + x)}{\Pr(X > s)} \\
&= \frac{\exp\{-\lambda(s + x)\}}{\exp\{-\lambda s\}} = e^{-\lambda x} = \Pr(X > x).
\end{aligned}$$

- The fact that this conditional probability is not a function of s is referred to as the **memoryless property**.

Memoryless (2)

- It is worth emphasizing with some concrete numbers what the memoryless property means, and what it does not. If, say, the lifetime of an electrical device, measured in days, follows an exponential distribution, and we have observed that it is still functioning at time $t = 100$ (the event $X > 100$), then the probability that it will last at least an additional 10 days (i.e., conditional on our observation) is the same as the *unconditional* probability that the device will last at least 10 days.
- What is *not* true is that the probability that the device, conditional on our observation, will last an additional 10 days, is the same as the unconditional probability of the device lasting at least 110 days. That is, it is *not* true that $\Pr(X > s + x \mid X > s) = \Pr(X > s + x)$.

Conditional Moments

- Moments (and expectations of other functions) of a conditional distribution are defined in a natural way:
- Let $\mathbf{Y} = (X_{m+1}, \ldots, X_n)$ so that $\mathbf{X} = (\mathbf{X}_m, \mathbf{Y})$, the expected value of function $g(\mathbf{X}_m)$ conditional on $\mathbf{Y}$ is given by

$$\mathbb{E}\left[g(\mathbf{X}_m) \mid \mathbf{Y}\right] = \int_{\mathbf{x}\in\mathbb{R}^m} g(\mathbf{X}_m)\, \mathrm{d}F_{\mathbf{X}_m|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}).$$

- Note that $\mathbb{E}\left[g(\mathbf{X}_m) \mid \mathbf{Y}\right]$ is **a function of y**.
- As such, it also makes sense to consider the expectation of $\mathbb{E}\left[g(\mathbf{X}_m) \mid \mathbf{Y}\right]$ **with respect to Y**.

Conditional Moments (2)

- In the continuous univariate setting,

$$\begin{aligned}\mathbb{E}_Y\mathbb{E}_{X\mid Y}\left[g(X)\mid Y\right] &= \int_{-\infty}^{\infty} f_Y(y)\int_{-\infty}^{\infty} g(x)\,\frac{f_{X,Y}(x,y)}{f_Y(y)}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)\,f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} g(x)\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} g(x)\,f_X(x)\,\mathrm{d}x = \mathbb{E}_X\left[g(X)\right].\end{aligned}$$

i.e.,

$$\mathbb{E}\mathbb{E}\left[g(X)\mid Y\right] = \mathbb{E}\left[g(X)\right],$$

which is referred to as the **Law of the iterated expectation**.

Example

Consider the bivariate density

$$f_{X,Y}(x,y) = e^{-y}\mathbb{I}_{(0,\infty)}(x)\,\mathbb{I}_{(x,\infty)}(y)\,.$$

- The marginal distribution of X is

$$f_X(x) = \mathbb{I}_{(0,\infty)}(x)\int_x^{\infty} e^{-y}\,\mathrm{d}y = e^{-x}\mathbb{I}_{(0,\infty)}(x)\,,$$

  so that $X \sim \mathrm{Exp}(1)$ and $\mathbb{E}[X] = 1$.
- For Y, note that the range of x is 0 to y so that

$$f_Y(y) = \int_0^{y} e^{-y}\,\mathrm{d}x = ye^{-y}$$

  so that $Y \sim \mathrm{Gam}(2,1)$ and $\mathbb{E}[Y] = 2$.

Example (2)

- The conditional density of $Y \mid X$ is

$$f_{Y|X}\left(y \mid X = x\right) = \frac{f_{X,Y}\left(x, y\right)}{f_X\left(x\right)} = e^{x-y}\mathbb{I}_{(0,\infty)}\left(x\right)\mathbb{I}_{(x,\infty)}\left(y\right),$$

  i.e., such that $0 < x < y$, with expected value

$$\mathbb{E}\left[Y \mid X\right] = \int_x^\infty y f_{Y|X}\left(y \mid x\right) \mathrm{d}y = e^x \int_x^\infty y e^{-y} \,\mathrm{d}y = x + 1,$$

  using the substitution $u = y$ and $\mathrm{d}v = e^{-y}\,\mathrm{d}y$.

- From the **law of the iterated expectation**,

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] = \mathbb{E}\left[X + 1\right] = 2$$

  as above.

Exercise

(From Ross, 1988, p. 286, based on an example in Emanuel Parzen, *Stochastic Processes*, 1962, page 50.)

- A miner is lost in a tunnel with 3 doors. Behind the first door is a tunnel which, after 3 hours, leads out of the mine.
- Behind the 2nd and 3rd doors are tunnels which, after 5 and 7 hours, respectively, lead back to the same position.
- Assume the doors all look the same and cannot be marked, and that the miner chooses randomly among them.
- How many hours do we expect the miner to require to get out of the mine?

Expected Shortfall

- An important measure in financial risk management is the *expected shortfall*. Before defining it, we state a general result:
- Let the p.d.f. and c.d.f. of r.v. R be f_R and F_R. Then the expected value of measurable function $g(R)$, given that $R \leq c$, is

$$\mathbb{E}[g(R) \mid R \leq c] = \frac{\int_{-\infty}^{c} g(r) f_R(r) \, \mathrm{d}r}{F_R(c)}.$$

- **Exercise I** Show that, for $R \sim \mathrm{N}(0,1)$ with p.d.f. ϕ and c.d.f. Φ, and a fixed $c < 0$, $\mathbb{E}[R \mid R \leq c] = -\phi(c) / \Phi(c)$.

Solution

Solution to Exercise I

Let $u = -r^2/2$. Then

$$\begin{aligned}
\mathbb{E}[R \mid R \le c] &= \frac{1}{\Phi(c)} \int_{-\infty}^{c} r\phi(r)\,\mathrm{d}r \\
&= \frac{1}{\Phi(c)} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} r \exp\left\{-\frac{1}{2}r^2\right\} \mathrm{d}r \\
&= \frac{1}{\Phi(c)} \frac{1}{\sqrt{2\pi}} \left(-\exp\left\{-\frac{1}{2}c^2\right\}\right) = -\frac{\phi(c)}{\Phi(c)}.
\end{aligned}$$

Expected Shortfall

- The expected shortfall is defined as

$$\mathrm{ES}_\theta(R) = \mathbb{E}[R \mid R \leq q_{R,\theta}] = \frac{1}{\theta}\int_{-\infty}^{q_{R,\theta}} r f_R(r)\,\mathrm{d}r,$$

  where R is a future period financial return and $q_{R,\theta}$ is the θ-quantile such that $\Pr(R \leq q_{R,\theta}) = \theta$ and θ is small, typically 1%.

- **Exercise II** Let Z be a location zero, scale one r.v., and let $Y = \sigma Z + \mu$ for $\sigma > 0$. Show that

$$\mathrm{ES}_\theta(Y) = \mu + \sigma \mathrm{ES}_\theta(Z),$$

  i.e., that ES preserves location-scale transformations.
  Hint: First show that $q_{Y,\theta} = \sigma q_{Z,\theta} + \mu$.

Expected Shortfall: Solution to Exercise II

First, we have

$$\begin{aligned} \Pr(Z \le q_{Z,\theta}) = \theta \quad &\Leftrightarrow \quad \Pr(\sigma Z + \mu \le \sigma q_{Z,\theta} + \mu) = \theta \\ &\Leftrightarrow \quad q_{Y,\theta} = \sigma q_{Z,\theta} + \mu. \end{aligned}$$

Then,

$$\begin{aligned} \mathrm{ES}_\theta(Y) &= \mathbb{E}[Y \mid Y \le q_{Y,\theta}] \\ &= \mathbb{E}[\sigma Z + \mu \mid \sigma Z + \mu \le \sigma q_{Z,\theta} + \mu] \\ &= \sigma \mathbb{E}[Z \mid Z \le q_{Z,\theta}] + \mu = \sigma\, \mathrm{ES}_\theta(Z) + \mu. \end{aligned}$$

Expected Shortfall: Solution to Exercise II: Alternate Proof

Recalling the location-scale density transformation $f_Y(y) = \sigma^{-1} f_Z((y-\mu)/\sigma)$, we get

$$\mathrm{ES}_\theta(Y) = \frac{1}{\theta}\int_{-\infty}^{q_{Y,\theta}} y f_Y(y)\,\mathrm{d}y = \frac{1}{\theta}\int_{-\infty}^{q_{Y,\theta}} y \frac{1}{\sigma} f_Z\left(\frac{y-\mu}{\sigma}\right)\mathrm{d}y.$$

Substituting $z = (y-\mu)/\sigma$, $y = \mu + \sigma z$, $\mathrm{d}y = \sigma \mathrm{d}z$,

$$\begin{aligned} \mathrm{ES}_\theta(Y) &= \frac{1}{\theta}\int_{-\infty}^{\frac{q_{Y,\theta}-\mu}{\sigma}} (\mu + \sigma z) f_Z(z)\,\mathrm{d}z \\ &= \frac{1}{\theta}\mu \int_{-\infty}^{q_{Z,\theta}} f_Z(z)\,\mathrm{d}z + \frac{1}{\theta}\sigma \int_{-\infty}^{q_{Z,\theta}} z f_Z(z)\,\mathrm{d}z \\ &= \mu + \sigma \mathrm{ES}_\theta(Z). \end{aligned}$$

Expected Shortfall: Exercise III

- Let Q_X be the quantile function of continuous r.v. X, i.e., $Q_X : (0,1) \to \mathbb{R}$ with $p \mapsto F_X^{-1}(p)$. Show that $ES_\theta(X)$ can be expressed as

$$ES_\theta(X) = \frac{1}{\theta} \int_0^\theta Q_X(p)\,\mathrm{d}p.$$

- This is a common form of expressing ES because a weighting function (called the risk spectrum or risk-aversion function) can be incorporated into the integral to form the so-called spectral risk measure.
- Use this result to construct another proof of Exercise II.

Expected Shortfall: Solution to Exercise III

Let $u = Q_X(p)$, so that $p = F_X(u)$ and $\mathrm{d}p = f_X(u)\mathrm{d}u$. Then, with $q_\theta = Q_X(\theta)$,

$$\int_0^\theta Q_X(p)\,\mathrm{d}p = \int_{-\infty}^{q_\theta} u f_X(u)\,\mathrm{d}u.$$

To verify this in Matlab, we use the $\mathrm{N}(0,1)$ case and run:

```
alpha=0.01; c=norminv(alpha);
ES1 = -normpdf(c)/normcdf(c)
ES2 = quadl(@norminv, 1e-7, alpha, 1e-7, 0) / alpha
```

Expected Shortfall: Solution to Exercise III

Solution: Using III to Show II

For $Y = \sigma Z + \mu$, we showed first in Exercise II that $q_{Y,\theta} = \sigma q_{Z,\theta} + \mu$. Thus,

$$\begin{aligned} \mathrm{ES}_{\theta}(Y) &= \frac{1}{\theta}\int_0^{\theta} Q_Y(p)\,\mathrm{d}p = \frac{1}{\theta}\int_0^{\theta} \left[\sigma Q_Z(p) + \mu\right]\mathrm{d}p \\ &= \sigma\frac{1}{\theta}\int_0^{\theta} Q_Z(p)\,\mathrm{d}p + \mu = \mu + \sigma \mathrm{ES}_{\theta}(Z). \end{aligned}$$

Lower Partial Moments

If it exists, the nth order *lower partial moment* with respect to reference point c is, for $n \in \mathbb{N}$,

$$\mathrm{LPM}_{n,c}(X) = \int_{-\infty}^{c} (c - x)^n f_X(x) \, \mathrm{d}x.$$

This is an important measure for financial portfolio risk with many advantages over the traditional measure (variance). It is related to the Expected Shortfall.

The LPM can be computed with numeric integration, though for both the normal and for fat-tailed distributions, choosing the lower bound on the integral can be problematic.

Lower Partial Moments

Applying the binomial theorem to $(c-x)^n$, we can write

$$\mathrm{LPM}_{n,c}(X) = \sum_{h=0}^{n} K_{h,c}\, T_{h,c}(X), \tag{14}$$

where we define

$$K_{h,c} = K_{h,c}(n) = \binom{n}{h} c^{n-h} (-1)^h$$

and

$$T_{h,c}(X) = \int_{-\infty}^{c} x^h f_X(x)\, \mathrm{d}x.$$

Now we just need "closed form" expressions for $T_{h,c}(X)$. With them, (14) can be quickly and accurately evaluated without the aforementioned numeric integration problem.

Exercise: Lower Partial Moments for Normal

For $Z \sim \mathrm{N}(0,1)$ and $c < 0$, calculation shows (let $u = z^2/2$ for $z < 0$) that, for $h \in \mathbb{N}$,

$$T_{h,c}(Z) = \frac{(-1)^h \, 2^{h/2-1}}{\sqrt{\pi}} \left[\Gamma\left(\frac{h+1}{2}\right) - \Gamma_{c^2/2}\left(\frac{h+1}{2}\right) \right],$$

where $\Gamma_x(a)$ is the incomplete gamma function.

In particular,

$$T_{0,c}(Z) = \Phi(c) \text{ and } T_{1,c}(Z) = -\phi(c).$$

First show $T_{0,c}(Z)$, then $T_{1,c}(Z)$, and finally the general expression for $T_{h,c}(Z)$ above.

Exercise: Lower Partial Moments for Student's t

For $X \sim T(v)$, with density

$$f_T(x; v) = K_v \left(1 + x^2/v\right)^{-\frac{v+1}{2}}, \quad K_v = \frac{v^{-\frac{1}{2}}}{B\left(\frac{v}{2}, \frac{1}{2}\right)},$$

we have (substitute $u = 1 + x^2/v$ for $x < 0$ and then $x = (u-1)/u$), for $h < v$,

$$T_{h,c}(X; v) = \frac{(-1)^h v^{h/2}}{2B\left(\frac{v}{2}, \frac{1}{2}\right)} \left[B\left(\frac{h+1}{2}, \frac{v-h}{2}\right) - B_w\left(\frac{h+1}{2}, \frac{v-h}{2}\right)\right],$$

where $w = \frac{c^2/v}{1+c^2/v}$ and B_w is the incomplete beta function. Show this.

In particular,

$$T_{0,c}(X; v) = F_X(c; v) = \Phi_v(c) \text{ and } T_{1,c}(X; v) = \phi_v(c)\left(v + c^2\right)/(1-v).$$

Independence

- Our first formal encounter with independence stated that r.v.s $X_1, \ldots, X_n$ are independent iff their joint density can be factored into the marginals as $f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$.
- An equivalent statement in the bivariate case can be given in terms of conditional r.v.s: X and Y are independent when $f_{X|Y}(x \mid y) = f_X(x)$ or

$$f_{X,Y}(x,y) = f_{X|Y}(x \mid y) f_Y(y) = f_X(x) f_Y(y) .$$

- This can be generalized to n r.v.s by requiring that $f_{\mathbf{Z}|(\mathbf{X} \setminus \mathbf{Z})} = f_{\mathbf{Z}}$ for all subsets $\mathbf{Z} = (X_{j_1}, \ldots X_{j_k})$ of $\mathbf{X}$.
- Thus, a set of r.v.s are mutually independent if their **marginal and conditional distributions coincide**.

Example

- If $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \mathrm{N}\left(\mu_i, \sigma_i^2\right)$, then their joint density is just the product of n normals, or

$$
\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{n} f_{\mathrm{N}}\left(x_i; \mu_i, \sigma_i^2\right) = \prod_{i=1}^{n} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right\} \\
&= \frac{1}{\sqrt{(2\pi)^n \prod_{i=1}^{n} \sigma_i^2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right\}.
\end{aligned}
$$

- As an important special case, if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$, then

$$
f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right\}.
$$

Computing Probabilities via Conditioning

- If A is an event of interest involving random variable X, the calculation $\Pr(A)$ can sometimes be easily computed by conditioning on X.
- In particular,

$$\Pr(A) = \int_{-\infty}^{\infty} \Pr(A \mid X = x) \, \mathrm{d}F_X(x).$$

- This is a natural generalization of the law of total probability. They coincide if X is discrete and we assign exclusive and exhaustive events B_i to all possible outcomes of X.
- As an (important) example, if X and Y are continuous random variables and event $A = \{X < aY\}$, then, conditioning on Y,

$$\begin{aligned}
\Pr(A) &= \Pr(X < aY) = \int_{-\infty}^{\infty} \Pr(X < aY \mid Y = y) \, f_Y(y) \, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} F_{X\mid Y}(ay) \, f_Y(y) \, \mathrm{d}y.
\end{aligned}$$

Computing Probabilities via Conditioning (2)

- More generally, if event $B = \{X - aY < b\}$, then

$$\begin{aligned} \Pr(B) &= \int_{-\infty}^{\infty} \Pr(X - aY < b \mid Y = y)\, f_Y(y)\, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} F_{X|Y}(b + ay)\, f_Y(y)\, \mathrm{d}y. \end{aligned}$$

- Note that $\Pr(B)$ is the c.d.f. of $X - aY$.
- As $\Pr(B)$ is the c.d.f. of $X - aY$, differentiating with respect to b gives the p.d.f. of $X - aY$ at b,

$$\begin{aligned} f_{X-aY}(b) &= \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}b} F_{X|Y}(b + ay)\, f_Y(y)\, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} f_{X|Y}(b + ay)\, f_Y(y)\, \mathrm{d}y, \end{aligned}$$

assuming that we can differentiate under the integral.

Computing Probabilities via Conditioning (3)

- If X and Y are independent, then $\Pr(B)$ and density $f_{X-aY}(b)$ simplify to

$$\int_{-\infty}^{\infty} F_X(b+ay)\, f_Y(y)\, \mathrm{d}y \quad \text{and} \quad \int_{-\infty}^{\infty} f_X(b+ay)\, f_Y(y)\, \mathrm{d}y,$$

respectively.

- With $a=-1$, the latter reduces to

$$f_{X+Y}(b) = \int_{-\infty}^{\infty} f_X(b-y)\, f_Y(y)\, \mathrm{d}y,$$

which is referred to as the **convolution** of X and Y.

Convolution of i.i.d. Cauchy Random Variables

- The p.d.f. of the sum S of two independent standard (i.e., location zero and scale one) Cauchy r.v.s can be computed via

$$f_{X+Y}(b) = \int_{-\infty}^{\infty} f_X(b-y) f_Y(y)\, \mathrm{d}y,$$

using $f_X(x) = f_Y(x) = \pi^{-1}\left(1+x^2\right)^{-1}$, or

$$f_S(s) = \frac{1}{\pi^2} \int_{-\infty}^{\infty} \frac{1}{1+(s-x)^2} \frac{1}{1+x^2}\, \mathrm{d}x.$$

Convolution of i.i.d. Cauchy Random Variables

- The calculation of this integral is not trivial. The Appendix of the book shows how to do this; it results in

$$f_S(s) = \frac{1}{\pi}\frac{1}{2}\frac{1}{1+(s/2)^2},$$

or $S \sim \mathrm{Cau}(0,2)$.

- A similar calculation shows that, if $X_i \overset{\text{ind}}{\sim} \mathrm{Cau}(0,\sigma_i)$, then

$$S = \sum_{i=1}^{n} X_i \sim \mathrm{Cau}(0,\sigma), \quad \sigma = \sum_{i=1}^{n} \sigma_i \ . \tag{15}$$

- An important special case of (15) is for $\sigma = 1/n$, so that the sample mean of n iid standard Cauchy r.v.s, $\bar{X} = n^{-1}(X_1 + \cdots + X_n)$, also follows a standard Cauchy distribution.

Example: Ratio of Independent Exponentials

- Let $X_i \overset{\text{ind}}{\sim} \mathrm{Exp}(\lambda_i)$, $i = 1, 2$ and define $R = X_1/X_2$.
- As $\Pr(X_2 > 0) = 1$, the distribution of R is

$$F_R(r) = \Pr(R \leq r) = \Pr(X_1 \leq rX_2) = \Pr(X_1 - rX_2 \leq 0).$$

- Recall that

$$\begin{aligned}\Pr(X < aY) &= \int_{-\infty}^{\infty} \Pr(X < aY \mid Y = y) f_Y(y)\, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} F_{X|Y}(ay) f_Y(y)\, \mathrm{d}y.\end{aligned}$$

Example: Ratio of Independent Exponentials

Using this,

$$\begin{aligned} F_R(r;\lambda_1,\lambda_2) &= \int_0^\infty F_{X_1}(rx_2)\, f_{X_2}(x_2)\ \mathrm{d}x_2 \\ &= \int_0^\infty \left(1-e^{-\lambda_1 r x_2}\right)\lambda_2 e^{-\lambda_2 x_2}\ \mathrm{d}x_2 \\ &= 1-\lambda_2\int_0^\infty e^{-x_2(\lambda_1 r+\lambda_2)}\ \mathrm{d}x_2 \\ &= 1-\frac{\lambda_2}{\lambda_1 r+\lambda_2} \\ &= \frac{\lambda_1 r}{\lambda_1 r+\lambda_2}. \end{aligned}$$

Example: Ratio of Independent Exponentials

- The density is

$$f_R(r; \lambda_1, \lambda_2) = \frac{\partial F_R(r)}{\partial r} = \frac{\lambda_1 \lambda_2}{(\lambda_1 r + \lambda_2)^2} \mathbb{I}_{(0,\infty)}(r)$$
$$= \frac{c}{(r+c)^2} \mathbb{I}_{(0,\infty)}(r),$$

  where

$$c = \frac{\lambda_2}{\lambda_1} = \frac{\mathbb{E}[X_1]}{\mathbb{E}[X_2]}.$$

- For $\lambda = \lambda_1 = \lambda_2$, this is just $f_R(r; \lambda) = (1+r)^{-2}$. This is similar (but not the same) as a Cauchy random variable.
- Observe that the mean (and higher moments) of R do not exist.

Ratio of Standard Normals

- We wish to show that, if X and Y are independent $\mathrm{N}(0,1)$ r.v.s, then $Z = X/Y$ follows a Cauchy distribution with $F_Z(z) = 1/2 + (\arctan z)/\pi$.
- To see this, write

$$\begin{aligned} F_Z(z) = \Pr(X/Y < z) = \Pr(X - zY < 0,\ Y > 0) \\ + \Pr(X - zY > 0,\ Y < 0), \end{aligned}$$

which is obtained by integrating the bivariate normal c.d.f. as

$$\begin{aligned} F_Z(z) &= \int_0^{\infty}\int_{-\infty}^{zy} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right)\,\mathrm{d}x\,\mathrm{d}y \\ &\quad + \int_{-\infty}^{0}\int_{zy}^{\infty} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right)\,\mathrm{d}x\,\mathrm{d}y \\ &=: I_1 + I_2. \end{aligned}$$

Ratio of Standard Normals

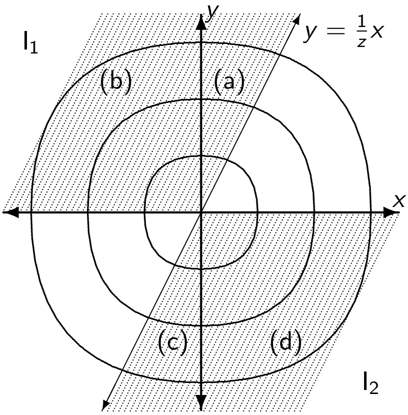
- Consider the case $0 < z < \infty$ or $0 < z^{-1} < \infty$ so that the range of integration is given in the left panel of the figure, where $I_1 = (a) + (b)$ and $I_2 = (c) + (d)$.
- From the spherical symmetry of the (zero-correlated) bivariate normal and the fact that it integrates to one, $(b) = (d) = 1/4$ and $(a) = (c) = \theta/2\pi$ so that

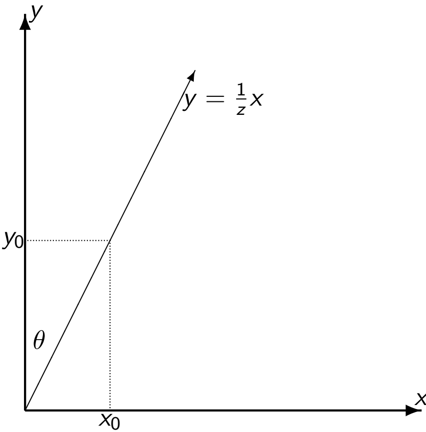
$$F_Z(z) = \frac{1}{2} + \frac{\theta}{\pi} = \frac{1}{2} + \frac{\arctan z}{\pi},$$

from the right panel with $\tan\theta = x$.
- A similar analysis holds for $z < 0$.

Ratio of Standard Normals

The Cauchy c.d.f. is given by $I_1 + I_2$ (left); $\tan\theta = x_0/y_0 = z$ (right)





Conditioning on Ordering Information

Let the lifetime of two electrical devices, X_1 and X_2, be r.v.s with joint p.d.f. f_{X_1,X_2}.

Interest centers on the distribution of the first device, given the information that the first device lasted longer than the second.

The conditional c.d.f. of X_1 given that $X_1 > X_2$ is

$$\begin{aligned} F_{X_1|(X_1>X_2)}(t) &= \Pr(X_1 \leq t \mid X_1 > X_2) = \frac{\Pr((X_1 \leq t) \cap (X_1 > X_2))}{\Pr(X_1 > X_2)} \\ &= \frac{\displaystyle\iint_{y<x,\ x\leq t} f_{X_1,X_2}(x,y)\,\mathrm{d}y\,\mathrm{d}x}{\displaystyle\iint_{y<x} f_{X_1,X_2}(x,y)\,\mathrm{d}y\,\mathrm{d}x} \\ &= \frac{\int_{-\infty}^{t}\int_{-\infty}^{x} f_{X_1,X_2}(x,y)\,\mathrm{d}y\,\mathrm{d}x}{\int_{-\infty}^{\infty}\int_{-\infty}^{x} f_{X_1,X_2}(x,y)\,\mathrm{d}y\,\mathrm{d}x} = \frac{I(t)}{I(\infty)}, \end{aligned} \tag{16}$$

where $I(t)$ is defined as the numerator expression in (16).

Conditioning on Ordering Information

If we now assume that X_1 and X_2 are independent, then

$$I(t) = \int_{-\infty}^{t} \int_{-\infty}^{x} f_{X_1}(x) f_{X_2}(y) \, \mathrm{d}y \, \mathrm{d}x = \int_{-\infty}^{t} f_{X_1}(x) F_{X_2}(x) \, \mathrm{d}x, \quad (17)$$

and further assuming that X_1 and X_2 are iid with common p.d.f. f_X, integrating by parts with $u = F_X(x)$ and $\mathrm{d}v = f_X(x)\,\mathrm{d}x$ gives

$$I(t) = [F_X(t)]^2 - \int_{-\infty}^{t} F_X(x) f_X(x) \, \mathrm{d}x = [F_X(t)]^2 - I(t),$$

or

$$I(t) = \frac{1}{2} [F_X(t)]^2, \quad (18)$$

so that

$$F_{X_1 \mid (X_1 > X_2)}(t) = \frac{[F_X(t)]^2 / 2}{[F_X(\infty)]^2 / 2} = [F_X(t)]^2 .$$

Conditioning on Ordering Information

If $X_i \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda)$, $F_{X_1|(X_1>X_2)}(t;\lambda) = \left[1 - e^{-\lambda t}\right]^2$.

If interest instead centers on the conditional c.d.f. of X_1 given that $X_1 < X_2$, then, similar to the above derivation and for $t < y$,

$$F_{X_1|(X_1<X_2)}(t) = \frac{\int_{-\infty}^{t}\int_{x}^{\infty} f_{X_1,X_2}(x,y)\,\mathrm{d}y\,\mathrm{d}x}{\int_{-\infty}^{\infty}\int_{x}^{\infty} f_{X_1,X_2}(x,y)\,\mathrm{d}y\,\mathrm{d}x} = \frac{J(t)}{J(\infty)},$$

and in the iid case, using (17) and (18) gives

$$\begin{aligned} J(t) &= \int_{-\infty}^{t} f_X(x)\,(1 - F_X(x))\,\mathrm{d}x \\ &= \int_{-\infty}^{t} f_X(x)\,\mathrm{d}x - \int_{-\infty}^{t} f_X(x)\,F_X(x)\,\mathrm{d}x \\ &= F_X(t) - \frac{1}{2}\left[F_X(t)\right]^2. \end{aligned}$$

Conditioning on Ordering Information

Thus,

$$F_{X_1|(X_1<X_2)}(t) = \frac{J(t)}{J(\infty)} = \frac{F_X(t) - \frac{1}{2}[F_X(t)]^2}{1-\frac{1}{2}} = 2F_X(t) - [F_X(t)]^2$$
$$= 1 - (1 - F_X(t))^2 .$$

If $X_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, $i = 1, 2$, then

$$F_{X_1|(X_1<X_2)}(t) = 2\left(1 - e^{-\lambda t}\right) - \left(1 - e^{-\lambda t}\right)^2 = 1 - e^{-2\lambda t},$$

showing that $X_1 \mid (X_1 < X_2) \sim \text{Exp}(2\lambda)$.

These results in the iid case can be generalized in the context of the study of *order statistics*.

Conditioning on Ordering Information

The Matlab code

```
lam1=1/10; lam2=1/10; s=10000; X=exprnd(1,s,1)/lam1;
Y=exprnd(1,s,1)/lam2; xly=X(find(X<Y)); hist(xly,40)
```

can be used to simulate the density of $X_1 \mid (X_1 < X_2)$ when X_1 and X_2 are independent.

Course Outline

Continuous Multivariate Transformations

- Earlier, we derived an expression for the p.d.f. of $Y = g(X)$ for continuous r.v. X as $f_Y(y) = f_X(x) \left| \mathrm{d}x / \mathrm{d}y \right|$, where $x = g^{-1}(y)$.
- The generalization to the multivariate case is more difficult to derive, but the result is straightforward to implement.
- Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an n-dimensional continuous r.v. and $\mathbf{g} = (g_1(\mathbf{x}), \ldots, g_n(\mathbf{x}))$ a continuous bijection which maps $\mathcal{S}_{\mathbf{X}}$, the support of $\mathbf{X}$, onto $\mathcal{S}_{\mathbf{Y}}$, a subset of $\mathbb{R}^n$.

Continuous Multivariate Transformations (2)

Then the p.d.f. of $\mathbf{Y} = (Y_1, \ldots, Y_n) = \mathbf{g}(\mathbf{X})$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \left|\det \mathbf{J}\right|,$$

where $\mathbf{x} = \left(g_1^{-1}(\mathbf{y}), \ldots, g_n^{-1}(\mathbf{y})\right)$ and

$$\mathbf{J} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_1^{-1}(\mathbf{y})}{\partial y_n} \\ \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_2} & & \frac{\partial g_2^{-1}(\mathbf{y})}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial g_n^{-1}(\mathbf{y})}{\partial y_2} & \cdots & \frac{\partial g_n^{-1}(\mathbf{y})}{\partial y_n} \end{pmatrix}$$

is the Jacobian of $\mathbf{g}$.

Observe that this reduces to the equation used in the univariate case.

Example: Convolution

- Let X and Y be continuous r.v.s with joint distribution $f_{X,Y}$ and define $S = X + Y$.
- The density f_S involves *convolution*, as was derived above.
- To derive the result using a bivariate transformation, a second, "dummy" variable is required, which can often be judiciously chosen so as to simplify the calculation.
- In this case, we take $T = Y$, which is both simple and such that $(s, t) = (g_1(x, y), g_2(x, y)) = (x + y, y)$ is a bijection.

Example: Convolution

- So, with $S = X + Y$ and $T = Y$, the inverse transformation is easily seen to be $(x, y) = \left(g_1^{-1}(s,t), g_2^{-1}(s,t)\right) = (s - t, t)$, so that $f_{S,T}(s,t) = |\det \mathbf{J}|\, f_{X,Y}(x,y)$, where

$$\mathbf{J} = \begin{bmatrix} \partial x/\partial s & \partial x/\partial t \\ \partial y/\partial s & \partial y/\partial t \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \quad |\det \mathbf{J}| = 1,$$

or $f_{S,T}(s,t) = f_{X,Y}(s - t, t)$. Thus,

$$f_S(s) = \int_{-\infty}^{\infty} f_{X,Y}(s-t, t)\, \mathrm{d}t,$$

as was shown earlier for X and Y independent.

Computational trick

It is sometimes computationally advantageous to use the fact that $|\mathbf{J}| = \left|\mathbf{J}^{-1}\right|^{-1}$, where

$$\mathbf{J}^{-1} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & & \frac{\partial g_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n(\mathbf{x})}{\partial x_1} & \frac{\partial g_n(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_n(\mathbf{x})}{\partial x_n} \end{pmatrix}.$$

In the univariate case this reduces to $\mathrm{d}y / \mathrm{d}x = 1 / (\mathrm{d}x / \mathrm{d}y)$.

Example of using $\left|\mathbf{J}^{-1}\right|$

- Let X and Y be cont. r.v.s with p.d.f. $f_{X,Y}$, and $P := XY$.
- Let $Q = Y$ so that the inverse transformation of $\{p = xy,\ q = y\}$ is $\{x = p/q,\ y = q\}$, and $f_{P,Q}(p,q) = |\det \mathbf{J}|\, f_{X,Y}(x,y)$, where

$$\mathbf{J} = \begin{bmatrix} \partial x/\partial p & \partial x/\partial q \\ \partial y/\partial p & \partial y/\partial q \end{bmatrix} = \begin{bmatrix} 1/q & -pq^{-2} \\ 0 & 1 \end{bmatrix}, \quad |\det \mathbf{J}| = \frac{1}{|q|},$$

  or $f_{P,Q}(p,q) = |q|^{-1} f_{X,Y}(p/q,\ q)$.
- Thus,

$$f_P(p) = \int_{-\infty}^{\infty} |q|^{-1} f_{X,Y}\left(\frac{p}{q}, q\right) \mathrm{d}q.$$

Notice also that $\mathbf{J}\mathbf{J}^{-1} = \mathbf{I}$, where

$$\mathbf{J}^{-1} = \begin{bmatrix} \partial p/\partial x & \partial p/\partial y \\ \partial q/\partial x & \partial q/\partial y \end{bmatrix} = \begin{bmatrix} y & x \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} q & p/q \\ 0 & 1 \end{bmatrix}, \ \left|\det \mathbf{J}^{-1}\right| = |q|.$$

Exercise 1

For $m > 2$, let $B \sim \mathrm{Beta}\left((m-1)/2, (m-1)/2\right)$ independent of $X \sim \chi^2_m$ with p.d.f.

$$f_X(x; m) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{m/2-1} e^{-x/2} \mathbb{I}_{(0,\infty)}(x).$$

Let $S = 2B - 1$, $Y = \sqrt{X}$, and $P = SY$. Derive an integral expression for $f_P(p; m)$ and, for $m = 3$, simplify it and show that, for $m = 3$, $P \sim \mathrm{N}(0,1)$.

Verify using numeric integration that P is $\mathrm{N}(0,1)$ for *any* value of $m > 2$. This is proven in, e.g., page 90 of Ellison (1964).[2]

[2] JASA 59, Two Theorems for Inferences about the Normal Distribution with Applications in Acceptance Sampling, pages 89-95.

Solution, 1/3

The p.d.f. of S is

$$\begin{aligned} f_S(s;m) &= \left|\frac{\mathrm{d}b}{\mathrm{d}s}\right| f_B(b) \\ &= \frac{1}{2}\frac{1}{B\left(\frac{m-1}{2},\frac{m-1}{2}\right)}\left(\frac{s+1}{2}\right)^{\frac{m-1}{2}-1}\left(1-\frac{s+1}{2}\right)^{\frac{m-1}{2}-1}\mathbb{I}_{(0,1)}\left(\frac{s+1}{2}\right) \\ &= \frac{2^{2-m}}{B\left(\frac{m-1}{2},\frac{m-1}{2}\right)}\left(1-s^2\right)^{\frac{m-3}{2}}\mathbb{I}_{(-1,1)}(s). \end{aligned}$$

For $m=3$, B is uniform, and $f_S(s;3)$ easily reduces to what we expect, $(1/2)\,I_{(-1,1)}(s)$.
For the density of Y,

$$\begin{aligned} f_Y(y) &= \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| f_X(x) = 2y\frac{1}{2^{m/2}\Gamma(m/2)}\left(y^2\right)^{m/2-1}e^{-y^2/2}\mathbb{I}_{(0,\infty)}\left(y^2\right) \\ &= \frac{2^{1-m/2}}{\Gamma(m/2)}y^{m-1}\exp\left\{-y^2/2\right\}\mathbb{I}_{(0,\infty)}(y). \end{aligned}$$

Solution, 2/3

Then the density of $P = SY$ is

$$
\begin{aligned}
f_P(p;m) &= \int_{-1}^{1} \frac{1}{s} f_S(s)\, f_Y(p/s)\, \mathrm{d}s \\
&= \frac{\Gamma(m-1)\, 2^{3(1-m/2)}}{\Gamma\left(\frac{m-1}{2}\right)\Gamma\left(\frac{m-1}{2}\right)\Gamma(m/2)} \\
&\quad \times \int_0^1 \frac{1}{s}\left(1-s^2\right)^{\frac{m-3}{2}} \left(\frac{p}{s}\right)^{m-1} \exp\left\{-(p/s)^2/2\right\} \mathrm{d}s,
\end{aligned}
$$

where the integral starts at zero because $f_Y(p/s)$ only has support on $(0,\infty)$.

Solution, 3/3

For $m = 3$, recall $\Gamma(3/2) = \sqrt{\pi}/2$, then substitute $u = p/s$ and then (for $p > 0$) $v = -u^2/2$ to get

$$\begin{aligned} f_P(p;3) &= \frac{1}{\sqrt{2\pi}} \int_0^1 \frac{1}{s} \left(\frac{p}{s}\right)^2 \exp\left\{-\frac{(p/s)^2}{2}\right\} \mathrm{d}s \\ &= \frac{1}{\sqrt{2\pi}} \int_p^\infty u \exp\left\{-\frac{u^2}{2}\right\} \mathrm{d}u = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-p^2/2} e^v \mathrm{d}v \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-p^2/2\right\}, \end{aligned}$$

so that, for $m = 3$, P is indeed standard normal.

The general expression does not seem to simplify, though numerically integrating it for any $m > 2$ shows that $P \sim \mathrm{N}(0,1)$.

Exercise 2

- Let $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(0,1)$, $i = 1, \ldots n$, let s and t be values such that $0 < s < t < 1$, and define $N_n(s) = \sum_{i=1}^{n} \mathbb{I}_{[0,s]}(X_i)$, i.e., $N_n(s)$ is the number of X_i which are less than or equal to s.
- Let $X = N_n(s)$, $Y = N_n(t)$ and $D = Y - X$. Prove that $(D \mid X = x) \sim \mathrm{Bin}(n - x, (t - s)/1 - s)$.

Hint to Exercise 2

What is $f_{X,Y}$ and f_X? From those, bivariate transformation to $f_{D,M}$ with $M = X$, then compute $f_{D|M}$.

Solution, 1/3

We first want the joint p.m.f. of X and Y, $f_{X,Y}(x,y)$.

From the definition of X and Y, x of the X_i are in $[0,s]$, and y values in $[0,t]$, or $y-x$ values in $(s,t]$, and the remaining $n-y$ are greater than t.

Because the X_i are i.i.d., the joint distribution of X and Y is just trinomial, i.e.,

$$f_{X,Y}(x,y;n,s,t)=\binom{n}{x,y-x,n-y}s^x(t-s)^{y-x}(1-t)^{n-y}\,\mathbb{I}(0\le x\le y\le n)$$

and the marginal of X is binomial,

$$f_X(x;n,s)=\binom{n}{x}s^x(1-s)^{n-s}\,\mathbb{I}(0\le x\le n).$$

Solution, 2/3

Let $D = Y - X$ and $M = X$, so that $X = M$ and $Y = D + M$. Then

$$f_{D,M}(d, m) = |\det \mathbf{J}| \, f_{X,Y}(x, y),$$

where

$$\mathbf{J} = \begin{bmatrix} \partial x / \partial d & \partial x / \partial m \\ \partial y / \partial d & \partial y / \partial m \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad |\det \mathbf{J}| = -1,$$

so that $f_{D,M}(d, m) = f_{X,Y}(m, d + m)$ or

$$\binom{n}{m, d, n - (d + m)} s^m (t - s)^d (1 - t)^{n-(d+m)} \, \mathbb{I}(0 \leq m \leq d + m \leq n).$$

Solution, 3/3

Thus, the conditional distribution of D given M is

$$f_{D|M}(d \mid m) = \frac{f_{D,M}(d,m)}{f_M(m)} \qquad \text{or}$$

$$\frac{\binom{n}{m}\binom{n-m}{d} s^m (t-s)^d (1-t)^{n-(d+m)}}{\binom{n}{m} s^m (1-s)^{n-m}} \mathbb{I}(0 \le m \le d+m \le n)\, \mathbb{I}(0 \le m \le n)$$

$$= \binom{n-m}{d} \frac{(t-s)^d (1-t)^{n-(d+m)}}{(1-s)^{n-m} \quad (1-s)^d (1-s)^{-d}} \mathbb{I}(0 \le d \le n-m)$$

$$= \binom{n-m}{d} \left(\frac{t-s}{1-s}\right)^d \left(1 - \frac{t-s}{1-s}\right)^{(n-m)-d} \mathbb{I}(0 \le d \le n-m).$$

That is, $(D \mid M = m) \sim \mathrm{Bin}(n-m, (t-s)/1-s)$, or
$(D \mid X = x) \sim \mathrm{Bin}(n-x, (t-s)/1-s)$, as was to be shown.

Example 9.3: Sums of i.i.d. Exponential is Gamma

- Let $X_i \overset{\text{iid}}{\sim} \operatorname{Exp}(\lambda)$, $i = 1, \ldots, n$, and define $Y_i = X_i$, $i = 2, \ldots, n$, and $S = \sum_{i=1}^{n} X_i$. This is a 1–1 transformation with $X_i = Y_i$, $i = 2, \ldots, n$, and $X_1 = S - \sum_{i=2}^{n} Y_i$. The inverse Jacobian is

$$\mathbf{J}^{-1} = \begin{bmatrix} \frac{\partial S}{\partial x_1} & \frac{\partial S}{\partial x_2} & \frac{\partial S}{\partial x_3} & \cdots & \frac{\partial S}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \frac{\partial y_n}{\partial x_3} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

with determinant 1, so that

$$\begin{aligned} f_{S,\mathbf{Y}}(s, \mathbf{y}) &= 1 \cdot f_{\mathbf{X}}(\mathbf{x}) = \lambda^n e^{-\lambda s} \mathbb{I}_{(0,\infty)}\left(s - \sum\nolimits_{i=2}^{n} y_i\right) \prod_{i=2}^{n} \mathbb{I}_{(0,\infty)}(y_i) \\ &= \lambda^n e^{-\lambda s} \mathbb{I}_{(0,\infty)}(s) \, \mathbb{I}_{(0,s)}\left(\sum\nolimits_{i=2}^{n} y_i\right) \prod_{i=2}^{n} \mathbb{I}_{(0,\infty)}(y_i) . \end{aligned}$$

Example 9.3: Sums of i.i.d. Exponential is Gamma

- By expressing the indicator functions as $\mathbb{I}_{(0,\infty)}(s)\,\mathbb{I}_{(0,s)}(y_n)\,\mathbb{I}_{(0,s-y_n)}(y_{n-1})\cdots\mathbb{I}_{(0,s-y_n-\cdots-y_3)}(y_2)$, the Y_i can be integrated out to give the density of S:

$$f_S(s) = \lambda^n e^{-\lambda s}\int_0^s\int_0^{s-y_n}\cdots\int_0^{s-y_n-\cdots-y_3} \mathrm{d}y_2\cdots\mathrm{d}y_{n-1}\,\mathrm{d}y_n\,\mathbb{I}_{(0,\infty)}(s)\,.$$

- Taking $n=5$ for illustration and treating quantities in square brackets as constants (set them to, say, t to help see things), we get

$$\int_0^s\int_0^{s-y_5}\int_0^{s-y_5-y_4}\int_0^{s-y_5-y_4-y_3} \mathrm{d}y_2\,\mathrm{d}y_3\,\mathrm{d}y_4\,\mathrm{d}y_5$$
$$=\int_0^s\int_0^{s-y_5}\int_0^{[s-y_5-y_4]}([s-y_5-y_4]-y_3)\,\mathrm{d}y_3\,\mathrm{d}y_4\,\mathrm{d}y_5$$
$$=\frac{1}{2}\int_0^s\int_0^{[s-y_5]}([s-y_5]-y_4)^2\,\mathrm{d}y_4\,\mathrm{d}y_5=\frac{1}{2}\frac{1}{3}\int_0^s(s-y_5)^3\,\mathrm{d}y_5=\frac{1}{2}\frac{1}{3}\frac{1}{4}s^4.$$

Example 9.3: Sums of i.i.d. Exponential is Gamma

- We conclude without a formal proof by induction that

$$f_S(s) = \frac{\lambda^n}{(n-1)!} e^{-\lambda s} s^{n-1} \mathbb{I}_{(0,\infty)}(s),$$

  or that $S \sim \mathrm{Gam}(n, \lambda)$.
- Notice that the above example also implies that, if $S_i \overset{\text{ind}}{\sim} \mathrm{Gam}(n_i, \lambda)$, then $\sum_{i=1}^{k} S_i \sim \mathrm{Gam}(n_\bullet, \lambda)$, where $n_\bullet = \sum_{i=1}^{k} n_i$.
- This is a fundamental result useful in various probabilistic and statistical contexts.

Examples 9.5 and 9.6 as Practice

- Calculate the joint distribution of $S = X + Y$ and $D = X - Y$ and their marginals for $X, Y \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda)$.
- Same, but for standard normal, i.e., the joint and marginals of $S = Z_1 + Z_2$ and $D = Z_1 - Z_2$ for $Z_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, $i = 1, 2$.

Computer Exercise

- Let $X_i \overset{\text{iid}}{\sim} \mathrm{Gam}(a, 1)$, $i = 1, 2$.
  Then $X_1 + X_2 \sim \mathrm{Gam}(2a, 1)$ and, for some $c > 0$,

$$S_1(c) := \Pr(X_1 + X_2 > c) = 1 - \bar{\Gamma}_c(2a),$$

  while

$$\begin{aligned} S_2(c) &:= \Pr(X_1 > c) + \Pr(X_2 > c) \\ &= 2 \Pr(X_1 > c) = 2\left(1 - \bar{\Gamma}_c(a)\right). \end{aligned}$$

- We wish to compare $S_1(c)$ and $S_2(c)$.
  Clearly, as $c \to 0$, $S_1(c) \to 1$, while $S_2(c) \to 2$, so that, as $c \to 0$, $S_1(c) < S_2(c)$.
  1. Verify numerically that, for a given a, there exists a c_0 such that, for all $c > c_0$, $S_1(c) > S_2(c)$.
  2. Write a function (of a) to numerically compute that value of c such that $S_1(c) = S_2(c)$.

Solution

1. For the first problem, the following code will produce the figure on the next slide. Changing the range of c to, say, 6 to 10 confirms that $S_1(c) > S_2(c)$ also for more extreme c.

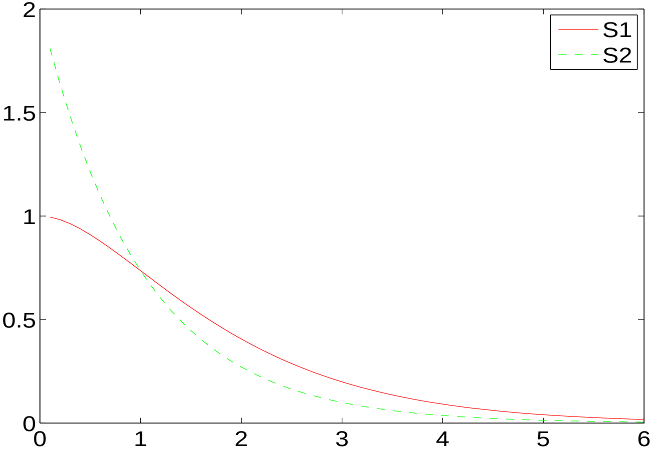
```
subplot(2,1,1) a=1; c=0.1:0.1:6; S1=1-gammainc(c,2*a);
S2=2*(1-gammainc(c,a)); plot(c,S1,'r-',c,S2,'g--')
legend('S1','S2'), set(gca,'fontsize',18)
```

2. For the second problem, the following function can be used.

```
function c = gamma_distribution_sums(a)
c=fzero(@(c) S_diff(c,a),[0 3*a]);

function d=S_diff(x,a)
S_1 =   1-gammainc(x,2*a);
S_2 =2*(1-gammainc(x,a));
d=S_2-S_1;
```

Solution (2)



Derivation of Student's t Distribution

Let $G \sim \mathrm{N}(0,1)$ independent of $C \sim \chi^2_n$ with joint p.d.f.

$$f_{G,C}(g,c) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}g^2\right\} \; \frac{1}{2^{n/2}\Gamma(n/2)} c^{n/2-1} \exp\left\{-\frac{c}{2}\right\} \mathbb{I}_{(0,\infty)}(c)$$

and define $T = G/\sqrt{C/n}$. With $Y = C$, the inverse transform is $G = T\sqrt{Y/n}$ and $C = Y$ so that

$$\mathbf{J} = \begin{pmatrix} \partial G/\partial T & \partial G/\partial Y \\ \partial C/\partial T & \partial C/\partial Y \end{pmatrix} = \begin{pmatrix} \sqrt{Y/n} & \cdot \\ 0 & 1 \end{pmatrix}, \quad |\mathbf{J}| = \sqrt{\frac{Y}{n}}$$

and

$$\begin{aligned} f_{T,Y}(t,y) &= |\mathbf{J}| \; f_{G,C}(g,c) \\ &= \frac{1}{\sqrt{2\pi}} \frac{2^{-n/2} n^{-1/2}}{\Gamma(n/2)} y^{n/2-1/2} \exp\left\{-\frac{1}{2}\left(t\sqrt{\frac{y}{n}}\right)^2 - \frac{y}{2}\right\} \mathbb{I}_{(0,\infty)}(y). \end{aligned}$$

Derivation of Student's t Distribution

Let $u = y\left(1+t^2/n\right)/2$ so that

$$y = \frac{2}{1+t^2/n}u \qquad \mathrm{d}y = \frac{2}{1+t^2/n}\,\mathrm{d}u$$

and

$$\begin{aligned}
I &= \int_0^\infty y^{n/2-1/2} \exp\left\{-\frac{y}{2}\left(1+\frac{t^2}{n}\right)\right\} \mathrm{d}y \\
&= \int_0^\infty \left(\frac{2}{1+t^2/n}u\right)^{n/2-1/2} \exp\{-u\}\ \frac{2}{1+t^2/n}\,\mathrm{d}u \\
&= \frac{2}{1+t^2/n}\left(\frac{2}{1+t^2/n}\right)^{n/2-1/2} \int_0^\infty u^{n/2-1/2} \exp\{-u\}\ \mathrm{d}u \\
&= \left(\frac{2}{1+t^2/n}\right)^{n/2+1/2} \Gamma\left(n/2+1/2\right).
\end{aligned}$$

Derivation of Student's t Distribution

- Using this, the marginal of T, i.e., $\int_0^\infty f_{T,Y}(t,y)\,dy$, simplifies to

$$f_T(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1+t^2/n\right)^{-(n+1)/2} = \frac{n^{-1/2}}{B(n/2,1/2)} \left(1+t^2/n\right)^{-(n+1)/2},$$

  showing that $T \sim t_n$.
- As G and C are independent, $\mathbb{E}[T]$ is

$$\mathbb{E}[T] = \mathbb{E}\left[\frac{G}{\sqrt{C/n}}\right] = \sqrt{n}\mathbb{E}[G]\,\mathbb{E}\left[C^{-1/2}\right] = 0,$$

  using the fact that $\mathbb{E}\left[C^{-1/2}\right]$ is finite.
- For the absolute value of T, using results from earlier examples,

$$\mathbb{E}|T| = \sqrt{n}\mathbb{E}|G|\,\mathbb{E}\left[C^{-1/2}\right] = \frac{\sqrt{n}\sqrt{2/\pi}\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\Gamma\left(\frac{n}{2}\right)} = \frac{\sqrt{n}\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n}{2}\right)}.$$

The F Distribution

- The ratio of two independent χ^2 r.v.s, each divided by their respective degrees of freedom (usually integers, although they need only be positive real numbers), follows an F distribution.
- That is, let $X_i \overset{\text{ind}}{\sim} \chi^2_{n_i}$, and let $Y_1 = (X_1/n_1)\,/\,(X_2/n_2)$. Then $Y_1 \sim \mathrm{F}\,(n_1, n_2)$.
- The derivation is similar to that for the Student's t distribution and is in the text.

Box Muller Transformation

- A computationally efficient way of generating normal r.v.s uses the transformation discovered by Box and Muller (1958).
- Let $U_i \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$ and define

$$\begin{aligned} X_1 &= g_1(U_1, U_2) = \sqrt{-2\ln U_1}\cos(2\pi U_2), \\ X_2 &= g_2(U_1, U_2) = \sqrt{-2\ln U_1}\sin(2\pi U_2). \end{aligned}$$

- It is straightforward to show (see Example 9.10) that $X_i \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$, $i = 1, 2$.
- Problem 9.7 shows how this transformation could have been discovered.

Gamma, Beta, Dirichlet

Let $\mathbf{X} = (X_1, \ldots, X_{n+1})$ with $X_i \overset{\text{ind}}{\sim} \mathrm{Gam}(\alpha_i, \beta)$, $\alpha_i, \beta \in \mathbb{R}$, and define $S = \sum_{i=1}^{n+1} X_i$, $\mathbf{Y} = (Y_1, \ldots, Y_n)$, $Y_i = X_i / S$, so that $\mathcal{S}_{\mathbf{X}} = \mathbb{R}_{>0}^{n+1}$ and

$$\mathcal{S}_{\mathbf{Y},S} = \left\{ (y_1, \ldots, y_n, s) : 0 < y_i < 1,\ \sum_{i=1}^{n} y_i < 1,\ s \in \mathbb{R}_{>0} \right\}.$$

Some work (see the text) shows (i) that

$$f_S(s) = \frac{\beta^\alpha}{\Gamma(\alpha)} s^{\alpha-1} e^{-\beta s} \mathbb{I}_{(0,\infty)}(s),$$

i.e., $S \sim \mathrm{Gam}(\alpha, \beta)$, (ii) that

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{n+1})} \prod_{i=1}^{n} y_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n} y_i\right)^{\alpha_{n+1}-1} \mathbb{I}_{(0,1)}(y_i)\ \mathbb{I}_{(0,1)}\left(\sum_{i=1}^{n} y_i\right),$$

which is referred to as the *Dirichlet* distribution and denoted $\mathbf{Y} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_{n+1})$, and (iii) that S and $\mathbf{Y}$ are independent.

Gamma, Beta, Dirichlet, cont.

- An important special case is $n = 1$, yielding $S = X_1 + X_2 \sim \mathrm{Gam}(\alpha_1 + \alpha_2, \beta)$ independent of $Y = X_1 / (X_1 + X_2)$ with distribution

$$f_Y(y) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} y^{\alpha_1 - 1} (1 - y)^{\alpha_2 - 1} \mathbb{I}_{(0,1)}(y),$$

  i.e., $Y \sim \mathrm{Beta}(\alpha_1, \alpha_2)$.

- The derivation also confirms that

$$B(a, b) = \int_0^1 y^{a-1} (1 - y)^{b-1} \, \mathrm{d}y = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}.$$

Gamma, Beta, Dirichlet, cont.

- Recall from Problem 7.20 that, if $X \sim \mathrm{F}(n_1, n_2)$, then

$$B = \frac{\frac{n_1}{n_2} X}{1 + \frac{n_1}{n_2} X} \sim \mathrm{Beta}\left(\frac{n_1}{2}, \frac{n_2}{2}\right), \tag{19}$$

  i.e.,

$$\Pr(B \leq b) = \Pr\left(X \leq \frac{n_2}{n_1} \frac{b}{1-b}\right),$$

  which was shown directly using a transformation.

- This relation can now be easily verified using the results from previous examples.
- For $n_1, n_2 \in \mathbb{N}$, let $G_i \overset{\text{ind}}{\sim} \chi^2_{n_i}$ or, equivalently, $G_i \overset{\text{ind}}{\sim} \mathrm{Gam}(n_i/2, 1/2)$. Then, we know that

$$F = \frac{G_1/n_1}{G_2/n_2} \sim \mathrm{F}(n_1, n_2).$$

Gamma, Beta, Dirichlet, cont.

- We can then express B from (19) as

$$B = \frac{\frac{n_1}{n_2} F}{1 + \frac{n_1}{n_2} F} = \frac{\frac{n_1}{n_2} \frac{G_1/n_1}{G_2/n_2}}{1 + \frac{n_1}{n_2} \frac{G_1/n_1}{G_2/n_2}} = \frac{\frac{G_1}{G_2}}{1 + \frac{G_1}{G_2}} = \frac{G_1}{G_1 + G_2}.$$

- This also holds for any scale factor common to the G_i, i.e.,

$$\text{if } G_i \overset{\text{ind}}{\sim} \mathrm{Gam}(\alpha_i, \beta), \text{ then } \frac{G_1}{G_1 + G_2} \sim \mathrm{Beta}(\alpha_1, \alpha_2).$$